Caractérisation génomique et transcriptomique du sexe et du système d'auto-incompatibilité diallélique chez l'espèce androdioïque Phillyrea angustifolia

Amélie Carré

UNIVERSITÉ DE LILLE

FACULTÉ DES SCIENCES ET TECHNOLOGIES

Thèse de doctorat pour obtenir le grade de

Docteur de l'Université de Lille

Ecole doctorale Sciences de la Matière, du Rayonnement et de l'Environnement

Caractérisation génomique et transcriptomique du sexe et du système d'auto-incompatibilité diallélique chez l'espèce androdioïque *Phillyrea angustifolia*

par

Amélie Carré

Soutenue publiquement le 22 novembre 2021

Isabelle Fobis-Loisy, Chargée de recherche CNRS, ENS Lyon	Rapportrice
Gabriel Marais, Directeur de recherche CNRS, Université Lyon 1	Rapporteur
Susana Coelho, Directrice MPI Tübingen	Examinatrice
Xavier Vekemans, Professeur des universités, Université de Lille	Président du jury
Pierre Saumitou-Laprade, Directeur de recherche CNRS, Université de Lille	Directeur de thèse
Vincent Castric, Directeur de recherche CNRS, Université de Lille	Membre invité

Remerciement

Il est souvent d'usage de commencer par expliquer comment on en est arrivé à faire une thèse. Je ne saurais pas dire pourquoi mais le doctorat m'avait toujours semblé une voix logique, cependant parfois il m'est arrivé de douter de ma capacité à aller jusqu'au bout, mes doutes s'étaient toujours estompés rapidement. Je trouve important de souligner un évènement particulier et de mettre en avant une personne sans qui je ne serais pas aujourd'hui en train d'écrire les remerciements de ma thèse. Lors de ma première expérience en laboratoire de recherche, alors que je n'étais qu'en première année de BTS biotechnologie, j'ai voulu tout arrêter. Tout, absolument tout, m'avait dégouté du monde la recherche, si être chercheur et travailler dans un laboratoire ressemblait à ça, je ne voulais pas faire ça. Alors toute ma gratitude va à **Harmony Alves dos Santos**, je la remercie d'avoir vu le problème sans que j'aie à en parler, je la remercie d'avoir trouvé une solution sans m'en parler et je la remercie de m'avoir donnée la possibilité et le courage de continuer.

Malgré cela j'avais gardé une certaine appréhension des laboratoires de recherche, et c'est grâce à mon stage de master 2 qu'elle s'est envolée. Je tiens donc à remercier **Marie-Christine Quillet** et **Vincent Castric** de m'avoir offert ma première opportunité de travail au sein du laboratoire EEP. Lorsqu'à la fin de mon master, **Pierre Saumitou-Laprade** m'a parlé du projet de thèse sur *Phillyrea* j'ai été profondément enthousiasmé et je le remercie de m'avoir offert cette opportunité de travail. Bien sûre ça serait mentir que de dire que ces années furent faciles que cela soit pour moi ou pour Pierre. Il y a eu des hauts et des bas, des avancés foudroyantes et des temps de stagnation angoissant, alors merci Pierre pour ta patience et merci un nouvelle fois à Vincent pour son aide précieuse lors de la rédaction.

J'ai eu la chance d'avoir un cadre de travail agréable et pour cela je remercie **Xavier Vekemans**, directeur du laboratoire, pour son accueil et sa bienveillance. **Phillippe Vernet** pour sa gentillesse à mon arrivé et ses discussions, parfois peut-être un peu longue. Ma gratitude se tourne aussi vers les équipes des plateformes techniques : **Cécile Godet, Christelle Blassiau, Laurence, Mathieu Genête, Clément Mazoyer** et **Sophie Gallina**, qui en plus de m'accompagner dans mon travail m'ont souvent offert un support moral important. Un grand merci aussi aux personnes avec qui j'ai pu échanger et auprès desquelles j'ai fait grandir mes connaissances et compétences. Je pense notamment aux membres de mes comités de thèses, aux collaborateurs et collègues du laboratoire. Il serait long d'en faire une

liste exhaustive donc j'espère qu'ils se reconnaîtront. Je remercie également les collaborateurs financiers **Climibio** et **France Olive**, qui m'ont permis de faire ce joli travail. Et s'ils lisent ces quelques lignes l'ensemble des membres de mon jury pour avoir accepté de juger mon travail et pour les précieuses discussions lors de ma soutenance.

Une thèse est un marathon, c'est une phrase qu'on m'a souvent répété au cours des années. C'est un travail qui n'est pas seulement éprouvant intellectuellement mais aussi moralement, elle ne s'arrête pas aux portes du laboratoire et peu entièrement vous engloutir. C'est pour cela que je souhaite offrir mes derniers remerciements à toutes les personnes qui m'ont soutenues, encouragées et permis de garder un semblant de vie sociale. Alors toute ma reconnaissance à **mes parents, Benjamin Fruit, Aurore et Josselin Durel, Leslie Faucher, Laura Henocq, Nicolas Burghraeve, Marina Voison, Natasha Demanicor, Alessandro Fisogni, Renato Bruno, Maryse Vanderplanck, Audrey Leveve, Estelle Barbot, Christophe Calarnou, Thomas Lesaffre, Christophe Van Brussel, Mathilde Latron et Bénédicte Felter**.

Introduction

1. Les systèmes sexuels et stratégies d'appariement

La reproduction chez les angiospermes, groupe majoritaire chez les végétaux terrestres, se caractérise par une diversité de systèmes qui peuvent coexister dans des conditions écologiques apparemment similaires (Barrett 1998). La compréhension des systèmes sexuels, de leur origine et de leur évolution se trouve à la croisée de nombreux domaines de recherche différents, notamment l'écologie, la biologie du développement et la génétique (Ming *et al.* 2011). Des termes clés, tels que « système reproducteur » (reproductive system), « système d'appariement» (mating system) et « système sexuel » (sexual system), sont parfois utilisés comme des synonymes mais le plus souvent dissociés (Diggle *et al.* 2011; Fernandes Cardoso *et al.* 2018). Le cadre général de l'utilisation de ces différents termes, tel qu'ils seront utilisés dans cette thèse, est résumé dans la Figure 1.

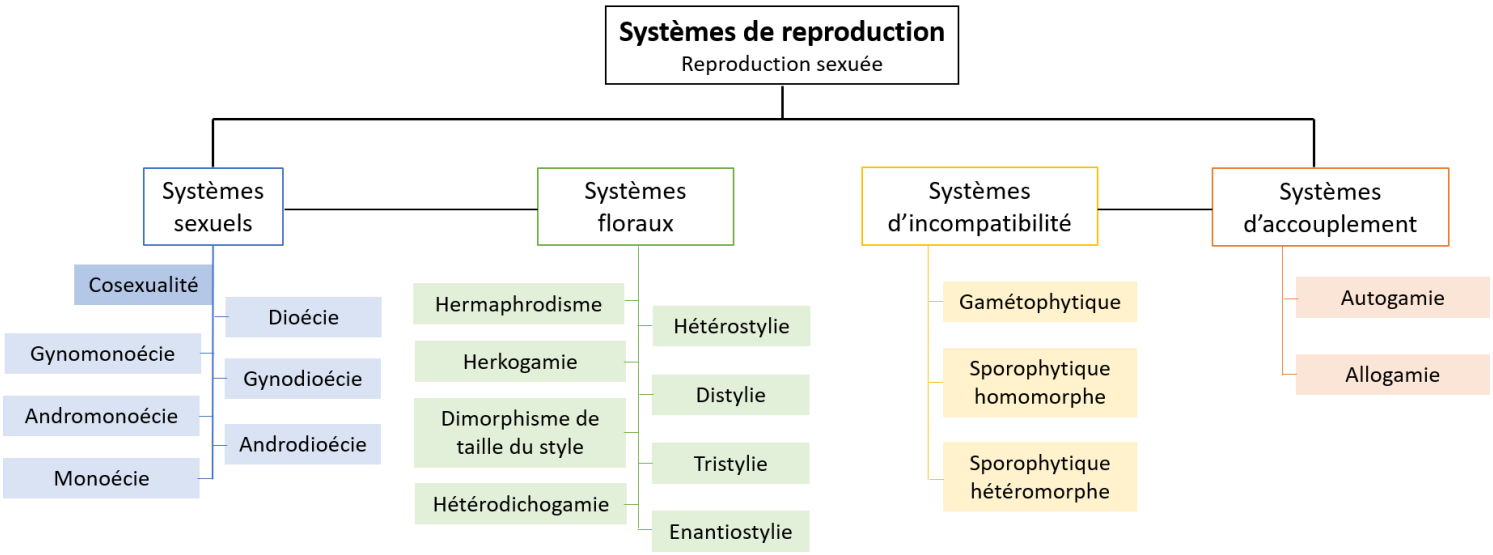


Figure 1 : Cadre général des différentes terminologies liées au système reproducteur des angiospermes (d'après Fernandes Cardoso *et al.* 2018).

Le terme de « système de reproduction » rassemble deux concepts souvent dissociés dans la littérature : la stratégie de reproduction (ou système d'appariement) et le système reproducteur (ou système sexuel). Les systèmes sexuels font référence à la distribution des structures morphologiques productrices de gamètes mâles et femelles sur et entre les individus (Diggle *et al.* 2011). Le système d'appariement est défini par la modalité de l'échange des gamètes et va de l'autogamie à l'allogamie stricte, selon le taux d'autofécondation. L'allogamie correspond à la fécondation croisée entre deux individus distincts. Ce mode de

reproduction évite les effets délétères de la dépression de consanguinité et favorise l'hétérozygotie, la variabilité génétique et les échanges génétiques. Les plantes ont développé divers mécanismes pour promouvoir l'allogamie, notamment le développement des systèmes sexuels.

1.1. Les systèmes sexuels

Une diversité spectaculaire de systèmes reproducteurs

Il existe une large gamme de systèmes sexuels pouvant s'apprécier à différentes échelles. Dans certains systèmes sexuels, les individus sont porteurs des deux fonctions (mâle et femelle), composant un seul phénotype sexuel visible à l'échelle de l'individu (par exemple la monoécie chez le saule, visible à l'échelle de l'individu) tandis que dans d'autres systèmes, les fonctions sexuelles sont séparées sur des individus différents et au moins deux phénotypes sexuels sont présents (comme la dioécie chez le bouleau, visible à l'échelle de la population) (Tableau 1).

Tableau 1 : résumé des systèmes sexuels pouvant être retrouvés chez les plantes. *les différents phénotypes floraux sont portés sur le même individu représenté par [], ** les différents phénotypes floraux sont portés par des individus différents représenté par [], avec ⚥fleurs hermaphrodites, ♂ fleurs mâle et ♀ fleurs femelles.

Un phénotype sexuel*	Au moins deux phénotypes sexuels**
Hermaphrodisme [⚥] Gynomonoécie [⚥♀] Andromonoécie [⚥♂] Monoécie [♀♂]	Dioécie [♀][♂] Gynodioécie[⚥][♀] Androdioécie[⚥][♂] Subgynodioécie[⚥♀][♀] Subandrodioécie[⚥♂][♂] Polygamie [♂][⚥♂][⚥][⚥♀][♀]

Le système sexuel qui prédomine chez les animaux est la restriction de la fonction mâle et femelle à des individus différents, mais il est étonnamment rare chez les plantes. Environ un dixième de toutes les angiospermes sont strictement dioïques ou monoïques (CHARLESWORTH 2002); les espèces hermaphrodites représentent 75% des espèces (GAUDE *et al.* 2001) tandis que les formes intermédiaires de dimorphisme sexuel, la gynodioécie et l'androdioécie, représentent environ 7 % des espèces (DELLAPORTA AND CALDERON-URREA 1993). De nombreux facteurs affectent le degré d'allogamie, notamment la répartition spatiale des sexes, la temporalité de la maturation des organes sexuels au sein de la fleur, ainsi que les vecteurs de

la pollinisation comme les insectes (entomophilie), le vent (anémophilie) ou l'eau (hydrophilie). L'allogamie est notamment favorisée par de nombreux mécanismes développés par les espèces à fleurs hermaphrodites, c'est-à-dire qui présentent à la fois les structures mâles et femelles au sein de la même fleur. Certaines espèces modifient le développement floral pour favoriser la pollinisation croisée, permettant de distinguer des catégories de partenaires réciproquement compatibles. Chez les espèces hétérostyles, qui sont à pollinisation entomophile, les individus produisent des fleurs morphologiquement différentes en ce qui concerne la longueur relative de leur style par rapport au niveau des anthères, comme illustré dans la Figure 2. Enfin certaines espèces ont développé des barrières temporelles de phénologie au sein de la fleur pour éviter l'autofécondation. Par exemple, la protogynie, courante chez les Brassicaceae, et la protandrie, courante chez les Asteraceae, entraînent la maturation asynchrone des organes sexuels femelles ou mâles (DELLAPORTA AND CALDERON-URREA 1993).

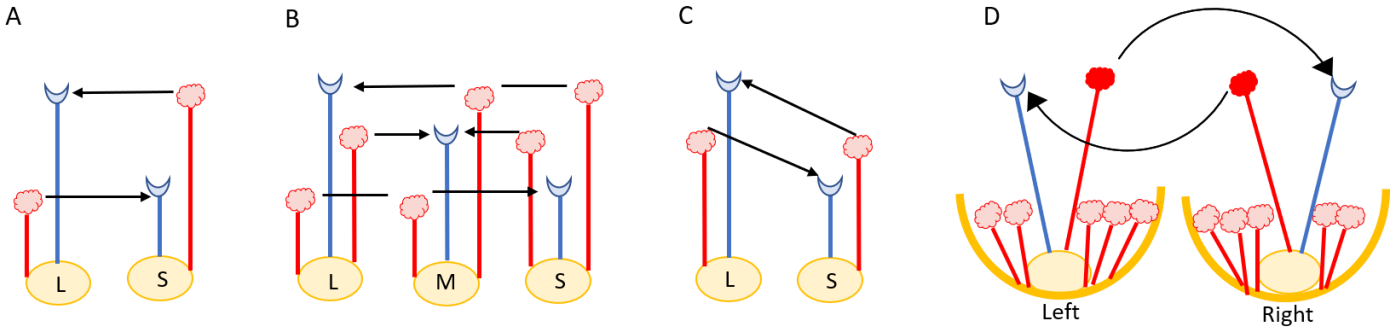


Figure 2 : Les quatre polymorphismes de style chez les plantes à fleurs. A. distylie, B. tristylie, C. dimorphisme de taille du stigmate, D. enantiostylie (d'après BARRETT *et al.* 2000).

De la cosexualité à l'unisexualité : les différentes voies d'évolution

Le terme « cosexuel » est utilisé lorsque les deux fonctions sexuelles sont portées au sein des mêmes individus, qu'elles soient présentes dans chaque fleur ou dans des fleurs séparées. De nombreuses espèces dioïques ont des ancêtres hermaphrodites suggérant une évolution récente des fleurs unisexuées. La faible fréquence et la distribution taxonomique dispersée des espèces dioïques suggèrent que la cosexualité est l'état ancestral des angiospermes (SAUQUET *et al.* 2017). Les déterminants génétiques permettant la séparation des sexes ont donc probablement évolué de manière répétée et indépendante dans différentes familles de plantes (CHARLESWORTH 2002 ; MING *et al.* 2011). Les conformations cosexuelles monomorphiques ou dimorphiques ont souvent été interprétées comme des états

intermédiaires dans une possible évolution vers la dioécie (CHARLESWORTH AND CHARLESWORTH 1978 ; RENNER AND RICKLEFS 1995) (Figure 3).

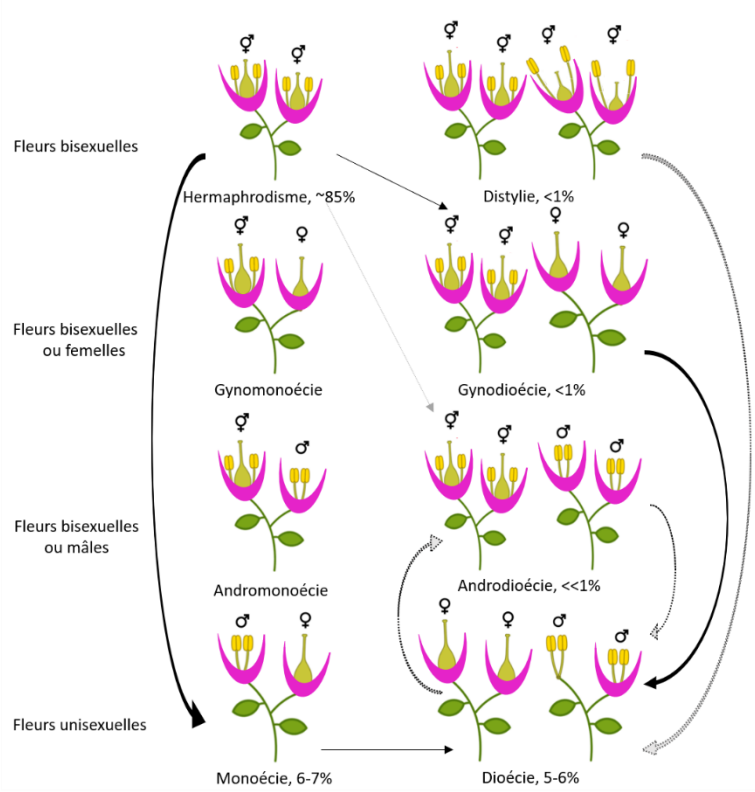


Figure 3 : Représentation schématique des diverses conformations des sexes monomorphiques ou dimorphiques dans les populations végétales et estimations du pourcentage d'espèces d'angiospermes correspondant à chaque conformation. Les flèches correspondent aux voies de transitions principales identifiées : en noir les voies majoritairement étudiées de la gynodioécie et de la monoécie-paradioécie, et en gris les voies pour lesquelles les études sont plus limitées (d'après KÄFER *et al.* 2017).

La transition vers la dioécie peut se faire *via* la voie de la gynodiéocie si les femelles augmentent en fréquence jusqu'à 50% dans la population et que les hermaphrodites déplacent entièrement leur allocation reproductrice vers la fonction mâle (CHARLESWORTH AND CHARLESWORTH 1978). Le déterminisme du sexe chez de nombreuses espèces gynodioïques est nucléo-cytoplasmique et fait intervenir des gènes de stérilité mâle cytoplasmiques (CMS) et des gènes nucléaires de restauration de la fertilité mâle (revue dans REBOUD AND ZEYL 1994). Dans le cas d'une CMS, les femelles peuvent envahir les populations hermaphrodites même avec un avantage faible. Les éléments cytoplasmiques responsables de la mutation de stérilité mâle ne sont pas désavantagés dans leur transmission à la descendance d'une femelle: la suppression de la voie mâle qu'ils n'utilisent pas n'affecte pas leur fitness; elle peut même la favoriser en cas de réallocation des ressources de la voie mâle vers la voie femelle (LEWIS 1941; LLOYD 1974). Le conflit génétique qui résulte de la différence de mode d'hérédité entre

information nucléaire (hérédité biparentale) et cytoplasmique (hérédité uniparentale maternelle) permet d'expliquer le "succès" de la gynodioécie chez les plantes. La fréquence des femelles peut être élevée dans les populations et la gynodioécie peut se maintenir dans le temps, ce qui faciliterait la transition vers la dioécie. A l'inverse, la propagation d'une mutation de stérilité femelle dans une population hermaphrodite, menant à l'établissement d'une population androdioïque, se produit dans des conditions plus strictes que la propagation d'une stérilité mâle menant à l'établissement d'une population gynodioïque. En effet, dans le cas de l'androdioecie les mâles doivent rivaliser avec les hermaphrodites pour la fécondation des ovules (CHARLESWORTH AND CHARLESWORTH 1978). La stérilité mâle empêche immédiatement l'autofécondation d'un individu porteur de la mutation et diminue la dépression de consanguinité des descendants de ces individus : plus le taux d'autofécondation et la dépression de de consanguinité seront élevés plus la mutation sera avantagée comparativement aux hermaphrodites. En revanche, dans le cas de la stérilité femelle, plus le taux d'autofécondation sera élevé moins il y aura d'ovules à féconder pour les mâles, et plus la stérilité femelle sera contrainte. La gynodioécie favorise donc efficacement l'allogamie, à la différence de l'androdioécie (LLOYD 1975). Les fréquences relatives de la gynodioécie et de l'androdioécie sont cohérentes avec la théorie (Figure 3): tandis que la première est commune et pourrait être l'origine de nombreuse transition vers la dioécie (DUFAY *et al.* 2014), la seconde est extrêmement rare et ne se trouve que dans une poignée d'espèces. De plus il a été montré que chez certaines de ces rares espèces, l'androdioécie résultait d'une restauration de la fonction mâle chez les femelles et correspondrait donc à des cas de rupture de la dioécie plutôt qu'à de vraies étapes de transitions vers la dioécie (CHARLESWORTH 2006).

Dans le cas d'une transition de la dioécie *via* la monoécie, aussi appelé "voie de la monoécie-paradioécie" (Figure 3), les individus développent d'abord des fleurs mâles (étamines) et femelles (pistillées) séparées, et la sélection favorise ensuite une tendance progressive de certains individus à accentuer leur fonction mâle et d'autres leur fonction femelle (KÄFER *et al.* 2017). Les hypothèses laissent à penser que le passage par la voie de la monoécie-paradioécie pourrait se faire en réponse à une sélection pour une spécialisation sexuelle accrue, et non en réponse à une sélection pour éviter la consanguinité, comme cela serait le cas lors du passage par la gynodioécie. La séparation des sexes dans cette voie est plus progressive que dans la voie de la gynodioécie : les fleurs femelles sont converties en

fleurs mâles sur certains individus, et les fleurs mâles en fleurs femelles sur d'autres, de sorte qu'il n'y a pas de réduction soudaine du rendement reproducteur (KÄFER *et al.* 2017).

Enfin, bien que la gynodiécie et la monoécie soient considérées comme les deux étapes intermédiaires alternatives les plus probables dans l'évolution de la dioécie à partir de l'hermaphrodisme, la dioécie peut également évoluer à partir de la distylie ou de l'hétérodichogamie, dans lesquelles les populations sont initialement polymorphes pour la séparation spatiale ou temporelle des sexes sur les plantes. Dans le cas du passage par la distylie, la dioécie peut évoluer lorsque l'une des deux morphologies (généralement celle avec des styles longs) évolue vers une spécialisation femelle croissante et que l'autre (généralement le morphe à styles courts) évolue vers une spécialisation mâle (PAILLER *et al.* 1998). Chez les espèces hétérodichogames, les populations comprennent des individus protandres et protogynes (généralement à des fréquences égales), où la moitié des individus fleurissent d'abord en mâle puis en femelle, et vice versa. La sélection peut donner lieu à l'évolution de la dioécie *via* un glissement progressif vers un phénotype mâle accru (par exemple chez les individus protandres) ou un phénotype femelle accru (chez les individus protogynes) (PANNELL AND VERDU 2006).

Déterminismes génétiques du sexe

Dans de nombreuses espèces, les phénotypes sexuels sont contrôlés par des régions génomiques dédiées, qui peuvent occuper de grandes portions de chromosomes, appelés chromosomes sexuels. Comme les autosomes, les chromosomes sexuels fonctionnent dans la majorité des cas par paires. L'un des sexes est qualifié d'homogamétique lorsqu'il porte les deux même types de chromosome sexuel tandis qu'il sera qualifié d'hétérogamétique s'il porte des chromosomes sexuels différents. Suivant le genre du sexe hétérogamétique, on parlera de systèmes XY (hétérogamétique mâle) ou de systèmes ZW (hétérogamétique femelle). Les humains et autres mammifères ont un système XY pour la détermination du sexe : c'est-à-dire que les femelles sont XX et les mâles sont XY, le chromosome Y contenant les facteurs nécessaires au développement d'un individu masculin. Chez les oiseaux, qui ont un système ZW pour la détermination du sexe, le contraire se produit : les mâles sont ZZ et c'est le chromosome W dans la combinaison ZW qui contient les facteurs responsables au développement des femelles. De nombreux insectes ont un système de détermination

sexuelle différent, basé sur le nombre de chromosomes sexuels appelé « système X0 » où le 0 indique l'absence de chromosomes sexuels. Chez les espèces ayant une phase haploïde indépendante dans leur cycle de vie, le sexe peut être déterminé chez les gamétophytes haploïdes par un système UV (femelles U, mâles V) et le sporophyte diploïde est alors hétérogamétique (UV), comme observé par exemple chez *Marchantia polymorpha* (YAMATO *et al.* 2007 ; MUYLE *et al.* 2017).

Les gènes impliqués dans le développement des fleurs sont répartis de manière aléatoire dans le génome et dispersés sur chaque chromosome. Des mutations dans de nombreux gènes de développement floral ont le potentiel de provoquer la stérilité mâle ou femelle, conduisant ainsi à une monoécie, une gynodioécie, une androdioécie ou et une dioécie (DELLAPORTA AND CALDERON-URREA 1993; MING *et al.* 2011). Deux types de mutations sont nécessaires pour établir la dioécie, l'une avortant les étamines (mâle stérile) et l'autre avortant les carpelles (femelle stérile). Les espèces dioïques ne peuvent développer des chromosomes sexuels que lorsque les deux gènes de détermination du sexe sont étroitement liés sur le même chromosome et ont une dominance complémentaire (CHARLESWORTH AND CHARLESWORTH 1978). La dominance complémentaire des deux gènes de détermination du sexe est nécessaire au fonctionnement des chromosomes sexuels. Par exemple, dans le système XY, la fonction femelle est contrôlée par des chromosomes XX homozygotes et la mutation de stérilité mâle doit être une mutation par perte de fonction (c'est-à-dire récessive). Le chromosome Y contient un allèle de fertilité mâle fonctionnel ainsi qu'une mutation de gain de fonction (c'est-à-dire dominante) à un locus différent qui supprime le développement des organes sexuels femelles (CHARLESWORTH AND CHARLESWORTH 1978 ; DELLAPORTA AND CALDERON-URREA 1993). Les études génétiques et génomiques approfondies ont permis de dégager des modèles généraux des étapes d'évolution des chromosomes sexuels. L'une des voies proposées se découpe en six étapes (Figure 4) (MING *et al.* 2011).

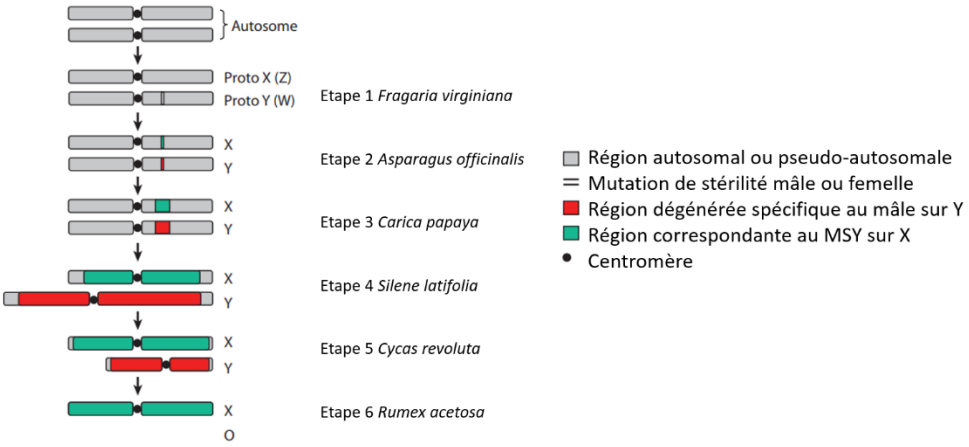


Figure 4. Les six étapes de l'évolution des chromosomes sexuels. Étape 1 : Mutation unisexe de deux gènes de détermination du sexe à dominance complémentaire. Étape 2 : La suppression de la recombinaison entre les deux gènes de détermination du sexe et le génotype YY est viable. Étape 3 : La suppression de la recombinaison s'est propagée aux régions voisines et une petite région spécifique au mâle de la région du chromosome Y (MSY) a évolué. Le génotype YY n'est pas viable. Étape 4 : Le MSY augmente en taille et dégénère en contenu génétique via l'accumulation d'insertions d'éléments transposables et de réarrangements intrachromosomiques. Les chromosomes X et Y deviennent hétéromorphes. Étape 5 : dégénérescence sévère du chromosome Y. La suppression de séquences d'ADN non fonctionnelles entraîne une réduction de la taille du chromosome Y. Étape 6 : La suppression de la recombinaison s'étend à l'ensemble du chromosome Y. Le chromosome Y est perdu et le système de détermination du sexe par ratio X/autosome a évolué (d'après MING *et al.* 2011).

Fonctionnement moléculaire, une grande diversité des mécanismes de déterminisme sexuel

Les chromosomes sexuels des plantes analysés à ce jour varient en termes d'âge, de taille et de contenu génétique global. Pour les gènes identifiés dans les plantes, certaines similitudes nécessaires existent : ils doivent être impliqués à un certain stade du développement des structures spécifiques au sexe (Figure 5). Dans certains cas, les gènes avec une expression tissu-spécifique peuvent être plus susceptibles d'acquérir un rôle dans la détermination du sexe (Figure 5). Cependant parfois il est compliqué de différentier les gènes de détermination du sexe des gènes à expression sexe-biaisée qui joue un rôle central dans l'établissement du sexe. Par exemple, chez le kiwi, *FrBy* est le gène de fertilité mâle lié à l'Y, mais *TDF1* montre également une expression biaisée chez les mâles (AKAGI *et al.* 2019). Dans plusieurs cas, une partie des gènes à expression sexe-biaisée et/ou de détermination du sexe joue un rôle dans la voie des cytokinines (par exemple, peuplier, saule, palmier dattier et kiwi AKAGI *et al.* 2018; TORRES *et al.* 2018 ; ALMEIDA *et al.* 2020; MÜLLER *et al.* 2020), qui est impliqué dans le développement floral, en particulier dans le carpelle et le gamétophyte femelle (revue dans WYBOUW AND DE RYBEL 2019). Un autre modèle notable émergent est le soutien empirique du modèle à deux gènes pour la dioécie. Chez l'asperge ou le kiwi par exemple (AKAGI *et al.* 2019; HARKESS *et al.* 2020), les régions déterminantes du sexe possèdent deux gènes impliqués

l'un dans la stérilité femelle et l'autre dans la stérilité mâle (Figure 5). Dans d'autres systèmes, un seul gène s'est avéré être un interrupteur déterminant le sexe, comme *ARR17* chez le peuplier et *OGI* chez le kaki (MÜLLER *et al.* 2020). Une observation générale sur la région de détermination du sexe est la suppression de recombinaison qui permet son maintien fonctionnel, souvent associée à des inversions chromosomiques (HOOPER *et al.* 2019; SHE *et al.* 2020). D'autres modèles communs frappants, comme la présence de motifs en palindromes, ont été trouvés dans les chromosomes sexuels des animaux et des plantes (SKALETSKY *et al.* 2003; ZHOU *et al.* 2020).

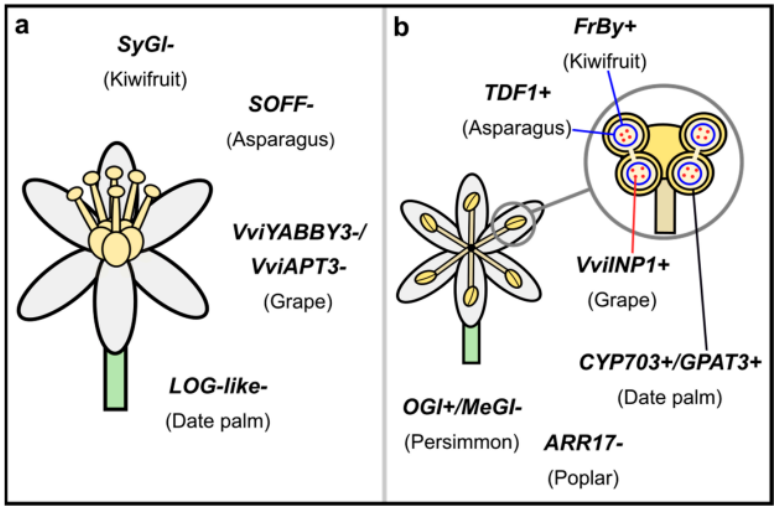


Figure 5 : Schéma résumant les gènes de détermination du sexe récemment mis en évidence. (a) Les gènes déterminant le sexe récemment identifié qui sont impliqués dans le développement des fonctions femelle comprennent *SyGI*, *SOFF* et *LOG-like*. Lorsque ces gènes sont exprimés (+) chez les mâles, cela supprime la fonction ou le développement des structures femelles. Cependant, le manque d'expression (-) chez les femelles permet le développement fonctionnel. Chez le raisin, on ne sait pas encore si *VviYABBY3* ou *VviAPT3* est le gène de la stérilité femelle. (b) Plusieurs gènes favorisant la fonction des étamines ont également été identifiés. *FrBy* et *TDF1* favorisent tous deux le développement du tapetum (en bleu) et *VviINP1* favorise le développement du pollen (en rouge). On ne sait pas encore si *CYP703* ou *GPAT3* est le gène déterminant le sexe mâle chez le dattier, mais les deux sont impliqués dans le développement du pollen et/ou des anthères. Chez le kaki et le peuplier, un seul gène est impliqué dans la détermination du sexe (*OGI* et *ARR17*, respectivement). Lorsque *MeGI* est exprimé, les fleurs développent des organes femelles fonctionnels, mais pas d'étamines. Cependant, lorsque l'*OGI* lié à l'Y est exprimé, il réprime *MeGI*, entraînant des étamines fonctionnelles. De même, chez les peupliers, l'expression d'*ARR17* entraîne la production de carpelles, mais le manque d'expression entraîne des étamines fonctionnelles (d'après CAREY *et al.* 2021).

Globalement, bien qu'on puisse dégager des tendances communes dans le fonctionnement des déterminants du sexe, il existe autant de différence que de similitudes. Notamment dans la façon selon laquelle la suppression de recombinaison se produit dans ces régions. Parfois, elle est associée à une hémizygotie ou un changement structurel comme les inversions (HOOPER *et al.* 2019; HARKESS *et al.* 2020), dans d'autres cas elle est liée à une accumulation d'éléments transposables (ALMEIDA *et al.* 2020). Ces caractéristiques rendent complexes les

approches expérimentales visant à identifier les régions déterminantes du sexe. Aujourd'hui, plusieurs méthodes bio-informatiques permettent d'identifier et de caractériser les chromosomes sexuels dans un large éventail d'espèces non modèles, comblant rapidement les nombreuses lacunes de notre connaissance des systèmes de chromosomes sexuels à travers l'arbre de la vie. À son tour, cet ensemble croissant de données facilite et alimente les efforts visant à mieux comprendre l'évolution des chromosomes sexuels. Des approches comme la cartographie génétique (pour identifier la région non recombinante liée au sexe), l'analyse de ségrégation (individus issus d'un croisement contrôlé dont le pédigré est connu) ainsi que des approches reposant sur la comparaison des couvertures génomique (exploitant la différence de ploïdie des chromosomes sexuels entre les mâles et les femelles) et d'analyse transcriptomique (visant à identifier des transcrits spécifiques des sexes) sont couramment utilisées (PALMER *et al.* 2019).

1.2. Les stratégies de reproduction

Outre leurs mécanismes de détermination du sexe, les angiospermes se distinguent par l'étendue de la variation de leurs stratégies de reproduction, et les modalités de leur contrôle sont elles aussi généralement mal connues. Bien que 75% des espèces d'angiospermes soient hermaphrodites, leurs stratégies de reproductions vont de l'autogamie à l'allogamie stricte, en passant par une large gamme de stratégies mixtes. Dans le cas d'une allogamie stricte, il est fréquent que des mécanismes génétiques d'évitement, appelés systèmes d'auto-incompatibilité (SI), contribuent à empêcher l'autofécondation. L'auto-incompatibilité est définie comme l'inaptitude pour une plante hermaphrodite fertile de produire un zygote par autofécondation (LUNDQVIST 1956; DE NETTANCOURT 1977). La présence des SI permet d'éviter l'autofécondation, et s'oppose ainsi aux effets délétères liés à la consanguinité qui se traduisent par une diminution de la fitness des descendants consanguins. L'importance évolutive des SI est considérable, comme en atteste le pourcentage élevé d'angiospermes qui possèdent de tels systèmes (plus de 50% des familles), et le fait que plusieurs SI soient apparus de manière indépendante au cours de l'évolution (GAUDE *et al.* 2001). Alors que dans de nombreuses espèces, les individus exprimant des spécificités de SI distinctes ne sont pas discernables morphologiquement (on parle de systèmes SI homomorphes), dans certains cas les mécanismes d'évitement de l'auto-fécondation sont associés à des variations

morphologiques observables. Les systèmes SI sont alors dits hétéromorphes, comme dans le cas de la distylie.

A partir des déterminants génétiques du phénotype du pollen, il est par ailleurs possible de classer les systèmes homomorphes en deux groupes : gamétophytique (GSI) par exemple chez les *Papaveraceae*, les *Solanaceae*, les *Poaceae* et les *Rosaseae* ou sporophytique (SSI) comme pour les *Brassicaceae*, les *Convolvulaceae* et les *Asteraceae*. Dans les systèmes gamétophytiques, le phénotype d'incompatibilité du pollen (gamétophyte) est déterminé par son propre génotype haploïde. Dans le cas des systèmes sporophytiques, le phénotype du pollen est déterminé par le génotype diploïde de l'individu producteur de gamète (sporophyte) (GAUDE *et al.* 2001).

Déterminismes génétiques du SI

La majorité des études moléculaires se sont concentrées sur les cas où le SI est sous le contrôle d'un seul locus multi-allélique, le locus S; une exception notable étant celle des *Poaceae* ayant deux locus S et Z (YANG *et al.* 2008). Ce locus contient généralement au moins deux gènes, l'un codant pour le déterminant mâle exprimé au niveau du grain de pollen et l'autre codant pour le déterminant femelle, exprimé dans le pistil. Les déterminants mâle et femelle sont polymorphes et hérités comme une seule unité. Les variants de ce complexe génique sont appelées haplotypes S. Les mécanismes de reconnaissance (interactions protéine-protéine spécifiques) entre les déterminants mâles et femelles porteurs du même haplotype S initient la réponse cellulaire d'auto-incompatibilité (TAKAYAMA AND ISOGAI 2005). Puisque deux gènes distincts contrôlent respectivement les spécificités du pistil et du pollen, ils doivent être étroitement liés et coévoluer en tant qu'une seule unité génétique afin de maintenir le SI. La recombinaison au niveau du locus S entraine une rupture de cette liaison étroite, provoquant la perte de l'auto-incompatibilité en générant des combinaisons haplotypiques auto-compatibles (WHEELER *et al.* 2009). De façon récurrente, les gènes déterminant les spécificités mâles et femelles sont trouvés sous la forme de paralogues dont le rôle est parfois essentiel à la fonction de reconnaissance tel que la duplication du déterminant mâle chez les *Solanaceae* (KUBO *et al.* 2015) mais dont la signification fonctionnelle peut également rester obscure comme la présence d'un gène (*SLG*) au rôle non essentiel chez Brassica (TAKAYAMA AND ISOGAI 2005). En outre, ces gènes sont parfois liés

génétiquement à un ensemble d'éléments moléculaires (petits ARNs non codants) régulant les relations de dominance/récessivité entre allèles (DURAND *et al.* 2014).

Bases génétiques et moléculaires des SI gamétophytiques

Solanaceae et Rosaceae

Dans le cas des Solanaceae et des Rosaceae, le système SI se manifeste par un système de reconnaissance du non-soi. Le locus S se compose d'au moins deux gènes codant pour les déterminants mâles et femelles. La S-RNase est le déterminant femelle, il s'agit de ribonucléases (RNase) sécrétées en grandes quantités dans la matrice extracellulaire du style. Dans un style pollinisé par du pollen incompatible, les S-RNases entrent dans le cytoplasme des tubes polliniques et fonctionnent comme une cytotoxine qui est active par défaut et inhibe l'allongement des tubes par dégradation de l'ARN (TAKAYAMA AND ISOGAI 2005). Le(s) déterminant(s) mâle lié(s) au gène de S-RNase code(nt) pour des protéines F-box et est appelé SLF pour « S-Locus F-box ». Il s'agit de sous-unités de reconnaissance de substrat du complexe *SCF* (*Skp1-Cullin1-F-box*) ubiquitine ligase *E3*, dont le rôle est l'ubiquitination (sorte de marquage) de protéines cibles pour les adresser au protéasome, assurant la dégradation (KUBO *et al.* 2015). Ces protéines produites par le pollen agissent par une action de détoxification en inactivant de manière spécifique l'ensemble des S-RNAses codées par les autres spécificités SI que celle portée par le pollen. Les pollens ne sont donc pas en mesure de détoxifier la S-RNAse des pistils des plantes qui les ont produits, empêchant de fait l'auto-fécondation. Chez les Rosaceae, le système *SLF* ou *SFB* (« *S-haplotype-specific F-Box* » chez *Prunus*), ou *SFBB* (« *S-locus F-Box Brothers* » chez *Pyrinae*) fonctionne avec un unique gène "pollen" inhibiteur général de S-RNase, tandis que chez les Solanaceae plusieurs gènes *SLF* sont présents au locus S et agissent comme un système collaboratif de reconnaissance du « non-soi », chacun interagissant avec un sous-ensemble d'allèles de S-RNase (KUBO *et al.* 2015).

Poaceae

La famille des Poaceae, quatrième plus grande famille de plantes à fleurs, englobe les cultures céréalières et fourragères. Des études sur *Secale cereale* par Lundqvist (1956) et sur *Phalaris coerulescens* par Hayman (1956) ont montré que le système SI est contrôlé gamétophytiquement par au moins deux locus multialléliques et indépendants, S et Z (YANG *et al.* 2008). Le phénotype d'incompatibilité du grain de pollen est déterminé par son génome

haploïde et dépend de la combinaison des allèles S et Z dans le grain de pollen. Malgré des efforts de recherche intenses au cours des six dernières décennies, les gènes sous-jacents S et Z restent presque inconnus. Cependant une étude récente a permis grâce à une combinaison de cartographie fine, de séquençage génomique, d'analyse transcriptomique et d'analyse de séquence comparative détaillée de faire du gène *LpSDUF247* un candidat sérieux pour être le déterminant pollen du système GSI chez *Lolium perenne L.* (MANZANARES *et al.* 2016).

Papaveraceae

Chez les Papaveraceae, le déterminant femelle est appelé *PrsS* (pour *Papaver rhoeas* style S). Il code pour une petite protéine sécrétée par les cellules des papilles du stigmate. L'application des protéines PrsS sur du pollen du même haplotype peut déclencher des réponses physiologiques telles que l'augmentation du calcium cytosolique, la phosphorylation de la pyrophosphatase inorganique, la dépolymérisation de l'actine et des microtubules, une augmentation des dérivés réactifs de l'oxygène et de l'oxyde nitrique, l'acidification cytosolique, et une fragmentation de l'ADN. Beaucoup de ces événements sont impliqués dans l'activation de la cascade de mort cellulaire programmée (apoptose). Le déterminant mâle *PrpS 7* (*P. rhoeas* pollen S) a également été identifié. Bien que la fonction de *PrpS* ne soit pas encore claire, les observations suggèrent que PrpS fonctionnerait comme un récepteur de surface pollinique qui interagit directement avec PrsS (DE GRAAF *et al.* 2012). Globalement chez Papaveraceae, l'auto-rejet (déclenché par une interaction spécifique entre PrsS et PrpS du même haplotype S) se traduit par la mort cellulaire programmé du pollen (FUJII *et al.* 2016).

Le SI sporophytique, l'exemple des Brassicaceae

Le système SSI, avec l'exemple des Brassicaceae, est encore aujourd'hui le plus largement étudié et documenté. Historiquement, la première protéine identifiée a été la protéine S-locus-glycoprotéine (SLG) chez les espèces du genre Brassica, initialement considéré comme le déterminant femelle. Le gène *SLG* n'est cependant pas présent chez tous les haplotypes S. Il a été mis en évidence que, bien qu'elle permette parfois d'améliorer la réaction d'incompatibilité, elle n'est pas essentielle à celle-ci (TAKAYAMA AND ISOGAI 2005). Le déterminant femelle, essentiel et suffisant pour déclencher la réponse d'incompatibilité, est le « S-locus receptor kinase » (SRK), un récepteur kinase transmembranaire localisé au niveau des papilles du stigmate. Le déterminant mâle est la « S-locus protein 11 » (SP11) ou « S-locus

cysteine rich » (SCR). Il code pour un petit peptide généralement sécrété par le tapis de l'anthère qui se localise dans le manteau du pollen. L'interaction moléculaire directe et spécifique entre SP11/SCR et SRK issus du même haplotype S induit la réponse d'incompatibilité par activation d'une cascade de signalisation complexe (FUJII *et al.* 2016).

Pour les systèmes SI sporophytiques, le phénotype de reconnaissance du pollen est contrôlé par le génotype du parent paternel diploïde, mais bien que la plupart des plantes soient hétérozygotes à ce locus, le phénotype de reconnaissance du pollen est généralement déterminé par un seul des deux allèles, en fonction des positions relatives des allèles dans une hiérarchie de dominance-récessivité. On s'attend à ce que la sélection favorise les éléments génétiques qui établissent une interaction de dominance-récessivité plutôt qu'une codominance, car les individus possédant un génotype codominant peuvent produire du pollen rejeté par des partenaires potentiels plus nombreux que dans un système dominant-récessif (LLAURENS *et al.* 2009; DURAND *et al.* 2014).

L'auto-incompatibilité, une équation à plusieurs inconnues

A travers ce panorama, j'ai voulu montrer que les mécanismes moléculaires contrôlant le SI dans les systèmes homomorphes sont remarquablement divers (Tableau 2). Alors que leur fonction ultime est partagée (reconnaissance et rejet de l'auto-pollen), les différentes familles d'Angiospermes utilisent pour réaliser cette fonction des voies physiologiques distinctes, témoin de leurs émergences indépendantes et convergentes au cours de l'évolution. Même si des schémas généraux communs peuvent être dégagés dans les fonctions moléculaires des déterminants des systèmes SI tels que le rôle de reconnaissance et d'interaction entre protéines spécifiques ou la localisation des protéines à la surface du pollen ou du stigmate. Cette remarquable diversité pose une difficulté importante dans la recherche des mécanismes d'auto-incompatibilité chez les espèces où ils ne sont pas connus.

Tableau 2 : Résumé des principaux systèmes SI et de leurs caractéristiques (d'après YANG *et al.* 2008)

	Homomorphique			
	Gamétophytique			Sporophytique
	Solanaceae/Rosaceae	Papaveraceae	Poaceae	Brassicaceae
Nombre de locus	1 (locus S)	1 (locus S)	2 (locus S et Z)	1 (locus S)
Rejet du pollen	Au début du tube pollinique	Surface du stigmate	Surface du stigmate	Surface du stigmate
Vitesse de réaction	lente	rapide	rapide	rapide
Déterminant femelle	S-RNase	PrsS	?	SRK/SLG
Déterminant mâle	SLF/SBP	PrpS	LpSDUF247 ?	SP11/SCR
Reconnaissance du « soi » ou du « non soi »	Non soi	soi	soi	soi

A l'inverse de cette diversité moléculaire des SI, des caractéristiques communes peuvent être clairement identifiées en ce qui concerne le polymorphisme des allèles S et les pressions de sélection permettant le maintien de cette diversité (CASTRIC AND VEKEMANS 2004). Le modèle de sélection qui s'applique de façon générale aux gènes déterminant l'auto-incompatibilité a été initialement développé par Wright (1939). Il stipule que les gènes qui gouvernent le SI devraient faire l'objet d'une sélection naturelle de forte intensité, de type fréquence-dépendante négative, agissant sur la fonction mâle. La première prédiction issue de ce modèle est l'apparition et le maintien d'une importante diversité allélique. En accord avec cette prédiction, plus de 30 et 50 haplotypes S ont par exemple été identifiés à l'échelle de l'espèce, respectivement chez *B.rapa* et *B.oleracea* (Brassicacées), environ une soixantaine d'haplotypes S chez *Arabidopsis halleri* ((GENETE *et al.* 2020)), tandis que chez *Papaver rhoeas* (Papaveracées) on estime à au moins 66 le nombre d'haplotypes S différents (TAKAYAMA AND ISOGAI 2005).

Plusieurs outils sont utilisés pour avancer dans la compréhension de ces systèmes et notamment pour tenter d'identifier les gènes à l'origine ou participant à la réponse d'auto-incompatibilité. Des gènes candidats peuvent être identifiés grâce à différentes approches comme des études protéomiques (FOOTE *et al.* 1994), des analyses transcriptomiques des tissus mâles et femelles (ZHOU *et al.* 2014; ZHANG *et al.* 2016) ou des approches de cartographies de traits quantitatifs (BERNACCHI AND TANKSLEY 1997). Une fois les gènes candidats sélectionnés il faut valider ou exclure la possibilité qu'ils soient réellement impliqués dans le système SI. L'«étalon-or» de la démonstration fonctionnelle consiste en une approche de transformation génétique pour soit inhiber le gène candidat soit à l'inverse l'introduire dans

un fond génétique dont il est absent (NASRALLAH *et al.* 2002). Elle n'est cependant possible que pour les espèces pour lesquelles les outils de transformation génétique sont disponibles (DURAND *et al.* 2014), ou de façon exceptionnelle lorsque les voies métaboliques concernées sont conservées à grande échelle phylogénétique (LIN *et al.* 2015) .

2.*Phillyrea angustifolia*, une espèce androdioïque particulière

2.1. Généralités sur *Phillyrea angustifolia*

Les Oléacées forment une famille composée de 24 genres et environ 600 espèces qui existe depuis le Miocène supérieur (JOHNSON 1957). Les distributions géographiques de ces espèces sont réparties sur l'ensemble du globe (Eurasie, Afrique, Amérique, Australie) et on trouve dans cette famille une grande diversité de systèmes sexuels (Figure 7) (WALLANDER AND ALBERT 2000). Les espèces du genre *Phillyrea* sont circum-méditerranéennes. *P. latifolia* s'étend sur tout le bassin méditerranéen (Grèce, Turquie, Israël), alors que la répartition de *P. angustifolia* est restreinte à l'ouest du bassin méditerranéen (SEBASTIAN 1956) .

Phillyrea angustifolia est un arbuste en forme de buisson dense et ramifié à feuillage persistant qui peut atteindre 4m de haut. C'est une espèce anémophile dont la floraison est abondante. Celle-ci se déroule sur une période de 3 à 4 semaines entre mars et avril. Les individus de filaire mettent 3 à 5 ans avant de produire des fleurs, et les individus ne fleurissent pas forcément tous les ans (LEPART AND DOMMEE 1992). Il s'agit d'une espèce androdioïque dont la fréquence de mâles moyenne dans les populations est de 50% (LEPART AND DOMMEE 1992), et peut même localement dépasser les 75% (HUSSE *et al.* 2013). Les fleurs mâles sont en fait femelle-stériles : l'ovaire est peu développé ou normal, et le stigmate est petit et avorté. Les fleurs mâles et hermaphrodites ont deux étamines de taille identique qui contiennent la même quantité de pollen avec une forte variabilité inter-individuelle (LEPART AND DOMMEE 1992). L'existence de populations androdioïques à fortes fréquences de mâles et sans compensation forte sur l'avantage mâle associé a été un paradoxe longtemps débattu. En effet, les conditions attendues pour le maintien d'une androdioécie fonctionnelle sont : un avantage mâle fort, des fréquences de mâles faibles et pas ou peu d'autofécondation (LEWIS AND CROWE 1958; LLOYD 1975). Clairement, *P. angustifolia* ne se conforme pas à cet attendu.

2.2. Le système d'auto-incompatibilité diallélique

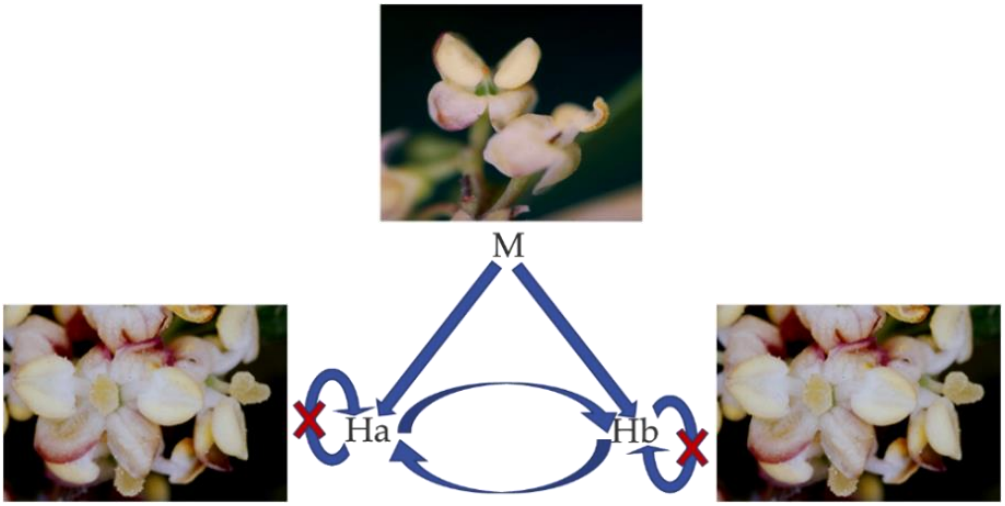


Figure 6 : Schéma des croisements compatible (flèches bleu) et incompatible (flèches barrées d'une croix rouge) en fonction du phénotype sexuel ou de l'haplotypes S chez la filaire.

Ce n'est que depuis 10 ans qu'une explication est venue lever ce paradoxe chez *P. angustifolia*. Il a été montré que le filaire possédait un système d'auto-incompatibilité homomorphe sporophytique diallélique (DSI : diallelic self-incompatibility) (SAUMITOU-LAPRADE *et al.* 2010), où les deux allèles ont été appelés S1 et S2. Sur la base d'observations de descendances de croisements contrôlés, un modèle génétique explicatif a été proposé (BILLIARD *et al.* 2015), où le premier groupe d'hermaphrodites Ha serait homozygote S1S1 et où le deuxième groupe d'hermaphrodites Hb serait hétérozygote S1S2 avec S2 dominant sur S1. De la même manière, l'hypothèse d'un locus du sexe indépendant du SI où les mâles seraient mM et les hermaphrodites mm, a été posé (BILLIARD *et al.* 2015). Les hermaphrodites peuvent être soit des Ha soit des Hb, et sont incompatibles lors de croisements intragroupes et compatibles lors de croisements intergroupes. A l'inverse, les mâles sont compatibles avec les deux groupes d'hermaphrodites, quel que soit leur génotype au locus S (Figure 6).

Dans un tel système, la compatibilité des mâles avec les deux groupes d'hermaphrodites compense le désavantage en fitness lié à la perte de leur fonction femelle : les mâles n'ont plus de désavantage face aux hermaphrodites, et un avantage mâle, même très faible, permet aux mâles de se maintenir dans les populations (PANNELL AND KORBECKA 2010 ; HUSSE *et al.* 2013). De plus l'analyse des descendances de croisements contrôlés (BILLIARD *et al.* 2015) a montré que si les mâles sont bien compatibles avec l'ensemble des hermaphrodites, en revanche la compatibilité des gamètes du donneur mâle varie selon le groupe d'incompatibilité auquel appartient le receveur hermaphrodite. En effet, les hermaphrodites Ha pollinisés par des

mâles produisent systématiquement des hermaphrodites et des mâles, tandis que les hermaphrodites Hb pollinisés par les mêmes mâles ne donnent que des mâles. L'allèle M aurait ainsi un triple effet pléiotrope : il serait tout d'abord responsable du non-développement du stigmate, mais il aurait également un effet épistatique sur le locus S, entrainant la suppression de l'expression du phénotype d'incompatibilité; enfin, il entraînerait une distorsion de ségrégation du sexe des descendants obtenus avec les hermaphrodites Hb. Il semblerait que ces deux derniers effets combinés au DSI des hermaphrodites puissent quantitativement expliquer la forte fréquence des mâles chez *Phillyrea* (BILLIARD *et al.* 2015).

2.3. Le DSI dans la famille des Oléacées

Dans la famille des Oléacées, il existe une hétérostylie ancestrale (TAYLOR 1945 ; WALLANDER AND ALBERT 2000) qui est donc associée à un DSI hétéromorphe (Figure 7). Chez l'espèce hétérostyle *Jasminum fructicans*, l'autofécondation et la fécondation intra-morphe est aussi impossible (DOMMEE *et al.* 1992), ce qui permet de dire qu'il existe un système moléculaire de reconnaissance des gamètes en plus de la barrière morphologique. Cette hétérostylie ancestrale a donné naissance à des espèces possédant des systèmes sexuels très divers (WALLANDER 2001), en association forte avec un évènement d'allo-tétraploïdisation (symbolisé par une étoile verte sur la Figure 7).

Il a été possible de transférer et d'appliquer le test stigmatique mis au point chez le filaire à d'autres espèces de la famille des Oléacées choisies pour leur système de reproduction et leur position dans la phylogénie (symbolisée par des étoiles rouges dans la Figure 7). *Fraxinus ornus* et *Fraxinus excelsior* sont respectivement androdioïque (DOMMEE *et al.* 1999) et trioïques (ALBERT *et al.* 2013), et auraient divergé de *P. angustifolia* il y a environ 20 millions d'années. *O. europea* est hermaphrodite et appartient à un phylum entièrement hermaphrodite ayant divergé plus récemment de *P. angustifolia*. Enfin, *Ligustrum vulgare* (le troène) est une espèce hermaphrodite située à la base des allo-tétraploïdes. Il a été mis en évidence que le pollen de *P. angustifolia* germe (ou ne germe pas) sur les stigmates de *F. ornus* en fonction du groupe d'incompatibilité auquel appartient le receveur *F. ornus.* La réciproque est vraie pour le pollen de *F. ornus* sur les stigmates de *P. angustifolia*. Le pollen des individus mâle du filaire germe aussi sur les stigmates des deux groupes d'hermaphrodites du frêne et réciproquement. Le

frêne possède donc un système DSI homologue, ce qui implique que ce mécanisme et sa fonctionnalité se sont maintenus dans le temps (VERNET *et al.* 2016).

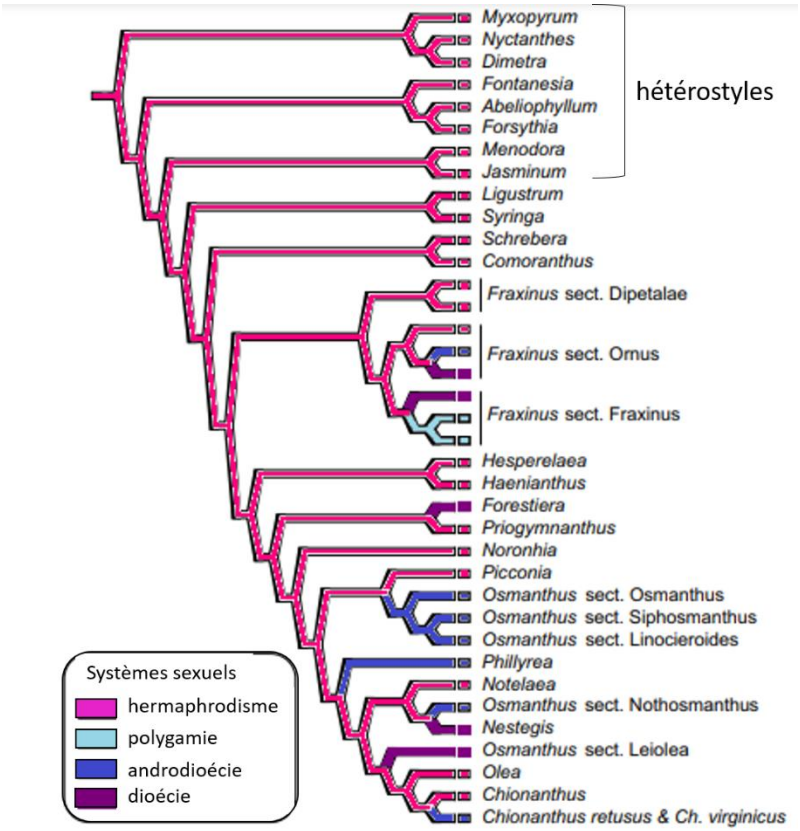


Figure 7 : Arbre phylogénétique de la famille des Oléacées (Wallander and Albert 2000, Wallander ms., and Wallander, Green and Harris, unpubl. Data). L'évènement de duplication de génome (étoile verte) à la base de la tribu des Oleae est représenté, sa date approximative d'apparition est estimée à 30 millions d'années (Unver et al. 2017).

Le DSI apparaissant stable dans le temps, la question qui se pose alors est : comment ce système résiste-il à l'apparition de nouveaux haplotypes S? Par modélisation, il a été montré que, sous un DSI, l'androdioécie pourrait jouer un rôle dans la résistance à l'apparition d'un mutant auto-compatible (BILLIARD *et al.* 2015). En effet la présence de mâles compatibles avec les deux groupes d'incompatibilité rend impossible la sélection de variants auto-compatibles. Ces variants, même s'ils sont compatibles avec les deux groupes, souffrent de dépression de consanguinité et ne peuvent se maintenir face aux mâles qui bénéficient du même avantage à la reproduction sans souffrir de dépression de consanguinité. Or l'émergence et le maintien de variants auto-compatibles est considéré comme la première étape vers l'apparition d'une nouvelle spécificité d'auto-incompatibilité (GERVAIS *et al.* 2011). A ce titre, l'androdioécie pourrait en retour contribuer à la stabilité du DSI dans le temps. Cependant la présence du DSI chez *O.europaea* (SAUMITOU-LAPRADE *et al.* 2017), pose la question théorique du maintien d'un

tel système sur le long terme dans un phylum qui ne compte que des espèces hermaphrodites depuis 30 millions d'années.

3. Le projet de thèse

La co-occurrence chez *Phillyrea angustifolia* de deux systèmes de reproduction particuliers et rares : une androdioécie fonctionnelle et un système sporophytique homomorphe d'auto-incompatibilité diallélique, font de cette espèce un modèle d'étude particulièrement intéressant pour ces deux traits. Le développement d'une approche de génomique chez une espèce arbustive non modèle nécessite de lever différents verrous technologiques, qu'ils soient liés (i) à l'absence de données de séquences sur les espèces étudiées ou (ii) à la durée de création de matériel biologique pertinent (plusieurs années nécessaires entre la production d'individus et leur floraison) et (iii) à l'architecture de l'arbre qui rend complexes le contrôle des pollinisations et la production de descendances en croisements contrôlés. La levée de ces verrous a permis de mettre en place le cadre favorable au travail effectué au cours de mon projet de thèse, dont le but était d'effectuer une caractérisation génomique et transcriptomique du sexe et du DSI chez le filaire.

Une situation particulièrement favorable pour une approche génomique chez *P. angustifolia*.

Notre laboratoire a entamé en 1991 une collaboration sur *P. angustifolia* avec le Centre d'Ecologie Fonctionnelle et Evolutive (CEFE-CNRS) à Montpellier. Au cours de cette collaboration il a été démontré que *P. angustifolia* était une véritable espèce androdioïque, et le premier modèle fondé sur une possible liaison génétique entre sexe et auto-incompatibilité a été proposé (VASSILIADIS *et al.* 2000; VASSILIADIS *et al.* 2002). A partir de 2003, le laboratoire a entamé un vaste programme de pollinisations contrôlées afin de déterminer le nombre de groupes d'incompatibilité présents dans l'espèce. Un protocole de tests stigmatiques a été développé, miniaturisé et rendu compatible avec l'étude de grands nombres d'individus. Afin de constituer des collections de pollen testeurs utilisables quel que soit le moment de floraison des individus, une méthodologie de congélation et de conservation à -80°C du pollen a développée, ce qui permet de tester la compatibilité entre individus et entre espèces dont la floraison est décalée dans le temps. Un programme de croisements contrôlés suivi de

validation des descendants par analyse de paternité a permis de disposer aujourd'hui de plus de 2500 individus de *P. angustifolia* issus d'une trentaine de croisements dont 1015 pour une seule descendance destinée à la cartographie génétique du sexe et de l'incompatibilité. Sur ce matériel biologique unique (et d'un type particulièrement rare dans les études développées sur des espèces arbustives), le laboratoire a développé une approche génomique pour identifier et décrire le fonctionnement des locus du sexe et de l'incompatibilité.

Comment sont déterminés le sexe et l'auto-incompatibilité chez le filaire ?

Au cours de ma thèse, j'ai tenté de caractériser le système d'auto-incompatibilité et le système sexuel chez *P. angustifolia* à travers trois expériences.

Dans un premier temps, une approche par cartographie génétique haute-densité a permis de tester l'hypothèse des deux locus génétiquement indépendants et d'étudier les types de ségrégation observés au locus S et au locus du sexe. Le but était de répondre à trois grandes questions : (i) est-ce que le locus du sexe suit un modèle de type XY où les mâles sont hétérozygotes et les hermaphrodites sont homozygotes pour l'haplotype récessif ? (ii) est-ce que le modèle suivi par le locus S est le même que celui du sexe, ou est-il similaire au déterminisme de l'héterostylie chez Primula (fonctionnement en hémizygotie) ? (iii) est-il possible d'estimer la taille des régions génomiques associées au locus S et au locus du sexe.

Dans un deuxième temps, cette approche a été couplée à une analyse transcriptomique dont le but était de trouver et caractériser les déterminants moléculaires responsables et/ou participant à la détermination des phénotypes d'auto-incompatibilité et du sexe. Tout d'abord, la comparaison de l'expression différentielle entre les transcrits d'individus mâles par rapport à ceux des hermaphrodites a permis de rechercher les transcrits potentiellement impliqués dans le développement du phénotype mâle et éventuellement dans la compatibilité particulière des mâles avec les deux groupes d'hermaphrodites. Par ailleurs, l'analyse de l'expression différentielle entre les transcrits des Ha et des Hb a permis de rechercher des transcrits à l'origine du -ou jouant un rôle dans- le déterminisme des phénotypes d'auto-incompatibilité.

Enfin, dans un troisième et dernier temps, une expérience de capture de gène par hybridation ciblée a été déployée sur les candidats mis en évidence par les analyses d'expression différentielle et de cartographie génétique. Cette dernière approche, effectuée

sur des individus issus d'un croisement contrôlé et d'une population naturelle, a pour but, à travers une analyse de la ségrégation des SNPs, de différencier les séquences qui sont liées génétiquement au locus S ou au locus du sexe et qui pourraient représenter des gènes « switch » potentiels de celles, non liées, qui seraient impliquées dans les cascades moléculaires d'établissement des différents phénotypes.

Dans ce manuscrit, j'ai choisi de détailler mon travail en trois parties qui ne suivent pas strictement les trois axes expérimentaux.

- Le premier chapitre est constitué de l'article dans lequel l'établissement de la cartographie génétique haute densité et le positionnement des locus du sexe et du SI est développé. Cet article a été recommandé par PCI Genomics.
- Dans le second chapitre, j'ai décidé de me concentrer sur la recherche des gènes candidats à l'origine de l'androdioécie chez *Phillyrea angustifolia* à travers la mise en évidence des différences transcriptomiques et génomiques entre mâles et hermaphrodites.
- Enfin le troisième chapitre, développe les résultats des expériences de transcriptomique et de capture par hybridation ciblée en lien avec la caractérisation du système d'auto-incompatibilité.

Peer Community In Genomics

RESEARCH ARTICLE

Genetic mapping of sex and self-incompatibility determinants in the androdioecious plant *Phillyrea angustifolia*

Amélie Carré[1], Sophie Gallina[1], Sylvain Santoni[2], Philippe Vernet[1], Cécile Godé[1], Vincent Castric[1], Pierre Saumitou-Laprade[1]

[1] CNRS, Univ. Lille, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France
[2] UMR DIAPC - Diversité et adaptation des plantes cultivées

Open Access
Open Data
Open Code
Open Peer-Review

This article has been peer-reviewed and recommended by

Peer Community in Genomics

https://doi.org/10.24072/pci.genomics.100011

ABSTRACT

The diversity of mating and sexual systems in angiosperms is spectacular, but the factors driving their evolution remain poorly understood. In plants of the Oleaceae family, an unusual self-incompatibility (SI) system has been discovered recently, whereby only two distinct homomorphic SI specificities segregate stably. To understand the role of this peculiar SI system in preventing or promoting the diversity of sexual phenotypes observed across the family, an essential first step is to characterize the genetic architecture of these two traits. Here, we developed a high-density genetic map of the androdioecious shrub *P. angustifolia* based on a F1 cross between a hermaphrodite and a male parent with distinct SI genotypes. Using a double restriction-site associated digestion (ddRAD) sequencing approach, we obtained reliable genotypes for 196 offspring and their two parents at 10,388 markers. The resulting map comprises 23 linkage groups totaling 1,855.13 cM on the sex-averaged map. We found strong signals of association for the sex and SI phenotypes, that were each associated with a unique set of markers on linkage group 12 and 18 respectively, demonstrating inheritance of these traits as single, independent, mendelian factors. The *P. angustifolia* linkage map shows robust synteny to the olive tree genome overall. Two of the six markers strictly associated with SI in *P. angustifolia* have strong similarity with a recently identified 741kb chromosomal region fully linked to the SI phenotype on chromosome 18 of the olive tree genome, providing strong cross-validation support. The SI locus stands out as being markedly rearranged, while the sex locus has remained relatively more collinear between the two species. This *P. angustifolia* linkage map will be a useful resource to investigate the various ways by which the sex and SI determination systems have co-evolved in the broader phylogenetic context of the Oleaceae family.

Keywords: mating systems, genetic mapping, diallelic self-incompatibility, sex determining region

Cite as: Carré A, Gallina S, Santoni S, Vernet P, Godé C, Castric V, Saumitou-Laprade P (2021) Genetic mapping of sex and self-incompatibility determinants in the androdioecious *plant Phillyrea angustifolia*. bioRxiv, 2021.04.15.439943, ver. 7 peer-reviewed and recommended by Peer community in Genomics. https://doi.org/10.1101/2021.04.15.439943

Posted: 18/07/2021

Recommenders: Tatiana Giraud and Ricardo Rodriguez de la Vega

Reviewers: two anonymous reviewers

Correspondence:

Pierre.Saumitou-Laprade@univ-lille.fr

Introduction

Modes of sexual reproduction are strikingly diverse across angiosperms, both in terms of the proportion of autogamous *vs.* allogamous matings and in terms of the distribution of male and female sexual functions within and among individuals (BARRETT 1998; SAKAI AND WELLER 1999; DIGGLE *et al.* 2011). The conditions under which this diversity could arise under apparently similar ecological conditions and have evolved rapidly -sometimes even within the same family- have been a topic of intense interest in evolutionary biology (BARRETT 1998). The control of self-fertilization and the delicate balance between its costs and benefits is considered to be a central force driving this diversity. Avoidance of self-fertilization is sometimes associated with observable phenotypic variations among reciprocally compatible partners. These variations can be morphological (e.g. distyly) or temporal (e.g. protandry, protogyny in the case of heterodichogamy), but in many cases the flowers show no obvious morphological or phenological variation, and self-fertilization avoidance relies on so-called "homomorphic" self-incompatibility (SI) systems. These systems are defined as the inability of fertile hermaphrodite plants to produce zygotes through self-fertilization (LUNDQVIST 1956; DE NETTANCOURT 1977), and typically rely on the segregation of a finite number of recognition "specificities" whereby matings between individuals expressing cognate specificities are not successful at producing zygotes. At the genetic level, the SI specificities most commonly segregate as a single multi-allelic mendelian locus, the S locus. This locus contains at least two genes, one encoding the male determinant expressed in pollen and the other encoding the female determinant expressed in pistils, with the male specificity sometimes determined by a series of tandemly arranged paralogs (KUBO *et al.* 2015). The male and female determinants are both highly polymorphic and tightly linked, being inherited as a single non-recombining genetic unit. In cases where the molecular mechanisms controlling SI could be studied in detail, they were found to be remarkably diverse, illustrating their independent evolutionary origins across the flowering plants (IWANO AND TAKAYAMA 2012). Beyond the diversity of the molecular functions employed, SI systems can also differ in their genetic architecture. In the Poaceae family for example, two independent loci (named S and Z) control SI (Yang, et al., 2008). In other cases, the alternate allelic specificities can be determined by presence-absence variants rather than nucleotide sequence variants of a given gene, such as *e.g.* in *Primula*

vulgaris, where one of the two reproductive phenotypes is hemizygous rather than heterozygous for the SI locus (LI *et al.* 2016).

In spite of this diversity of molecular mechanisms and genetic architectures, a common feature of SI phenotypes is that they are all expected to evolve under negative frequency-dependent selection, a form of natural selection favoring the long-term maintenance of high levels of allelic diversity (WRIGHT 1939). Accordingly, large numbers of distinct SI alleles are commonly observed to segregate within natural and cultivated SI species (reviewed in CASTRIC AND VEKEMANS 2004). There are notable exceptions to this general rule, however, and in some species only two SI specificities seem to segregate stably. Most often in such diallelic SI systems, the two SI specificities are in perfect association with morphologically distinguishable floral phenotypes. In distylous species, for instance, two floral morphs called "pin" (L-morph) and "thrum" (S-morph) coexist (BARRETT 1992; BARRETT 2019). In each morph, the anthers and stigma are spatially separated within the flowers, but located at corresponding, reciprocal positions between the two morphs. Additional morphological differences exist, with S-morph flowers producing fewer but larger pollen grains than L-morph flowers (DULBERGER 1992). These morphological differences are believed to enhance the selfing avoidance conferred by the SI system but also to increase both male and female fitnesses (BARRETT 1990; BARRETT 2002; KELLER *et al.* 2014), although it is not clear which of SI or floral morphs became established in the first place (CHARLESWORTH AND CHARLESWORTH 1979).

The Oleacea family is another intriguing exception, where a diallelic SI system was recently found to be shared across the entire family (VERNET *et al.* 2016). In this family of trees, the genera *Jasminum* (2*n* = 26), *Fontanesia* (2*n* = 26) and *Forsythia* (2*n*= 28) are all heterostylous and are therefore all expected to possess a heteromorphic diallelic SI system; in *Jasminum fruticans* self- and within-morph fertilization are unsuccessful (DOMMÉE *et al.* 1992). The ancestral heterostyly gave rise to species with hermaphrodite (e.g. *Ligustrum vulgare*, *Olea europaea*), androdioecious (e.g. *P. angustifolia*, *Fraxinus ornus*), polygamous (e.g. *Fraxinus excelsior*) and even dioecious (e.g. *Fraxinus chinensis*) sexual systems, possibly in association with a doubling of the number of chromosomes (2*n*= 46 in the Oleeae tribe) (TAYLOR 1945; WALLANDER AND ALBERT 2000). Evaluation of pollen germination success in controlled *in vitro* crossing experiments (whereby fluorescence microscopy is used to score the growth of pollen tubes reaching the style through the stigma; referred to below as the "stigma test") revealed the existence of a previously unsuspected homomorphic diallelic SI in one of these species, *P.*

angustifolia (SAUMITOU-LAPRADE *et al.* 2010). In this androdioecious species (i.e. in which male and hermaphrodite individuals coexist in the same populations), hermaphrodite individuals form two morphologically indistinguishable groups of SI specificities that are reciprocally compatible but incompatible within groups, whereas males show compatibility with hermaphrodites of both groups (SAUMITOU-LAPRADE *et al.* 2010). This "universal" compatibility of males offsets the reproductive disadvantage they suffer from lack of their female function, such that the existence of the diallelic SI system provides a powerful explanation to the long-standing evolutionary puzzle represented by the maintenance of high frequencies of males in this species (PANNELL AND KORBECKA 2010; SAUMITOU-LAPRADE *et al.* 2010; BILLIARD *et al.* 2015; PANNELL AND VOILLEMOT 2015). Extension of the stigma test developed in *P. angustifolia* to other species of the same tribe including *L. vulgaris* (DE CAUWER *et al.* 2020), *F. ornus* (VERNET *et al.* 2016) and *O. europaea* (SAUMITOU-LAPRADE *et al.* 2017; DUPIN *et al.* 2020), demonstrated that all species exhibited some form of the diallelic SI system, but with no consistent association with floral morphology. Cross-species pollination experiments even showed that pollen from *P. angustifolia* was able to trigger a robust SI response on *O. europaea* and the more distant *F. ornus* and *F. excelsior* stigmas (the reciprocal is also true). This opens the question of whether the homomorphic diallelic SI determinants are orthologs across the Oleeae tribe, even in the face of the variety of sexual polymorphisms present in the different species. More broadly, the link between determinant of the homomorphic diallelic SI in the Oleeae tribe and those of the heteromorphic diallelic SI in the ancestral diploid, largely heterostylous species, remains to be established (BARRETT 2019). Understanding the causes of the long-term maintenance of this SI system and exploring its consequences on the evolution of sexual systems in hermaphrodite, androdioecious, polygamous or dioecious species of the family represents an important goal. The case of *P. angustifolia* is particularly interesting because it is one of the rare instances where separate sexes decoupled from mating types can be studied in a single species (CHARLESWORTH 1978).

A first step toward a better understanding of the role of the diallelic SI system in promoting the sexual diversity in Oleaceae is to characterize and compare the genetic architecture of the SI and sexual phenotypes. At this stage, however, the genomic resources for most of these non-model species remain limited. In this context, the recent sequencing efforts (UNVER *et al.* 2017; JIMÉNEZ-RUIZ *et al.* 2020) and the genetic mapping of the SI locus in a biparental population segregating for SI groups in *Olea europaea* (MARIOTTI *et al.* 2020) represent major

breakthroughs in the search for the SI locus in Oleaceae. They have narrowed down the SI locus to an interval of 5.4cM corresponding to a region of approximately 300kb, but it is currently unknown whether the same region is controlling SI in other species. In *P. angustifolia,* based on a series of genetic analysis of progenies from controlled crosses, Billiard *et al.* (2015) proposed a fairly simple genetic model, where sex and SI are controlled by two independently segregating diallelic loci. Under this model, sex would be determined by the "M" locus at which a dominant *M* allele codes for the male phenotype (*i.e. M is a* female-sterility mutation leading e.g. to arrested development of the stigma) and a recessive *m* allele codes for the hermaphrodite phenotype. The S locus would encode the SI system and comprise a dominant allele *S2* and a recessive allele *S1.* The model thus hypothesizes that hermaphrodites are homozygous *mm* at the sex locus, and fall into two groups of SI specificities, named H_a and H_b carrying the *S1S1* and *S1S2* genotypes at the S locus, respectively (their complete genotypes would thus be *mmS1S1* and *mmS1S2* respectively). The model also hypothesizes three male genotypes (M_a: *mMS1S1*, M_b: *mMS1S2*, and M_c*: mMS2S2*). In addition, BILLIARD *et al.* (2015) experimentally showed that, while males are compatible with all hermaphrodites, the segregation of sexual phenotypes varies according to which group of hermaphrodites they sire: the progeny of H_a hermaphrodites pollinated by males systematically consists of both hermaphrodites and males with a consistent but slight departure from 1:1 ratio, while that of H_b hermaphrodites pollinated by the very same males systematically consists of male individuals only. These segregation patterns suggests a pleiotropic effect of the *M* allele, conferring not only female sterility and universal pollen compatibility, but also a complete male-biased sex-ratio distortion when crossed with one of the two groups of hermaphrodites and a more subtle departure from 1:1 ratio when crossed with the other group of hermaphrodites (BILLIARD *et al.* 2015). The latter departure, however, was observed on small progeny arrays only, and its magnitude thus comes with considerable uncertainty.

In this study, we developed a high-density genetic map for the non-model tree *P. angustifolia* using a ddRAD sequencing approach and used it to address three main questions related to the evolution of its peculiar reproductive system. First, are the SI and sex phenotypes in *P. angustifolia* encoded by just two independent loci, as predicted by the most likely segregation model of BILLIARD *et al.* (2015)? Second, which genomic regions are associated with the SI and sex loci, and what segregation model do the SI and sex-associated

loci follow (i.e. which of the males or hermaphrodites, and which of the two SI phenotypes are homozygous *vs.* heterozygous at either loci, or are these phenotypes under the control of hemizygous genomic regions?). Third, what is the level of synteny between our *P. angustifolia* genetic map and the recently published Olive tree genome (UNVER *et al.* 2017; MARIOTTI *et al.* 2020), both globally and specifically at the SI and sex-associated loci?

Material and Methods

Experimental cross and cartography population

In order to get both the SI group and the sexual phenotype (males vs hermaphrodites) to segregate in a single progeny array, a single maternal and a single paternal plant were chosen among the progenies of the controlled crosses produced by Billard *et al.*(2015). Briefly, a H_a maternal tree (named 01.N-25, with putative genotype mmS1S1) was chosen in the progeny of a (H_a x M_a) cross. It was crossed in March 2012 to a M_b father (named 13.A-06, putative genotype mM S1S2) chosen in the progeny of a (H_a x M_c) cross, following the protocol of Saumitou-Laprade *et al.* (2010). Both trees were maintained at the experimental garden of the "Plateforme des Terrains d'Expérience du LabEx CeMEB," (CEFE, CNRS) in Montpellier, France. F1 seeds were collected in September 2012 and germinated in the greenhouse of the "Plateforme Serre, cultures et terrains expérimentaux," at the University of Lille (France). Seedling paternity was verified with two highly polymorphic microsatellite markers (VASSILIADIS *et al.* 2002), and 1,064 plants with confirmed paternity were installed in May 2013 on the experimental garden of the "Plateforme des Terrains d'Expérience du LabEx CeMEB," (CEFE, CNRS) in Montpellier. Sexual phenotypes were visually determined based on the absence of stigma for 1,021 F1 individuals during their first flowering season in 2016 and 2017 (absence of stigma indicates male individuals). Twenty-one progenies did not flower and 22 died during the test period. The hermaphrodite individuals were assigned to an SI group using the stigma test previously described in Saumitou-Laprade *et al.* (2010; SAUMITOU-LAPRADE *et al.* 2017).

DNA extraction, library preparation and sequencing

In 2015, *i.e.* the year before sexual phenotypes were determined and stigma tests were performed, 204 offspring were randomly selected for genomic library preparation and genotyping. Briefly, DNA from parents and progenies was extracted from 100 mg of frozen young leaves with the Chemagic DNA Plant Kit (Perkin Elmer Chemagen, Baesweller, DE, Part # CMG-194), according to the manufacturer's instructions. The protocol was adapted to the use of the KingFisher Flex™ (Thermo Fisher Scientific, Waltham, MA, USA) automated DNA purification workstation. The extracted DNA was quantified using a Qubit fluorometer (Thermo Fisher Scientific, Illkirch, France). Genome complexity was reduced by double

digestion restriction associated DNA sequencing (ddRAD seq) (PETERSON *et al.* 2012) using two restriction enzymes: *PstI*, a rare-cutting restriction enzyme sensitive to methylation recognizing the motif CTGCA/G, and *MseI*, a common-cutting restriction enzyme (recognizing the motif T/TAA). The libraries were constructed at the INRAE - AGAP facilities (Montpellier, France). Next-generation sequencing was performed in a 150-bp paired-ends-read mode using three lanes on a HiSeq3000 sequencer (Illumina, San Diego, CA, USA) at the Get-Plage core facility (Genotoul platform, INRAE Toulouse, France).

GBS data analysis and linkage mapping

Illumina sequences were quality filtered with the *process_radtags* program of Stacks v2.3 (CATCHEN *et al.* 2011) to remove low-quality base calls and adapter sequences. We followed the Rochette & Catchen protocol (ROCHETTE AND CATCHEN 2017) to obtain a *de novo* catalog of reference loci. Briefly, the reads were assembled and aligned with a minimum stack depth of 3 (–m=3) and at most two nucleotide differences when merging stacks into loci (–M=2). We allowed at most two nucleotide differences between loci when building the catalog (–n=2). Both parental and all offspring FASTQ files were aligned to the *de novo* catalog using Bowtie2 v2.2.6 (LANGMEAD AND SALZBERG 2012), the option 'end-to-end' and 'sensitive' were used for the alignment. At this step, one .bam file was obtained per individual to construct the linkage map with Lep-MAP3 (RASTAS 2017). A custom python script was used to remove SPN markers with reads coverage <5. After this step, the script calls Samtools v1.3.1 and the script *pileupParser2.awk* (limit1=5) to convert *.bam* files to the format used by Lep-MAP3. We used the *ParentCall2* module of Lep-MAP3 to select loci with reliable parental genotypes by considering genotype information on parents and offspring. The *Filtering2* module was then used to remove non-informative and distorted markers (dataTolerance = 0.0000001). The module *SeparateChromosomes2* assigned markers to linkage groups (LGs), after test, where the logarithm of odds score (LodLimit) varied from 10 to 50 in steps of 5 then from 20 to 30 in steps of 1 and the minimum number of SNP markers (sizeLimit) per linkage group from 50 to 500 in steps of 50 for each of the LodLimit. The two parameters, lodLimit = 27 and sizeLimit = 250, were chosen as the best parameters to obtain the 23 linkage groups (as expected in members of the Oleoideae subfamily; (WALLANDER AND ALBERT 2000). A custom python script removed loci with SNPs mapped on two or more different linkage groups. The last module

OrderMarkers2 ordered the markers within each LG. To consider the slight stochastic variation in marker distances between executions, the module was run three times on each linkage group, first separately for the meiosis that took place in each parent (sexAveraged = 0) and then averaged between the two parents (sexAveraged = 1). To produce the most likely final father and mother specific maps and a final sex-averaged maps (DE-KAYNE AND FEULNER 2018), we kept for each map the order of markers that had the highest likelihoods for each linkage group. In the end of some linkage groups, we removed from the final genetic map markers that were clearly outliers i.e. that had orders of magnitude more recombination to any marker than the typical average (Table 1). The original map is provided in Figure S1.

Sex and SI locus identification

To identify the sex-determination system in *P. angustifolia* we considered two possible genetic models. First, a "XY" male heterogametic system, where males are heterozygous or hemizygous (XY) and hermaphodites are homozygous (XX). Second, a "ZW" hermaphrodite heterogametic system, where hermaphodites are heterozygous or hemizygous (ZW) and males are homozygous (ZZ). We applied the same logic to the SI determination system, as segregation patterns (BILLIARD *et al.* 2015) suggested that SI possibly also has a heterogametic determination system, with homozygous H_a and heterozygous H_b. In the same way as for sex, it is therefore possible to test the different models (XY, ZW or hemizygous) to determine which SNPs are linked to the two SI phenotypes.

Based on this approach, we identified sex-linked and SI-linked markers on the genetic map by employing SEX-DETector, a maximum-likelihood inference model initially designed to distinguish autosomal from sex-linked genes based on segregation patterns in a cross (MUYLE *et al.* 2016). Briefly, a new alignment of reads from each individual on the loci used to construct the linkage map was done with bwa (LI AND DURBIN 2009). This new alignment has the advantage of retrieving more SNPs than used by LepMap3, as SNPs considered as non-informative by LepMap3 can still be informative to distinguish among sex- or SI-determination systems by SEX-DETector. The alignment was analyzed using Reads2snp (default tool for SEX-DETector) (TSAGKOGEORGA *et al.* 2012) with option -par 0. We ran Reads2snp without the -aeb (account for allelic expression bias) option to accomodate for the use of genomic rather than RNA-seq data. For each phenotype (H_a vs. H_b and males vs. hermaphrodites), SEX-DETector

was run for both a XY and a ZW model with the following parameters: -detail, -L, -SEM, -thr 0.8, -E 0.05. For each run, SEX-DETector also calculates the probability for X (or Z)-hemizygous segregation in the heterozygous haplotypes. To compensate for the heterogeneity between the number of males (83) and hermaphrodites (113) in our progeny array, each model was tested three times with sub-samples of 83 hermaphrodites obtained by randomly drawing from the 113 individuals. We retained SNPs with a ≥80% probability of following an XY (or ZW) segregation pattern, with a minimum of 50% individuals genotyped and less than 5% of the individuals departing from this model (due to either genotyping error or crossing-over).

Synteny analysis with the olive tree

To study synteny, we used basic local alignment search tool (BLAST) to find regions of local similarity between the *P. angustifolia* ddRADseq loci in the linkage map and the *Olea europea var. sylvestris* genome assembly (UNVER *et al.* 2017). This assembly is composed of 23 main chromosomes and a series of 41,233 unanchored scaffolds for a total of 1,142,316,613 bp. Only loci with a unique hit with at least 85% identity over a minimum of 110 bp were selected for synteny analysis. Synteny relationships were visualized with *circos-0.69-6* (KRZYWINSKI *et al.* 2009). Synteny between linkage groups of *P. angustifolia* and the main 23 *O. europea* chromosomes was established based on the number of markers with a significant BLAST hit. At a finer scale, we also examined synteny with the smaller unanchored scaffolds of the assembly, as the history of rearrangement and allo-tetraploidization is likely to have disrupted synteny.

Results

Phenotyping progenies for sex and SI groups

As expected, our cartography population segregated for sex and SI phenotypes, providing a powerful resource to genetically map these two traits. Among the 1,021 F1 individuals that flowered during the two seasons of phenotyping, we scored 619 hermaphrodites and 402 males, revealing a biased sex ratio in favor of hermaphrodites (khi^2= 46.12, p-value=$1.28x10^{-11}$). Stigma tests were successfully performed on 613 hermaphrodites (6 individuals flowered too late to be included in a stigma test), revealing 316 H_a and 297 H_b, i.e. an equilibrated segregation of the two SI phenotypes (khi^2=1.22, p-value= 0.27). The random subsample of 204 F1 progenies chosen before the first flowering season for ddRAD-seq analysis (see below) followed similar phenotypic proportions. Only 196 of the 204 progenies ended up flowering, revealing 83 males and 113 hermaphrodites, among which 60 belonged to the H_a group and 53 to the H_b group.

Linkage mapping

The two parents and the 196 offspring that had flowered were successfully genotyped using a ddRAD-seq approach. Our stringent filtering procedure identified 11,070 loci composed of 17,096 SNP markers as being informative for Lep-MAP3. By choosing a LOD score of 27, a total of 10,388 loci composed of 15,814 SNPs were assigned to, and arranged within, 23 linkage groups in both sex-averaged and sex-specific maps (Table 1).

The linkage groups of the mother map were on average larger (78.88 cM) than the linkage groups of the father map (73.40 cM) and varied from 22.73 cM to 112.38 cM and from 35 cM to 121.94 cM respectively (Table 1, Figure S1). The total map lengths were 1586.57 cM, 1688.16 cM and 1814.19 cM in the sex-averaged, male and female maps, respectively. The length of the linkage groups varied from 23.90 cM to 110.69 cM in the sex-averaged map, with an average of 683 SNPs markers per linkage group (Table 1).

Table 1. Comparison of the sex-averaged, male and female linkage maps. The values in this table are computed without the outliers SNP markers at the extremity of the linkage groups.

			Sex-averaged map			Paternal map			Maternal map		
Linkage group	Number of SNPs	Number of SNPs (without outliers)	LG length (cM)	SNPs/cM	average intermarker distance	LG Length (cM)	SNPs/cM	average intermarker distance	LG Length (cM)	SNPs/cM	average intermarker distance
1	854	839	75.78	11.07	0.09	79.30	10.58	0.09	92.11	9.11	0.11
2	633	621	23.90	25.98	0.10	35.00	17.74	0.12	22.73	27.32	0.11
3	676	676	74.50	9.07	0.11	61.94	10.91	0.09	85.42	7.91	0.13
4	535	535	68.07	7.86	0.13	71.63	7.47	0.13	69.82	7.66	0.13
5	502	494	56.02	8.82	0.11	50.58	9.77	0.10	67.84	7.28	0.14
6	877	877	96.89	9.05	0.11	90.81	9.66	0.10	103.63	8.46	0.12
7	609	601	64.14	9.37	0.11	68.93	8.72	0.11	64.99	9.25	0.11
8	486	479	62.71	7.64	0.13	91.89	5.21	0.19	119.25	4.02	0.25
9	408	406	63.28	6.42	0.16	56.06	7.24	0.14	71.04	5.72	0.18
10	1365	1361	110.69	12.30	0.08	121.95	11.16	0.09	112.38	12.11	0.08
11	793	783	91.66	8.54	0.12	80.40	9.74	0.10	108.84	7.19	0.14
12	973	969	77.12	12.56	0.08	91.52	10.59	0.09	88.09	11.00	0.09
13	849	848	77.40	10.96	0.09	77.29	10.97	0.09	80.77	10.50	0.10
14	566	565	62.12	9.10	0.11	72.25	7.82	0.13	71.39	7.91	0.13
15	750	747	76.92	9.71	0.10	82.98	9.00	0.11	96.01	7.78	0.13
16	591	589	53.53	11.00	0.09	56.93	10.35	0.10	69.56	8.47	0.12
17	613	613	76.52	8.01	0.13	70.58	8.69	0.12	83.94	7.30	0.14
18	660	659	76.22	8.65	0.12	77.26	8.53	0.12	81.99	8.04	0.12
19	806	806	69.29	11.63	0.09	91.10	8.85	0.11	79.49	10.14	0.10
20	547	531	56.26	9.44	0.11	67.56	7.86	0.13	62.91	8.44	0.12
21	479	476	54.86	8.68	0.12	69.49	6.85	0.15	54.14	8.79	0.11
22	550	544	57.05	9.54	0.11	55.64	9.78	0.10	61.44	8.85	0.11
23	690	684	61.64	11.10	0.09	67.10	10.19	0.10	66.42	10.30	0.10
average	687	682	68.98	10.28	0.11	73.40	9.46	0.11	78.88	9.29	0.12

Sex and SI locus identification

We found evidence that a region on linkage group 18 (LG18) was associated with the SI phenotypes, with Hb hermaphrodites having heterozygous genotype, akin to a XY system. Indeed, when comparing H_a and H_b, among the 38,998 SNPs analyzed by SEX-DETector, 496 had a probability of following an XY pattern ≥0.80. We then applied two stringent filters by

retaining only SNPs that had been genotyped for more than 50% of the offspring (n=211), and for which less than 5% of the offspring departed from the expected genotype under a XY model (n=23). Six of these 23 SNPs, distributed in 4 loci, followed a segregation pattern strictly consistent with a XY model. These four loci are tightly clustered on the linkage map and define a region of 1.230 cM on LG18 (Figure 1) in the sex-averaged map. Relaxing the stringency or our thresholds, this region also contains five loci that strictly follow an XY segregation but with fewer than 50% of offsprings successfully genotyped, as well as six loci with autosomal inheritance, possibly corresponding to polymorphisms accumulated within allelic lineages associated with either of the alternate SI specificities. Using the same filtering scheme, none of the SNPs was found to follow a ZW pattern.

For the comparison of male and hermaphrodites, an average of 44,565 SNPs were analyzed by SEX-DETector across the three subsamples, among which an average of 438 had a probability of following an XY pattern ≥0.80. We applied the same set of stringent filters and retained an average of 171 SNPs having been genotyped for at least 50% of the offspring, among which 41 had less than 5% of the offspring departing from the expected genotype under a XY model and were shared across the three subsets. Thirty-two of these SNPs followed a segregation pattern strictly consistent with a XY model. These 32 markers, corresponding to 8 loci, are distributed along a region of 2.216 cM on linkage group 12 (LG12, Figure 1) in the sex-averaged map. Relaxing the stringency or our thresholds, this region also contains five loci that strictly follow an XY segregation pattern but with fewer than 50% of offspring successfully genotyped, as well as 17 loci consistent with autosomal inheritance, possibly corresponding to polymorphisms accumulated within allelic lineages associated with either of the alternate sex phenotypes. Again, no SNP was found to follow a ZW pattern. This provides evidence that this independent region on LG12 is associated with sex, with a determination system akin to a XY system where males have the heterogametic genotype.

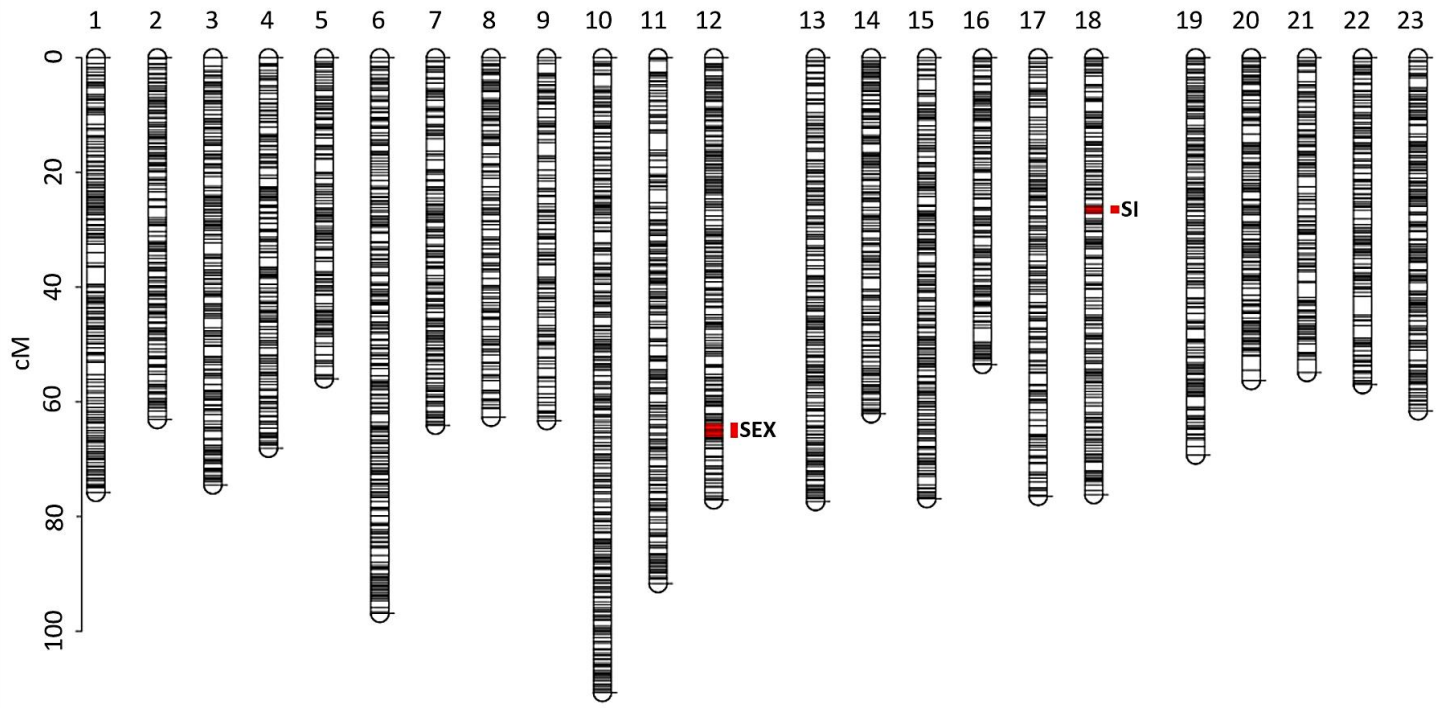


Figure 1. *Phillyrea angustifolia* sex-averaged linkage map showing the grouping and position of 15703 SNPs. The length of each of the 23 linkage groups is indicated by the vertical scale in cM. The markers strictly linked to sex and self-incompatibility (SI) phenotypes are shown in red. Markers that were clearly outliers at the end of some linkage groups were removed (see Table1, Figure S1).

Synteny analysis with the olive tree

About half (49%) of the 10,388 *P. angustifolia* loci used for the genetic map had a significant BLAST hit on the olive tree genome. Overall, the relative position of these hits was highly concordant with the structure of the linkage map. Indeed, the vast majority (79.7%) of loci belonging to a given linkage group had non-ambiguous matches on the same olive tree chromosome. Loci that did not follow this general pattern did not cluster on other chromosomes, suggesting either small rearrangements or mapping/assembly errors at the scale of individual loci. The order of loci within the linkage groups was also well conserved with only limited evidence for rearrangements (Figure 2, Figure 3), suggesting that the two genomes have remained largely collinear.

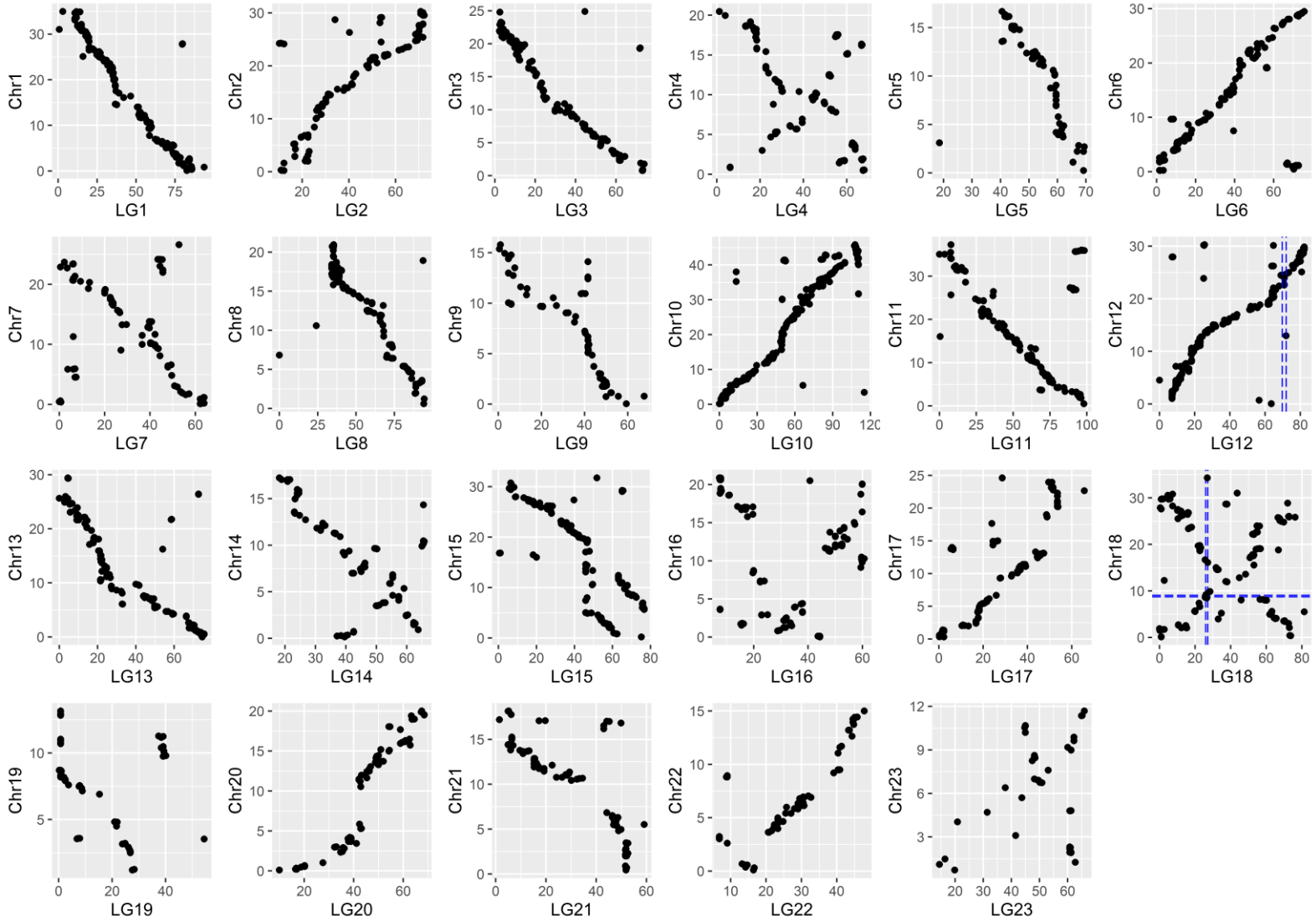


Figure 2. Visualization of chromosome-scale synteny by comparing the location of markers along the *P. angustifolia* linkage groups (LG, scale in cM) with the location of their best BLAST hit along the homologous olive tree chromosome (Chr, scale in Mbp). The vertical lines on LG12 and LG18 indicate the position of markers strictly associated with sex and SI phenotypes in *P. angustifolia*, respectively. The horizontal line on Chr18 indicates the chromosomal region containing the SI locus in *Olea europaea* according to Mariotti *et al.* (2020).

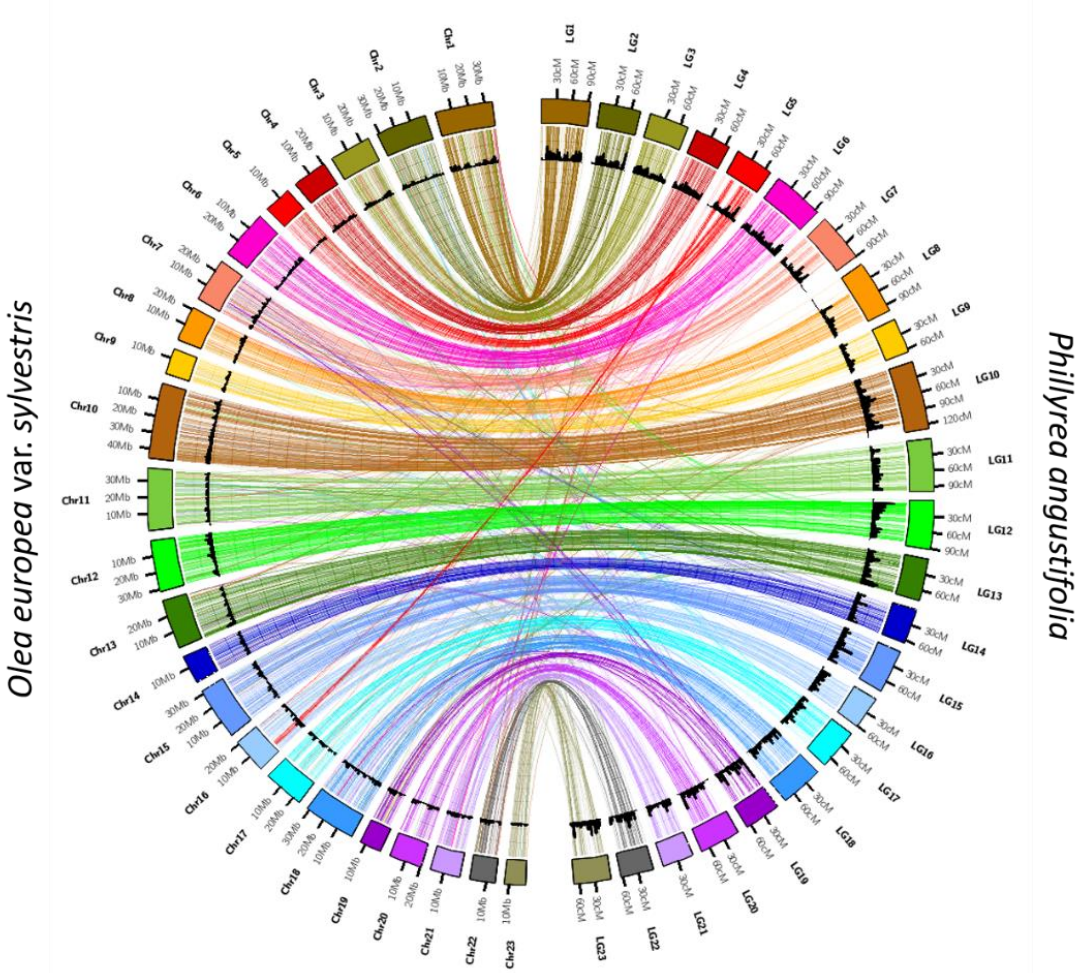


Figure 3. Synteny plot identifying homologous *P. angustifolia* linkage groups (LG, scale in cM) with olive tree chromosomes (Chr, scale in Mb). Lines connect markers in the *P. angustifolia* linkage map with their best BLAST hit in the *O. europea* genome and are colored according to the linkage group. Variation of the density of loci in bins of 3.125cM along linkage groups and 1 Mbp along chromosomes is shown in the inner circle as a black histogram.

We then specifically inspected synteny between the linkage groups carrying either the sex or the SI locus and the olive tree genome (Figure 4). Synteny was good for LG12, the linkage group containing the markers associated with the sex phenotype. Among the 645 loci of LG12, 365 have good sequence similarity in the olive tree genome. Eighty eight percent had their best hits on the same chromosome of the olive tree (chromosome 12 per our numbering of the linkage groups), and the order of markers was largely conserved along this chromosome. Six loci contained in the region associated with sex on LG12 had hits on a single 1,940,009bp region on chromosome 12. This chromosomal interval contains 82 annotated genes in the olive tree genome (Table S1). In addition, eight loci in the sex region had their best hits on a series of five smaller scaffolds (Sca393, Sca1196, Sca1264, Sca32932, Sca969) that could not be reliably anchored in the main olive tree assembly but may nevertheless also contain

candidates for sex determination. Collectively, these scaffolds represent 1.849.345bp of sequence in the olive tree genome and contain 57 annotated genes (Table S1).

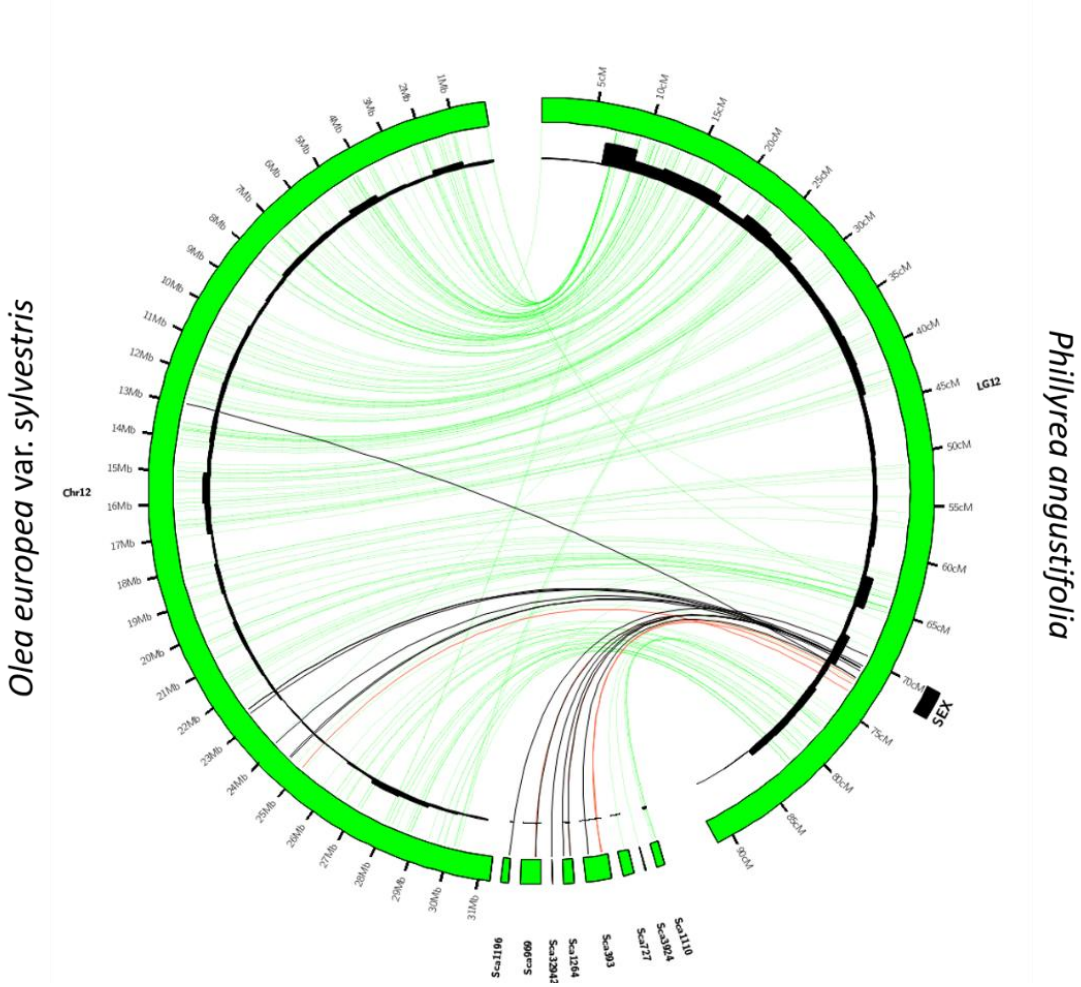


Figure 4. Synteny plot between the *P. angustifolia* linkage group 12 (scale in cM) and the olive tree chromosomes 12 and a series of unanchored scaffolds (scale in Mb). Lines connect markers in the *P. angustifolia* linkage map with their best BLAST hit in the *O. europea* genome. Green lines correspond to markers with autosomal inheritance. Black lines correspond to markers which strictly cosegregate with sex phenotypes (males *vs.* hermaphrodites). Red lines correspond to markers with strong but partial (95%) association with sex. Variation of the density of loci in bins of 3.125cM along linkage groups and 1 Mbp along chromosomes is shown in the inner circle as a black histogram.

Synteny was markedly poorer for markers on LG18, the linkage group containing the markers associated with the SI specificity phenotypes (Figure 5). Of the 440 loci on LG18, 203 had non-ambiguous BLAST hits on the olive tree genome. Although a large proportion (89%) had their best hits on chromosome 18, the order of hits along that chromosome suggested a large number of rearrangements. This more rearranged order was also observed for the six markers that were strictly associated with SI in *P. angustifolia*. Two of them had hits on a single region of 741,403bp on the olive tree genome. This region contains 32 annotated genes (Table S2) and contains two markers that were previously found to be genetically associated with SI

directly in the olive tree by Mariotti *et al.* (2020). Three markers more loosely associated with SI in *P. angustifolia* had hits on a more distant region on chromosome 18 (19,284,909-19,758,630Mb). The three other strongly associated markers all had hits on scaffold 269, which contains 15 annotated genes and represents 545,128bp. Nine other loci strongly or loosely associated with SI had hits on a series of seven other unanchored scaffolds (Sca1199, Sca1200, Sca1287, Sca1579, Sca213, Sca327, Sca502) that collectively represent 96 annotated genes (Table S2) and 2,539,637bp.

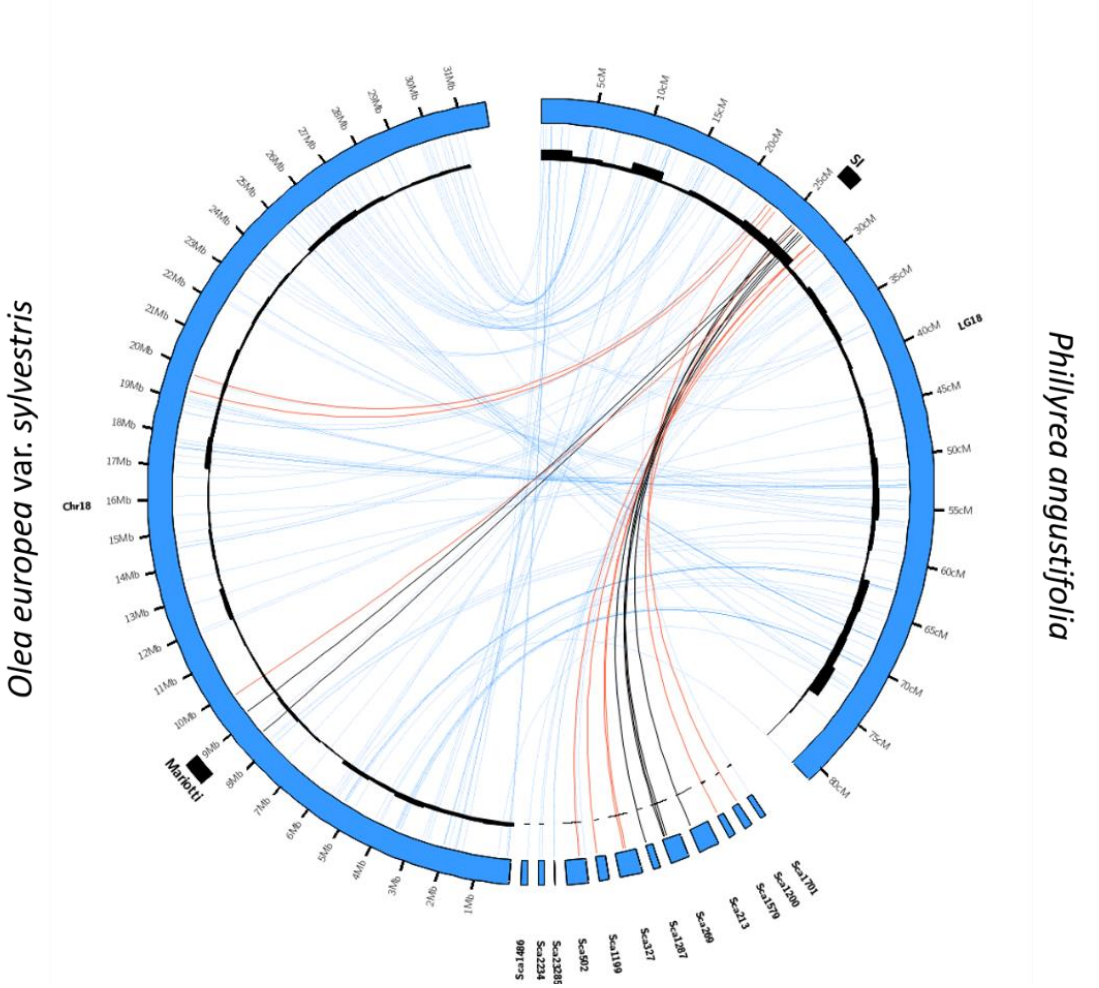


Figure 5. Synteny plot between the *P. angustifolia* linkage group 18 (scale in cM) and the olive tree chromosomes 18 and a series of unanchored scaffolds (scale in Mb). Lines connect markers in the *P. angustifolia* linkage map with their best BLAST hit in the *O. europea* genome. Blue lines correspond to markers with autosomal inheritance. Black lines correspond to markers which strictly cosegregate with SI phenotypes (Ha vs. Hb). Red lines correspond to markers with strong but partial (95%) association with SI. The region found to be genetically associated with SI in the olive tree by Mariotti et al. (2020) is shown by a black rectangle. Variation of the density of loci in bins of 3.125cM along linkage groups and 1 Mbp along chromosomes is shown in the inner circle as a black histogram.

Discussion

Until now, studies have mostly relied on theoretical or limited genetic segregation analyses to investigate the evolution of sexual and SI phenotypes in *P. angustifolia* (VASSILIADIS *et al.* 2002; SAUMITOU-LAPRADE *et al.* 2010; HUSSE *et al.* 2013; BILLIARD *et al.* 2015). In this study, we created the first genetic map of the androdioecious species *P. angustifolia* and identified the genomic regions associated with these two important reproductive phenotypes. The linkage map we obtained shows strong overall synteny with the olive tree genome, and reveals that sex and SI phenotypes segregate independently from one another, and are each strongly associated with a different genomic region (in LG18 and LG12, respectively).

The SI linked markers on LG18 are orthologous with the genomic interval recently identified by Mariotti *et al.* (2020) as the region controlling SI in the domesticated olive tree, providing strong reciprocal support that the determinants of SI are indeed located in this region. Interestingly, we observed a series of shorter scaffolds that could not previously be anchored in the main assembly of the olive tree genome but match genetic markers that are strictly linked to SI in *P. angustifolia.* These unanchored scaffolds provide a more complete set of genomic sequences that will be important to consider in the perspective of identifying the (currently elusive) molecular determinants of SI in these two species. We note that poor assembly of the S-locus region (MARIOTTI *et al.* 2020) was expected given the considerable levels of structural rearrangements typically observed in SI- and more generally in the mating type-determining regions (GOUBET *et al.* 2012; BADOUIN *et al.* 2015), making *P. angustifolia* a useful resource to map the SI locus in the economically important species *O. europeae*.

Our observations also provide direct support to the hypothesis that the determinants of SI have remained at the same genomic position at least since the two lineages diverged, 30 to 40 Myrs ago (BESNARD *et al.* 2009; OLOFSSON *et al.* 2019). Stability of the genomic location of SI genes has been observed in some Brassicaceae species, where the *SRK-SCR* system maps at orthologous positions in the Arabidopsis and Capsella genuses (GUO *et al.* 2011). In other Brassicaceae species, however, the SI system is found at different genomic locations, such as in Brassica and Leavenworthia. In the former, the molecular determinants have remained the same (also a series of *SRK-SCR* pairs, (IWANO *et al.* 2014), but in the latter SI seems to have evolved d*e novo* from exaptation of a pair of paralogous genes (CHANTHA *et al.* 2013; CHANTHA *et al.* 2017). Together with the fact that *P. angustifolia* pollen is able to trigger a robust SI response on *O. europaeae* stigmas (SAUMITOU-LAPRADE *et al.* 2017), our results provide strong

support to the hypothesis that the *P. angustifolia* and *O. europaeae* SI systems are homologous. Whether mating type determinants occupy orthologous genomic regions in different species and rely on the same molecular players has also been discussed in oomycetes by Dussert *et al.* (2020).

Several approaches could now be used to refine the mapping of SI in *P. angustifolia,* and ultimately zero in on its molecular determinants. One possibility would require fine-mapping using larger offspring arrays, starting from our cross for which only a fraction of all phenotyped individuals were genotyped. Beyond the analysis of this controlled cross, evaluating whether the association of the SI phenotype still holds for markers within a larger set of accessions from diverse natural populations will constitute a powerful fine-mapping approach. Since the SI phenotypes seem to be functionally homologous across the Oleeae tribe (VERNET *et al.* 2016), the approach could, in principle, be extended to more distant SI species of the family like *L. vulgare* or *F. ornus*. Identification of sequences that have remained linked over these considerable time scales would represent excellent corroborative evidence to validate putative SI candidates. In parallel, an RNA-sequencing approach could be used to identify transcripts specific to the alternate SI phenotypes.

While comparison to the closely related *O. europeae* genome is a useful approach for the mapping of SI in *P. angustifolia*, it is *a priori* of limited use for mapping the sex-determining region, since the olive tree lineage has been entirely hermaphroditic for at least 32.22 Myrs (confidence interval: 28-36 Myrs) (FigS1 in OLOFSSON *et al.* 2019). Detailed exploration of the genomic region in the olive tree that is orthologous to the markers associated with sexual morphs in *P. angustifolia* is however interesting, as it may either have anciently played a role in sex determination and subsequently lost it, or alternatively it may contain quiescent sex-determining genes that have been activated specifically in *P. angustifolia*. At a broader scale, mapping and eventually characterizing the sex locus in other androdioecious species such as *F. ornus* could indicate whether the different instances of androdioecy in the family represent homologous phenotypes or independent evolutionary emergences.

Identifying the molecular mechanisms of the genes controlling SI and sex and tracing their evolution in a phylogenetic context would prove extremely useful. First, it could help understand the strong functional pleiotropy between sex and SI phenotypes, whereby males express universal SI compatibility (SAUMITOU-LAPRADE *et al.* 2010). In other words, males are able to transmit the SI specificities they inherited from their parents, but they do not express them themselves even though their pollen is fully functional. This intriguing feature of the SI

system was key to solve the puzzle of why *P. angustifolia* maintains unusually high frequencies of males in natural populations (HUSSE *et al.* 2013), but the question of how being a male prevents expression of the SI phenotype in pollen is still open. A possibility is that the *M* allele of the sex locus contains a gene interacting negatively either with the pollen SI determinant itself or with a gene of the downstream response cascade. Identifying the molecular basis of this epistasis will be an interesting next step. Second, another intriguing feature of the system is segregation distortion, which is observed at several levels. BILLIARD *et al.* (2015) observed complete segregation bias in favor of males among the offspring of H_b hermaphrodites sired by males. Here, by phenotyping >1,000 offspring of a H_a hermaphrodite sired by a M_b male, we confirmed that this cross also entails a departure from Mendelian segregation, this time in favor of hermaphrodites, albeit of a lesser magnitude. Although the generality of this observation still remains to be determined by careful examination of the other possible crosses (H_a hermaphrodites x M_a and M_c males), it is clear that segregation distortion is a general feature of this system, as was already observed in other sex determination systems causing departures from equal sex ratios (e.g. KOZIELSKA *et al.* 2010). Beyond identification of the mechanisms by which the distortions arise, pinpointing the evolutionary conditions leading to their emergence will be key to understanding the role they may have played in the evolution of this reproductive system.

More broadly, while sex and mating types are confounded in many species across the tree of life and cannot be distinguished, the question of when and how sex and mating types evolve separately raises several questions. The evolution of anisogamy (and hence, sexual differentiation) has been linked to that of mating types (CHARLESWORTH 1978). In volvocine algae for instance, the mating-type locus in isogamous species is orthologous to the pair of U/V sex chromosomes in anisogamous/oogamous species, suggesting that the sex-determination system derives from the mating-type determination system (GENG *et al.* 2014). From this perspective the Oleaceae family is an interesting model system, where a SI system is ancestral, and in which some species have evolved sexual specialization that is aligned with the two SI phenotypes (e.g. in the polygamous *F. excelsior* males belong to the H_a SI group and can only mate with hermaphrodites or females of the H_b group, and the sexual system of *F. excelsior* can be viewed as subdioecy (SAUMITOU-LAPRADE *et al.* 2018). In other species, sexual phenotypes are disjoint from SI specificities and led to the differentiation of males and hermaphrodites. For instance, in the androdiecious *P. angustifolia* and probably *F. ornus*, the male determinant is genetically independent from the SI locus but fully linked to a genetic

determinant causing the epistatic effect over SI (BILLIARD *et al.* 2015; VERNET *et al.* 2016). Yet other species have remained perfect hermaphrodites and have no trace of sexual differentiation whatsoever (*O. europeae*). Understanding why some species have followed one evolutionary trajectory while others have followed another will be an exciting avenue for future research (BILLIARD *et al.* 2011).

Data accessibility

Fastq files for all 204 offspring and both parents are deposited in the NCBI BioProject PRJNA724813.

Supplementary material

All scripts used can be accessed at https://github.com/Amelie-Carre/Genetic-map-of-Phillyrea-angustifolia.

Acknowledgements

We thank Sylvain Bertrand and Fantin Carpentier for technical help for the phenotyping and Jos Käfer for scientific discussion and help in applying Sex-DETector to our material. We thank Jacques Lepart†, Mathilde Dufay, Pierre Olivier Cheptou, Xavier Vekemans, Sylvain Billiard and Bénédicte Felter for scientific discussions. Field and laboratory work for phenotyping were done at the Platform Terrains d'Expériences (Labex CeMEB ANR-10-LABX-0004-CeMEB) of the Centre d'Ecologie Fonctionnelle et Evolutive (CEFE, CNRS) with the help of Thierry Mathieu and David Degueldre and at the platform Serres cultures et terrains expérimentaux of the Lille University with the help of Nathalie Faure and Angélique Bourceaux. We are grateful to Marie-Pierre Dubois for providing access to the microscopy facilities at the SMGE (Service des Marqueurs Génétiques en Ecologie) platform (CEFE). This work was funded by the French National Research Agency through the project 'TRANS' (ANR-11-BSV7- 013-03) and by a grant from the European Research Council (NOVEL project, grant #648321). A.C was supported by a doctoral grant from the French ministry of research. The authors also thank the Région Hauts-de-France, and the Ministère de l'Enseignement Supérieur et de la Recherche (CPER Climibio), and the European Fund for Regional Economic Development for their financial support. We also thank the HPC Computing Mésocentre of the University of Lille which provided us with the computing grid. We thank Tatiana Giraud, Ricardo C. Rodríguez de la Vega and two anonymous referees for constructive reviews of an

earlier version. Version 7 of this preprint has been peer-reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100011).

Conflict of interest disclosure and Author Contributions

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. All authors contributed to the study presented in this paper. PS-L and PV developed, designed and oversaw the study; they coordinated the cross and carried out the phenotyping and stigma tests. CG performed the seedling paternity analysis. SS performed DNA extraction, library preparation and organized sequencing. AC and SG constructed the data analysis pipeline and AC, SS, PS-L and VC interpreted the results and wrote the manuscript.

Mise en évidence des différences transcriptomiques et génomiques entre mâles et hermaphrodites : une recherche des gènes candidats à l'origine de l'androdioécie chez *Phillyrea angustifolia*.

Amélie Carré[1], Sophie Gallina[1], Sylvain Santoni[2], Philippe Vernet[1], Cécile Godé[1], Clément Mazoyer[1], Vincent Castric[1], Pierre Saumitou-Laprade[1]

[1] CNRS, Univ. Lille, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France

[2] UMR DIAPC - Diversité et adaptation des plantes cultivées

Résumé

La grande variété des systèmes sexuels chez les angiospermes est spectaculaire, mais comprendre leur déterminisme génétique et des facteurs contrôlant leur évolution reste un enjeu majeur en biologie évolutive et en génomique. Dans la famille des Oleacées, l'espèce *Phillyrea angustifolia* représente un organisme modèle passionnant. Il s'agit d'une espèce androdioïque, c'est-à-dire que des individus mâles et hermaphrodites coexistent dans les populations de cet arbuste. L'évolution et le maintien de l'androdioécie sans contrepartie évidente sur la fonction mâle des hermaphrodites représente un véritable paradoxe dans un contexte où le maintien de systèmes purement hermaphrodite, dioïque ou gynodioïque sont plus courants ou faciles à expliquer. Dans ce chapitre, nous avons développé deux approches afin d'étudier les différences transcriptomiques et génomiques qui existent entre les individus mâles et les individus hermaphrodites chez le filaire. La première approche par séquençage de transcrits sur des boutons floraux nous a permis de mettre en évidence que 0,15% des transcrits avaient une expression sexe-biaisée en faveur mâle et 0,45% des transcrits présentaient une sur-expression chez les hermaphrodites. En couplant ces résultats à ceux obtenus lors de la cartographie génétique haute densité, nous avons établi une liste de séquences que nous avons capturé par hybridation ciblée chez 95 individus de filaire. L'analyse de la ségrégation des SNPs dans ces séquences nous a permis de dresser une liste non exhaustive de 26 séquences candidates pour la détermination du sexe chez *P. angustifolia*. Ces résultats constituent une base importante pour les futures études sur la détermination génétique et l'identification des mécanismes moléculaires impliqués dans l'établissement de l'androdioécie chez le filaire.

Introduction

Environ 10% des Angiospermes ont des fleurs unisexuées, et cette caractéristique est associée à un large éventail de stratégies sexuelles qui impliquent diverses combinaisons de fleurs femelles, mâles et hermaphrodites au niveau de la plante et de la population (BARRETT 2002). Deux modèles distincts de variation entre les sexes existent. Dans le premier modèle, les populations présentent un « monomorphisme de genre » c'est-à-dire que tous les individus possèdent les fonctions mâle et femelle (*ie.* hermaphrodisme, andromonoécie, monoécie). Dans le second modèle, les populations présentent un « dimorphisme de genre », c'est-à-dire qu'une bimodalité de genre est observable (LLOYD 1980). Ces populations sont composées de deux morphes sexuels distincts qui peuvent être strictement unisexués, auquel cas le système sexuel est connu sous le nom de dioécie, tandis que dans d'autres cas un morphe est hermaphrodite et l'autre est femelle (gynodioécie) ou mâle (androdioécie) (BARRETT 2002).

Plusieurs voies sont communément admises dans l'évolution de l'hermaphrodisme vers la dioécie, impliquant la transition du monomorphisme de genre au dimorphisme *via* des mutations de stérilité qui donnent naissance à des fleurs unisexuées (BARRETT 2002). Dans la voie de la gynodioécie, une population initialement hermaphrodite est d'abord partiellement envahie par des mutations de stérilité mâle. Ces mutations sont favorisées lorsqu'elles sont codées par des éléments cytoplasmiques car ceux-ci sont généralement transmis de façon exclusive par la voie femelle et ne sont pas affectées par la perte de la voie mâle (LEWIS 1941; LLOYD 1974). Le conflit qui résulte des différences de mode de transmission entre les génomes cytoplasmique (à transmission maternelle) et nucléaire (à transmission biparentale) peut aboutir au maintien de fréquences élevées de femelles dans les populations gynodioïques, ce qui facilite l'étape suivante de transition vers la dioécie. Cette dernière étape peut se faire soit par l'invasion de mâles, soit par une sélection pour augmenter la fonction mâle chez les hermaphrodites (KÄFER *et al.* 2017). La deuxième voie implique le passage à la dioécie *via* l'androdioécie. Dans ce scénario, dont les protagonistes sont nucléaires, les premiers modèles théoriques étaient univoques sur le fait que le « morphe » mâle ne peut se maintenir dans les populations que si ces individus compensent la perte de leur fonction femelle par une augmentation substantielle de leur capacité à se reproduire par la fonction mâle (LLOYD 1975 ; CHARLESWORTH AND CHARLESWORTH 1978). De fait, le nombre d'espèces androdioïques est très faible à l'échelle des angiospermes, et est encore plus limité si on considère le fait que la fonction mâle des individus hermaphrodites dans de nombreuses espèces

“morphologiquement” androdioïques est si réduite que ces espèces se comportent “fonctionnellement” comme des espèces dioïques. On parle dans ce cas de dioécie cryptique (MAYER AND CHARLESWORTH 1991). Plusieurs cas d’espèces fonctionnellement androdioïques existent cependant bien. Chez *Datisca glomerata* (Datiscaceae), des études génétiques ont mis en évidence qu’un seul locus nucléaire dominant chez les mâles contrôle le phénotype sexuel (WOLF *et al.* 2001). Il en est de même chez *Mercurialis annua* (Euphorbiaceae), une autre espèce androdioïque intensivement étudiée, où un seul locus semble régir l'expression sexuelle (PANNELL 1997a). Bien que ces deux exemples correspondent à des cas d’androdioécie réellement fonctionnelle (les hermaphrodites ont bien une reproduction effective par leur voie mâle), les données phylogénétiques montrent qu’elles résultent en fait d’une transition inverse, de la dioécie vers l’androdioécie (PANNELL 1997a ; ZHANG *et al.* 2006 ; DELPH 2009; KÄFER *et al.* 2017), et non de l’hermaphrodisme vers la dioecie, comme classiquement considéré. Enfin, il existerait une troisième voie permettant l’évolution de la dioécie, cette fois à partir de la distylie. La distylie correspond à la coexistence au sein d’une espèce de deux morphes sexuels distincts, tous deux hermaphrodites. L’hétéromorphie la plus commune chez les espèces distyles correspond à des variations des longueurs relatives des anthères et des styles, et est généralement associée à un système d’auto-incompatibilité permettant uniquement les fécondations entre morphes. Cette voie a été documentée chez trois familles : les Boraginaceae, les Menyanthaceae et les Rubiaceae, chez lesquelles la transition impliquerait une spécialisation croissante du genre des morphes à style long et à style court, les convertissant dans la plupart des cas en plantes femelles et mâles, respectivement (PAILLER *et al.* 1998). L'unisexualité peut donc évoluer par diverses voies (BARRETT 2002), ce que confirme la grande diversité des déterminants du genre lorsqu’ils ont pu être étudiés (LEBEL-HARDENACK AND GRANT 1997 ; WOLF *et al.* 2001), mais la fréquence de ces différentes voies et les processus sous l’effet desquels elles peuvent être empruntées restent à ce jour mal appréciés.

Chez les Angiospermes, la dioécie peut être associée à l’existence de chromosomes sexuels, qui ont été documentés chez un nombre croissant d'espèces végétales (DELLAPORTA AND CALDERON-URREA 1993 ; CAREY *et al.* 2021). Pour les espèces dioïques qui expriment le sexe gamétique au stade diploïde, comme chez les plantes à graines, les chromosomes sexuels sont appelés XY ou ZW selon le sexe hétérogamétique (mâle ou femelle respectivement) (CAREY *et al.* 2021). Chez l’asperge (*Asparagus)*, le sexe est déterminé par des chromosomes sexuels homomorphes dans lesquels les mâles (XY) sont du sexe hétérogamétique (BRACALE *et al.* 1991). L'asperge serait "mâle dominant" et le chromosome sexuel contiendrait des

déterminants génétiques mâle-activateur-femelle-suppresseur similaires à ceux postulés pour *Silene dioica* (DELLAPORTA AND CALDERON-URREA 1993). En plus de ces gènes majeurs de détermination du sexe, des modificateurs génétiques influenceraient la dégénérescence du style (BRACALE *et al.* 1991). Ces systèmes de détermination peuvent être labiles. Par exemple, dans le genre *Silene,* la dioécie et les chromosomes sexuels ont évolué plusieurs fois : certaines espèces ont des chromosomes sexuels de type XY tandis que d'autres sont de types ZW (revue dans MING *et al.* 2011).

L'hypothèse génétique majoritaire chez les espèces hétérostyles est que la distylie serait gouvernée par un seul locus diallélique (S/s), pour lequel le morphe à styles courts (morphe S) serait hétérozygote (Ss) et le morphe à styles longs (morphe L) serait homozygote (ss) (LEWIS AND JONES 1992). L'analyse génétique sur *Primula*, a initialement conduit à un modèle dit de « supergène » du locus S selon lequel une région chromosomique composée de locus distincts contrôlerait différentes composantes du morphe, dont l'association génétique serait maintenue par une liaison physique étroite et un ensemble de réarrangements chromosomiques (BARRETT AND SHORE 2008 ; CHARLESWORTH 2015b; BRENNAN 2017). Des approches de séquençage de génome et de transcriptome, de protéomique, de mutagenèse et de cartographie génétique, ont remis en cause ce modèle initial et permis de mieux appréhender l'architecture génétique du locus S chez *Primula*. L'haplotype S dominant du morphe S comprend une région hémizygote de plusieurs gènes absents de l'haplotype s du morphe L. (USHIJIMA *et al.* 2012; YASUI *et al.* 2012; KAPPEL *et al.* 2017; COCKER *et al.* 2018; SHORE *et al.* 2019). Plutôt qu'un supergène diallélique avec des allèles dominants et récessifs, il y aurait donc un groupe de liaison hémizygote au locus S composé de gènes qui contrôlent la distylie. Il y aurait donc une liaison étroite entre les gènes contrôlant le développement des morphes floraux et les gènes contrôlant la compatibilité génétique entre morphes entre les individus chez les espèces hétérostyles (ces derniers n'ont, pour l'instant, pas été identifiés). Un modèle chromosomique a été proposé par Kappel et al. (2017), selon lequel l'évolution du supergène au locus S serait dû à une duplication segmentaire et un réarrangement impliquant la perte de certains gènes et la néo-fonctionnalisation d'autres gènes (BARRETT 2019). Dans le cas d'une évolution de la dioécie à partir de la distylie, il serait donc envisageable que la région génétique associée conserve un fonctionnement hémizygote.

Enfin, un dernier mode de détermination du sexe est observé dans le genre *Rumex*, sous-genre *Acefosa*, dans lequel c'est le rapport X/autosomes qui semble essentiel (PARKER AND CLARK 1991). Les femelles sont XX (2n=14) et les mâles XY_1Y_2 (2n = 15, Y_1 et Y_2 représentent

deux chromosomes Y non-homologues) ; cependant, les plantes diploïdes avec les génotypes XXY et XXY_1Y_2 sont des femelles fertiles. Chez les polyploïdes, un rapport X/autosomes supérieur à 1 conduit au développement de femelles, tandis qu'un rapport X/autosomes de 0,5 ou moins conduit au développement de mâles. Les chromosomes Y de *Rumex* sont nécessaires à la fertilité du pollen mais pas au développement des étamines, et contrairement à *S. dioica* ils n'inhibent pas le développement du gynécée (DELLAPORTA AND CALDERON-URREA 1993). Ce fonctionnement rappelle celui observé chez certaines espèces de *Drosophila* et *Caenorhabditis*, où le principal déterminant du sexe est le rapport X/autosome (HODGKIN 1990).

Dans ce contexte de grande diversité des modalités génétiques du déterminisme sexuel, la famille des Oleacées apparaît comme un modèle particulièrement intéressant et complexe. Dans cette famille, il existe une hétérostylie ancestrale (TAYLOR 1945 ; WALLANDER AND ALBERT 2000) qui est associée à une auto-incompatibilité (DOMMEE *et al.* 1992). Cette hétérostylie ancestrale a donné naissance à des espèces possédant des systèmes sexuels très divers (WALLANDER 2001), en association forte avec un doublement du nombre de chromosomes (2n=46 dans la tribu des Oléées) (TAYLOR 1945; WALLANDER AND ALBERT 2000). Ainsi, on retrouve des espèces hermaphrodites (ex. *Ligustrum vulgare*, *Olea europaea*), androdioïques (ex. *P. angustifolia*, *Fraxinus ornus*), polygames (ex. *Fraxinus excelsior*) et même dioïques (ex. *Fraxinus chinensis*). Une transition des sexes à partir de la distylie comme documentée chez les Rubiaceae par Pailler et *al.* (1998) est ici légitimement envisageable. Au sein de cette famille, *P. angustifolia* représente un organisme modèle particulièrement attractif car cette espèce présente deux systèmes de compatibilité sexuelle distincts et rares (SAUMITOU-LAPRADE *et al.* 2010). Il s'agit tout d'abord d'une espèce androdioïque où les mâles et les hermaphrodites coexistent dans les populations, mais dont le fonctionnement est *a priori* différent d'autres espèces androdioïques. Par exemple, chez *Datisca glomerata* (PHILBRICK AND RIESEBERG 1994) et *Mercurialis annua* (PANNELL 1997b ; PANNELL *et al.* 2014), il y a une compensation très nette dans la voie mâle de la perte de la fonction femelle avec une réallocation des ressources qui se traduit par une augmentation (4 à 10 fois plus) de la quantité de pollen produite chez les mâles. Chez le filaire à l'inverse, des études détaillées ont montré que la quantité de pollen produite par les mâles n'est que marginalement supérieure à celle des hermaphrodites (VASSILIADIS *et al.* 2000; VASSILIADIS *et al.* 2002), et bien inférieure à la quantité qui serait théoriquement nécessaire, selon les modèles classiques, pour rendre compte des fréquences de mâles observées dans les populations (50% et plus) . Il s'agit ensuite

d'une espèce au sein de laquelle , le maintien des mâles dans les populations de filaire est expliqué par la présence d'un système d'auto-incompatibilité diallélique homomorphe (SAUMITOU-LAPRADE *et al.* 2010) indépendant du sexe mais jouant un rôle central dans les modalités d'appariement (Billiard et al 2015). En effet, chez cette espèce les mâles sont compatibles avec l'ensemble des hermaphrodites et leur pollen bénéficie donc d'un avantage de compatibilité "universelle" à l'échelle de l'espèce (SAUMITOU-LAPRADE *et al.* 2010). Dans le contexte d'un système d'auto-incompatibilité à seulement deux allèles, cette compatibilité universelle permet aux mâles de féconder l'ensemble des hermaphrodites, et de compenser ainsi complètement la perte de leur fonction femelle.

Chez le filaire, le modèle génétique le plus probable implique l'existence de deux locus bialléliques indépendants, l'un codant pour l'auto-incompatibilité et l'autre pour le sexe (Billard *et al.* (2015)), et a été validé dans le premier chapitre de cette thèse (CARRE *et al.* 2021). Le locus du sexe, localisé sur le LG12, suit une ségrégation génétique de type XY où les allèles sont appelés m et M, et où les mâles sont hétérozygotes (mM) tandis que les hermaphrodites sont homozygotes (mm) pour un ensemble de marqueurs génétiques. Cependant, en raison des filtres de couverture de séquençage appliqués, un fonctionnement partiel en hémizygotie aurait pu ne pas être détecté dans cette analyse, et il n'est pas possible d'exclure totalement ce cas de figure (CARRE *et al.* 2021). De plus, la cartographie a été réalisée sur la base de la ségrégation de courtes séquences en nombre relativement important (marqueurs GBS) mais représentant tout de même un sous-échantillonnage très incomplet du génome entier et des régions génomiques d'intérêt. Il est donc indispensable de densifier les marqueurs étudiés, en se focalisant en particulier sur les régions codantes présentes pour espérer restreindre la région d'intérêt et identifier de possibles gènes candidats. Le locus codant pour l'auto-incompatibilité est quant à lui localisé sur le LG18. Il suit lui aussi une ségrégation génétique de type XY selon laquelle les hermaphrodites Hb sont hétérozygotes S1S2 (ou hémizygotes) et les hermaphrodites Ha sont homozygotes S1S1 (CARRE *et al.* 2021). Les trois génotypes au locus S possibles pour les individus mâles sont S1S1 (Ma), S1S2 (Mb) et S2S2 (Mc). L'allèle M du locus du sexe serait un « supergène » à effets pléiotropes. Cet allèle conférerait tout d'abord la stérilité femelle, mais serait également associé à deux propriétés particulières des mâles: leur compatibilité avec tous les hermaphrodites indépendamment de leur génotype au locus S ainsi que la distorsion de ségrégation dont la pénétrance est conditionnelle au génotype de l'hermaphrodite (biais léger de 60% en faveur des descendants hermaphrodites lors des croisements sur Ha, mais biais très fort de 100% en faveur des descendants mâles lors des

croisements sur Hb) (BILLIARD *et al.* 2015 ; CARRE *et al.* 2021). L'indépendance du sexe et du système d'incompatibilité nous permet d'étudier ces deux traits séparément avec la même démarche expérimentale de départ.

Ce deuxième chapitre est consacré à l'étude du phénotype sexuel (mâles *vs.* hermaphrodites), l'étude du phénotype d'auto-incompatibilité faisant l'objet du troisième chapitre. Il comporte deux étapes. Dans un premier temps nous avons mis en œuvre une analyse transcriptomique comparant un ensemble d'individus mâles à un ensemble d'individus hermaphrodites de *P. angustifolia* dans le but d'identifier les déterminants moléculaires responsables et/ou participant à la détermination du phénotype sexuel. En l'absence d'un génome de référence pour cette espèce, nous profitons de la relativement bonne synténie globale avec le génome de l'olivier que nous avons mise en évidence dans le premier chapitre (23 groupes de liaison correspondant aux 23 chromosomes), pour vérifier si les gènes différentiellement exprimés co-localisent avec la position du locus du sexe établie grâce à la cartographie de *Phillyrea* (CARRE *et al.* 2021). Cependant, l'olivier étant une espèce sans sexes séparés, il est vraisemblable que son génome ne contienne pas les éléments génétiques permettant l'encodage du sexe tel qu'il se manifeste chez le filaire. Cette première approche par synténie ne permet donc pas de positionner avec certitude ces gènes sur la carte génétique du filaire; par conséquent elle n'est pas suffisante pour établir de façon définitive leur association génétique avec le phénotype sexuel. Pour nous affranchir de ces limitations, nous avons mis en œuvre une approche de capture de séquences génomiques par hybridation ciblée afin de densifier fortement la quantité d'information génétique sur les régions génomiques d'intérêt et incluant l'ensemble les transcrits issus de l'analyse différentielle. Cette approche nous permet d'examiner, directement chez *P. angustifolia*, la ségrégation des SNPs contenus dans l'ensemble de ces séquences. Cette approche nous a permis de valider ou d'exclure de façon efficace la liaison génétique des séquences ciblées avec le phénotype sexuel et de raffiner la liste des candidats potentiels.

Matériel et méthodes

1. Analyse transcriptomique

Matériel biologique

Le matériel biologique utilisé dans cette étude a pour origine deux populations (génération 0) constituées chacune d'une centaine d'individus de *P. angustifolia* et situées dans le sud de la France à proximité du littoral méditerranéen (LEPART AND DOMMEE 1992). Il s'agit d'une part d'une friche agricole abandonnée en 1970 située à la Gardiole (Fabrègues ; 43°28' 36''N ; 3°45'37''E) à proximité de Montpellier et d'autre part d'une prairie située à la Tour du Valat (43°29'52''N ; 4°40'55''E) en Camargue. A partir de graines récoltées après fécondation libre sur onze hermaphrodites localisés dans la population naturelle de Fabrègue, une population artificielle (génération 1) de 242 individus a été créée sur le terrain expérimental du CEFE-CNRS (Montpellier). Dans neuf descendances, huit mères et neuf pères ont été sélectionnés pour réaliser des croisements contrôlés et générer 15 descendances (génération 2) dont l'analyse a permis de produire le modèle génétique explicatif de la stérilité femelle et de l'auto-incompatibilité chez *P. angustifolia* (BILLIARD *et al.* 2015). Dans deux descendances de la génération 2, un hermaphrodite (01N25) et un mâle (104A-06) supposés être respectivement homozygote et hétérozygote aux locus du sexe et de l'auto-incompatibilité, ont été croisés afin de produire une descendance (génération 3) utilisée pour l'approche de cartographie génétique à haute densité du premier chapitre de cette thèse (CARRE *et al.* 2021). Trois ensembles de phénotypes (Ha, Hb et M) ont été sélectionnés pour former les trois groupes de l'analyse transcriptomique. Pour constituer ces groupes, 12 individus Ha, 13 Hb et 12 mâles ont été choisis parmi les individus de la génération 2. Des boutons floraux correspondant à plusieurs stades de maturation (3 fleurs ouvertes juste avant déhiscence des anthères et 4 boutons fermés) ont été prélevés en mars pour chaque individu.

Les échantillons ont été broyés dans de l'azote liquide et l'ARN cellulaire total a été extrait à l'aide du kit *Spectrum Plant Total RNA* (Sigma, Inc., USA) avec un traitement à la DNAse. La concentration en ARN a d'abord été mesurée à l'aide d'un spectrophotomètre NanoDrop ND-1000 puis sur un spectrofluorimètre Tecan Genius avec le protocole *Quant-iT™ RiboGreen®* (Invitrogen, USA). La qualité de l'ARN a été évaluée en analysant 1 µL de chaque échantillon d'ARN à l'aide d'une puce RNA 6000 Pico sur un Bioanalyzer 2100 (Agilent Technologies, Inc., USA). Les échantillons présentant un « RNA Integrity Number » (RIN) supérieur à huit ont été jugés acceptables selon le protocole *Illumina TruSeq RNA*.

Collecte des données RNAseq

Le kit *TruSeq RNA sample Preparation v2* (Illumina Inc., USA) a été utilisé selon le protocole du fabricant avec les modifications suivantes. Les molécules d'ARNm contenant des poly-A ont été purifiées à partir de 1 ug d'ARN total en utilisant des billes magnétiques attachées à un oligo poly-T. L'ARNm purifié a été fragmenté par ajout du tampon de fragmentation et a été chauffé à 94°C dans un thermocycleur pendant 4 min. Le temps de fragmentation de 4 min a été utilisé pour produire des fragments de 250-500 pb. L'ADNc du premier brin a été synthétisé en utilisant des amorces aléatoires pour éliminer le biais général vers l'extrémité 3' du transcrit. La synthèse de l'ADNc du deuxième brin, la réparation des extrémités, *A-tailing* et la ligation de l'adaptateur ont été effectuées conformément aux protocoles fournis par le fabricant. Les matrices d'ADNc purifiées ont été enrichies par 15 cycles de PCR pendant 10 s à 98°C, 30 s à 65°C et 30 s à 72°C en utilisant les amorces PE1.0 et PE2.0 et avec l'ADN polymérase Phusion (NEB, USA). Chaque banque d'ADNc indexée a été vérifiée et quantifiée à l'aide d'une puce à ADN 100 sur un Bioanalyseur 2100 puis également mélangée par dix (à partir d'échantillons différents). La bibliothèque finale a ensuite été quantifiée par PCR en temps réel à l'aide le kit *KAPA Library Quantification Kit* for Illumina Sequencing Platforms (Kapa Biosystems Ltd, SA), ajustée à 10 nM dans de l'eau puis envoyée à la plateforme Get-PlaGe (plateforme GenoToul, INRA Toulouse, France http://www.genotoul.fr) pour le séquençage.

Les banques d'ADNc mixtes finales ont été regroupées à l'aide du kit *TruSeq PE Cluster Kit v3*, selon le protocole *Illumina PE_Amp_Lin_Block_V8.0*, puis ont été chargées sur l'instrument Illumina HiSeq 2000 en suivant les instructions du fabricant. La chimie de séquençage utilisée était la v3 (FC-401-3001, *kit TruSeq SBS*) avec le protocole indexé en paired-end de 2 x 100 cycles. Les analyses d'images et l'appel de base ont été effectués à l'aide du logiciel de contrôle HiSeq (HCS 1.5.15) et du composant d'analyse en temps réel (RTA 1.13.48). Le démultiplexage a été effectué à l'aide de CASAVA 1.8.1 (Illumina) pour produire des fichiers de séquences pairées contenant les lectures (reads) pour chaque échantillon au format Illumina FASTQ.

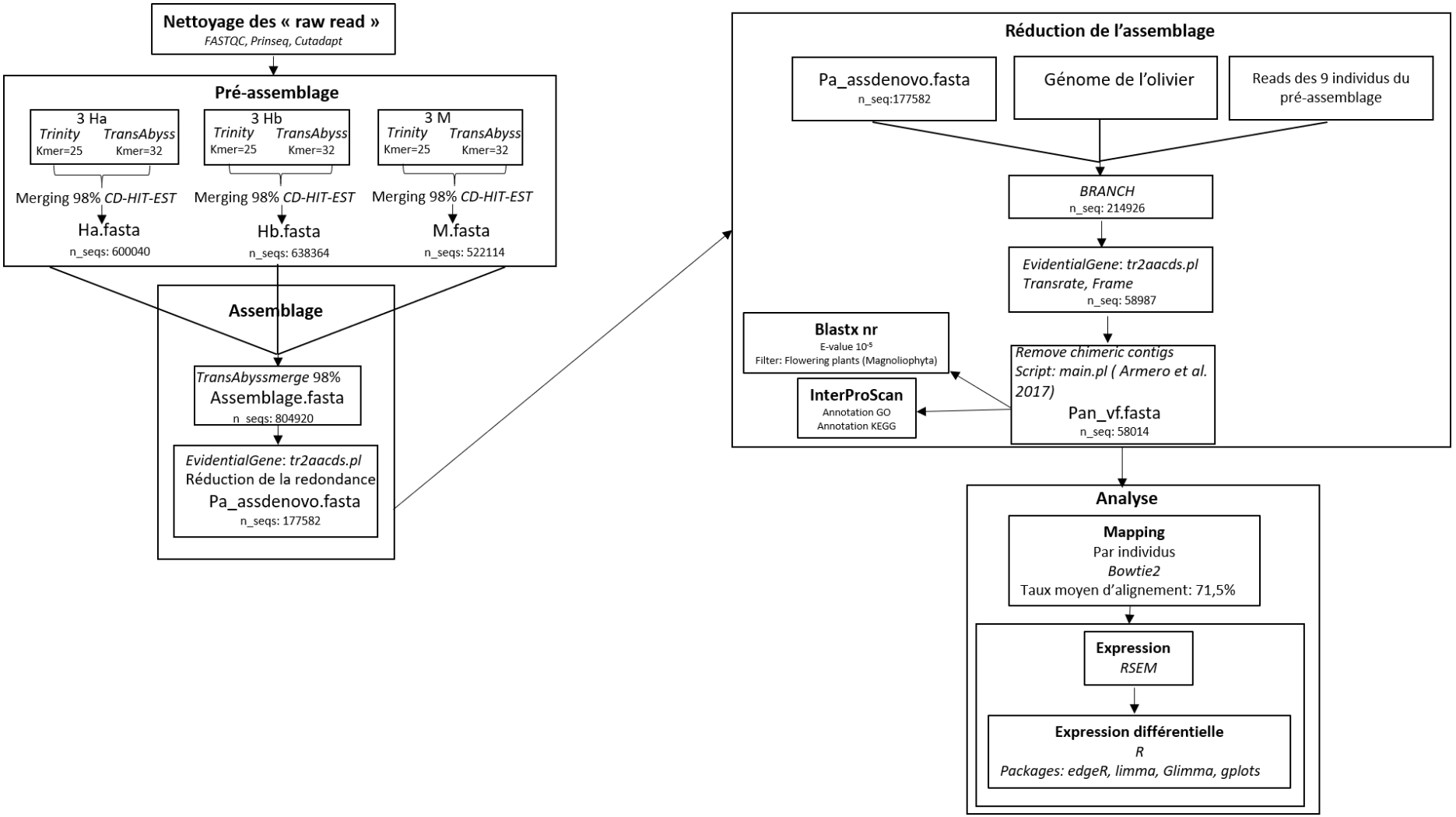


Figure 2.1 : Organigramme du pipeline d'analyse du transcriptome de *P. angustifollia* : assemblage *de novo*, annotation et recherche de contigs d'intérêt par analyse d'expression différentielle

Pipeline d'assemblage *de novo*

Les lectures ont été nettoyés et filtrés à l'aide des outils CutAdapt (MARTIN 2011) (option : *--overlap=30*) et Prinseq (SCHMIEDER AND EDWARDS 2011) (option : *-min_len 80 -trim_tail_left 5 -trim_tail_right 5 -lc_method entropy -lc_threshold 70*).

Le protocole d'assemblage *de novo* a été développé en s'inspirant du pipeline d'Evangelistella *et al*. (2017) et est composé de trois étapes (Figure 2.1). La première est une étape de pré-assemblage pour laquelle deux outils de reconstruction avec deux valeurs de *K-mer* différentes sont utilisés. Le choix de Trinity v2.5.1 (GRABHERR *et al.* 2011) et Trans-Abyss v1.5.5 (ROBERTSON *et al.* 2010) est fait sur la base des résultats de l'étude comparative des programmes de reconstruction *de novo* de Wang and Gribskov (2017). Les paramètres par défaut des outils sont gardés : Trinity *K-mer = 25*, taille minimum des contigs = 200 pb et Trans-Abyss *K-mer = 32,* taille minimum des contigs = 100 pb.

Afin de préserver un maximum de variation inter-individuelle tout en minimisant le bruit de fond, les pré-assemblages ont été effectués sur trois individus de chacun des groupes. Dans cette première étape, neuf transcriptomes ont donc été reconstruits, dont la qualité a été évaluée à l'aide du logiciel Transrate v1.0 (SMITH-UNNA *et al.* 2016). Le but de cet outil est de détecter les erreurs majeures qui pourraient biaiser les reconstructions, et de donner plusieurs mesures de qualité et statistiques sur les assemblages. Transrate a été utilisé plusieurs fois le long du pipeline afin de contrôler l'amélioration du transcriptome au fur et à mesure des étapes d'assemblage. Une fois l'étape de vérification passée, pour chacun des individus, les reconstructions obtenues par Trinity et Trans-Abyss ont été rassemblées sur la base de l'identité nucléotidique des contigs (0.98) en utilisant l'outil CD-HIT-EST (LI AND GODZIK 2006). Nous obtenons donc trois pré-assemblages.

La seconde étape, que l'on nomme « Assemblage », commence après la vérification de la qualité des trois pré-assemblages (Figure 2.1). Il s'agit de rassembler les reconstructions à l'aide de l'option « Trans-Abyss merge » ce qui permet d'obtenir une première reconstruction *de novo* complète qui est ensuite traitée avec le pipeline EvidentialGene tr2aacds (GILBERT 15 Dec 2013). La première étape du pipeline consiste à rechercher le cadre de lecture de chacun des contigs afin de ne garder que les séquences codantes (CDS) qui sont traduites en séquences d'acides aminés. Ensuite les séquences redondantes sont éliminées (au seuil d'identité 0.98) afin de garder le meilleur CDS pour chacun des contigs que l'on nommera par la suite unitig.

L'étape de réduction de l'assemblage (Figure 2.1), a été construite à partir du protocole et des scripts développés par Armero *et al.* (2017). La première étape consistant à utiliser l'outil BRANCH (BAO *et al.* 2013) qui comporte deux étapes. L'outil commence par aligner les lectures d'ARN contre les unitigs de l'assemblage *de novo*, puis il aligne les unitigs et les lectures restants (qui n'ont pas réussi à s'aligner sur l'assemblage) sur les contigs d'un génome étroitement lié -dans notre cas celui de l'olivier (UNVER *et al.* 2017)- à l'aide d'une version modifiée du programme d'alignement BLAT (KENT 2002). Par la suite, il identifie les exons et les jonctions d'épissages dans les piles de lectures contre ces contigs. Les régions d'empilement répondant à certaines exigences de longueur minimale et de couverture de read sont considérées comme des exons, et si elles sont faiblement couvertes par des alignements espacés et des signaux de jonction d'épissage, elles seront considérées comme des introns. En plus des exons contenus dans les unitigs initiaux, cette étape identifie de nouveaux exons candidats qui sont souvent manqués dans les assemblages de transcriptome *de novo*, principalement en raison d'une couverture inégale. Dans un second temps, BRANCH étend les exons à l'aide des informations de jonction d'épissage précédemment obtenues.

La dernière étape de réduction a permis, à l'aide du logiciel FrameDP v1.2.2 (GOUZY *et al.* 2009) et des scripts : *tr2aacds.pl* du pipeline EvidentialGene (GILBERT 15 Dec 2013) et *main.pl* du pipeline d'Armero *et al.* (ARMERO *et al.* 2017), d'identifier les unitigs redondants et/ou potentiellement chimériques sur la base de leur traduction en séquence protéique. C'est-à-dire que les unitigs traduits de *P. angustifolia* ont été alignés sur la référence protéomique de l'olivier. Un alignement était retenu si l'identité était d'au moins 80% et la couverture du polypeptide cible était d'au moins 70%. Les alignements ont été ordonnés selon la couverture du polypeptide cible, l'identité de l'alignement et la couverture de la protéine de référence, dans cet ordre d'importance pour garder le meilleur alignement.

Evaluation de la complétude de la reconstruction *de novo* du transcriptome de boutons floraux, annotation et classification fonctionnelle

Le logiciel BUSCO -Benchmarking Universal SingleCopy Orthologs- (SIMAO *et al.* 2015) a été utilisé afin d'évaluer la qualité et l'exhaustivité de la reconstruction *de novo.* Il s'agit d'une mesure quantitative de la complétude de l'assemblage basée sur la présence de copies uniques d'orthologues quasi-universels. L'assemblage du transcriptome de boutons floraux du filaire est comparé à l'ensemble de données d'Embryophyta, contenant 1 440 groupes

BUSCO de 31 espèces. Cet outil permet de dire si chacun des groupes est retrouvé de manière complète, dupliquée, fragmentée ou s'il est manquant.

L'outils Blast2GO 5 (CONESA *et al.* 2005) a été utilisé pour annoter l'assemblage de référence de *P. angustifolia*. Premièrement, les contigs ont été alignés à l'aide de l'outil d'alignement local (BLAST) contre la base de données NCBI nr « flowering plants » (taxa : 3398, Magnoliophyta) avec l'option blastx (E-value < 1.10E-5). Cette option permet de comparer les contigs traduits par tous les cadres de lectures trouvés à une base de données de séquences protéiques. Ensuite, l'étape de « mapping » permet de retrouver les « gene ontology terms » (GO) associés à chacun des Hits obtenus par la recherche BLAST. Dans un second temps, on utilise l'outil en ligne public EMBL-EBI InterPro afin de rechercher des signatures InterPro dans les unitigs. La fonctionnalité d'annotation d'InterPro de l'outils Blast2GO permet ainsi de retrouver des informations sur les domaines/motifs présents dans les séquences et de leur attribuer l'annotation GO correspondante.

De l'alignement des lectures à l'analyse d'expression différentielle

Bowtie2 v 2.3.3.1 (LANGMEAD AND SALZBERG 2012) a été utilisé pour réaligner les lectures sur le transcriptome de référence précédemment obtenu. Les options "*end-to-end*", "*sensitive*" et "*mp 1,1*" ont été utilisées pour l'alignement. Après cette étape, un fichier Bam est obtenu par individu, qui peut ensuite être utilisé pour l'estimation du niveau d'expression des transcrits. La quantification de l'expression chez chacun des individus est ensuite effectuée avec le logiciel RSEM v 1.3.0 (LI AND DEWEY 2011). RSEM est un outil qui peut quantifier les abondances de gènes et d'isoformes à partir de données RNA-Seq sans qu'il soit nécessaire de fournir un génome de référence. Comme il ne repose pas sur l'existence d'une référence, il est particulièrement adapté pour la quantification avec des assemblages de transcriptome *de novo*.

Les packages *EdgeR* (ROBINSON *et al.* 2010), *limma* (RITCHIE *et al.* 2015), *Glimma* (SU *et al.* 2017) et *gplots* (WARNES *et al.* 2009) ont été utilisés avec le logiciel R V3.5.0 pour suivre les étapes d'analyse décrites par Chen Y. *et al* (2016). La première étape consiste à calculer un facteur de normalisation (*method by trimmed mean of M value*), un pour chaque échantillon, afin d'éliminer les biais de composition entre les différentes librairies.

L'analyse d'expression différentielle (DE), entre les hermaphrodites et les mâles, a été faite à l'aide du pipeline de quasi-vraisemblance d'*EdgeR*. Il s'agit en premier lieu d'appliquer la

fonction *glmTreat* de *EdgeR* qui permet d'implémenter un test d'expression différentielle par rapport à un seuil de « *fold-change »* fixé. Nous avons choisi une valeur du paramètre DE *fold change* de 1.5, qui est classiquement utilisée dans la littérature (CHEN *et al.* 2016). Dans un second temps, la fonction *decideTestsDGE* -qui consiste en une correction de tests multiples basée sur la méthode Benjamini-Hochberg- permet de contrôler le « taux de fausses découvertes » (FDR). Le FDR choisi est de 1%. Les comparaisons se faisant sur la distribution du niveau d'expression intra-groupe, certains individus peuvent tirer vers le haut cette expression. Afin de limiter ce biais, une dernière étape de tri est effectuée, les contigs n'étant pas exprimés chez 25% individus (soit 3 chez les mâles et 6 chez les hermaphrodites) ou plus dans les groupes d'intérêts ont été retirés. Les séquences issues de cette première analyse serviront à la définition des séquences cibles de l'expérience de capture de séquences par hybridation développée dans la prochaine partie. Ainsi, capturer chez plus d'individus les séquences correspondant à l'ensemble des unitigs d'intérêt permettra : (i) de déterminer si la présence/absence d'expression pour certaines séquences est d'origine génétique ou seulement transcriptomique, (ii) d'étudier la structuration génétique (SNPs) en fonction des phénotypes.

Nous avons ensuite cherché à identifier les propriétés fonctionnelles des gènes candidats obtenus, en particulier des unitigs dont l'expression ou la surexpression est spécifique d'un phénotype. Dans un souci de clarté de la présentation, nous nous sommes concentrés sur les unitigs différentiellement exprimés dont la valeur de Log-fold-change est la plus extrême (pour un FDR≤0.001).

Positionnement sur le génome de l'olivier des unitigs différentiellement exprimés

Nous avons utilisé le logiciel BLAST (KENT 2002) pour positionner les unitigs différentiellement exprimés et déterminer leur position sur le génome d'*Olea europaea* var. *sylvestris* (UNVER *et al.* 2017) afin d'en déduire leur co-localisation possible par rapport à la région orthologue liée au sexe du LG12 chez *P. angustifolia (CARRE et al. 2021)*. Pour les loci présents sur le LG12 issus de la cartographie, seuls ceux ayant un hit unique avec au moins 85% d'identité sur un minimum de 110 pb ont été sélectionnés pour l'analyse de synténie. Nous avons positionné les unitigs d'intérêt issus des analyses d'expression différentielle sur le génome de l'olivier en prenant la localisation du meilleur hit avec au moins 85% d'identité sur

un minimum de 300 pb. Les relations de synténie ont ensuite été visualisées avec l'outil circos-0.69-6 (KRZYWINSKI *et al.* 2009).

2. Capture de séquences par hybridation ciblée

Du choix du matériel biologique au séquençage

L'expérience de capture de séquences a porté sur un total de 95 individus d'origines différentes. D'une part, les deux parents (Ha et M) et 65 descendants (18 Ha, 18 Hb, 29 mâles) ont été choisis parmi les individus du croisement de cartographie (CARRE *et al.* 2021). D'autre part, 28 individus (11 Ha, 10 Hb et 7 mâles) ont été choisis au sein de populations naturelles, dont 27 dans la population de Fabrègue et un dans la population de Camargue afin d'évaluer la robustesse de l'association SNPs-phénotype dans le contexte d'une base génétique élargie.

Définition des sondes de captures

Grâce à la synténie entre *P. angustifolia* et *O. europaea* (var. *sylvestris*) (CARRE *et al.* 2021), nous avons défini une première série de séquences cibles à capturer chez le filaire. La région associée au sexe (LG12) chez le filaire définit, sur le génome de l'olivier, une région de 2 195 656pb sur le chromosome 12. A l'aide des annotations disponibles sur le génome de l'olivier, l'intégralité des séquences codantes de cet intervalle ont été sélectionnées comme cibles pour définir des sondes de capture. Par ailleurs, plusieurs marqueurs strictement associés au sexe trouvent une homologie soit sur un autre chromosome (Chr3) soit sur plusieurs scaffold (sca1196, sca1264, sca727) non ancrés dans le génome de l'olivier. Les séquences codantes encadrant la zone définie sur le chromosome 3 et l'intégralité des séquences codantes annotées au sein de ces scaffolds ont également été sélectionnées comme cibles. La définition des régions à cibler sur le génome de l'olivier ayant été réalisée à un stade précoce de l'analyse du jeu de données de cartographie, des divergences existent entre les régions ciblées dans cette expérience et les régions définies comme strictement liées au sexe dans le premier chapitre (CARRE *et al.* 2021). Ce qui explique l'absence des scaffold 393 et 969, la présence du scaffold 727 et d'une portion du chromosome 3 ainsi que la taille plus importante de la région ciblée du chromosome 12. Au total, 114 séquences cibles représentant un total de 258 957pb ont été définies à partir du génome de l'olivier.

Sur la base des analyses transcriptomiques, nous avons par ailleurs inclus l'ensemble des unitigs différentiellement exprimés entre les mâles et les hermaphrodites ayant un DE *fold change* de 1,5 et un FDR≤0,01 et ayant passé les filtres de redondance, de pourcentage de GC et de qualité imposés pour la définition des sondes par l'entreprise *MyBaits*. En cas de redondance avec des séquences cibles définies sur le génome de l'olivier, seule la séquence génétique de l'olivier (plus grande que la séquence transcriptomique du filaire) a été retenue pour la définition des sondes de capture. Au total, 61 cibles composées des unitigs sur-exprimés chez les mâles et 211 cibles représentant les unitigs sur-exprimés chez les hermaphrodites ont été définies. L'ensemble de ces 272 cibles couvrent 391 601 pb du transcriptome du filaire.

Une partie de l'expérience consistant à capturer sur le filaire des séquences définies à partir du génome de l'olivier (divergence moyenne entre le génome de l'olivier et le transcriptome du filaire <95%), nous avons privilégié un design de sondes courtes afin d'augmenter nos chances de capturer les séquences d'intérêt, mêmes modérément divergentes. Nous avons donc fait le choix d'un kit personnalisé *"myBaits Custom DNA-Seq ®"*, avec 32 309 sondes de 80 nucléotides avec une couverture de 3X c'est-à-dire que les sondes étaient définies tous les 20nt le long des cibles.

Préparation des banques d'ADN

Pour les parents et les descendants du croisement, l'ADN de 100 mg de jeunes feuilles congelées a été extrait avec le kit Chemagic DNA Plant (Perkin Elmer Chemagen, Baesweller, DE, Part # CMG-194), selon les instructions du fabricant. L'ADN des individus issus des populations a été extrait à partir de 50 mg de feuilles fraîches en tube individuel selon le protocole disponible en Annexe 2.1 . Ensuite le protocole Chemagic DNA Plant a été adapté à l'utilisation de la station de travail automatisée de purification d'ADN KingFisher Flex™ (Thermo Fisher Scientific, Waltham, MA, USA) et l'ADN extrait a été quantifié à l'aide d'un fluorimètre Qubit (Thermo Fisher Scientific, Illkirch, France).

La préparation des banques d'ADN a été faite avec le kit NEXTFLEX® Rapid DNA seq kit version 2.0. Brièvement les étapes consistent en une fragmentation de 20 cycles de 30sec par sonication sur RotorGene, une ligation d'index unique par individus (NEXTFLEX® Unique Dual Index Barcodes), une amplification par PCR et une double sélection de taille permettant de ne garder que les fragments compris entre 250pb et 450pb. Une vérification sur puce DNA HS est

effectuée après la sonication et sur puce Agilent HS à la fin de la préparation des banques (Annexe 2.2). Le multiplexage en deux pools de 48 individus a été effectué en mélangeant à molarité égale (10ng d'ADN de chaque échantillon). Puis les librairies ont été enrichies en suivant le protocole du kit MYbaits (v.4.01) et regroupées en un seul échantillon ajusté à 2nM dans 25µL. Les librairies regroupées ont été soumises à un séquençage pairé selon le protocole Illumina MiSeq à l'aide du kit de séquençage v3 à 2x300 cycles (plateforme LIGAN, Lille).

Des données brutes au "variant calling"

Les adaptateurs ont été d'abord supprimés des séquences brutes par Trimmomatic version 0.39 (BOLGER *et al.* 2014) à l'aide des paramètres suivants : PE LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100.

Afin de pouvoir aligner les séquences et effectuer le « variant calling » à partir d'une séquence de référence le plus proche possible du filaire (en l'absence d'un génome de référence pour cette espèce), nous avons commencé par reconstruire de novo les séquences ciblées. Les séquences cibles ont été traitées à l'aide du pipeline HybPiper v1.3.1 (JOHNSON *et al.* 2016). Le script *distribute_targets.py* permet d'abord de regrouper les lectures correspondant aux séquences cibles de référence en utilisant BWA (LI AND DURBIN 2009), pour ensuite assembler *de novo* ces lectures en contigs (un par individus) à l'aide de SPAdes (www.github.com/ablab/spades). HybPiper comprend des scripts Python (*reads_first.py* : option *--evalue 1e-5*, *intronerate.py* : options *–merge et retrieve_sequences.py* : option *dna*) qui permettent d'extraire les contigs correspondant à la même cible des résultats du pipeline.

Afin de créer une séquence consensus, qui permettra de remplacer la séquence cible initiale, nous avons aligné la cible de référence avec l'ensemble des contigs *de novo* lui correspondant à l'aide de Clustal Omega v1.2.4 (SIEVERS *et al.* 2011) avec les options *–full* et *–percent-id*. Afin d'identifier les individus divergents ou les paralogues, nous avons créé une matrice de distances deux à deux entre contigs *de novo* de chaque individu. Nous avons considéré une séquence comme chimérique si la moyenne de ses distances est supérieure à la distance moyenne entre séquences moins la déviation standard moyenne totale. Une fois les séquences divergentes filtrées nous avons recréé un fichier d'alignement avec Clustal Omega puis utilisé l'outil *Spruceup* version 19.02.2020 (options : no guide_tree, cutoffs:0.98). *Spruceup* compare les séquences entre individus dans une fenêtre glissante ce qui permet d'identifier et de rogner les portions de séquences trop divergentes qui pourraient résulter

d'erreurs de séquençage ou de paralogues (BOROWIEC 2019). Il a enfin été possible de générer une séquence consensus par cible à l'aide de l'outils *cons* du logiciel EMBOSS v6.6.0 avec l'option : *-plurality 1* (RICE *et al.* 2000).

Nous avons utilisé, à partir de cette étape, les séquences consensus comme nouvelle référence. Pour chacun des individus nous avons aligné les lectures filtrées à l'aide de l'outil bowtie2 v2.4.1 (LANGMEAD AND SALZBERG 2012) options : *--minins 0 --maxins 1000* puis nous avons filtré les lectures dupliquées avec l'outil picard v2.21.4 (http://broadinstitute.github.io/picard/) et les options *MarkDuplicates REMOVE_DUPLICATES=true*. Nous avons pu ensuite détecter les SNPs et les indels par individus grâce à l'outil GATK 4.1.4.1 (DE SUMMA *et al.* 2017) (options *: HaplotypeCaller -ERC GVCF*). Nous avons rassemblé les variants de tous les individus avec l'option *CombineGVCFs* puis avons généré les génotypes avec l'option *GenotypeGVCFs -stand-call-conf 10*. Nous avons filtré ensuite les fichiers .vcf obtenus à l'aide de l'outils vcftools v0.1.16 (DANECEK *et al.* 2011) et des options *: --remove-indels --min-alleles 2 --max-alleles 2 --minDP 5 --recode --recode-INFO-all*, afin de conserver uniquement les SNPs bi-alléliques ayant une couverture d'au moins 5 lectures.

Méthodes d'analyse de ségrégation des SNPs

Afin d'identifier les SNPs présentant une forte hétérogénéité de fréquence des génotypes entre les groupes comparés pour ce site (Hermaphrodite vs. Mâle), nous avons utilisé l'indice de fixation F_{ST}. Les séquences présentant au moins un SNPs ayant un F_{ST} supérieur ou égal à 0,3 ont été sélectionnées (la valeur de 0,3 correspond à l'attendu sous un strict système XY où un groupe n'est constitué que d'hétérozygotes et l'autre groupe est fixé pour un allèle). Les séquences contenant des SNPs présentant une forte hétérogénéité de couverture entre les groupes de type présence/absence (par exemple aucun hermaphrodite couvert et mâles couverts ou inversement) ont aussi été gardées afin de tenter d'identifier des situations d'hémizygotie. Nous avons ensuite utilisé les scripts *genotype_plot.R* et *combine_plot R* (https://github.com/JimWhiting91/genotype_plot) qui permettent de visualiser et regrouper les génotypes.

Encadré 1 : Analyse de la qualité d'assemblage d'un transcriptome par Transrate

Les scores Transrate mesurent la qualité de l'assemblage *de novo*. Un score est produit pour l'ensemble de l'assemblage, et pour chaque contig. Le processus de notation utilise les lectures qui ont été utilisées pour générer l'assemblage.

Transrate Assembly Score (TSA)

Le TSA permet de comparer plusieurs assemblages réalisés avec les mêmes lectures. Le score est conçu de manière à ce qu'un score élevé corresponde à un assemblage plus précis sur le plan biologique. Le TSA est calculé comme la moyenne géométrique de tous les contig scores multipliée par la proportion de lectures d'entrée.

Contig score

Le contig score est considéré comme une mesure permettant de savoir si le contig est une représentation précise, complète et non redondante d'un transcrit qui était présent dans l'échantillon séquencé. Il se compose de quatre éléments :
(i)Une mesure qui permet de savoir si chaque base a été appelée correctement. Ceci est estimé à l'aide de la distance d'édition moyenne par base, c'est-à-dire combien de changements devraient être apportés à une lecture couvrant une base avant que la séquence de cette lecture et la région couverte du contig ne concordent parfaitement.
(ii)Une mesure qui estime si chaque base fait vraiment partie de la transcription. Ceci est estimé en déterminant si les lectures fournissent une couverture suffisante pour chaque base.
(iii)La probabilité que le contig soit dérivé d'un seul transcrit (plutôt que de morceaux de deux ou plusieurs transcrits). Ceci est mesuré comme la probabilité que la couverture de lectures soit mieux modélisée par une seule distribution de Dirichlet, plutôt que par deux ou plusieurs distributions.
(iv)La probabilité que le contig soit structurellement complet et correct. Ceci est estimé comme la proportion de paires de lectures attribuées qui sont en accord avec la structure et la composition du contig, qui à son tour est calculée en classant les alignements de paires de lectures.

Read mapping metrics (RMM)

Ces indicateurs de qualité sont directement basés sur l'alignement des lectures utilisées pour l'assemblage sur les contigs assemblés. Ce sont les indicateurs les plus utiles, les lectures contiennent une mine d'informations spécifiques à l'organisme séquencé, et ces informations peuvent être utilisées pour évaluer la confiance dans chaque base et contig de l'assemblage. Parmi les différents RMM généré, deux sont particulièrement intéressant. Le premier est la proportion de « bon contigs » c'est-à-dire la proportion de contigs où les lectures pairées sont alignés de manière complète sur le même contig, dans le bon sens et de manière non chevauchante sur les extrémités du contig. Le second indicateur donnant une bonne vision de la qualité d'un assemblage est la proportion de contigs segmenté c'est-à-dire la proportion de contigs qui a une chance d'être segmenté supérieure ou égale à 50%.

Résultats

1. Analyse transcriptomique : une recherche de gènes à expression sexe-biaisée chez une espèce androdioïque

Assemblage *de novo* du transcriptome de boutons floraux de *P.angustifolia*

Les trois reconstructions spécifiques des groupes Ha, Hb et M obtenues après le rassemblement des deux assemblages indépendants effectués avec Trinity (GRABHERR *et al.* 2011) et TransAbyss (ROBERTSON *et al.* 2010) étaient composées respectivement de 600 040, 638 364 et 522 114 contigs (Figure 2.1). La fusion, sur la base de l'identité nucléotidique des contigs (0.98), de ces trois reconstructions a permis d'obtenir un pré-assemblage composite de 804 920 contigs (taille moyenne=501 bp) qui a été réduit à 177 582 unitigs (taille moyenne=836 bp) dans l'étape d'assemblage (Tableau 2.1) sur la base de l'identité des séquences traduites (0.98).

Tableau 2.1. Résumé des principaux scores évaluant la qualité des assemblages obtenu avec l'outils Transrate

		Pré-assemblage	**Assemblage**	**Assemblage +BRANCH**	**Assemblage final**
Statistique	Nombre de séquences	804920	177582	214926	58014
	Taille minimale	100	100	100	100
	Taille maximale	16928	16911	16911	12274
	Taille moyenne	501.58	836.17	861.42	1103.15
	Unitigs avec ORF (%)	54.87	65.12	68.78	70.67
Qualité	Nombre de lectures test	33171507	33171507	33171507	33171507
	Lectures alignées (%)	14	53	52	71
	Bons contigs (%)	22	45	38	69
	Bases non couvertes (%)	77	45	55	13
	Contigs segmentés (%)	1	8	7	6
	Transrate Assembly Score	0.003	0.045	0.0325	0.1488

La dernière étape du pipeline, qui utilise le génome de l'olivier comme guide pour rassembler des unitigs issus des mêmes gènes ou raccrocher des lectures qui n'avaient pas pu l'être lors de la reconstruction *de novo* puisque non chevauchantes avec les unitigs, a commencé par augmenter le nombre de contigs à 214 926 (taille moyenne=861 bp). Cette dernière étape à finalement permis de réduire l'assemblage de référence à 58 014 unitigs (taille moyenne=1103 bp). Au fur et à mesure des étapes d'assemblage, les différentes statistiques et indicateurs de qualité (Tableau 2.1) montrent une optimisation de l'assemblage (Encadré 1). En effet, si on compare le pré-assemblage à l'assemblage final, on divise par 6 le

pourcentage de bases non couvertes (13% contre 77%), on augmente par 3 le pourcentage de bon contigs (69% contre 22%) et le Transrate Assembly Score est 50 fois supérieur (0.1488 contre 0.003).

Le second indicateur de qualité de l'assemblage final est son score BUSCO (SIMAO *et al.* 2015). Parmi les 1440 groupes recherchés, 87,2% sont retrouvés de manière complète (dont 77,8% en copie unique et 9,4% de manière dupliquée), 2,2% sont fragmentés et 10,6% sont manquants. La recherche d'homologie, pour l'assemblage total, par blastx et l'outils EMBL-EBI InterPro a permis d'annoter 79.79% de l'assemblage. Parmi les 58 014 unitigs de l'assemblage final, 45 917 séquences ont trouvé une homologie dans la base de données des Magnoliophyta, dont 75% ont été annotées par homologie avec un des génomes de l'olivier disponibles (*Olea europaea* var. *sylvestris* et *Olea europaea* subsp. *europaea*), ce qui permettra par la suite de fournir une base d'annotation des unitigs d'intérêt. Globalement, ces indicateurs montrent que le transcriptome que nous avons obtenu pour *P. angustifolia* est de qualité très satisfaisante, comparable à celle d'études récentes sur des espèces non-modèles (CRUZ *et al.* 2016; EVANGELISTELLA *et al.* 2017).

Analyse d'expression différentielle : sélection des transcrits pour la capture

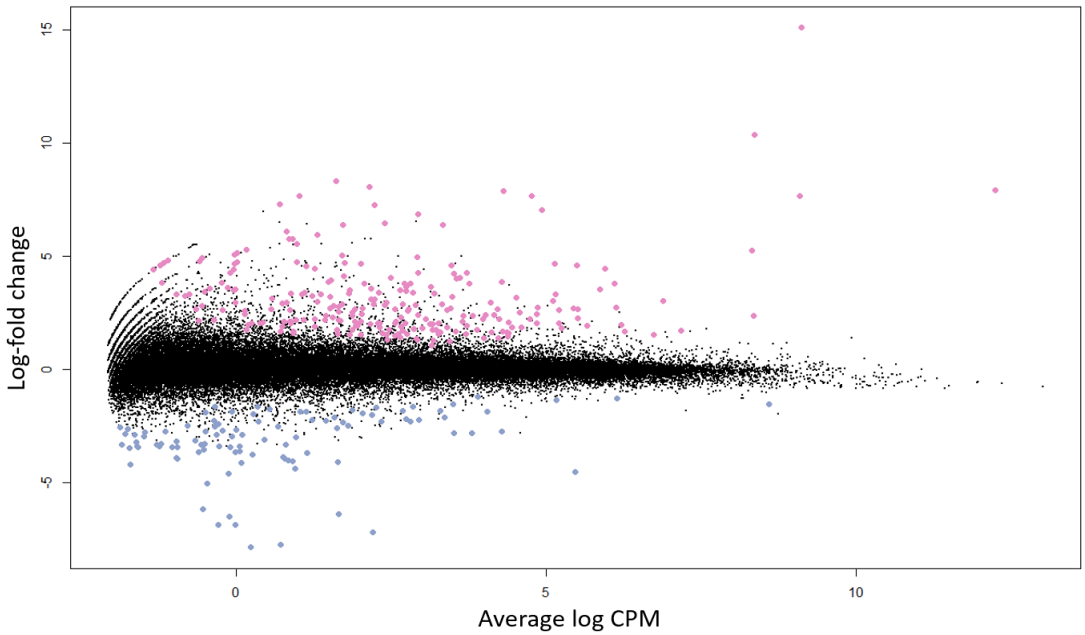


Figure 2.2 : « Mean-difference » (MD) plot du log-fold-change chez les hermaphrodites par rapport aux « count per million » (CPM) moyen (log2) chez les mâles. Chaque point représente un unitig, en violet les unitigs surexprimés chez les hermaphrodites et en bleu les unitigs surexprimés chez les mâles, sur la base d'un log-fold-change de 1,5 et un FDR≤0,01.

Le taux de réalignement des lectures de chacun des individus sur la référence est relativement élevé, avec une moyenne de 71,5%. La comparaison entre les niveaux

d'expression chez les mâles et les hermaphrodites met en évidence 75 unitigs significativement surexprimés chez les mâles et 228 unitigs significativement surexprimés chez les hermaphrodites (Figure 2.2; Log-fold-change >=1.5 ; FDR≤0.01). Un nombre plus élevé de gènes surexprimés chez les hermaphrodites était attendu, étant donné que le phénotype mâle correspond à un arrêt précoce du développement du gynécée, et aboutit donc à l'absence d'un certain nombre de types cellulaires. Les niveaux d'expression de chacun des unitigs d'intérêt sont visualisables dans les heatmaps en Annexe 2.3 et 2.4.

Positionnement des unitigs différentiellement exprimés sur le génome de l'olivier

Dans un premier temps, nous avons cherché à exploiter la relativement bonne synténie entre les génomes de l'olivier et du filaire pour comparer la position des transcrits issus de l'analyse d'expression différentielle à celle de la région génomique associée au sexe (chapitre 1, CARRE *et al.* 2021)). Parmi les 75 unitigs sur-exprimés chez les mâles par rapport aux hermaphrodites, 72% (n=54) trouvent une position dans le génome de l'olivier (Figure 2.3 A). La majorité d'entre eux (n=38) se trouvent sur l'un des 23 chromosomes assemblés, et le reste (n=16) se répartit dans 14 scaffolds non ancrés dans l'assemblage principal. Lorsque l'on regarde plus précisément le positionnement des unitigs d'intérêts sur le génome de l'olivier comparativement à la position probable du sexe définie chez *P. angustifolia* (CARRE *et al.* 2021), un seul (evg89402) colocalise strictement sur le génome de l'olivier (Figure 2.3 B), et deux sont à proximité immédiate (evg87500 et evg87372).

Parmi les 228 unitigs sur-exprimés chez les hermaphrodites par rapport aux mâles, 89% (n=203) trouvent une position dans le génome de l'olivier (Figure 2.4 A). L'olivier étant une espèce hermaphrodite, cette proportion plus élevée peut refléter un développement floral plus proche des *P. angustifolia* hermaphrodites que des *P. angustifolia* mâles. La majorité de ces transcrits sur-exprimés chez les hermaphrodites (n=125) se localise sur l'un des 23 chromosomes et le reste (n=78) se répartit dans 66 scaffolds. On observe 16 unitigs qui se positionnent sur le chromosome 12 de l'olivier, dont quatre unitigs (evg147054, evg178602, evg87871 et evg70363) co-localisant avec la position probable du sexe définie chez *P. angustifolia* (CARRE *et al.* 2021) (Figure 2.4 B). Aucun des unitigs différentiellement exprimés n'est associé à l'un des 8 scaffolds présentant une homologie avec les marqueurs liés au sexe sur le LG12.

(A)

(B)

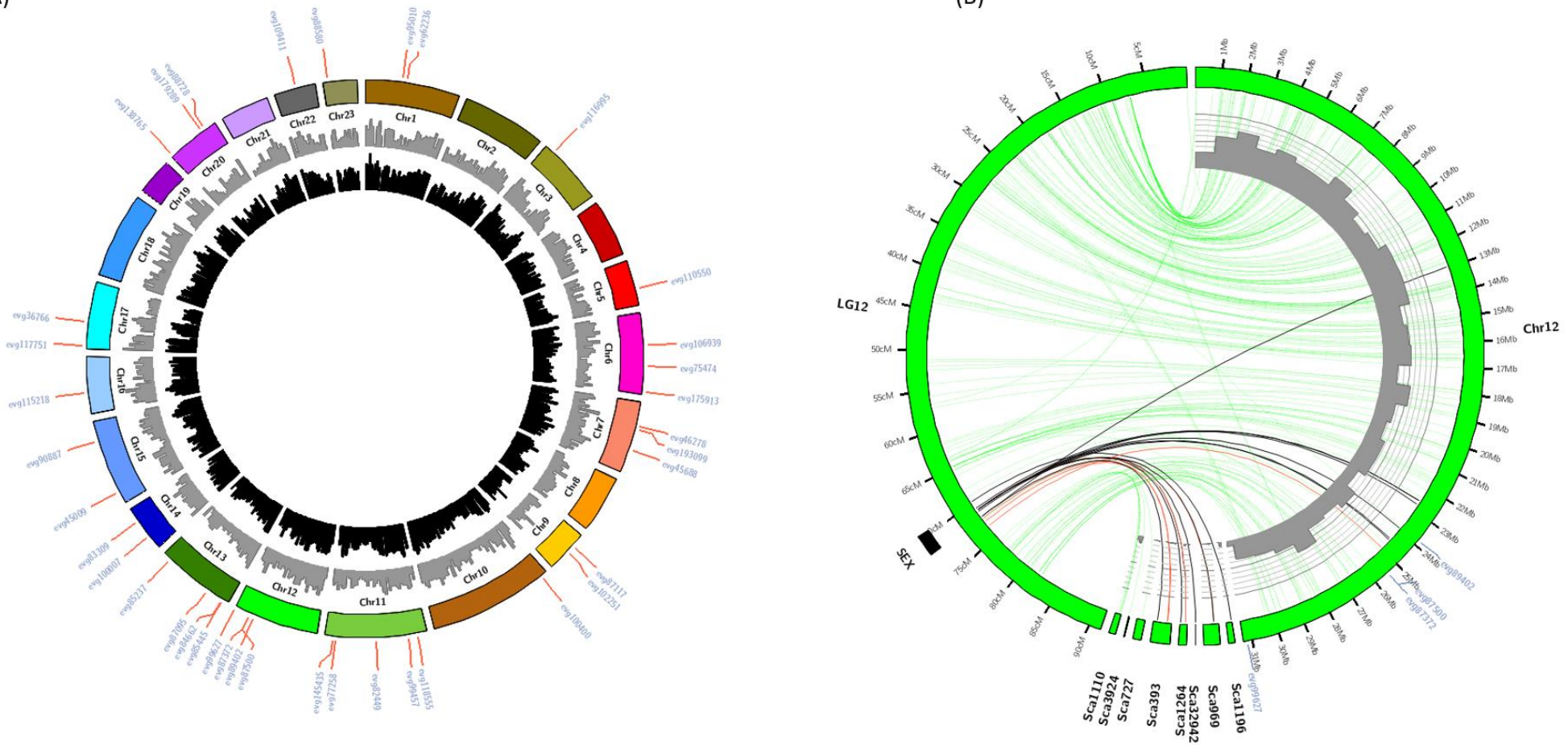


Figure 2.3 : A. Positionnement des unitigs sur-exprimés chez les individus mâles sur les 23 chromosome de l'olivier, les histogrammes noirs représentent le nombre de gènes annotés sur le génome de l'olivier et les histogrammes gris représentent le nombre totald'unitigs du filaire pouvant se positionner sur le génome de l'olivier par « bins » de 1Mpb ; B. Synteny plot entre le groupe de liaison de *P. angustifolia* (échelle en cM) et le chromosome 12 de l'olivier et une série scaffolds non ancrés (échelle en Mb), échelle 1 Mbp=3.125 cM. Les lignes relient les marqueurs de la carte de liaison de *P. angustifolia* avec leur meilleur hit BLAST dans le génome d'*O. europea*. Les lignes vertes correspondent aux marqueurs à transmission autosomique. Les lignes noires correspondent à des marqueurs qui coségrègent strictement avec les phénotypes sexuels (mâles vs hermaphrodites). Les lignes rouges correspondent aux marqueurs qui présentent une association forte mais partielle (95 %) avec le sexe. Les histogrammes gris représentent le nombre d'unitigs total du filaire pouvant se positionner sur le génome de l'olivier par « bins » de 1Mpb. Les unitigs surexprimés chez les individus mâles se positionnant sur le chromosome 12 de l'olivier sont indiqués en bleu.

(A)

(B)

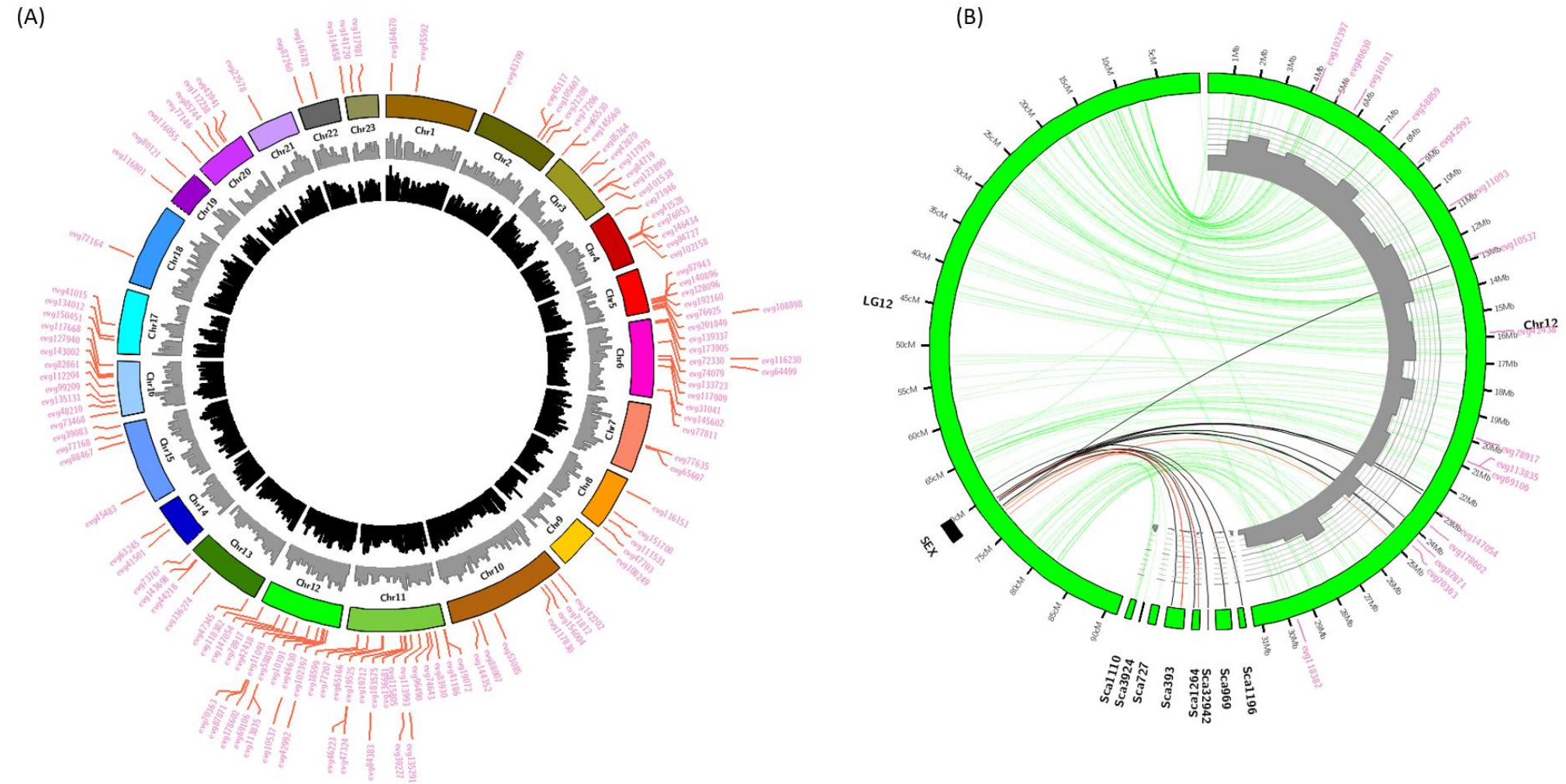


Figure 2.4 : A. Positionnement des unitigs sur-exprimés chez les individus hermaphrodites sur les 23 chromosome de l'olivier, les histogrammes noirs représentent le nombre de gènes annotés sur le génome de l'olivier et les histogrammes gris représentent le nombre d'unitigs du filaire pouvant se positionner sur le génome de l'olivier par « bins » de 1Mpb ; B. Synteny plot entre le groupe de liaison 12 de *P. angustifolia* (échelle en cM) et le chromosome 12 de l'olivier et une série scaffolds non ancrés (échelle en Mb), échelle 1 Mbp=3.125 cM. Les lignes relient les marqueurs de la carte de liaison de *P. angustifolia* avec leur meilleur hit BLAST dans le génome d'*O. europea*. Les lignes vertes correspondent aux marqueurs à transmission autosomique. Les lignes noires correspondent à des marqueurs qui coségrègent strictement avec les phénotypes sexuels (mâles vs hermaphrodites). Les lignes rouges correspondent aux marqueurs qui présentent une association forte mais partielle (95 %) avec le sexe. Les histogrammes gris représentent le nombre d'unitigs total du filaire pouvant se positionner sur le génome de l'olivier par « bins » de 1Mpb. Les unitigs surexprimés chez les individus hermaphrodites se positionnant sur le chromosome 12 de l'olivier sont indiqués en rose.

Au total, notre approche a permis d'identifier cinq unitigs (evg89402, evg147054, evg178602, evg87871 et evg70363) qui sont à la fois différentiellement exprimés et dont la position génomique probable inférée chez l'olivier correspond à celle du déterminisme du sexe chez le filaire. Ces unitigs sont à ce stade nos meilleurs candidats au contrôle des phénotypes sexuels. Le reste des unitigs différentiellement exprimés sont localisés dans d'autres régions du génome, et sont plutôt des candidats pour la participation à la cascade de processus aboutissant en aval au développement des phénotypes sexuels sous le contrôle des candidats précédents.

Prédiction fonctionnelle des unitigs à expression sexe-spécifique

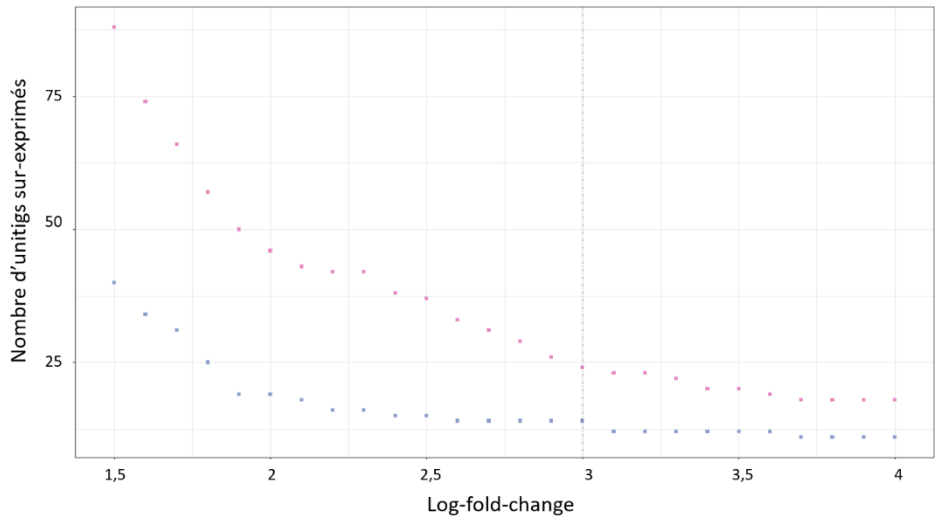


Figure 2.5 : Distributions du nombre d'unitigs différentiellement exprimés chez les mâles (bleu) ou chez les hermaphrodites (rose) en fonction de la valeur de Log-fold-change pour un FDR≤0,001.

Nous avons ensuite cherché à identifier les propriétés fonctionnelles des gènes candidats obtenus, en particulier des unitigs dont l'expression est très contrastée, voire spécifique d'un phénotype. Afin de clarifier la présentation, nous avons choisi de nous concentrer pour cette analyse sur la queue de la distribution des Log-fold-change (Figure 2.5), en ne retenant que les unitigs différentiellement exprimés caractérisés par un Log-fold-change supérieur ou égal à 3, et nous avons examiné le détail de leurs patrons d'expression et la prédiction des rôles fonctionnels de ces unitigs.

Ce seuil plus strict correspond à 14 unitigs "fortement" surexprimés chez les individus mâles par rapport aux hermaphrodites. Parmi ces unitigs, cinq ont une expression (quasi)spécifique des individus mâles (evg99627, evg87117, evg94345, evg101109 et evg87372, Figure 2.6 A). Cependant, étant donné le caractère incomplet de l'annotation du

génome de l'olivier, trois de ces unitigs (evg87117, evg94345 et evg101109) n'ont aucune catégorisation fonctionnelle. Deux autres unitigs à expression quasi-spécifique des mâles et possédant une localisation potentiel proche de la région liée au sexe sur le chromosome 12 sont catégorisés comme facteur de transcription (evg99627) et protéine de liaison ciblée (evg87372). On observe que cinq autres unitigs (evg83309, evg87500, evg89402, evg191695 et evg175913, Figure2.6 A) qui ont un profil d'expression homogène et fort chez tous les individus mâles et qui s'exprime faiblement chez la quasi-totalité des hermaphrodites. Deux de ces unitigs sont prédits comme étant positionnés sur le chromosome 12 de l'olivier, et plus particulièrement dans la région strictement liée au sexe pour evg89402.

Ce seuil plus strict (Log-fold-change de 3 et FDR≤0.001) correspond également à un total de 24 unitigs surexprimés chez les hermaphrodites par rapport aux mâles (Figure 2.7 A). Cinq unitigs (evg23584, evg141405, evg58859, evg74079 et evg125229) ont une expression totalement ou quasi-spécifique des hermaphrodites. Deux de ces unitigs (evg23584 et evg141405) ne trouvent pas d'homologie dans les bases de données et n'ont donc pas de prédiction de fonction. Les trois derniers unitigs ont soit une activité catalytique (*hydrolase or oxidoreductase activity*) soit une fonction de régulation (*response to auxin*) (Figure 2.7 B). Seuls les unitigs evg58859 et evg 113835 se positionnent sur le chromosome 12, mais à des positions assez éloignées de la région du sexe.

(A)

Séquence	localisation	prédiction fonctionnelle	GO annotation
evg83309	chr14	transcription factor HHO5-like	F:GO:0003677: DNA binding; F:GO:0003700:DNA-binding transcription factor activity; P:GO:0009058: regulation of transcription, DNA-templated; P:GO:0034641: cellular nitrogen compound metabolic process
evg107617	sca1558	berberine bridge enzyme-like 18	F:GO:0016491:oxidoreductase activity; F:GO:0071949:FAD binding
evg87500	chr12	uncharacterized protein	F:GO:0003677:DNA binding; F:GO:0046872:metal ion binding
evg89402	chr12	uncharacterized protein	no GO terms
evg191695	NA	---NA---	no GO terms
evg175913	chr6	Receptor-like protein kinase precursor	no GO terms
evg109411	chr22	uncharacterized protein	F:GO:0005484:SNAP receptor activity; P:GO:0006890:retrograde vesicle-mediated transport Golgi to endoplasmic reticulum
evg100400	chr10	K(+) efflux antiporter 1, chloroplastic	C:GO:0009536:plastid; P:GO:0015979:photosynthesis; F:GO:0022857:transmembrane transporter activity
evg90887	chr15	auxin-induced protein 15A-like	P:GO:0009733:response to auxin
evg99627	chr12	homeobox-leucine zipper protein ATHB-54-like	F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription DNA-templated; F:GO:0043565:sequence-specific DNA binding
evg87117	chr11	Hypothetical predicted protein	no GO terms
evg94345	NA	---NA---	no GO terms
evg101109	NA	---NA---	no GO terms
evg87372	chr12	pentatricopeptide repeat-containing protein At1g62350-like	F:GO:0005515:protein binding

(B)

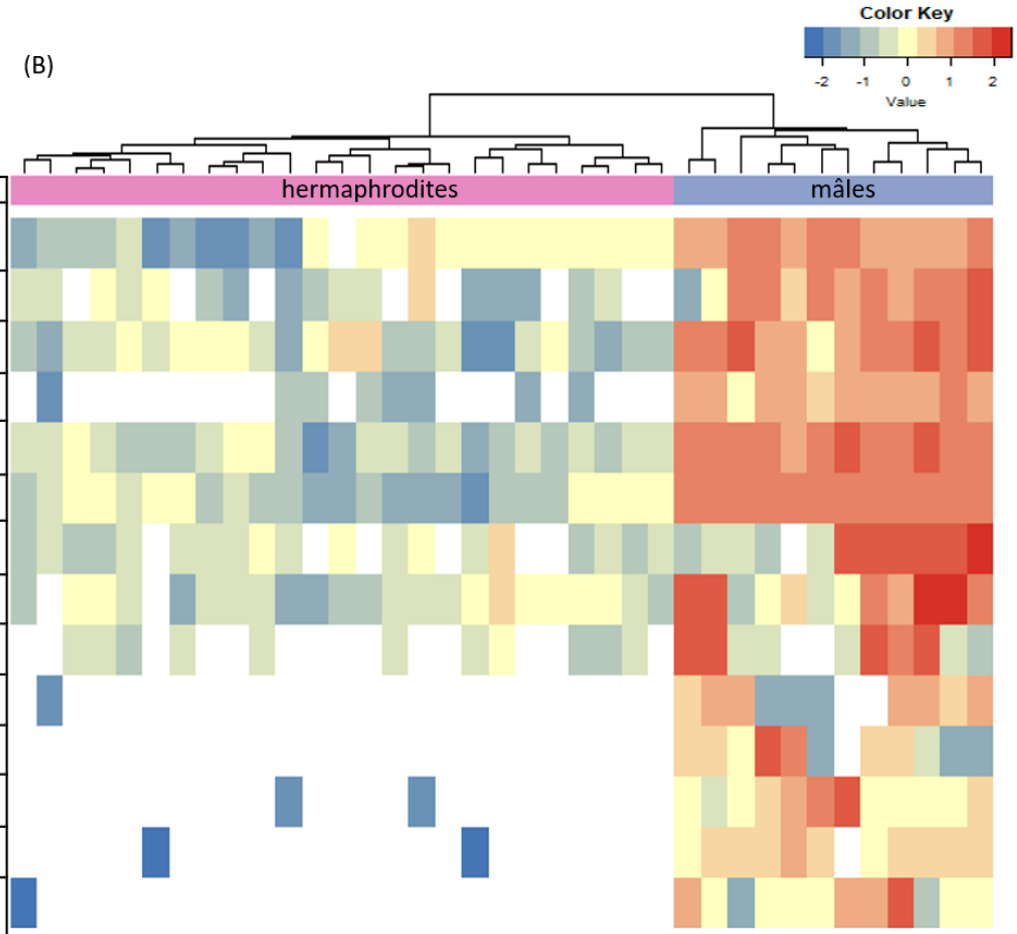


Figure 2.6 : A. Table séquence, localisation probable sur le génome de l'olivier, prédiction fonctionnelle et annotation GO, chaque ligne correspond à un unitigs dont le pattern d'expression est mis en regard sur le heatmap B. heatmap unitigs sur-exprimés chez les individus mâles (bleu) par rapport aux hermaphrodites (rose), chaque colonne représente un individu et chaque ligne un unitigs. Les couleurs représentent la divergence d'un gène particulier dans un échantillon particulier, par rapport à la valeur moyenne de ce gène sur tous les échantillons, centré réduit sur 0 en unités d'écarts types (bleu peu exprimé, jaune expression moyenne, rouge fortement exprimé).

(A)

Séquence	localisation	prédiction fonctionnelle	GO annotation
evg81085	sca1205	---NA---	no GO terms
evg113835	chr12	---NA---	no GO terms
evg74643	chr11	probable glycosyltransferase At5g03795	P:GO:0006486:protein glycosylation; F:GO:0016757:glycosyltransferase activity
evg76925	chr5	protein RALF-like 19	no GO terms
evg111531	chr8	transcription factor bHLH75-like	P:GO:0006355:regulation of transcription, DNA-templated;F:GO:0046983:protein dimerization activity
evg112991	sca978	caffeic acid 3-O-methyltransferase-like	F:GO:0008171:O-methyltransferase activity; F:GO:0046983:protein dimerization activity
evg43666	sca1590	glutamate receptor 3.6-like	C:GO:0005575:cellular_component
evg145602	chr6	transcription factor bHLH75-like isoform X3	P:GO:0006355:regulation of transcription, DNA-templated; F:GO:0046983:protein dimerization activity
evg113019	sca1590	ABA transporter	C:GO:0016021:integral component of membrane; F:GO:0016887:ATP hydrolysis activity; F:GO:0042626:ATPase-coupled transmembrane transporter activity
evg195089	NA	zeatin O-glucosyltransferase-like	F:GO:0005515:protein binding; F:GO:0008194:UDP-glycosyltransferase activity; F:GO:0016758:hexosyltransferase activity
evg140896	chr5	Replication factor C, subunit RFC3	F:GO:0003677:DNA binding; P:GO:0006260:DNA replication
evg46223	chr11	sm-like protein LSM3B	P:GO:0006397:mRNA processing; P:GO:0022607:cellular component assembly; P:GO:0034655:nucleobase-containing compound catabolic process
evg128790	sca858	RING-H2 finger protein ATL1-like	C:GO:0005575:cellular_component
evg115805	chr11	putative LisH domain-containing protein C1711.05-like protein	no GO terms
evg85264	chr3	pathogenesis-related protein 5-like	C:GO:0005576:extracellular region; P:GO:0006950:response to stress
evg125229	sca2411	probable 2-oxoglutarate-dependent dioxygenase AOP1 isoform X2	F:GO:0016491:oxidoreductase activity; P:GO:0055114:obsolete oxidation-reduction process
evg102670	NA	beta-glucosidase-like	F:GO:0004553:hydrolase activity, hydrolyzing O-glycosyl compounds; P:GO:0005975:carbohydrate metabolic process
evg101538	chr3	delta(12)-fatty-acid desaturase FAD2-like	P:GO:0006629:lipid metabolic process; F:GO:0016717:oxidoreductase activity; P:GO:0055114:obsolete oxidation-reduction process
evg78038	NA	flowering time control protein FPA-like isoform X1	F:GO:0003676:nucleic acid binding
evg123890	chr3	basic form of pathogenesis-related protein 1-like	no GO terms
evg74079	chr6	glucan endo-1,3-beta-glucosidase 14-like	F:GO:0004553:hydrolase activity, hydrolyzing O-glycosyl compounds; P:GO:0005975:carbohydrate metabolic process
evg58859	chr12	auxin-responsive protein SAUR23	P:GO:0009733:response to auxin
evg141405	sca806	uncharacterized protein	no GO terms
evg23584	sca2101	uncharacterized protein	no GO terms

(B)

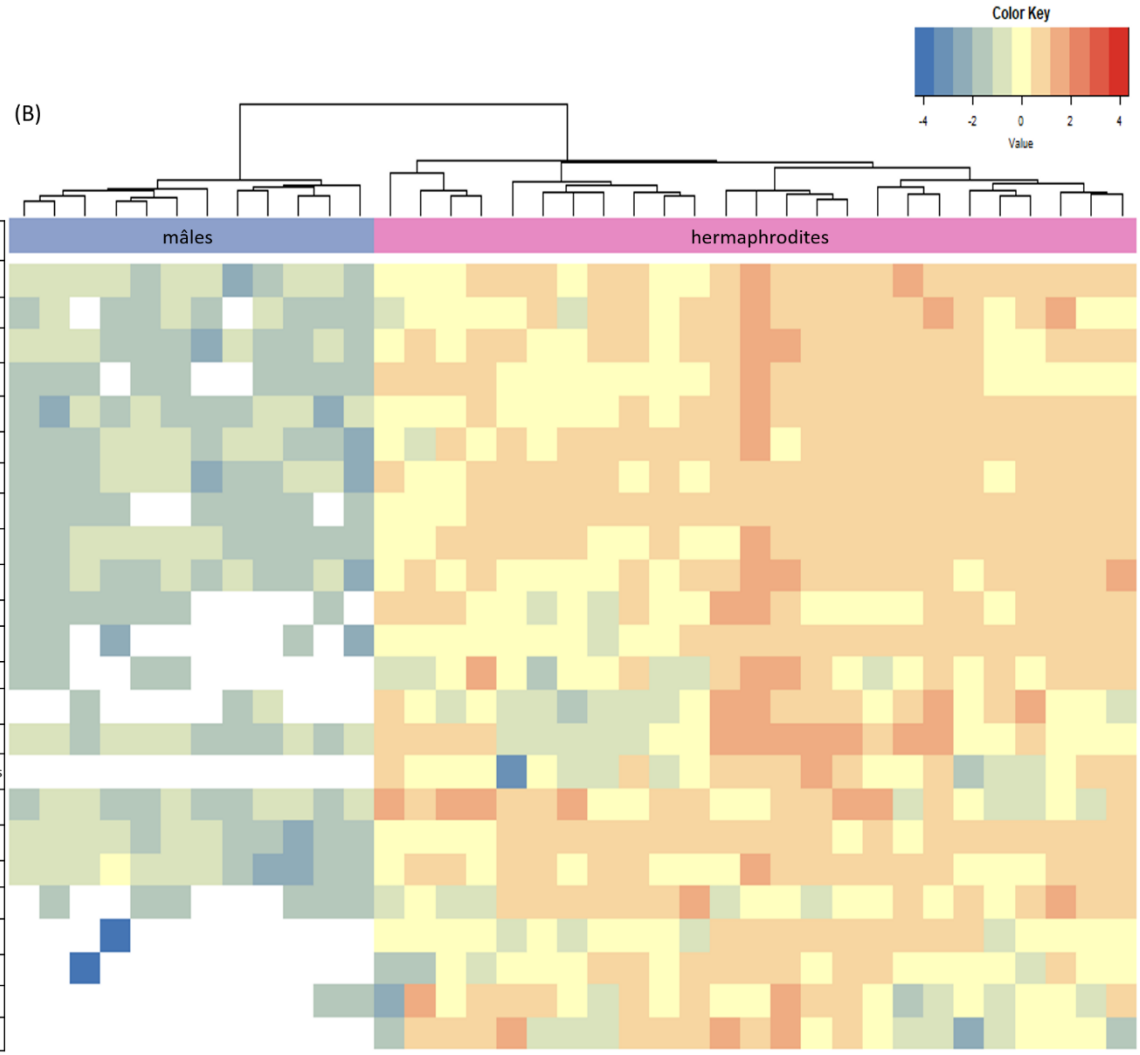


Figure 2.7 : A . Table séquence, localisation probable sur le génome de l'olivier, prédiction fonctionnelle et annotation GO, chaque ligne correspond à un unitigs dont le pattern d'expression est mis en regard sur le heatmap B. heatmap unitigs sur-exprimés chez les individus hermaphrodites (rose) par rapport aux mâles (bleu), chaque colonne représente un individu et chaque ligne un unitigs. Les couleurs représentent la divergence d'un gène particulier dans un échantillon particulier, par rapport à la valeur moyenne de ce gène sur tous les échantillons, centré réduit sur 0 en unités d'écarts types (bleu peu exprimé, jaune expression moyenne, rouge fortement exprimé).

2. L'approche de capture met en évidence des séquences candidates pour le déterminisme du sexe

En l'absence de génome de référence pour le filaire, le positionnement direct des transcrits différentiellement exprimés sur la carte génétique du filaire est impossible. Bien que la synténie entre les génomes du filaire et de l'olivier soit élevée, elle n'est pas parfaite, ce qui constitue une limite inhérente à l'approche que nous avons utilisée pour obtenir les résultats précédents. De plus la définition de la région liée au sexe sur le génome de l'olivier à partir de la cartographie effectuée dans le premier chapitre ne permet d'estimer la taille de celle-ci qu'à partir d'un seul croisement. Enfin, l'olivier étant une espèce hermaphrodite, il est possible que l'architecture génétique de cette région ne soit pas la même chez le filaire.

Pour tenter de pallier ces difficultés et densifier les informations sur des gènes potentiellement liés au sexe (de manière transcriptomique ou génomique), nous avons mis en place une expérience de capture par hybridation ciblée à partir de séquences de références que nous avons définies par deux approches. D'une part, des régions ont été définies sur le génome de l'olivier ("Ole") à partir des résultats des pré-analyses de la cartographie génétique du filaire. Les marqueurs GBS définis comme étant liés au phénotype du sexe ont été positionnés sur le génome de l'olivier, ce qui a permis de borner des intervalles sur le chromosome 12 et plusieurs scaffolds non ancrés dans ce génome (Figure 2.9 B). Nous avons défini des séquences cibles sur l'ensemble des gènes annotés contenus dans ces intervalles et répondant aux contraintes d'une capture par hybridation ciblée. D'autre part, des séquences cibles ont été définies à partir de l'ensemble des transcrits différentiellement exprimés issus de l'analyse de l'expérience de « *RNA sequencing* » sur les boutons floraux de *P. angustifolia* ("RNA"). Outre la possibilité de s'abstraire de l'hypothèse de synténie entre les deux espèces, un avantage de cette approche est que l'ensemble des transcrits différentiellement exprimés ont pu être analysés ici, incluant des transcrits spécifiques du filaire pour lesquels des orthologues n'avaient pu être identifiés dans le génome de l'olivier.

Reconstruction *de novo* des séquences cibles de références

Lorsque les séquences cibles avaient été définies à partir du génome de l'olivier (indiquées "Ole" dans les figures 2.8 et 2.10), les séquences consensus reconstruites *de novo* chez le filaire par notre pipeline intégrant Hybpiper étaient en moyenne à 89,47% identiques à la

séquence cible d'origine de l'olivier (Figure 2.8 A), ce qui correspond à l'accumulation de différences nucléotidiques depuis la séparation des deux espèces. Ces nouvelles séquences consensus correspondant à la séquence génomique du filaire permettent d'augmenter sensiblement la proportion moyenne de couverture de chaque séquence qui passe de 0,475 à 0,576 (Figure 2.8 B) et plus légèrement la profondeur moyenne (c'est-à-dire le nombre de lectures s'alignant à chaque site) qui passe de 5,57X à 6X (Figure 2.8 C).

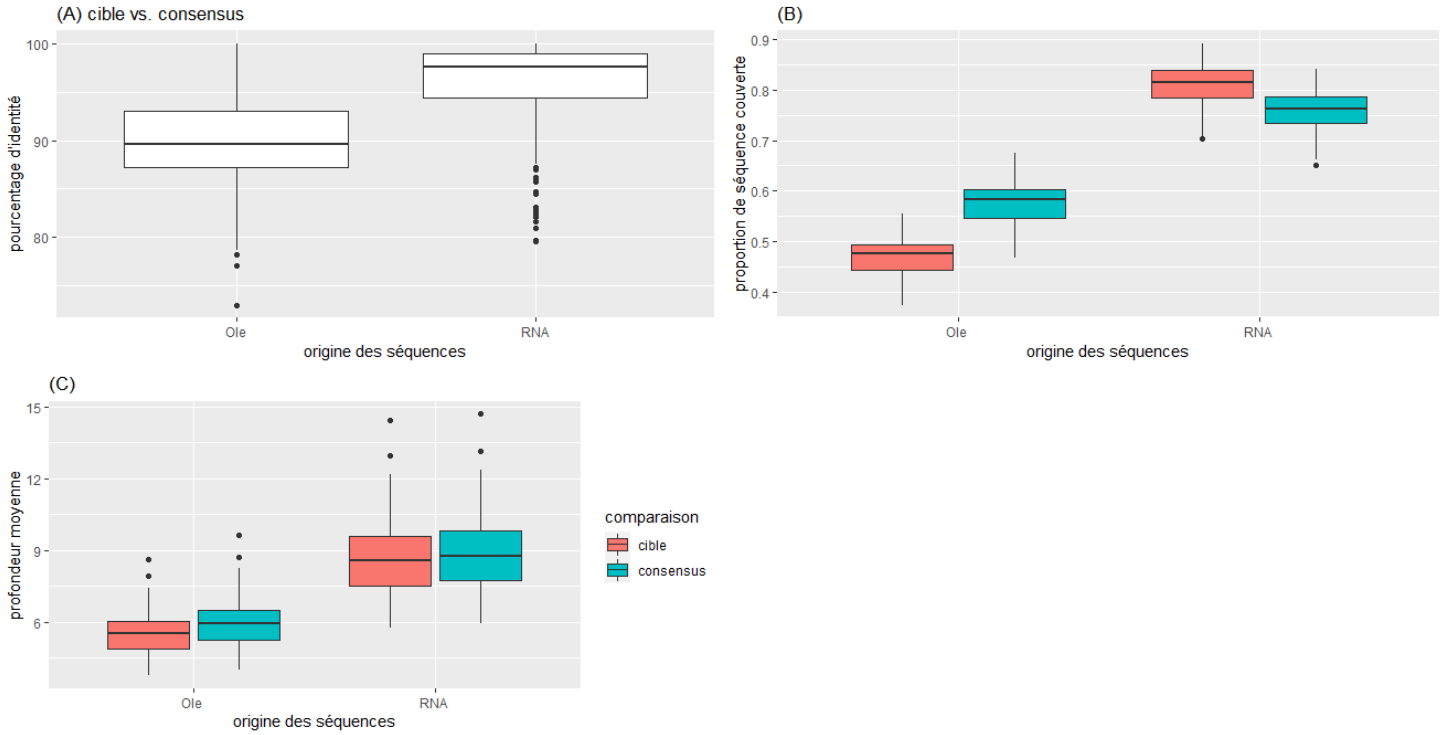


Figure 2.8 : A. distribution du pourcentage d'identité entre les séquences nucléotidiques cibles utilisées pour définir les sondes de capture et les séquences consensus reconstruites *de novo* chez le filaire en fonction de l'origine des séquences *ie.* transcriptome du filaire (RNA) ou génome de l'olivier (Ole). B. comparaison de la proportion de séquence couverte lors d'un alignement brut des lectures de séquençage sur les séquences nucléotidiques cibles utilisées pour définir les sondes de capture et les séquences consensus reconstruites *de novo* chez le filaire en fonction de l'origine des séquences. C. comparaison de la profondeur moyenne des séquences entre les séquences nucléotidiques cibles utilisées pour définir les sondes de capture et les séquences consensus reconstruites *de novo* chez le filaire en fonction de l'origine des séquences.

La tendance est différente pour les séquences cibles définies directement à partir du transcriptome du filaire. En effet, les séquences consensus reconstruites *de novo* divergent peu par rapport à la séquence d'origine (Figure 2.8 A), et peu d'amélioration était attendue ici. A l'inverse, nous avons observé que ces nouvelles séquences consensus intégraient de courtes séquences en amont et en aval des régions ciblées, qui peuvent correspondre à des séquences introniques et des portions de séquences en amont et en aval des gènes. L'alignement des lectures de reséquençage peut être difficile sur ces portions plus variables et, en conséquence, on observe une légère diminution de la proportion de couverture de

chaque séquence (Figure 2.8 B), passant en moyenne de 0,809 à 0,757. Ces variations n'ont là aussi qu'une faible incidence sur la variation de profondeur moyenne (8,72X à 8,91X (Figure 2.8 C)). Dans ce cas, l'intérêt du passage par HybPyper est donc discutable, mais dans un souci d'homogénéité du traitement des données, nous avons choisi de conserver cette étape. Globalement, ces profondeurs de séquençage restent faibles à ce stade, indiquant que l'interprétation des SNPs identifiés devra rester prudente.

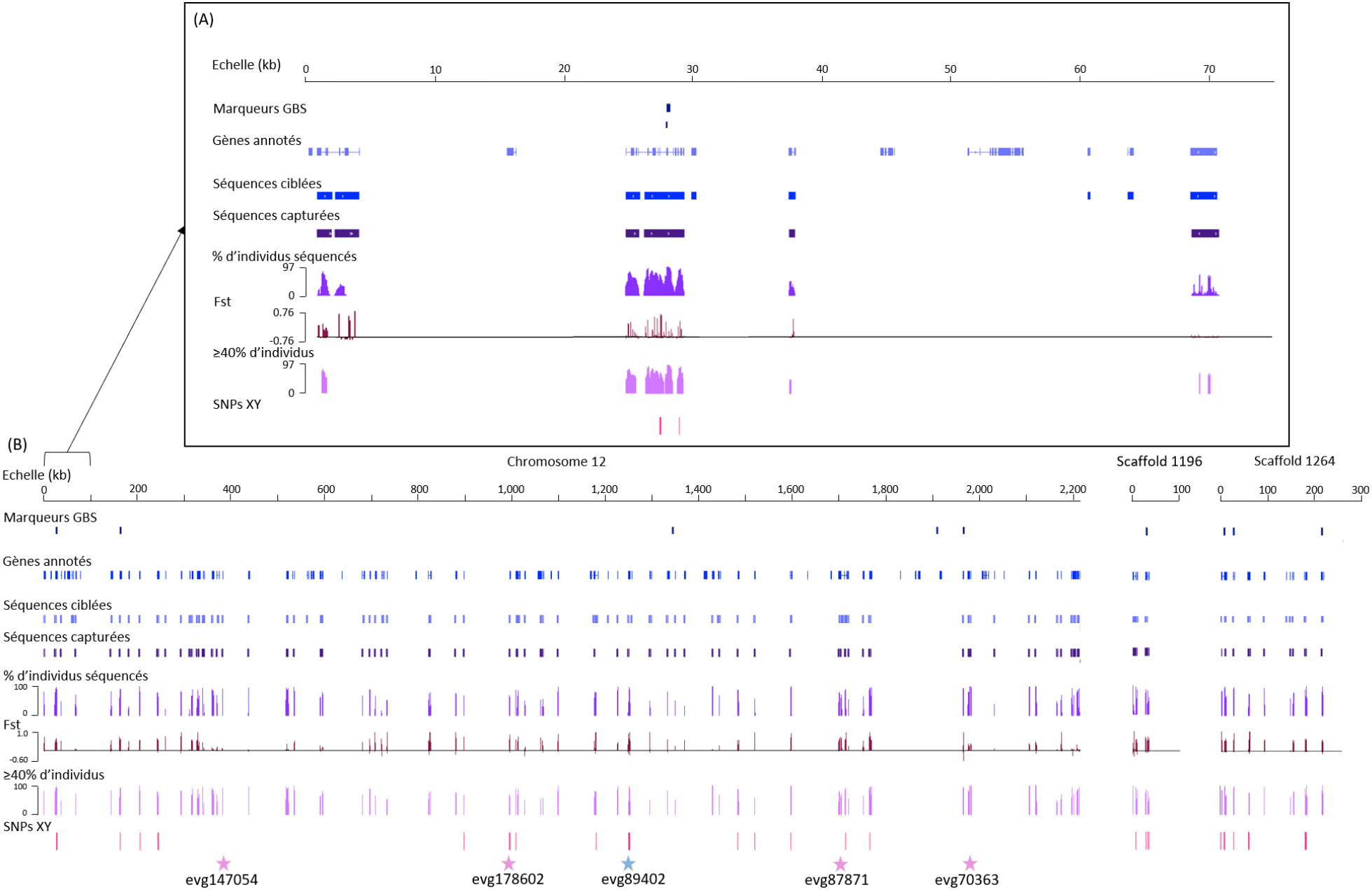

(A)
Echelle (kb)
0
10
20
30
40
50
60
70
Marqueurs GBS
Gènes annotés
Séquences ciblées
Séquences capturées
% d'individus séquencés
97
0
Fst
0.76
-0.76
≥40% d'individus
97
0
SNPs XY
(B)
Chromosome 12
Scaffold 1196
Scaffold 1264
Echelle (kb)
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
0
100
0
100
200
300
Marqueurs GBS
Gènes annotés
Séquences ciblées
Séquences capturées
% d'individus séquencés
100
0
Fst
1.0
-0.60
≥40% d'individus
100
0
SNPs XY
evg147054
evg178602
evg89402
evg87871
evg70363

Figure 2.9 : Schéma des différentes étapes de l'expérience de capture de séquence pour les séquences liées au SI définies sur le génome de l'olivier (chromosome18, scaffolds 327, 269 et 1287). L'échelle est en kb, la position des marqueurs GBS définis comme strictement liés au locus du SI est représentée sur la première ligne. Dans l'ordre sont indiquées la position des gènes annotés chez l'olivier, la position des séquences ciblées (c'est-à-dire celles pour lesquelles des sondes ont pu être définies) et la position des séquences capturées et reconstruites *de novo* par notre pipeline. En violet est indiqué le pourcentage d'individus génotypés pour chaque SNPs identifié, en rouge les valeurs de Fst pour chacun de ces SNPs et en rose le pourcentage d'individus génotypés pour les SNPs ayant au moins 40% d'individus génotypés. La position de SNPs détectés comme étant hétérozygote chez tous les individus mâles et homozygotes chez tous les individus hermaphrodites génotypés est indiqué par un trait rose dans la section « SNPs XY ». Les positions des unitigs sur-exprimés chez les individus mâles et chez les individus hermaphrodites ont été reportés par des étoiles bleues et roses respactivement.

Analyse de ségrégation des SNPs liés aux phénotypes du sexe

Tableau 2.2 : Résumé des premières étapes d'analyse de l'expérience de capture ciblé en fonction de l'origine des séquences (*i.e.* transcriptome du filaire (RNA) ou génome de l'olivier (Ole)).

	Ole	**RNA**	**Total**
Séquences ciblées	114	272	386
Séquences capturées	99	272	371
Nombre total de SNPs	6222	10932	17154
Séquences avec Fst≥0.3	66	44	110
Nombre de SNP avec Fst≥0.3	748	100	848

Nous avons ensuite examiné la répartition des SNPs identifiés dans les deux catégories de séquences ciblées ("Ole" et "RNA"). Parmi les 114 séquences (258 957pb) initialement ciblées comme pouvant être liées au déterminisme du sexe chez le filaire d'après les analyses effectuées dans le chapitre 1 ("Ole"), une majorité (n= 99) a pu être capturée et reconstruite *de novo* avec succès par notre pipeline (Table 2.2, Figure 2.9 A). Ces séquences couvrent un total de 178 454pb et ont permis d'identifier 6 222 SNPs. L'analyse des valeurs de Fst par SNPs entre les mâles et les hermaphrodites a permis de mettre en évidence 748 SNPs, répartis sur 66 séquences ayant une valeur de Fst entre mâles et hermaphrodites supérieure ou égale à 0,3 (Figure 2.9 B). Ces 748 SNPs fortement différenciés représentent 12% du total des 6 222 SNPs caractérisés, ce qui tout à la fois confirme globalement la liaison génétique de la région génomique concernée avec le phénotype du sexe, et permet d'exclure de façon relativement stringente certaines des séquences ciblées. De la même façon, l'ensemble des unitigs différentiellement exprimés entre les mâles et les hermaphrodites qui ont été ciblés (272 séquences ; 391 601pb), ont pu être capturés et reconstruits *de novo* par notre pipeline d'analyse. Les séquences consensus couvrent 398 436pb et ont permis d'identifier au total 10 932 SNPs. Seulement 100 SNPs, se répartissant dans 44 séquences, ont un Fst supérieur ou égal à 0,3. A l'inverse, ces 100 SNPs fortement différenciés ne représentent que 0,9% des 10 932 SNPs caractérisés, confirmant de façon non ambiguë que la majorité des transcrits

différentiellement exprimés ségrègent à des positions génomiques indépendantes du locus du sexe.

Sur les SNPs ayant une valeur de Fst≥0.3, un filtre final de couverture à plus de 40 % d'individus génotypés a été effectué (Figure 2.9). Comme l'illustre la Figure 2.9, de nombreux SNPs ne sont plus suffisamment couverts pour être retenus. Après observation de la ségrégation des SNPs restants, 26 séquences, couvrant un total de 69 728pb, contenant au moins un SNPs suivant strictement le modèle XY (avec les hermaphrodites homozygotes et les mâles hétérozygotes) ont pu être retenues. La représentation graphique des génotypes individuels à chaque séquence est présentée en Annexe 2.5.

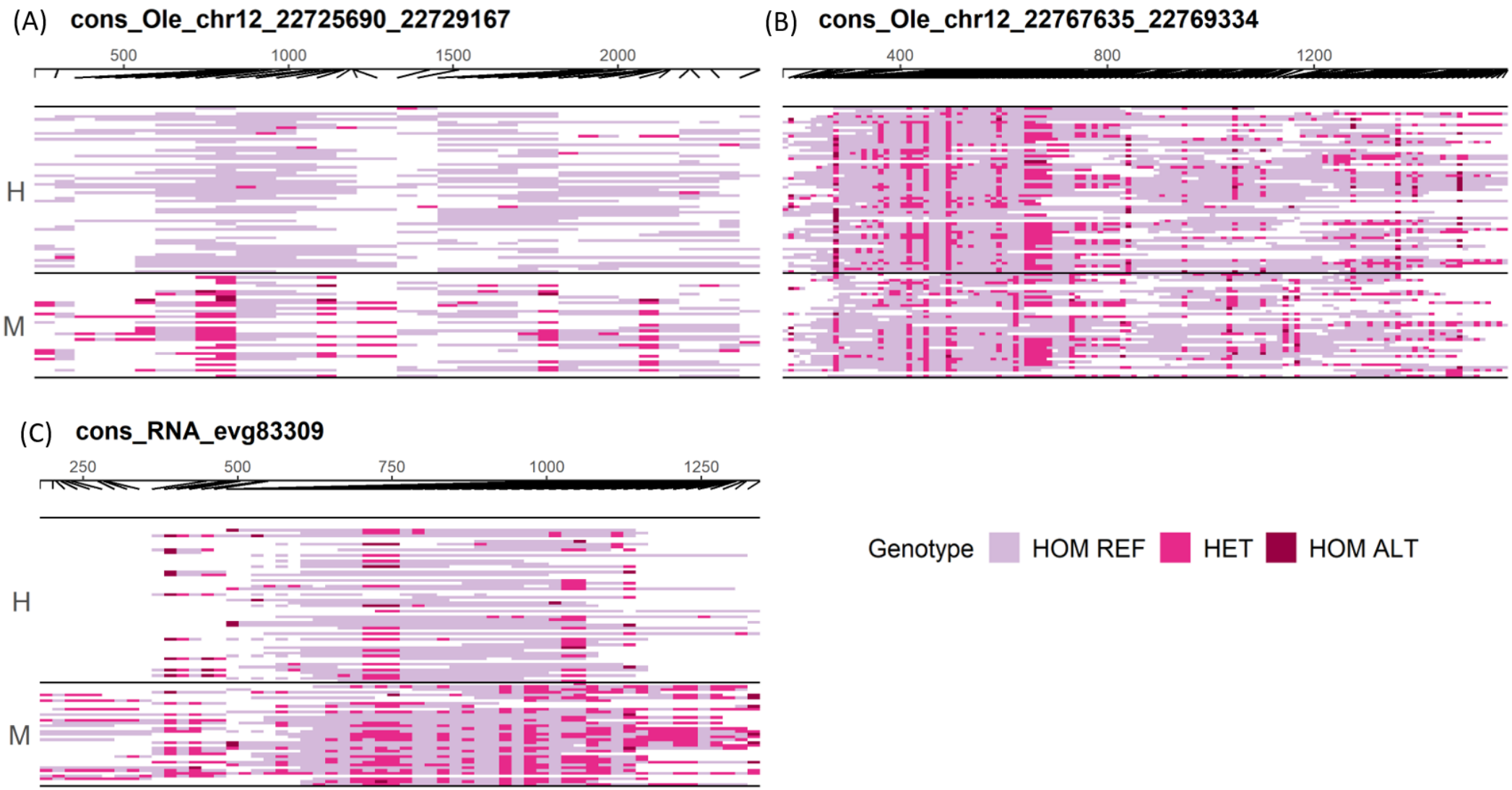


Figure 2.10 : Représentation graphique des génotypes pour 3 séquences (A, B, C) permettant d'illustrer les différents profiles obtenues lors de l'expérience de capture.

Parmi ces 26 séquences, 23 se répartissent sur les régions liées au sexe définies à partir du génome de l'olivier (Figure 2.9 B) et trois sont des séquences provenant des analyses d'expression différentielle. Trois séquences illustrant les types de profils de génotypes obtenus sont visibles dans la Figure 2.10. On y observe l'état allélique de chacun des SNPs en fonction de leur position sur la séquence (axe horizontal). Chaque individu est représenté par une ligne, et les individus sont regroupés par phénotype (axe vertical). Il est important de prendre en compte que l'absence de génotype à une position pour un individu donné peut avoir plusieurs explications. Premièrement, la stringence du filtre de couverture que nous avons utilisé pour le *SNP calling* (≥5 lectures) a pu supprimer l'information à cette position, ce qui a souvent été observé pour les séquences issues du génome de l'olivier dont la profondeur de couverture moyenne n'était que de 6X (Figure 2.10 A). Deuxièmement, cette portion de séquence a pu ne pas être capturée chez l'individu, ce qui arrive régulièrement sur les extrémités des séquences. Ce cas de figure est facilement reconnaissable, notamment sur des séquences particulièrement bien couvertes comme illustré dans la Figure 2.10 B. Enfin, une troisième possibilité est que la portion de séquence ne soit tout simplement pas présente chez l'individu en question. Dans ce cas, il faut s'intéresser à la taille et la localisation de cette portion ainsi qu'au nombre d'individus qui présentent un profil d'absence afin de ne pas la confondre avec une des deux explications précédentes. La Figure 2.10 C illustre ce cas de figure: dans la séquence evg83309, 70 SNPs se répartissent sur un total de 1 531pb. On y observe une portion de séquence absente pour tous les individus hermaphrodites sur les 9 premiers SNPs, qui correspond à une absence totale des 350 premiers nucléotides.

Les SNPs suivant un modèle de ségrégation de type XY ont une répartition non homogène le long des régions définies sur le chromosome 12 et les scaffolds 1196 et 1264 de l'olivier (Figure 2.9). Le scaffold 727 et la portion du chromosome 3 qui avaient été inclus dans l'expérience de capture ne présentent aucun SNPs strictement liés au phénotype du sexe, ce qui concorde avec les résultats définitifs de la cartographie génétique (ils avaient été inclus sur la base d'une version préliminaire de la cartographie). Les positions des transcrits sur-exprimés chez les individus mâles et chez les individus hermaphrodites sont indiquées par des étoiles bleues et roses, respectivement, sur la Figure 2.9. On observe que les unitigs evg89402 (en bleu) et evg178602 (en rose) se positionnent sur des séquences pour lesquels des SNPs strictement liés au sexe ont bien été mis en évidence. En amont de ces deux séquences, evg147054 se positionne sur une séquence dont la profondeur est satisfaisante, mais qui ne possède aucun SNP lié au sexe et ayant de très faibles valeurs de Fst. L'unitig evg87871, bien

que se positionnant dans une portion dont les séquences sont fortement associées aux phénotypes du sexe, possède une couverture insuffisante pour conclure sur la liaison stricte des SNPs. Le dernier unitig le long de cette portion de chromosome , evg70363, se positionne quant à lui sur une séquence bien couverte mais ne possédant aucun SNP liées au sexe et n' ayant que de très faibles valeurs de Fst, ce qui suggère qu'elle se trouve à l'extrémité de la région spécifique du sexe (cette hypothèse est corrélée avec la position des marqueurs GBS strictement associés au sexe dans le croisement contrôlé).

Prédiction fonctionnelle des candidats aux déterminisme du sexe

L'annotation fonctionnelle du génome de l'olivier reste relativement incomplète, et de nombreux gènes prédits ne sont associés à aucune annotation fonctionnelle. Il est cependant possible d'utiliser ces annotations partielles pour tenter de dégager de façon exploratoire des indications sur le rôle des gènes mis en évidence précédemment (Tableau 2.3).

Parmi les 26 séquences identifiées comme ayant une association génétique importante avec le phénotype du sexe, 17 séquences ont une classification GO de fonction moléculaire (Figure 2.11). Trois de ces séquences ne peuvent être classifiées que de manière généraliste (séquences 3, 4 et 5) dans un rôle de liaison protéique. Les 14 autres séquences se répartissent dans une des 5 grandes classes de fonction moléculaire : activité de transport, activité catalytique, fonction de régulation moléculaire, activité de liaison et activité de facteur de transcription. La classe la plus représentée est celle d'activité catalytique avec neuf séquences, intervenant majoritairement dans les procédés métaboliques des lipides et des glucides. Parmi les autres classes, on retrouve des séquences codant pour des rôles de transport transmembranaire (séquences 3), de liaison à l'ADN ou de facteurs de transcription (séquences 13, 14, 24 et 27). Par ailleurs, trois séquences du chromosome 12 (séquences 10, 12 et 15; Tableau 2.3) ne trouvent ni prédiction par homologie (uncharacterized protein) ni classification fonctionnelle dans les bases de données (GO ou KOG). Quatre autres séquences qui n'ont pas de classification GO, ont au contraire une prédiction fonctionnelle (séquences 1, 18, 22, 25 Table 2.3), et à l'inverse quatre autres possèdent une classification GO sans prédiction fonctionnelle (séquences 7, 11, 14 et 24 Table 2.3). Globalement, à ce stade où les candidats restent nombreux, il est difficile d'aller plus en profondeur dans l'analyse fonctionnelle.

Tableau 2.3: prédiction fonctionnelle et classification GO (Gene ontologie) des différentes séquences liées aux déterminisme du sexe. C (composant cellulaire), F (fonction moléculaire) et P (processus biologique) ou classification KOG (euKaryotic Orthologous Groups). Les séquences 1 à 23 sont identifiées par leur position sur le chromosome 12 ou un scaffold de l'olivier (Figure 2.9B). La séquence 26 présentée entre parenthèses correspond à une redondance avec la séquence 10.

Séquence	Description (Blast2GO)	GO (ou KOG) classification
1. cons_chr12_22588680_22591779	DNA-binding BIN4 isoform X1	no GO terms
2. cons_chr12_22725690_22729167	oligopeptide transporter 6	F:GO:0022857:oligopeptide transmembrane transporter activity
3. cons_chr12_22767635_22769334	pentatricopeptide repeat-containing At2g13600-like	F:GO:0003674:protein binding
4. cons_chr12_22805886_22807765	WD repeat-containing protein YMR102C-like isoform X2	F:GO:0003674:protein binding
5. cons_chr12_22808225_22809489	WD repeat-containing YMR102C-like isoform X1	F:GO:0003674:protein binding
6. cons_chr12_23442858_23446438	caffeoylshikimate esterase	KOG1455
7. cons_chr12_23559333_23560692	uncharacterized protein LOC111406130	C:GO:0005575:membrane; F:GO:0016757:glycosyltransferase activity
8. cons_chr12_23572521_23581409	5'-3' exoribonuclease 3-like isoform X1	P:GO:0006464:protein dephosphorylation; F:GO:0016791:protein tyrosine phosphatase activity
9. cons_chr12_23740799_23748128	protein C2-DOMAIN ABA-RELATED 4-like isoform X3	KOG1030
10. cons_chr12_23813672_23816896	uncharacterized protein LOC111396017 isoform X2	no GO terms
11. cons_chr12_24046784_24050139	uncharacterized protein LOC111406815	P:GO:0006629:lipid metabolic process; F:GO:0008970:phospholipase A1 activity
12. cons_chr12_24083331_24084421	uncharacterized protein LOC111406147	no GO terms
13. cons_chr12_24158884_24163726	polyadenylate-binding -interacting 11-like	F:GO:0003676:nucleic acid binding (KOG0131)
14. cons_chr12_24274400_24282295	uncharacterized protein LOC111406816	C:GO:0000127:transcription factor TFIIIC complex; F:GO:0004402:histone acetyltransferase activity; F:GO:0005515:protein binding; P:GO:0006384:transcription initiation from RNA polymerase III promoter
15. cons_chr12_24328158_24330747	uncharacterized protein LOC111406151 isoform X2	no GO terms
16. cons_sca1196_1526_9575	ADP-ribosylation factor 1-like 2	F:GO:0003924:GTPase activity; F:GO:0005525:GTP binding
17. cons_sca1196_29015_31954	serine/threonine-protein phosphatase 4 regulatory subunit 2 isoform X2	F:GO:0019888:protein phosphatase regulator activity; C:GO:0030289:protein phosphatase 4 complex
18. cons_sca1196_32214_35114	reticulon-like protein B21 isoform X3	no GO terms
19. cons_sca1264_37398_39757	hypothetical protein SASPL_130752	P:GO:0005975:carbohydrate metabolic process; F:GO:0016868:intramolecular transferase activity, phosphotransferases
20. cons_sca1264_44217_48985	hypothetical protein SADUNF_Sadunf10G0083900	
21. cons_sca1264_62982_66036	glucan endo-1,3-beta-glucosidase 3-like isoform X1	P:GO:0005975:carbohydrate metabolic process; F:GO:0016798:hydrolase activity, hydrolyzing O-glycosyl compounds
22. cons_sca1264_95262_98312	probable peroxygenase 4	no GO terms
23. cons_sca1264_215234_216903	lipid phosphate phosphatase 2-like	P:GO:0006644:phospholipid metabolic process; F:GO:0042577:lipid phosphatase activity
24. cons_RNA_evg87500	uncharacterized protein LOC113694833	F:GO:0003677:DNA binding; F:GO:0046872:metal ion binding
25. cons_RNA_evg99457	duf724 domain-containing 3	no GO terms
27. cons_RNA_evg83309	transcription factor HHO5-like	F:GO:0003677: DNA binding; F:GO:0003700:DNA-binding transcription factor activity; P:GO:0009058: regulation of transcription, DNA-templated; P:GO:0034641: cellular nitrogen compound metabolic process

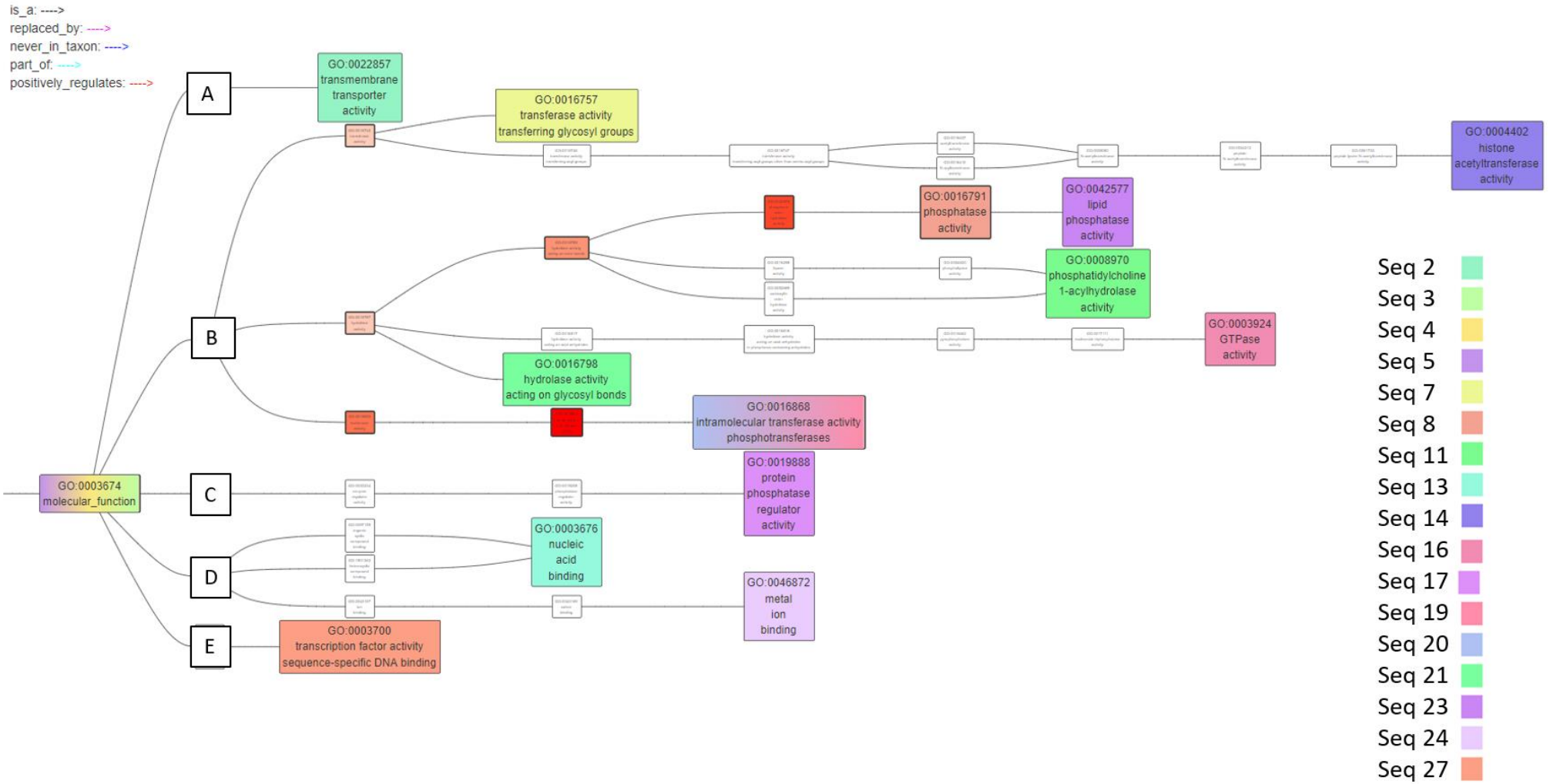


Figure 2.11 : Visualisation de la position des séquences d'intérêt possédant une annotation GO dans les grandes catégories fonctionnelles : A. activité de transport, B. activité catalytique, C. fonction de régulation moléculaire, D. activité de liaison, E. activité de facteur de transcription ; générée à l'aide de l'outils NaviGO (Wei *et al.* 2017)

Discussion

Gènes à expression sexe-biaisée dans une espèce androdioïque

Jusqu'à récemment, les études s'étaient principalement appuyées sur des analyses de ségrégation pour définir un modèle génétique pour le sexe et l'incompatibilité chez *P. angustifolia* (VASSILIADIS *et al.* 2002; SAUMITOU-LAPRADE *et al.* 2010; HUSSE *et al.* 2013; BILLIARD *et al.* 2015). À partir de la première cartographie génétique de l'espèce androdioïque *P. angustifolia*, nous avons mis en évidence que le locus du sexe, est localisé sur le LG12 et suit une ségrégation génétique de type XY ou hémizygote (CARRE *et al.* 2021). La première approche développée dans ce chapitre nous a permis de reconstruire *de novo* le premier transcriptome de référence du filaire, et ainsi de déterminer si l'expression de certains transcrits était structurée par sexes.

De nombreuses espèces ont des sexes séparés, et plusieurs études ont cherché à identifier les gènes montrant des différences d'expression entre morphes sexuels. Dans la plupart des cas, les individus montrent une spécialisation sexuelle complète et les morphes sexuels correspondent à des mâles et des femelles. Chez l'humain par exemple, Shen *et al.* (2017) ont montré que jusqu'à 3,8% des gènes étaient différentiellement exprimés entre hommes et femmes. La comparaison des niveaux d'expression des transcrits entre les individus mâles et les individus hermaphrodites chez le filaire a mis en évidence de nombreux unitigs différentiellement exprimés. Ainsi, 0,15% et 0,45% des unitigs sont surexprimés chez les mâles et chez les hermaphrodites respectivement, soit un total de 0,6% de gènes différentiellement exprimés. Nos résultats sont similaires à ce qui est observé chez les espèces dioïques de *Leucadendron* ayant les plus faibles proportions de gènes différentiellement exprimés (de 0,1% jusqu'à 2,5% selon les espèces) entre sexe (SCHARMANN *et al.* 2021) mais largement inférieurs aux valeurs rapportées chez *Mercurialis annua* (COSSARD *et al.* 2019) et *Silene latifolia* (ZEMP *et al.* 2016), où l'expression d'environ 2% des gènes diffère selon les sexes. Outre la difficulté à comparer des résultats d'études ayant utilisé des seuils de détection potentiellement différents, ces dernières études ont été effectuées sur les tissus somatiques (feuilles) d'espèces dioïques, contrairement à notre analyse sur *P. angustifolia* qui a été effectuée sur des boutons floraux à différents stade de maturation. Dans notre d'espèce androdioïque les hermaphrodites portent des tissus reproducteurs mâles parfaitement fonctionnels, et seule la présence d'un stigmate les différencient d'un mâle (qui possède par ailleurs des ovaires). Il est donc cohérent que les proportions d'unitigs à expression sexe-

biaisée soit faibles, et ce particulièrement chez les mâles. Les gènes à expression sexe-biaisés montrent généralement des taux d'évolution de leur séquence codante et de leurs niveaux d'expression particulièrement élevés par rapport au reste du génome, ce qui est généralement interprété comme l'effet de la sélection sexuelle (HARRISON *et al.* 2015). Il pourrait être intéressant de réaliser cette analyse dans le cas du filaire, où la différenciation sexuelle est partielle chez les hermaphrodites, et non complète comme dans le cas plus classique des espèces dioïques.

Seule une poignée de gènes à expression sex-biaisée est génétiquement liée au sexe

En profitant de la bonne synténie globale entre l'olivier et le filaire, nous avons pu déterminer que cinq unitigs (evg89402, evg147054, evg178602, evg87871 et evg70363) différentiellement exprimés entre les sexes co-localisaient avec la région identifiée comme étant liée au sexe. Cette liste n'est pas exhaustive, car à ce stade environ un tiers des unitigs se positionnaient sur des scaffolds non ancrés dans l'assemblage principal de l'olivier, et jusqu'à 15% des unitigs n'ont pas pu être replacés sur le génome de l'olivier. Cette localisation au sein de la région associée au sexe peut traduire un rôle direct de ces gènes dans le déclenchement du déterminisme du sexe, mais à ce stade, sur la base des annotations fonctionnelles disponibles, trois de ces unitigs (evg89402, evg147054 et evg70363) n'ont pas d'annotation, tandis que les deux autres (evg178602 et evg87871) joueraient un rôle dans la croissance cellulaire. Sur cette base, il reste difficile de porter des conclusions plus détaillées à ce stade quant aux fonctions biologiques qui pourraient être impliquées.

La vaste majorité des gènes à expression sexe-biaisée ne semble donc pas être liée génétiquement aux déterminants des phénotypes sexuels. Cette observation est cohérente avec d'autres approches d'analyse transcriptomique entre sexes, qui ont montré que si une partie des gènes sexe-biaisés était bien associée aux chromosomes sexuels, une majorité a tendance à être associée aux autosomes (ZEMP *et al.* 2016; PALMER *et al.* 2019). Ces gènes autosomaux différentiellement exprimés entre sexes peuvent participer à la mise en œuvre développementale du dimorphisme sexuel sous le contrôle des gènes génétiquement liés au sexe.

Afin d'augmenter la quantité d'information sur les unitigs issus de l'analyse transcriptomique comparative et sur les régions génomiques mises en évidence dans la cartographie, nous avons mis en place une expérience de capture de gènes ciblés.

Globalement, cette approche nous a permis de réduire de moitié la taille des régions liées au sexe positionnées sur le chromosome 12 et sur plusieurs scaffolds de l'olivier. Cependant, pour une partie des séquences l'information a été masquée par les filtres de couverture appliqués lors des analyses. Une profondeur de séquençage plus importante est à envisager pour les expériences suivantes.

Elargissement: plus d'individus, plus de populations, plus d'espèces

A ce stade, le nombre de séquence liées au sexe chez *Phillyrea* reste importante et l'arrangement de ces séquences entre elles n'est pas connu. Afin de restreindre encore le nombre de séquences chez le filaire, il est envisageable d'introduire plus d'individus dans les analyses de capture de gènes. Une première série d'individus à inclure regrouperait les individus de filaire du croisement de cartographie (chapitre 1) afin de procéder à une cartographie en bonne et due forme des SNPs identifiés sur les séquences capturées. Par ailleurs, il pourrait être intéressant d'étendre l'analyse à des individus d'autres populations afin d'évaluer la robustesse des associations identifiées. Les populations espagnoles seraient particulièrement pertinentes à intégrer. En effet, les individus mâles de ces populations semblent plus différenciés sexuellement que ceux des populations françaises. Pour rappel, les mâles utilisés pour nos expériences ont une absence de stigmate mais ne présentent pas d'altération du reste du gynécée: les ovaires contiennent des ovules qui semble normaux tandis que les individus espagnols auxquels nous avons eu accès (une population de Cadix) ne possèdent qu'un gynécée très réduit (ni stigmates ni ovaire ne sont observés), il est donc envisageable que ces populations soient plus avancées dans la transition vers des sexes séparés phénotypiquement et génétiquement.

Nous avons développé le pipeline de reconstruction *de novo* pour les séquences capturées par hybridations ciblées dans le but de pouvoir par la suite l'appliquer à d'autres espèces et ainsi étudier l'évolution des régions associées au sexe dans la famille des oléacées. En corrigeant certaines des limites de l'approche tel que la couverture lors du séquençage afin d'atteindre une profondeur moyenne d'au moins 15X, il est envisageable de pouvoir faire une étude phylogénique de cette région en introduisant des individus d'espèces hermaphrodites (ex. *Ligustrum vulgare*, *Olea europaea*), androdioïques (ex. *P. latifolia*, *Fraxinus ornus*), polygames (ex. *Fraxinus excelsior*) et dioïques (ex. Fraxinus chinensis).

Un génome pour le filaire

Les régions associées au déterminisme du sexe sont connues pour être de larges portions chromosomiques non-recombinantes (CAREY *et al.* 2021). Afin de mieux comprendre l'architecture génétique de la région liée au sexe chez le filaire et d'étudier l'évolution de cette région au sein de la famille, des génomes de référence permettraient d'avoir une meilleure information sur l'arrangement des gènes entre eux. De plus, pouvoir replacer sur le génome du filaire les séquences pour le moment positionnées sur des scaffold chez l'olivier serait une étape majeure dans cette étude. Depuis récemment, un nouvel assemblage du génome de l'olivier est disponible (JIMENEZ-RUIZ *et al.* 2020) ainsi qu'un génome de référence pour *Fraxinus excelsior* (SOLLARS *et al.* 2017) et *Fraxinus ornus* (KELLY *et al.* 2020). A ce jour cependant, un génome de *P. angustifolia* n'est toujours pas disponible. Le séquençage complet du génome du double hétérozygote (mM/S1S2), ayant servi de père au croisement contrôlé de la population utilisée pour la cartographie génétique haute densité du filaire (CARRE *et al.* 2021), est actuellement en cours par une collaboration de l'équipe. Etant donnée la complexité attendue des deux régions concernées (sexe et auto-incompatibilité) et l'ampleur des réarrangements structuraux suspectée, cet assemblage devra se faire sur la base d'une méthode de séquençage produisant des lectures de longs fragments. Si l'assemblage des régions d'intérêt de cet individu s'avère compliqué, il est envisagé de faire un séquençage complet du génome d'un individu hermaphrodite double homozygote (mm/S1S1). La disponibilité prochaine de cette nouvelle ressource pourrait permettre de lever de nombreuses incertitudes, notamment quant à la position génomique réelle des unitigs d'intérêt et l'arrangement des séquences liées au sexe entre elles chez le filaire.

Comment détecter un fonctionnement en hémizygotie ?

L'information de structure globale du génome chez *Phillyrea* pourrait nous permettre de statuer sur l'éventuelle fonctionnement en hémizygotie du sexe chez cette espèce. En effet, dans notre étude de nombreuses limites s'opposaient à la détection d'un système hémizygote. Tout d'abord dans l'analyse transcriptomique, la détection d'un transcrit totalement spécifique d'un phénotype, bien que possible, nécessitait que l'extraction de l'ARN soit faite à l'instant propice pour chacun des individus. En effet dans le cas d'un gène totalement spécifique une expression de courte durée même de faible intensité peut être suffisante pour avoir un impact majeur sur le phénotype. Le fait d'avoir mélangé des boutons floraux à

plusieurs stades de développement pour chacun des individus a pu créer un effet de dilution si l'expression du/des gènes se faisai(en)t à un seul de ces stades. Une séparation par stade de développement et/ou éventuellement par tissus (dans notre cas en ne gardant que le gynécée composé des ovaires, du style et du stigmate chez les hermaphrodites et seulement des ovaires et d'un style avorté chez les mâles) aurait pu permettre d'être plus précis dans la détection de transcrits spécifiques (ZHOU *et al.* 2015; PEI *et al.* 2017). La deuxième limite à la détection de séquence hémizygote dans notre expérience est le passage par le génome de l'olivier pour capturer les séquences codantes potentiellement liées au sexe. Bien que l'utilisation du génome de l'olivier soit parfaitement légitime à la vue de la proximité phylogénique des deux espèces (30 to 40 Myrs ago BESNARD *et al.* 2009; OLOFSSON *et al.* 2019), l'olivier est une espèce hermaphrodite qui ne possède donc pas de région déterminante du sexe comme elle pourrait être présente chez une espèce présentant un dimorphisme sexuel entre individus. La comparaison des assemblages de génomes d'individus mâles et hermaphrodites pourrait permettre de repérer des fragments chromosomiques spécifiques de l'un ou l'autre sexe.

Contributions

Tous les auteurs ont contribué à l'étude présentée dans ce chapitre. PS-L et PV ont développé, conçu et supervisé l'étude de séquençage de transcrits ; ils ont coordonné et réalisé les tests de phénotypage du sexe. SS a effectué l'extraction d'ARN, la préparation des bibliothèques et le séquençage du transcriptome. AC a construit le pipeline d'analyse transcriptomique sous la supervision de CM. AC, PS-L et VC ont conçu le design de l'expérience de capture de gène. AC et CG ont effectué les constructions des banques de capture à partir des extractions d'ADN effectuées par CG et SS. AC et SG ont construit le pipeline d'analyse des données de capture et AC, PS-L et VC ont interprété les résultats et rédigé le manuscrit.

Recherche des différences transcriptomiques et génomiques entre les deux groupes d'auto-incompatibilité : une étape vers la compréhension du DSI chez *Phillyrea angustifolia*

Amélie Carré[1], Sophie Gallina[1], Sylvain Santoni[2], Philippe Vernet[1], Cécile Godé[1], Clément Mazoyer[1], Vincent Castric[1], Pierre Saumitou-Laprade[1]

[1] CNRS, Univ. Lille, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France

[2] UMR DIAPC - Diversité et adaptation des plantes cultivées

Résumé

Le déterminisme génétique et les facteurs contrôlant l'évolution des systèmes d'appariement et sexués représentent des enjeux majeurs en biologie évolutive et en génomique. Dans la famille des Oleaceae, l'espèce androdioique *Phillyrea angustifolia* représente un organisme modèle passionnant, car l'évolution et le maintien de l'androdioécie sans contrainte sur la fonction mâle des hermaphrodites représente un véritable paradoxe. Une première réponse a été apportée par la mise en évidence d'un système homomorphe d'auto-incompatibilité diallélique (DSI) qui sépare les hermaphrodites en deux groupes (Ha et Hb). Ce système est mystérieux car les systèmes d'auto-incompatibilité sont généralement très hautement multi-alléliques, car les allèles rares sont avantageux, ce qui favorise d'habitude l'émergence de nouveaux allèles. Dans ce chapitre nous avons développé deux approches afin d'étudier les différences transcriptomiques et génomiques existantes entre les individus Ha et les individus Hb chez le filaire. La première approche par séquençage de transcrits sur des boutons floraux nous a permis de mettre en évidence que 0,09% des transcrits présentaient une expression différentielle entre les groupes. En couplant ces résultats à ceux obtenus lors de la cartographie génétique haute densité nous avons établi une liste de séquences que nous avons capturées par hybridation ciblée chez 59 individus de filaires. L'analyse de la ségrégation des SNPs dans ces séquences nous a permis d'exclure la liaison génétique avec les phénotypes d'auto-incompatibilité pour la majorité des séquences ciblées. La région liée au SI, s'avère être extrêmement complexe, beaucoup de séquences n'ont pu être capturées dans l'approche d'hybridation ciblée. Nous discutons du fait que certains transcrits ayant une expression spécifique à chacun des groupes de compatibilité restent des candidats intéressants pour comprendre la cascade moléculaire intervenant dans le déterminisme du phénotype d'auto-incompatibilité.

Introduction

Chez les plantes à fleurs, une manière de privilégier l'allogamie dépend de la distribution et de la fonction des structures morphologiques productrices de gamètes sur et entre les individus, autrement dit des systèmes sexuels (DIGGLE *et al.* 2011). Cependant, 75% des espèces d'angiospermes possèdent simultanément les organes mâles et femelles dans la même structure, la fleur hermaphrodite (GAUDE AND CABRILLAC 2001), ce qui rend la possibilité d'autofécondation particulièrement forte. Or, chez de nombreuses espèces l'autofécondation induit une consanguinité élevée qui peut s'avérer néfaste, et des stratégies de reproduction constituant des barrières à l'autofécondation ont évolué chez les champignons, les plantes à fleurs, les algues vertes et les oomycètes (BAUMANN *et al.* 2000; BILLIARD *et al.* 2012 ; DUSSERT *et al.* 2020). Parmi ces mécanismes, les systèmes d'auto-incompatibilité (en anglais "self-incompatibility", SI) sont définis comme l'inaptitude pour une plante hermaphrodite fertile de produire un zygote par autofécondation (LUNDQVIST 1956; DE NETTANCOURT 1977). On estime à plus de 50% la proportion de familles d'angiospermes possédant un système SI (GAUDE AND CABRILLAC 2001) et pour qui ces systèmes auraient évolué de manière indépendante au moins 35 fois dans l'histoire des angiospermes (BORIS IGIC *et al.* 2008; IWANO AND TAKAYAMA 2012). Compte tenu de l'importance fondamentale des types d'appariement dans les cycles de vie et l'évolution, leur déterminisme moléculaire a été largement étudié, en particulier chez les plantes et les champignons. Les types d'accouplement sont contrôlés par des mécanismes qui peuvent être différents, même au sein d'une même famille (BILLIARD *et al.* 2012; FUJII *et al.* 2016; DUSSERT *et al.* 2020).

Ces mécanismes d'évitement de l'autofécondation sont parfois associés à des variations phénotypiques observables chez les partenaires réciproquement compatibles. Ces variations peuvent être morphologiques, comme dans le cas d'espèces distyles où les SI sont qualifiés d'hétéromorphe, mais dans de nombreux cas les fleurs ne présentent aucune variation morphologique ou phénologique évidente, et l'évitement de l'autofécondation repose sur ce qu'on appelle les SI homomorphes. Ces systèmes reposent généralement sur la ségrégation d'un nombre fini de « spécificités » de reconnaissance par lesquelles les pollinisations entre individus exprimant des spécificités identiques ne produisent pas de zygote (KUBO *et al.* 2015). Au niveau génétique, les SI homomorphes sont généralement sous le contrôle d'un seul locus S multi-allélique contenant au moins deux gènes, l'un codant le déterminant mâle porté par le pollen et l'autre codant le déterminant femelle exprimé par les pistils. La spécificité mâle

est parfois déterminée par une série de paralogues disposés en tandem (KUBO *et al.* 2015). Les déterminants mâles et femelles sont hérités comme une seule unité génétique non recombinante (CASTRIC AND VEKEMANS 2004). Classiquement, les SI homomorphes sont classés en deux types sur la base des modes de contrôle génétique de la spécificité mâle. Dans les SI gamétophytiques (GSI), le pollen haploïde détermine lui-même la spécificité S (par exemple chez les Papaveraceae et Solanaceae). En revanche, dans les SI sporophytiques (SSI), le génotype des tissus donneurs diploïdes détermine la spécificité S du pollen (par exemple chez les Brassicaceae). Par conséquent, dans les SSI, les interactions alléliques (dominance/récessivité) sont d'une importance critique pour la détermination de la spécificité S (DURAND *et al.* 2014). A noter que les SI peuvent également différer dans leur architecture génétique. Dans la famille des Poacées par exemple, deux locus indépendants (nommés S et Z) contrôlent le SI (YANG *et al.* 2008). Dans la plupart des cas, les SI se caractérisent par une diversité allélique très importante (CASTRIC AND VEKEMANS 2004), de nombreuses spécificités alléliques ségrégeant au sein des populations naturelles (p.e. jusqu'à 54 spécificités chez A. halleri,GENETE *et al.* 2020). Cette diversité allélique importante est la résultante du modèle de sélection qui s'applique de façon générale aux gènes déterminant l'auto-incompatibilité (WRIGHT 1939) qui entraîne le fait que les gènes qui gouvernent le SI devraient faire l'objet d'une sélection naturelle de forte intensité, de type fréquence-dépendante négative, agissant sur la fonction mâle.

Il existe donc de nombreux SI homomorphes aux fonctionnements moléculaires et aux déterminismes génétiques variés dont le point commun est le polymorphisme des spécificités S. Il existe des exceptions notables à cette règle générale et, chez certaines espèces, seules deux spécificités S semblent ségréger de manière stable. Le plus souvent dans de tels SI dialléliques, les deux spécificités S sont en parfaite association avec des phénotypes floraux morphologiquement distinguables. Chez les espèces distyles par exemple, deux formes florales appelées « pin » (morphe L pour "long style") et « thrum » (morphe S pour "short style") coexistent (BARRETT 1992; BARRETT 2019). Dans chaque morphe, les anthères et le stigmate sont spatialement séparés à l'intérieur de la fleur, et les fleurs de morphologie S produisent moins de pollen mais des grains plus gros que les fleurs de morphologie L (DULBERGER 1992). Ces différences morphologiques sont généralement interprétées comme améliorant l'évitement de l'autofécondation conférée par le système SI. Chez l'espèce hétérostyle *Primula vulgaris* les spécificités alléliques sont déterminées par des variants de présence-absence d'un fragment chromosomique plutôt que par des variants de séquence

nucléotidique d'un gène donné. Ce fragment de 278kb, contient cinq gènes totalement liés formant un « supergène » qui contrôle les tailles du style, du filet des anthères et des grains de pollen et très probablement, bien que le déterminant moléculaire n'ait pas été déterminé à ce jour, la spécificité SI. Ce fragment est uniquement présent chez les individus de morphes thrum et les individus pin sont homozygotes pour son absence. En d'autres termes, le phénotype reproducteur S est hémizygote plutôt qu'hétérozygote pour le locus S tandis que phénotype L ne possède pas le fragment chromosomique en question (LI *et al.* 2016).

L'auto-incompatibilité dans la famille des Oléacées pose un ensemble de questions spécifiques non résolues à ce jour. Une partie des espèces de cette famille est hétérostyle et possède un système d'auto-incompatibilité en lien avec les différents morphes floraux, et ces espèces représentent probablement l'état ancestral au sein de la famille. Par exemple chez *Jasminum fruticans*, l'autofécondation et la fécondation intra-morphe est impossible (DOMMEE *et al.* 1992), ce qui suggère qu'il existe un système moléculaire de reconnaissance des gamètes en plus de la barrière morphologique. La branche dérivée de cette tribu est marquée à sa base par un évènement d'allo-tétraploïdisation associé à une perte de l'entomophilie (TAYLOR 1945; WALLANDER AND ALBERT 2000; UNVER *et al.* 2017). Chez les espèces issues de cet événement d'allo-tétraploïdisation et phylogénétiquement très différenciées (OLOFSSON *et al.* 2019), on observe une grande diversité de systèmes sexuels avec la présence à des fréquences élevées de systèmes habituellement rares dans le reste du règne végétal. Enfin, chez les espèces *Phillyrea angustifolia* et *Fraxinus ornus,* toutes deux androdioïques (DOMMEE *et al.* 1999), chez *Fraxinus excelsior*, une espèce trioïque (ALBERT *et al.* 2013), chez *O. europea (l'olivier) et Ligustrum vulgare (le troène),* deux espèces hermaphrodites, un système SI inhabituel a été mis en évidence. En effet, il a été montré que ces espèces possédaient un système SI homomorphe sporophytique et diallélique (DSI : diallelic self-incompatibility) où la spécificité de reconnaissance des deux allèles S1 et S2 a été conservée entre le frêne, l'olivier et le filaire d'une part (SAUMITOU-LAPRADE *et al.* 2010; VERNET *et al.* 2016; SAUMITOU-LAPRADE *et al.* 2017) et entre le troène et le lilas d'autre part (Vernet et Saumitou-Laprade com. Pers). L'existence de ce DSI a permis de résoudre l'énigme du maintien d'une fréquence élevée de mâles dans les populations naturelles de filaire (SAUMITOU-LAPRADE *et al.* 2010). Sur la base d'observations de descendances de croisements contrôlés, Billiard *et al.* (2015) ont proposé un modèle génétique rendant compte des relations complexes entre locus S et locus du sexe. Les hermaphrodites constituent deux groupes: un premier groupe d'hermaphrodites Ha serait homozygote S1S1 et un second groupe d'hermaphrodites Hb serait hétérozygote S1S2 (avec

S2 dominant sur S1). Selon ce modèle les hermaphrodites seraient incompatibles au sein de chaque groupe mais compatibles entre groupes. En revanche les mâles transmettent les allèles d'auto-incompatibilité à leur descendance mais ne les expriment pas, au sens où ils sont compatibles avec les deux groupes d'hermaphrodites quel que soit leur propre génotype au locus S.

Dans le chapitre 1 (CARRE *et al.* 2021), la cartographie génétique haute-densité que nous avons réalisée sur *P. angustifolia* a permis de valider l'hypothèse de l'indépendance des locus du sexe et de l'auto-incompatibilité proposée par Billard *et al.* (2015). Nous avons montré que le locus S, localisé sur le LG18, suivait une ségrégation génétique de type XY, selon laquelle les hermaphrodites Hb sont hétérozygotes (ou hémizygotes) et les hermaphrodites Ha sont homozygotes pour un ensemble de marqueurs génétiques au sein de cette région génomique. La synténie avec l'olivier a permis de montrer que nos résultats concordent avec les analyses sur le SI de l'olivier faites par Mariotti et *al.* (2020). Cependant, il a aussi été mis en évidence que cette région était complexe et probablement remaniée depuis la divergence entre l'olivier et le filaire (CARRE *et al.* 2021).

Ce dernier chapitre sera consacré à l'étude de l'auto-incompatibilité chez les hermaphrodites (Ha vs Hb). Tout d'abord nous avons mis en place une approche d'analyse transcriptomique de l'expression différentielle entre les Ha et les Hb, afin d'identifier des transcrits à l'origine du -ou jouant un rôle dans le- déterminisme des phénotypes d'auto-incompatibilité. Nous avons ensuite fait l'hypothèse que la bonne synténie globale avec l'olivier (23 groupes de liaison correspondant aux 23 chromosomes), pourrait nous permettre de déterminer si les gènes différentiellement exprimés co-localisent avec la position du locus du SI établie grâce à la cartographie de *P. angustifolia* (CARRE *et al.* 2021). Sur la même trame que celle que nous avons suivie dans le chapitre 2, nous avons alors tenté d'augmenter la quantité d'information sur les régions d'intérêt en mettant en œuvre une approche de capture de séquences par hybridation ciblée. Cette approche combinée a permis d'identifier un ensemble de transcrits différentiellement exprimés entre les spécificités Ha et Hb, mais elle s'est avérée globalement plus difficile que pour l'étude du sexe (chapitre 2), possiblement en raison de l'ampleur des remaniements chromosomiques de la région et de la complexité apparente de la région liée au SI.

Matériel et méthodes

1. Analyse transcriptomique

Analyse d'expression différentielle

L'intégralité du matériel biologique, la méthode de collecte des données de séquençage d'ARN ainsi que le pipeline d'assemblage *de novo* ont été détaillés dans le chapitre précédent. Dans ce chapitre, seuls les individus dont on connaissait le phénotype d'incompatibilité, c'est-à-dire les 12 individus Ha et 13 individus Hb ont été pris en compte. Comme dans le chapitre précédent, l'analyse d'expression différentielle entre les Ha et les Hb a été faite à l'aide du pipeline de quasi-vraisemblance d'*EdgeR* (ROBINSON *et al.* 2010). Les deux groupes d'hermaphrodites ne différant théoriquement que par leur phénotype d'incompatibilité, nous avons choisi une valeur du paramètre DE *fold change* de 1,2 et un FDR de 5%, afin de pouvoir sélectionner les transcrits à intégrer à l'expérience de capture de gènes ciblés. Une dernière étape de tri a été effectuée, afin de ne garder que les unitigs sur-exprimés chez 75% des individus du groupe d'intérêt.

Nous avons ensuite cherché à identifier les propriétés fonctionnelles des gènes les plus susceptibles d'intervenir dans la détermination du phénotype d'incompatibilité, en particulier des unitigs dont l'expression ou la surexpression est spécifique d'un phénotype. Nous nous sommes alors concentrés sur les unitigs différentiellement exprimés dont la valeur de Log-fold-change était la plus extrême (pour un FDR≤0.001).

Positionnement sur le génome de l'olivier des unitigs différentiellement exprimés chez le filaire

Nous avons utilisé le logiciel BLAST (KENT 2002) pour positionner les unitigs différentiellement exprimés et déterminer leur position sur le génome d'*Olea europaea* var. *sylvestris* (UNVER *et al.* 2017) par rapport à la région orthologue liée au SI du LG18 chez *P. angustifolia* (CARRE *et al.* 2021). Pour les locus présents sur le LG18 seuls ceux ayant un hit unique avec au moins 85% d'identité sur un minimum de 110 pb ont été sélectionnés pour l'analyse de synténie. Pour les unitigs d'intérêt issus de l'analyse d'expression différentielle, leur positionnement sur le génome de l'olivier a été effectué en choisissant la localisation du meilleur hit avec au moins 85% d'identité sur un minimum de 300 pb. Les relations de synténie ont ensuite été visualisées avec l'outil circos-0.69-6 (KRZYWINSKI *et al.* 2009).

2. Capture de séquences par hybridation ciblée

Matériel biologique et origines des ADN :

Pour rappel, l'expérience de capture de séquences a porté sur un total de 95 individus (Ha, Hb et M) d'origines différentes. Pour l'étude du SI, seuls les individus dont le groupe de compatibilité était connu ont été pris en compte. Ainsi, la mère hermaphrodite Ha [S1S1] et le père mâle [S1S2] du croisement ayant servi à l'établissement de la cartographie génétique ainsi que 36 de leur descendants (18 Ha, 18 Hb) ont été conservés (CARRE *et al.* 2021). Vingt hermaphrodites (11 Ha et 10 Hb) de la population naturelle de Fabrègue (la population d'origine des deux parents du croisement) ont été analysés pour déterminer la robustesse de la liaison génétique dans une base génétique élargie.

Définition des sondes de captures

Grâce à la synténie entre *P. angustifolia* et *O. europaea* (var. *sylvestris*) (CARRE *et al.* 2021), il a été possible de définir, une série de séquences cibles à capturer chez le filaire. La région associée au SI (LG18) chez le filaire définit, sur le génome de l'olivier, une région de 741 403 pb sur le chromosome 18. A l'aide des annotations disponibles sur le génome de l'olivier, l'intégralité des séquences codantes de cet intervalle ont été sélectionnées comme cibles pour définir des sondes de capture. Par ailleurs, plusieurs marqueurs associés au SI trouvaient une homologie soit sur un autre chromosome (Chr10) soit sur plusieurs scaffold (sca269, sca327, sca1287) non ancrés dans le génome de l'olivier. L'intégralité des séquences codantes annotées au sein de ces scaffolds ont également été sélectionnées comme cibles. Au total, 72 séquences cibles représentant 154 417pb ont été définies à partir du génome de l'olivier.

Sur la base des analyses transcriptomiques, nous avons par ailleurs inclus l'ensemble des unitigs différentiellement exprimés entre les Ha et les Hb ayant un DE *fold change* de 1,2 et un FDR≤0,05 (correspondant à un seuil relativement relâché) qui ont passé les filtres de redondance, de pourcentage de GC et de qualité imposés pour la définition des sondes par l'entreprise *MyBaits*. Au total, 53 séquences cibles composées des unitigs sur-exprimés chez les Ha et 49 cibles représentant les unitigs sur-exprimés chez les Hb ont été définies. L'ensemble de ces 102 cibles couvrent 138 363pb du transcriptome du filaire.

L'extraction des ADN, la préparation des banques d'ADN, la capture des séquences cibles ont suivi les modalités décrites dans le chapitre 2 (cf p68-70 et Annexe 2.1 et 2.2) et le

séquençage a été réalisé par la plateforme LIGAN, Lille. Le même pipeline de reconstruction *de novo* de séquences consensus a été appliqué aux cibles définies pour la capture des régions potentiellement liées au SI qu'aux cibles définies pour le sexe.

Méthodes d'analyse de ségrégation des SNPs

Comme dans le chapitre 2, nous avons calculé les index de fixation FST (WEIR AND COCKERHAM 1984) pour chacun des SNPs. De nouveau, une forte valeur de FST indiquera une forte hétérogénéité de fréquence des génotypes entre les groupes comparés pour ce site. Les séquences présentant une forte structuration et au moins un SNPs ayant un FST supérieur ou égal à 0,3 ont été sélectionnées. Les séquences contenant des SNPs présentant une hétérogénéité de couverture entre les groupes de type présence/absence (par exemple aucun Ha couvert et tous les Hb couverts ou inversement) ont aussi été gardés pour tenter de découvrir d'éventuelles situations d'hémizygotie. Nous avons ensuite utilisé les scripts *genotype_plot.R* et *combine_plot.R* qui permettent de visualiser et regrouper les génotypes (https://github.com/JimWhiting91/genotype_plot).

Résultats

I. Analyse transcriptomique: une recherche des gènes candidats impliqués dans la détermination des spécificités d'auto-incompatibilité

Analyse d'expression différentielle : sélection des transcrits pour la capture

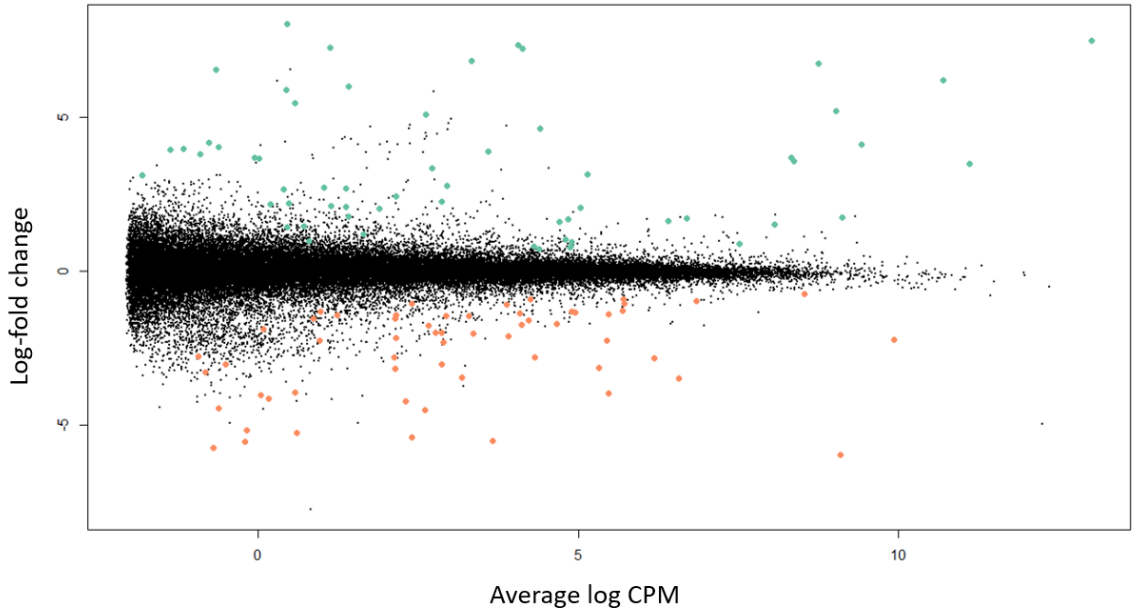


Figure 3.1 : « Mean-difference » (MD) plot du log-fold-change chez les Ha par rapport aux CPM moyens (log2) chez les Hb. Chaque point représente un unitig, en vert les unitigs sur-exprimés chez les Ha et en orange les unitigs sur-exprimés chez les Hb, pour un log-fold-change de 1,2 et un FDR≤0,05.

La comparaison entre les niveaux d'expression chez les Ha et les Hb met en évidence 53 unitigs significativement surexprimés chez les Ha par rapport aux Hb et 52 unitigs significativement surexprimés chez les Hb par rapport aux Ha (Figure 3.1 ; Log-fold-change ≥1,2 ; FDR≤0,05). Les niveaux d'expressions de chacun des unitigs d'intérêt sont visualisables dans les heatmap des Annexe 3.1 et 3.2.

Positionnement sur le génome de l'olivier des unitigs différentiellement exprimés chez le filaire

Sur les 53 unitigs sur-exprimés chez les individus Ha, 42 ont trouvé une position sur le génome de l'olivier (en vert sur la Figure 3.2 A). La majeure partie d'entre eux (n= 22) se trouvent sur l'un des 23 chromosomes assemblés et les autres (n=20) se positionnent sur 17 scaffolds non ancrés dans l'assemblage principal. Un seul unitig semble se localiser sur le chromosome 18 de l'olivier (evg145668, en vert sur la Figure 3.2 B), mais sa position est relativement éloignée de celle du SI. Nous n'identifions donc pas d'unitig qui serait à la fois

différentiellement exprimé et génétiquement lié au SI, donc pas de candidat direct au contrôle des spécificités d'auto-incompatibilité.

De la même façon, sur les 51 unitigs sur-exprimés chez les individus Hb par rapport aux individus Ha, 42 ont trouvé une position sur le génome de l'olivier (en orange sur la Figure 3.2 A). La majorité d'entre eux (n= 28) se trouvent sur l'un des 23 chromosomes assemblés et le reste (n=14) se placent sur 14 scaffold non ancrés dans l'assemblage principal. On observe qu'un seul de ces unitigs se positionne sur le chromosome 18 de l'olivier (en orange sur la Figure 3.2 B), cependant là encore sa position est relativement éloignée de celle du SI. Au final, aucun des unitigs d'intérêt surexprimé soit chez les Ha soit chez les Hb ne se positionne sur l'un des scaffolds contenant des marqueurs liés au SI, et de manière plus générale très peu de transcrits issus du transcriptome du filaire (histogramme gris dans le Figure 3.2 B) trouvent une homologie sur ces scaffolds, cette observation est cohérente avec les annotations disponibles pour ces scaffolds indiquant qu'ils contiennent peu de gènes .

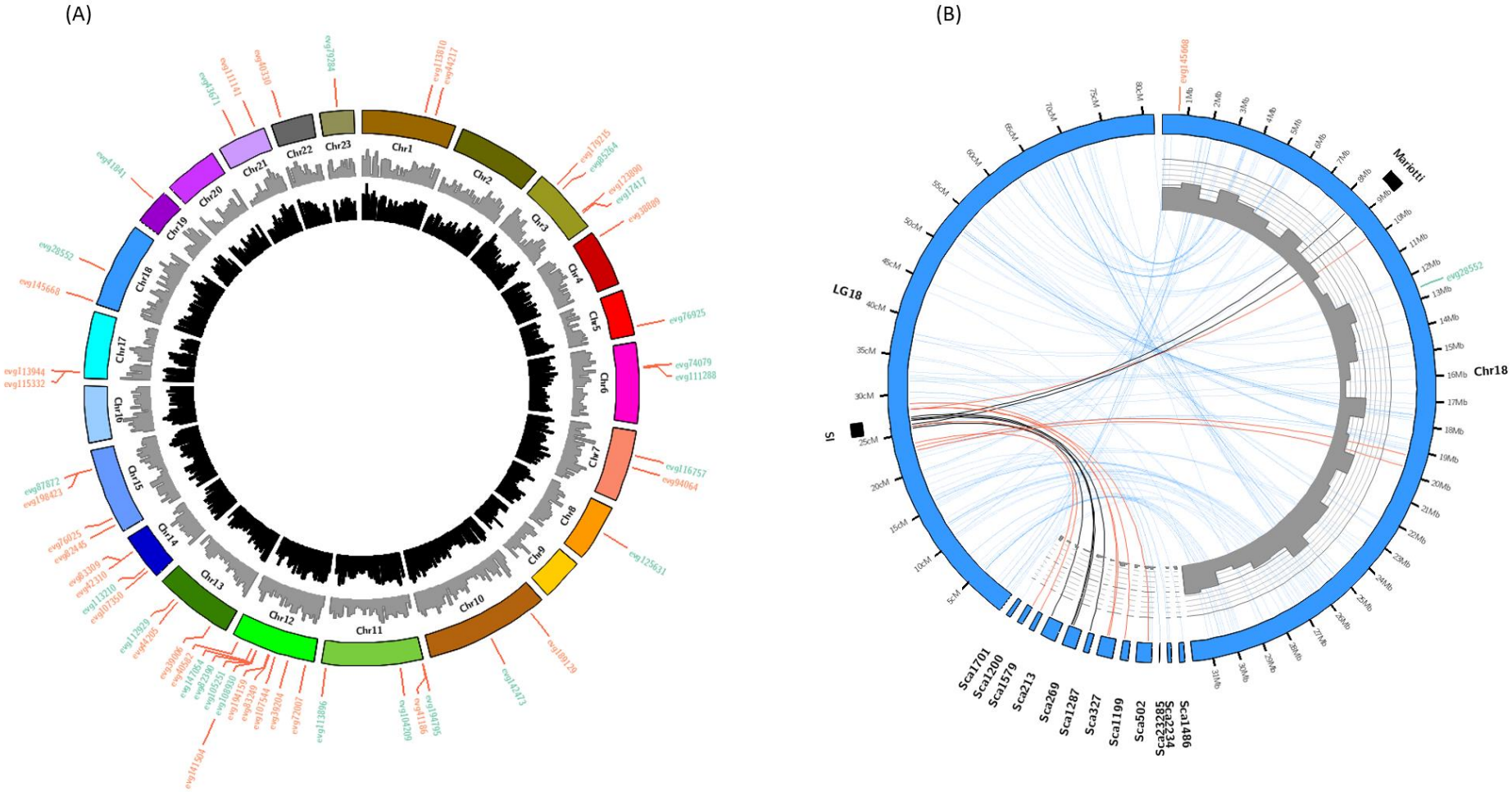


Figure 3.2 : A. Positionnement des unitigs sur-exprimés chez les individus Ha (vert) et Hb (orange) sur les 23 chromosomes de l'olivier. Les histogrammes noirs représentent le nombre de gènes annotés sur le génome de l'olivier et les histogrammes gris représentent le nombre total d'unitigs du filaire pouvant se positionner sur le génome de l'olivier par « bins » de 1Mpb . B. Synteny plot entre le groupe de liaison 18 de *P. angustifolia* (échelle en cM) et le chromosome 18 de l'olivier et une série scaffolds non ancrés (échelle en Mb), échelle 1 Mbp=3.125 cM. Les lignes relient les marqueurs de la carte de liaison de *P. angustifolia* avec leur meilleur hit BLAST dans le génome d'*O. europaea*. Les lignes bleues correspondent aux marqueurs à transmission autosomique. Les lignes noires correspondent à des marqueurs qui co-ségrègent strictement avec les phénotypes d'incompatibilités (Ha vs Hb). Les lignes rouges correspondent aux marqueurs qui présentent une association forte mais partielle (95 %) avec le SI. Les histogrammes gris représentent le nombre d'unitigs total du filaire pouvant se positionner sur le génome de l'olivier par « bins » de 1Mpb. Les unitigs sur-exprimés chez les individus Ha se positionnant sur le chromosome 18 de l'olivier sont indiqués en vert et les unitigs sur-exprimés chez les individus Hb en orange. La région trouvée génétiquement associée au SI dans l'olivier par Mariotti et al. (2020) est représentée par un rectangle noir.

Prédiction fonctionnelle des unitigs à expression SI-spécifique

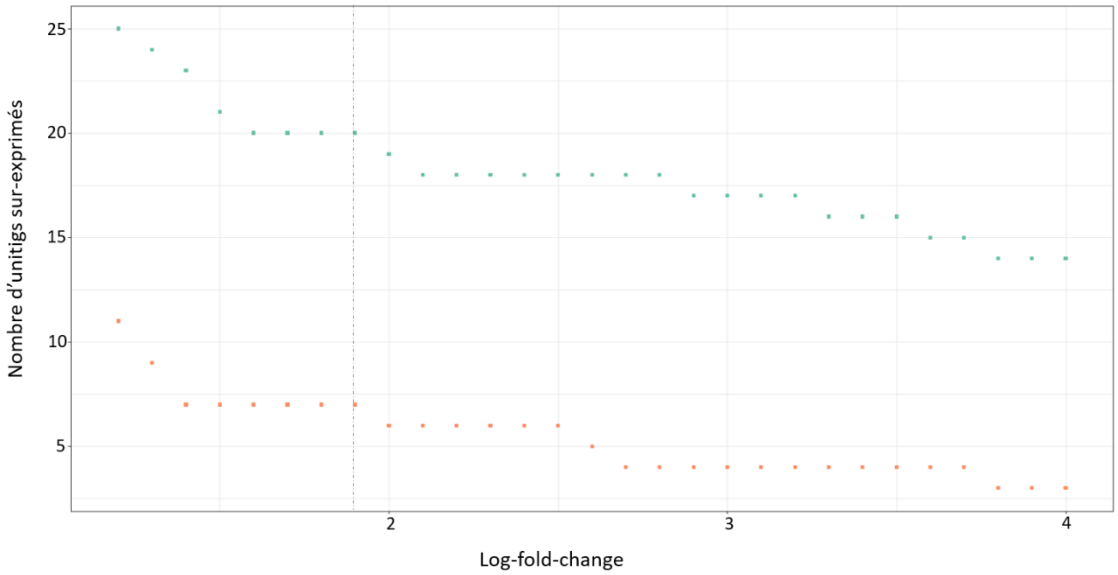


Figure 3.3 : Distributions du nombre d'unitigs différentiellement exprimés chez les Ha (vert) et chez les Hb (orange) en fonction de la valeur de Log-fold-change pour un FDR≤0,001.

Nous avons ensuite cherché à identifier les propriétés fonctionnelles des gènes candidats obtenus, en particulier des unitigs dont l'expression ou la surexpression est spécifique d'un phénotype. En examinant la distribution des Log-fold-change (Figure 3.3), nous avons choisi de nous concentrer sur les unitigs différentiellement exprimés caractérisés par un Log-fold-change supérieur ou égal à 1,9 pour examiner le détail de leurs patrons d'expression et les prédictions de leurs rôles fonctionnels.

A ce seuil plus strict, un total de 7 unitigs sur-exprimés chez les Hb par rapport aux Ha sont mis en évidence (Figure 3.4). Bien que leur niveau d'expression soit très élevé, aucun de ces unitig n'a une expression totalement spécifique des individus Hb. Deux de ces unitigs n'ont ni d'annotation GO ni prédiction fonctionnelle, un unitig (evg82589) code pour une protéine de liaison à l'ARN et les quatre autres (evg81150, evg42310, evg83309 et evg44434) ont une prédiction de facteur de transcription (HHO5). Deux de ces facteurs de transcription (evg42310 et evg83309) se localisent à une même position du chromosome 14 de l'olivier et un sur le scaffold1337. Notre pipeline de reconstruction *de novo* a gardé ces trois unitigs comme des contigs uniques que les pourcentages d'identité nucléotidique et protéomique ne permettaient pas de rassembler, mais nos paramètres de blast sur le génome de l'olivier tendent à laisser penser qu'il pourrait s'agir de deux gènes paralogues dont l'un n'a pas pu être ancré sur un des 23 chromosomes. De façon intéressante, l'unitig evg83309 avait

également été mis en évidence dans l'analyse du sexe. Cet unitig était sur-exprimé chez les mâles par rapport à l'ensemble des hermaphrodites, possédait des SNPs hétérozygotes spécifiques des mâles et environ 20% de la séquence était totalement absente chez les hermaphrodites. Ces dernières observations nous font émettre l'hypothèse que les séquences de cet unitig capturées chez les mâles et les hermaphrodites sont deux paralogues dont un serait spécifique des mâles et l'autre des hermaphrodites avec une sur-expression chez les individus Hb.

Au seuil plus strict (Log-fold-change de 1.9 et FDR≤0.001), 20 unitigs sont à l'inverse mis en évidence comme étant surexprimés chez les individus Ha par rapport aux individus Hb (Figure 3.4). On observe quatre unitigs ayant une prédiction fonctionnelle de facteur de transcription (HHO3) se localisant sur la même portion du scaffold 393 et montrant un profil d'expression spécifique (evg116212 et evg5530) ou quasi-spécifique (evg109875 et evg143084) des hermaphrodites Ha. Ici encore le pipeline d'assemblage *de novo* a gardé ces unitigs comme des unitigs différents, mais les paramètres de blast sur le génome de l'olivier suggèrent qu'il pourrait s'agir du même gène. On observe par ailleurs 13 unitigs qui n'ont pas d'annotation GO, dont quatre (evg86569, evg105251, evg104681 et evg104209) qui n'ont pas non plus de prédiction fonctionnelle. Trois unitigs (evg85264, evg76925, evg111440) coderaient pour des fonctions diverses (« pathogenesis-related protein », « proteinase inhibitor » et « protein RALF-like 19 »). L'unitig evg85264, qui code pour une protéine de reconnaissance de pathogène et est localisé sur le chromosome 3, est aussi très intéressant car, à l'inverse de evg83309 (surexprimé chez Hb et chez les mâles), il avait été mis en évidence comme globalement plus exprimé chez l'ensemble des hermaphrodites par rapport aux mâles. Les six autres unitigs (evg73651, evg105772, evg108930, evg82390, evg142473 et evg73649) n'ayant pas d'annotation GO ont une prédiction fonctionnelle de composés polliniques allergènes majeurs et sont présents dans le génome du frêne (« Lig v 1 ») et l'olivier (« Ole e 1 »). Les trois derniers unitigs sur-exprimés chez les individus Ha possèderaient des rôles enzymatiques de transport (evg112929 et evg143353) ou de synthèse (evg89364).

(A)

Séquence	Localisation	Prédiction fonctionnelle	GO annotation
evg38889	Chr4	---NA---	no GO terms
evg82589	NA	DNA RNA-binding Alba	F:GO:0003676:nucleic acid binding
evg38934	Sca437	uncharacterized AAA	no GO terms
evg81150	Sca1337	transcription factor HHO5-like isoform X1	F:GO:0003677:DNA binding
evg42310	Chr14	transcription factor HHO5-like	F:GO:0003677:DNA binding
evg83309	Chr14	transcription factor HHO5-like	F:GO:0003677:DNA binding
evg44434	NA	transcription factor HHO5-like	F:GO:0003677:DNA binding
evg73651	Sca3308	major pollen allergen Ole e 1-like	no GO terms
evg105772	NA	major pollen allergen Lig v 1	no GO terms
evg108930	Chr12	Ole e 1 protein	no GO terms
evg82390	Chr12	major pollen allergen Ole e 1-like	no GO terms
evg142473	Chr10	major pollen allergen Lig v 1-like	no GO terms
evg73649	NA	major pollen allergen Ole e 1-like	no GO terms
evg86569	Sca1235	Hypothetical predicted protein	no GO terms
evg105251	Chr12	---NA---	no GO terms
evg104681	Sca274	---NA---	no GO terms
evg76925	Chr5	protein RALF-like 19	no GO terms
evg85264	Chr3	pathogenesis-related protein 5-like	no GO terms
evg104209	Chr11	uncharacterized protein	no GO terms
evg112929	Chr13	putative ABC transporter B family member 8	F:GO:0005524:ATP binding; F:GO:0042626: ATPase-coupled transmembrane transporter activity;
evg143353	Sca889	cytochrome P450 81D11-like	F:GO:0005506:iron ion binding; F:GO:0016705:oxidoreductase activity; F:GO:0020037:heme binding
evg109875	Sca393	transcription factor HHO3-like	F:GO:0003677:DNA binding
evg89364	Sca1933	cellulose synthase G3	F:GO:0016760:cellulose synthase (UDP-forming) activity; P:GO:0030244:cellulose biosynthetic process
evg143034	Sca393	transcription factor HHO3-like	F:GO:0003677:DNA binding
evg111440	Sca1175	proteinase inhibitor PSI-1.2-like	no GO terms
evg116212	Sca393	transcription factor HHO3-like	F:GO:0003677:DNA binding
evg5530	NA	transcription factor HHO3-like	F:GO:0003677:DNA binding

(B)

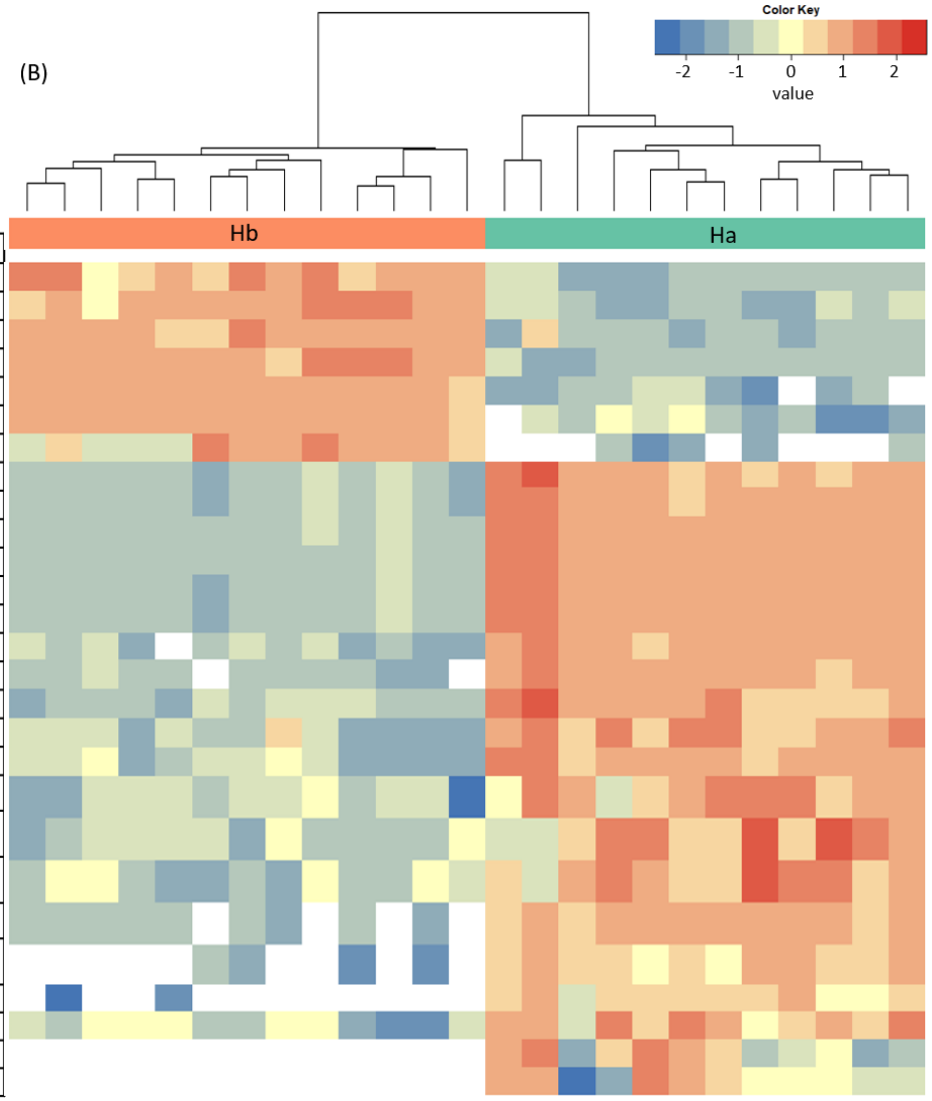


Figure 3.4 : Figure 3.4 : Caractéristiques des unitigs d'intérêt identifiés à partir de l'analyse transriptomique A. Tableau présentant les références des unitigs, leurs localisations probables sur le génome de l'olivier, et leurs prédictions fonctionnelles et annotations GO, chaque ligne correspond à un unitigs dont le pattern d'expression est mis en regard sur le heatmap. B. heatmap des unitigs différentiellement sur-exprimés entre les individus Hb et Ha; chaque colonne représente un individu et chaque ligne un unitigs. Les couleurs représentent la divergence d'expression d'un gène particulier dans un échantillon particulier, par rapport à la valeur moyenne de ce gène sur tous les échantillons, centré réduit sur 0 en unités d'écarts types (bleu peu exprimé, jaune expression moyenne, rouge fortement exprimé).

2. L'approche par capture de séquences se révèle difficile pour le SI

Reconstruction *de novo* des séquences cibles de références

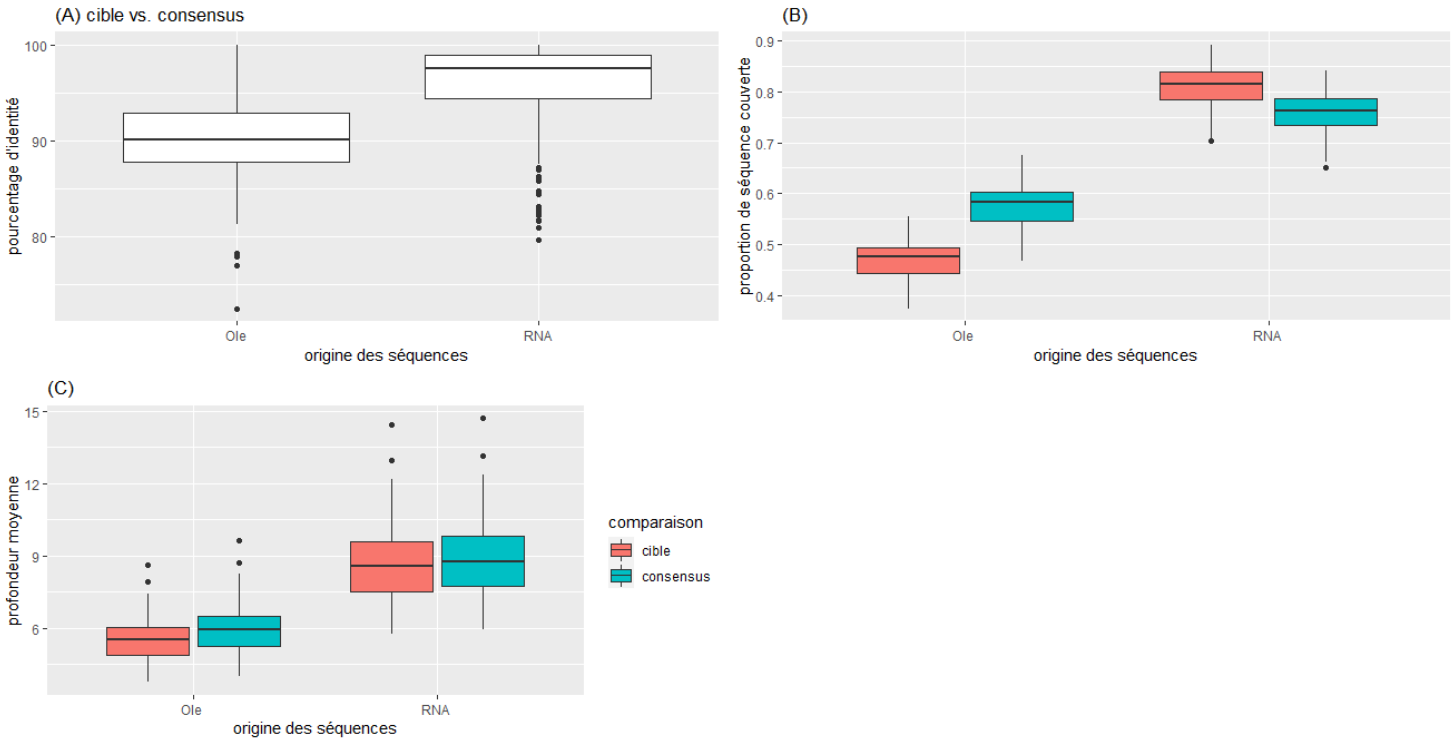


Figure 3.5 : Descriptif qualitatif de la reconstruction *de novo* des séquences cibles de référence en fonction de leur origine (*i.e.* transcriptome du filaire (RNA) ou génome de l'olivier (Ole)). A. distribution du pourcentage d'identité entre les séquences nucléotidiques cible utilisées pour définir les sondes de capture et les séquences consensus reconstruites *de novo*. B. comparaison de la proportion de séquence couverte lors d'un alignement brut des lectures de séquençage sur les séquences nucléotidiques cibles utilisées pour définir les sondes de capture et les séquences consensus reconstruites *de novo* chez le filaire. C. comparaison de la profondeur moyenne des séquences entre les séquences nucléotidiques cibles utilisées pour définir les sondes de capture et les séquences consensus reconstruites *de novo* chez le filaire.

Lorsque les séquences cibles sont définies à partir du génome de l'olivier (notées "Ole" dans la Fig. 3.5), les séquences consensus reconstruites *de novo* chez le filaire par notre pipeline intégrant Hybpiper sont en moyenne à 89,98% identiques aux séquences cibles d'origine (Figure 3.5 A). Ces nouvelles séquences consensus correspondant à la séquence génomique du filaire permettent d'augmenter la proportion moyenne de couverture d'environ 10% (Figure 3.5 B) et la profondeur moyenne de 5X à 6X (Figure 3.5 C).

Les séquences consensus reconstruites *de novo* divergent peu par rapport à la séquence d'origine lorsque celles-ci étaient issues du transcriptome du filaire (notées "RNA", Figure 3.5 A). Ces nouvelles séquences consensus correspondent maintenant à des séquences génomiques qui peuvent intégrer des séquences introniques et des portions de séquences situées en amont et en aval des gènes. En conséquence, comme dans le chapitre 2, on observe

une légère diminution de la proportion de couverture de chaque séquence (Figure 3.5.B), et une très faible variation de la profondeur moyenne (8,72 à 8,91 (Figure 3.5 C)).

Analyse de ségrégation des SNPs lié aux phénotypes du sexe

Tableau 3.1 : Résumé des premières étapes d'analyse de l'expérience de capture ciblée en fonction de l'origine des séquences (*i.e.* transcriptome du filaire (RNA) ou génome de l'olivier (Ole)).

	Ole	RNA	Total
Séquences ciblés	72	102	174
Séquences capturés	49	101	150
Nombre de SNP total	3142	6551	9693
Séquences avec Fst≥0.3	8	10	18
Nombre de SNP avec Fst≥0.3	9	16	25

Au total, sur les 72 séquences (154 417pb) initialement ciblées, seules 49 ont pu être capturées et correctement reconstruites *de novo* par notre pipeline. Ces séquences couvrent un total de 104 855pb et comptent 3142 SNPs. Parmi les 51 séquences présentes dans la région du SI définie par notre cartographie sur le chromosome 18 de l'olivier, seules 31 ont pu être capturées (Figure 3.6). L'analyse des valeurs de Fst par SNPs entre les Ha et les Hb a permis de mettre en évidence 9 SNPs répartis sur 8 séquences ayant une valeur supérieure ou égale à 0,3 (Figure 3.6).

En ce qui concerne les 102 unitigs (138 363pb) différentiellement exprimés entre les Ha et les Hb qui ont été ciblés, 101 séquences ont été capturées et reconstruites *de novo* par notre pipeline d'analyse (deux séquences ont été fusionnées). Les séquences consensus couvrent un total de 141 369pb et comptent au total 6551 SNPS. Seulement 16 SNPs sont mis en évidence par l'analyse des valeurs de Fst, se répartissant sur 10 séquences.

Malgré quelques valeurs de Fst élevées, aucun SNP ne correspond ni à un modèle où tous les Ha seraient homozygotes et tous les Hb seraient hétérozygotes (ou inversement) ni à un modèle où soit les Ha soit les Hb seraient hémizygotes (Figure 3.6). Cette analyse ne remet pas en cause les résultats précédents qui avait localisé le SI dans la région génomique du chromosome 18 de l'olivier qui a été ciblée. En effet, nous observons que 85% des séquences non capturées se trouvent dans cette région et représentent 40% des séquences ciblées dans cette région. Il est donc envisageable que nous n'ayons simplement pas capturé les gènes liés au système SI et que ces derniers se trouvent être trop divergents entre le filaire et l'olivier où simplement mal assemblés dans la référence génomique de l'olivier utilisée.

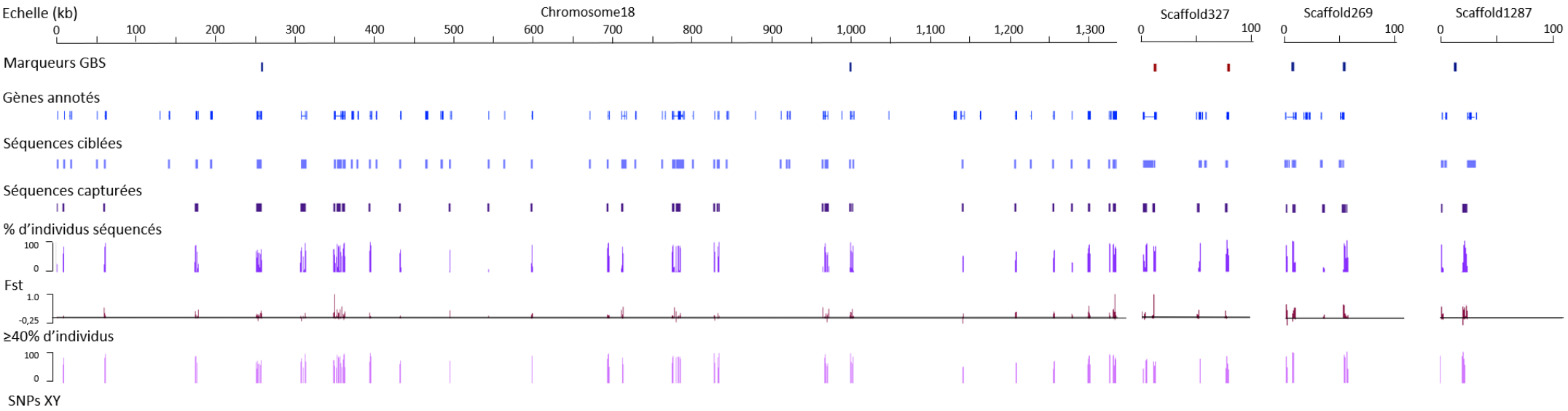


Figure 3.6 : Schéma des différentes étapes de l'expérience de capture de séquence pour les séquences liées au SI définies sur le génome de l'olivier (chromosome18, scaffolds 327, 269 et 1287). L'échelle est en kb, la position des marqueurs GBS définis comme strictement liés au locus du SI est représentée sur la première ligne. Dans l'ordre sont indiquées la position des gènes annotés chez l'olivier, la position des séquences ciblées (c'est-à-dire celles pour lesquelles des sondes ont pu être définies) et la position des séquences capturées et reconstruites *de novo* par notre pipeline. En violet est indiqué le pourcentage d'individus génotypés pour chaque SNPs identifié, en rouge les valeurs de Fst pour chacun de ces SNPs et en rose le pourcentage d'individus génotypés pour les SNPs ayant au moins 40% d'individus génotypés.

Discussion

A la recherche de gènes candidats impliqués dans la détermination des spécificités d'auto-incompatibilité

Dans ce chapitre, nous avons dans un premier temps comparé au niveau transcriptomique les deux groupes d'hermaphrodites, morphologiquement identiques, qui ne diffèrent théoriquement que par leur phénotype d'auto-incompatibilité. Nous avons mis en évidence une sur-expression de 0,09% des unitigs dans chacun des groupes, ainsi qu'une expression totalement spécifique de certains unitigs (evg5530 et evg116212) chez les individus Ha, mais aucun unitig spécifique des individus Hb. A ce stade, et sur la base de la synténie entre l'olivier et le filaire, nous n'avons pu identifier aucun unitig différentiellement exprimé entre les deux groupes d'hermaphrodites et co-localisant avec la région identifiée comme étant liée au SI (CARRE *et al.* 2021). Cependant, environ 20% des unitigs n'ont pu être replacés sur le génome de l'olivier, et environ 40% des unitigs se positionnent sur des scaffolds non ancrés dans l'assemblage principal de l'olivier. Parmi les unitigs mis en évidence, certains se détachent par leur prédiction fonctionnelle, notamment ceux codant pour des protéines de surface polliniques, les facteurs de transcription totalement spécifiques des individus Ha et celui impliqué dans la fonction de reconnaissance de pathogènes. En effet, il a été mis en évidence dans plusieurs systèmes SI que des mécanismes de reconnaissance de pathogènes avaient été détournés en systèmes de reconnaissance du « non-soi » dans la réaction d'auto-incompatibilité (MONDRAGON-PALOMINO *et al.* 2017 ; KODERA *et al.* 2021). Le parallèle avec d'autres études transcriptomiques sur les systèmes d'auto-incompatibilité est ici difficile. En effet, ces études sont généralement plus ciblées soit au niveau des tissus en séparant par exemple stigmate, ovaire et feuille (ZHOU *et al.* 2014) soit en se concentrant sur les réactions moléculaires d'incompatibilités en comparant par exemple l'expression transcriptomique lors d'une réaction compatible et incompatible au niveau du stigmate (KODERA *et al.* 2021) et/ou en comparant des mutants auto-compatibles à des individus auto-incompatibles de la même espèce (CARUSO *et al.* 2012 ; MA *et al.* 2017).

L'étude génétique du SI rendue difficile par sa complexité

Afin d'augmenter la quantité d'information sur les unitigs issus de l'analyse transcriptomique comparative et sur les régions génomiques mises en évidence dans la cartographie, nous avons mis en place une expérience de capture de gènes ciblés. Cette approche s'est révélée particulièrement complexe. Tout d'abord, une proportion importante des séquences ciblées n'a pas été capturée chez le filaire. Ainsi, près de 40% des séquences présentent dans la région du chromosome 18 de l'olivier mises en évidence comme étant liées au SI (MARIOTTI *et al.* 2020 ; CARRE *et al.* 2021) n'ont pas été capturées à l'aide des sondes. Pour une partie des autres séquences, l'information a été masquée par les filtres de couverture qu'il était nécessaire d'appliquer pour un appel de variants de bonne qualité étant donné la faible profondeur de séquençage obtenue. Une profondeur de séquençage plus importante aurait dû être envisagée et devra impérativement être appliquée pour les expériences suivantes, en doublant au minimum l'effort de séquençage par individu. Lors de l'analyse de synténie entre les chromosomes de l'olivier et les groupes de liaisons du filaire dans le premier chapitre, nous avions déjà mis en évidence que la relation entre le chromosome 18 et le LG18 contenant les marqueurs associés au locus du SI était moins linéaire que celle des autres groupes de liaison, indiquant la présence de réarrangements de grande ampleur. Nous avions également relevé que la plupart des marqueurs associées au SI chez le filaire trouvaient une homologie dans des scaffolds non ancrés dans le génome de l'olivier, traduisant vraisemblablement la difficulté à assembler cette région.

A l'intersection des déterminants moléculaires du sexe et du SI ?

Malgré ces difficultés, une observation intéressante est que certains unitigs sur-exprimés chez les individus Ha ou chez les individus Hb ont des positions potentielles proches des marqueurs associés au sexe sur le génome de l'olivier. De plus, deux unitigs (evg83309 qui est sur-exprimé chez les individus Hb et evg85264 qui sur-exprimé chez les individus Ha), avaient déjà été mis en évidence au chapitre 2 lors de l'analyse transcriptomique comparant des individus mâles et hermaphrodites. Le premier unitig avait une sur-expression chez les individus mâles et le second chez l'ensemble des hermaphrodites. Ces observations sont intéressantes dans le contexte de la pléiotropie de fonction du locus du sexe et des interactions complexes entre les locus du sexe et de l'incompatibilité relevées par les analyses

génétiques (BILLIARD *et al.* 2015). En effet, en plus de modifier le type sexuel des individus, le locus du sexe est capable de rendre compatibles des individus possédant des génotypes incompatibles au locus S : par exemple, les mâles Ma (mM S1S1) peuvent féconder les hermaphrodites Ha (mm S1S1) alors qu'ils partagent l'allèle S1. Par ailleurs, une des fonctions du « supergène » mâle est également d'empêcher la transmission de gamètes entre individus possédant au locus S des génotypes compatibles : par exemple la fécondation d'un hermaphrodite Hb (mm S1S2) par un mâle Ma (mM S1S1) produit une descendance entièrement mâle, ce qui signifie que les gamètes (mS1) produits par ce mâle ne permettent pas la fécondation alors qu'ils sont parfaitement viables. Une explication parcimonieuse de ces deux observations serait de supposer que le caractère mâle permet de «mimer» la spécificité Hb, ce qui nous rapproche de l'observation faite dans l'analyse transcriptomique d'une expression commune de gènes par les hermaphrodites Hb et les mâles.

La prochaine étape: assembler un génome pour le filaire

Au final, il est difficile à ce stade d'aller plus loin dans l'analyse génétique du SI ou d'émettre des hypothèses sur les origines de la complexité de cette région génomique. Durant cette thèse, des progrès considérables ont été réalisés dans les méthodes de séquençage d'ADN, rendant l'assemblage de génome enfin accessible pour des espèces non-modèles telles que le filaire. La manière la plus efficace de progresser sur l'architecture du SI pourrait être de profiter de ces développements en séquençant le génome du filaire. Le séquençage complet du génome du double hétérozygote (mM/S1S2), ayant servi de père au croisement contrôlé de la population utilisée pour la cartographie génétique haute densité du filaire (CARRE *et al.* 2021), est actuellement en cours par une collaboration de l'équipe. Si l'assemblage des régions d'intérêt de cet individu s'avère compliqué, il est envisagé de faire un séquençage complet du génome d'un individu hermaphrodite double homozygote (mm/S1S1). Une autre possibilité serait d'envisager le séquençage d'un mâle homozygote S2 qui pourrait être pertinente dans l'étude du locus du SI.

Pour conclure, bien que ces expériences ne nous aient pas permis de mettre en évidence de séquences candidates au déterminisme du DSI chez *P. angustifolia,* nous savons que la région génomique ciblée était la bonne (MARIOTTI *et al.* 2020 ; CARRE *et al.* 2021) et nous avons pu exclure la liaison génétique des séquences que nous avons capturées. Il est donc

envisageable que les gènes clés du DSI se trouvent dans les 40% de séquences non capturées. Nous avons donc recueilli des informations cruciales qui nous aideront dans l'avancement de la compréhension de ce système.

Contributions

Tous les auteurs ont contribué à l'étude présentée dans ce chapitre. PS-L et PV ont développé, conçu et supervisé l'étude de « RNA-sequencing » ; ils ont coordonné et réalisé les tests stigmatiques de phénotypage des groupes d'incompatibilité. SS a effectué l'extraction d'ARN, la préparation des librairies et le séquençage du transcriptome. AC a construit et optimisé le pipeline d'analyse transcriptomique sous la supervision de CM. AC, PS-L et VC ont conçu le design de l'expérience de capture de gène. AC et CG ont effectué les constructions des banques de capture à partir des extractions d'ADN effectuées par CG et SS. AC et SG ont construit le pipeline d'analyse des données de capture. AC, PS-L et VC ont interprété les résultats et rédigé le manuscrit.

Discussion et perspectives

Deux régions génomiques distinctes et de petite taille sont associées au sexe et à l'auto-incompatibilité

Au cours de cette thèse, nous avons créé la première cartographie génétique haute-densité de l'espèce androdioïque *P. angustifolia* et identifié les régions génomiques associées à ses deux phénotypes reproducteurs importants. Nous avons ainsi répondu au premier objectif de thèse, qui était de valider l'hypothèse génétique d'indépendance des deux locus qui avait été posée par Billard *et al.* en 2015 sur la base de croisements contrôlés d'effectifs limités. Nous avons confirmé que les locus du sexe et de l'auto-incompatibilité sont situés sur deux groupes de liaison distincts (LG12 et LG18 respectivement) et correspondent à des systèmes de type XY (CARRE *et al.* 2021). Un tel système de type XY représente un mécanisme de détermination génétique rare pour l'auto-incompatibilité (où plus de spécificités ségrègent généralement), et présente des similitudes avec le contrôle des types d'appariement (mating types) chez les oomycètes (BADOUIN *et al.* 2015). Une comparaison avec le génome de l'olivier, espèce étroitement apparentée, a permis d'estimer les tailles respectives des régions du sexe et du SI à environ 1 849 kb et 741kb. Ces estimations restent incomplètes, étant donné que d'une part de nombreux marqueurs associés à chacun des phénotypes se positionnement sur des scaffolds qui ne sont pas ancrés dans l'assemblage principal du génome de l'olivier, et que d'autre part de nombreux réarrangements ont eu le temps de s'accumuler depuis la divergence entre olivier et filaire. Ceci est particulièrement vrai pour le locus dit "du sexe", qui n'a par définition pas de fonction de déterminisme sexuel chez l'espèce hermaphrodite qu'est l'olivier. Malgré ces incertitudes, il est cependant notable que la taille de ces deux régions reste relativement limitée. La structure génomique des locus d'auto-incompatibilité reste mal connue dans la plupart des espèces, mais il est rare qu'ils occupent des chromosomes entiers et ils sont généralement cantonnés à de petites portions chromosomiques, comme observé ici. Une exception concerne les oomycètes où, comme chez le filaire, seules deux spécificités ségrègent, mais chez qui des chromosomes presque entiers, très réarrangés, peuvent devenir associés aux types sexuels (BADOUIN *et al.* 2015). La raison pour laquelle ces deux systèmes suivent des voies distinctes reste à ce jour inconnue.

La région de détermination du sexe s'étend fréquemment sur l'essentiel des chromosomes sexuels chez certaines espèces, ne laissant qu'une portion pseudo-autosomale de petite taille. Ce n'est clairement pas le cas chez le filaire, où il semble que cette région de détermination du sexe ne concerne qu'une fraction très limitée de tout le groupe de liaison 12. Il est possible qu'à l'avenir, si l'androdioécie reste stable chez l'espèce, voire si elle réalise une transition vers la dioécie, cette région s'étende, comme prédit dans les modèles classiques d'évolution des chromosomes sexuels. La question des mécanismes évolutifs par lesquels cette région s'étend reste ouverte (ABBOTT *et al.* 2017), et il serait intéressant d'explorer si l'androdioécie plutôt que la dioécie représente une situation suffisamment asymétrique en termes de sélection sexuellement antagoniste pour être favorable à ce phénomène d'extension. Par ailleurs, des modèles récents ont proposé que l'extension des chromosomes sexuels pourrait être simplement la conséquence du maintien de combinaisons hétérozygotes aux gènes à la lisière des régions de détermination du sexe et pseudo-autosomale (JAY *et al.* 2021). De ce strict point de vue, dioécie et androdioécie représentent des situations identiques, et on pourrait s'attendre à ce que la sélection favorise l'extension de la région de détermination du sexe. Le fait qu'elle soit restée petite à ce jour pourrait résulter de deux scénarios distincts. D'une part, elle pourrait refléter son émergence récente, la séparation des sexes observée chez *P. angustifolia* n'étant pas observée chez son proche parent l'olivier. A l'inverse, cet état pourrait remonter à une date plus ancienne, la séparation des sexes s'étant alors perdue chez les autres espèces, et la petite taille de la région correspondrait à un état stable. L'enjeu serait alors de comprendre pourquoi la sélection naturelle ne favorise pas son extension. L'étude de l'évolution de cette région génomique à l'échelle de la famille entière représente une perspective prometteuse pour résoudre cette question.

Forces et limites de l'approche de capture de séquences employée

Les approches par hybridation ciblée permettent de capturer et séquencer des régions d'intérêt variables à partir d'espèces et de souches taxonomiquement divergentes. Cela leur permet d'être utilisées sur des organismes non modèles dépourvus de génome ou de transcriptome séquencé. Cependant, ces approches nécessitent une connaissance préalable de l'espèce étudiée, qui est indispensable pour définir les sondes de capture. L'établissement de la cartographie génétique du filaire, l'analyse transcriptomique et la mise à disposition d'un

génome de référence pour l'espèce phylogénétiquement proche qu'est l'olivier ont constitué des préalables qui nous ont permis de mettre en place cette approche.

Nous nous attendions à quelques difficultés et perte d'information sur les séquences que nous avons définies à partir du génome de l'olivier, principalement en raison de la spécificité des sondes de capture et de la divergence entre l'olivier et le filaire. Globalement, d'après les observations d'homologie que nous avions pu faire entre le génome de l'olivier utilisé et les marqueurs de notre cartographie GBS (CARRE *et al.* 2021) ainsi qu'avec le transcriptome reconstruit *de novo* du filaire, nous attendions une perte d'environ 15% des informations ciblées. Concernant les séquences définies pour l'études du sexe nous avons obtenu un rendement parfaitement satisfaisant de 87% des séquences ciblées capturées. Pour les séquences définies pour l'étude du SI cependant, seules 68% des séquences ciblées ont pu être capturées, ce qui est largement inférieur, et suggère que l'ampleur de la divergence des séquences de cette région est importante. Malgré ces limites, nous avons pu étudier de nombreux SNPs sur les séquences codantes ciblées, qu'elles aient été identifiées sur la base de leur liaison génétique au phénotype (du sexe ou d'auto-incompatibilité) ou de leur expression différentielle (entre sexes ou groupes d'auto-incompatibilité). Ainsi, sur l'ensembledes séquences capturées issues de la liaison génétique ou de l'analyse transcriptomique seulement 26 (représentant collectivement 69 728pb) avaient au moins un SNP montrant une association stricte (XY) avec le phénotype du sexe. Deux de ces séquences sont particulièrement intéressantes, car en plus de posséder des SNPs strictement liés au sexe, elles co-localisent avec des transcrits ayant une sur-expression soit chez les mâles (evg89402) soit les hermaphrodites (evg178602). Nous n'avons cependant pu identifier aucune prédiction fonctionnelle ou catégorisation GO pour ces deux séquences, ce qui rend compliqué une analyse plus détaillée à ce stade. Sur l'intégralité des séquences codantes ciblées, liées génétiquement au phénotype du SI (154 417pb) (CARRE *et al.* 2021) ou d'après l'analyse transcriptomique (138 363pb), aucune des séquences ayant pu être analysées ne présente d'association génétique avec ce phénotype. A ce stade, bien qu'il n'ait pas été possible durant cette thèse de mettre en évidence de SNPs liés au phénotype d'incompatibilité dans des régions codantes, la cartographie génétique (CARRE *et al.* 2021) et les récentes avancées sur le sujet chez l'olivier (MARIOTTI *et al.* 2020) ont montré que nous en étions proches. D'autres outils, plus directs tels que l'assemblage de génomes, seront nécessaires pour progresser vers l'identification des déterminants des spécificités SI. Un assemblage du génome du père double

hétérozygote (mM S1S2) du croisement contrôlé ayant servi à l'établissement de la cartographie génétique est actuellement en cours. Le cas échéant, si les régions génomiques associées au déterminisme du sexe et de l'auto-incompatibilité s'avèrent complexes à reconstruire, il est envisageable de séquencer un double homozygote (Ha), voire un homozygote pour l'allèle S2 au locus de SI (individus mâles Mc).

Des difficultés plus importantes pour l'étude de la région du SI que pour celle du sexe

Les chromosomes sexuels, les chromosomes de type sexuel et les supergènes en général sont répandus dans la nature. Ces structures sont définies par de vastes régions de suppression de recombinaison englobant plusieurs gènes, et elles contrôlent des polymorphismes emblématiques, tels que le dimorphisme sexuel ou le polymorphisme de couleur, dans de nombreux organismes, y compris les humains (BERGERO AND CHARLESWORTH 2009 ; SCHWANDER *et al.* 2014; CHARLESWORTH 2015a ; ABBOTT *et al.* 2017). Ces locus représentent des défis majeurs pour l'analyse génomique en raison de leurs propriétés évolutives particulières (VEKEMANS *et al.* 2021). Nous avons cependant observé un fort contraste entre les deux régions que nous avons étudiées, et les difficultés rencontrées ont été substantiellement plus marquées pour l'étude de la région du SI que celle du sexe. Globalement, alors que nous avions observé une bonne colinéarité entre l'arrangement des marqueurs au sein des groupes de liaison du filaire par rapport à leur localisation sur le génome de l'olivier (incluant le LG12 contenant les déterminants du sexe), la synténie entre le LG18 (contenant le locus du SI) et le chromosome 18 de l'olivier échappait à cette tendance, indiquant de forts réarrangements. La région génomique contenant les déterminants de l'auto-incompatibilité semble globalement plus remaniée et moins bien assemblée que le reste du génome, et les difficultés rencontrées lors de l'expérience de capture vont dans ce sens. Ce contraste entre un locus du sexe relativement bien conservé et un locus du SI plus fortement remanié peut être mis en lien avec la mise en place de ces deux phénotypes au sein de la famille des oleaceae. En effet, alors que le SI est ancestral et semble avoir perduré sous sa forme diallélique depuis l'origine de la famille, il est possible que la séparation des sexes telle qu'elle se manifeste chez le filaire soit récente, et ne soit possiblement pas encore associée à des remaniements plus substantiels. Là encore, l'étude de cette région à l'échelle

du genre et de la famille des oleaceae dans son ensemble semble incontournable pour tester cette hypothèse.

Des interactions croisées entre sexe et SI

Notre étude représente par ailleurs la première analyse transcriptomique d'une espèce androdioïque fonctionnelle possédant un DSI. La configuration expérimentale nous a permis de reconstruire *de novo* un transcriptome de référence, qui représente une ressource génomique inédite pour cette espèce non-modèle, et de comparer par la suite l'expression des gènes entre les sexes et les groupes d'auto-incompatibilité. La comparaison des niveaux d'expression des transcrits entre les individus mâles et les individus hermaphrodites chez le filaire a mis en évidence de nombreux unitigs différentiellement exprimés. Ainsi, 0,15% et 0,45% des unitigs sont surexprimés chez les mâles et chez les hermaphrodites respectivement. En profitant de la bonne synténie globale entre l'olivier et le filaire, nous avons pu déterminer que cinq unitigs (evg89402, evg147054, evg178602, evg87871 et evg70363) différentiellement exprimés entre les sexes co-localisaient avec la région identifiée comme étant liée au sexe. La vaste majorité des gènes à expression sexe-biaisée ne semble donc pas être liée génétiquement aux déterminants des phénotypes sexuels. En comparant les niveaux d'expression entre les individus Ha et les individus Hb, nous avons mis en évidence une sur-expression de 0,09% des unitigs dans chacun des groupes, ainsi qu'une expression totalement spécifique de certains unitigs (evg5530 et evg116212) chez les individus Ha. Par ailleurs, aucun des unitigs différentiellement sur-exprimés entre les deux groupes d'hermaphrodites ne co-localise avec la région identifiée comme étant liée au SI (Carre *et al.* 2021). On observe cependant que certains unitigs sur-exprimés chez les individus Ha ou chez les individus Hb ont des positions potentielles proches de marqueurs associés au sexe sur le génome de l'olivier. De plus deux unitigs, evg83309 (un facteur de transcription qui est sur-exprimé chez les individus Hb) et evg85264 (un gène codant pour une protéine de reconnaissance de pathogène qui est sur-exprimé chez les individus Ha) ont aussi été mis en évidence comme étant sur-exprimés chez les individus mâles et chez les individus hermaphrodites respectivement. Nous avons enfin observé que plusieurs transcrits sur-exprimés chez les individus Ha avaient une localisation proche de celle de marqueurs du filaire liés au phénotype du sexe sur le génome de l'olivier. Ces transcrits codent pour des facteurs de transcription (position potentielle sur le scaffold 393) et des protéines de surface pollinique

(position potentielle légèrement en amont de la région liée au sexe sur le chr12). Ces interactions croisées entre les régions contrôlant l'androdioécie et le DSI sont particulièrement intéressantes à la lumière des interactions génétiques qui existent entre ces deux traits (les mâles sont d'une part compatibles avec tous les hermaphrodites quel que soit leur génotype au locus d'incompatibilité (S1S1, S1S2, S2S2), une d'autre part une fécondation d'un individu Hb par un mâle donne une descendance à 100% composée de mâle, tandis qu' un biais inverse en faveur des hermaphrodites (60% vs. 40%) a été mis en évidence pour un croisement Mb (mmS1S2) X Ha (CARRE *et al.* 2021)). L'utilisation des marqueurs du sexe, inférés des SNPs spécifiques mis en évidence, permettra de ne plus devoir attendre la floraison des individus pour connaitre leur phénotype sexuel, ce qui facilitera l'analyse à plus grande échelle sur les biais de ségrégation de descendance, notamment concernant les croisements Ha X Ma (mM S1S1) et Ha X Mc (mM S2S2). Une étape ultérieure sera de déterminer le rôle de ces transcrits différentiellement exprimés dans ces interactions génétiques: les transcrits spécifiques des mâles sont-ils, et si oui comment, impliqués dans leur phénotype de compatibilité universelle? Comment les transcrits spécifiques des Ha déterminent-ils la direction et l'ampleur de la distorsion de ségrégation du pollen des mâles?

Enfin, ces interactions croisées entre sexe et SI sont intéressantes dans le cadre de l'observation récente faite par De Cauwer *et al.* (2021) chez *Ligustrum vulgare*, une autre espèce de la famille des oléacées possédant le DSI. Cette espèce hermaphrodite et entomophile se caractérise par le maintien d'une fraction d'hermaphrodites ayant perdu l'auto-incompatibilité, et bénéficiant de ce fait d'un avantage de compatibilité universelle. Cette propriété rappelle la capacité des mâles de *P. angustifolia* à féconder l'ensemble des hermaphrodites. Une hypothèse attractive serait que les mâles de *P. angustifolia* dérivent d'une situation similaire, où des individus dont le pollen est universellement compatible se spécialisent vers la fonction mâle et contournent ainsi les désavantages liés à la dépression de consanguinité pour remplacer les hermaphrodites auto-compatibles (VAN DE PAER *et al.* 2015). Accéder à l'analyse génomique des régions du sexe et de l'auto-incompatibilité chez *Ligustrum* pourrait permettre de tester cette hypothèse.

Bibliographie

Abbott, J. K., A. K. Nordén and B. Hansson, 2017 Sex chromosome evolution: historical insights and future perspectives. Proc Biol Sci 284. https://dx.doi.org/10.1098/rspb.2016.2806

Akagi, T., I. M. Henry, H. Ohtani, T. Morimoto, K. Beppu *et al.*, 2018 A Y-Encoded Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response Regulator in Kiwifruit. The Plant cell 30: 780-795. https://dx.doi.org/10.1105/tpc.17.00787

Akagi, T., S. M. Pilkington, E. Varkonyi-Gasic, I. M. Henry, S. S. Sugano *et al.*, 2019 Two Y-chromosome-encoded genes determine sex in kiwifruit. Nature Plants 5: 801-809. https://dx.doi.org/10.1038/s41477-019-0489-6

Albert, B., M.-É. Morand-Prieur, S. Brachet, P.-H. Gouyon, N. Frascaria-Lacoste *et al.*, 2013 Sex expression and reproductive biology in a tree species, Fraxinus excelsior L. Comptes Rendus Biologies 336: 479-485. https://dx.doi.org/10.1016/j.crvi.2013.08.004

Almeida, P., E. Proux-Wera, A. Churcher, L. Soler, J. Dainat *et al.*, 2020 Genome assembly of the basket willow, Salix viminalis, reveals earliest stages of sex chromosome expansion. BMC biology 18: 78. https://dx.doi.org/10.1186/s12915-020-00808-1

Armero, A., L. Baudouin, S. Bocs and D. This, 2017 Improving transcriptome de novo assembly by using a reference genome of a related species: Translational genomics from oil palm to coconut. PloS one 12: e0173300. https://dx.doi.org/10.1371/journal.pone.0173300

Badouin, H., M. E. Hood, J. Gouzy, G. Aguileta, S. Siguenza *et al.*, 2015 Chaos of Rearrangements in the Mating-Type Chromosomes of the Anther-Smut Fungus Microbotryum lychnidis-dioicae. Genetics 200: 1275-1284. https://dx.doi.org/10.1534/genetics.115.177709

Bao, E., T. Jiang and T. Girke, 2013 BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences. Bioinformatics 29: 1250-1259. https://dx.doi.org/10.1093/bioinformatics/btt127

Barrett, S. C., 2002 The evolution of plant sexual diversity. Nat Rev Genet 3: 274-284. https://dx.doi.org/10.1038/nrg776

Barrett, S. C. H., 1990 The evolution and adaptive significance of heterostyly. Trends in Ecology & Evolution 5: 144-148. https://doi.org/10.1016/0169-5347(90)90220-8

Barrett, S. C. H., 1992 Heterostylous Genetic Polymorphisms: Model Systems for Evolutionary Analysis, pp. 1-29 in *Evolution and Function of Heterostyly*, edited by S. C. H. Barrett. Springer Berlin Heidelberg, Berlin, Heidelberg. https://dx.doi.org/10.1007/978-3-642-86656-2_1

Barrett, S. C. H., 1998 The evolution of mating strategies in flowering plants. Trends in Plant Science 3: 335-341. http://dx.doi.org/10.1016/S1360-1385(98)01299-0

Barrett, S. C. H., 2019 'A most complex marriage arrangement': recent advances on heterostyly and unresolved questions. New Phytologist 224: 1051-1067. https://doi.org/10.1111/nph.16026

Barrett, S. C. H., L. K. Jesson and A. M. Baker, 2000 The Evolution and Function of Stylar Polymorphisms in Flowering Plants. Annals of Botany 85: 253-265. http://dx.doi.org/10.1006/anbo.1999.1067

Barrett, S. C. H., and J. S. Shore, 2008 New insights on heterostyly: Comparative Biology, Ecology and Genetics, pp. 3-32 in *Self-Incompatibility in Flowering Plants*, edited by S.-V. B. Heidelberg.

Baumann, U., J. Juttner, X. Bian and P. Langridge, 2000 Self-incompatibility in the Grasses. Annals of Botany 85: 203-209. http://dx.doi.org/10.1006/anbo.1999.1056

Bergero, R., and D. Charlesworth, 2009 The evolution of restricted recombination in sex chromosomes. Trends Ecol Evol 24: 94-102. https://dx.doi.org/10.1016/j.tree.2008.09.010

Bernacchi, D., and S. D. Tanksley, 1997 An interspecific backcross of Lycopersicon esculentum x L. hirsutum: linkage analysis and a QTL study of sexual compatibility factors and floral traits. Genetics 147: 861-877. https://dx.doi.org/10.1093/genetics/147.2.861

Besnard, G., R. Rubio de Casas, P.-A. Christin and P. Vargas, 2009 Phylogenetics of Olea (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: Tertiary climatic shifts and lineage differentiation times. Annals of Botany 104: 143-160. https://dx.doi.org/10.1093/aob/mcp105

Billiard, S., L. Husse, P. Lepercq, C. Gode, A. Bourceaux *et al.*, 2015 Selfish male-determining element favors the transition from hermaphroditism to androdioecy. Evolution 69: 683-693. https://dx.doi.org/10.1111/evo.12613

Billiard, S., M. López-Villavicencio, B. Devier, M. E. Hood, C. Fairhead *et al.*, 2011 Having sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating types. Biological Reviews 86: 421-442. https://doi.org/10.1111/j.1469-185X.2010.00153.x

Billiard, S., M. LÓPez-Villavicencio, M. E. Hood and T. Giraud, 2012 Sex, outcrossing and mating types: unsolved questions in fungi and beyond. J Evol Biol 25: 1020-1038. https://doi.org/10.1111/j.1420-9101.2012.02495.x

Bolger, A. M., M. Lohse and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120. https://dx.doi.org/10.1093/bioinformatics/btu170

Boris Igic, Russell Lande and Joshua R. Kohn, 2008 Loss of Self-Incompatibility and Its Evolutionary Consequences. Int J Plant Sci 169: 93-104. https://dx.doi.org/10.1086/523362

Borowiec, M. L., 2019 Spruceup: fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. Journal of Open Source Software 4: 1635. https://dx.doi.org/10.21105/joss.01635

Bracale, M., E. Caporali, M. G. Galli, C. Longo, G. Marziani-Longo *et al.*, 1991 Sex determination and differentiation in Asparagus officinalis L. Plant Science 80: 67-77. https://dx.doi.org/10.1016/0168-9452(91)90273-B

Brennan, A. C., 2017 Distyly supergenes as a model to understand the evolution of genetic architecture. American Journal of Botany 104: 5-7. https://doi.org/10.3732/ajb.1600363

Carey, S., Q. Yu and A. Harkess, 2021 The Diversity of Plant Sex Chromosomes Highlighted through Advances in Genome Sequencing. Genes (Basel) 12. https://dx.doi.org/10.3390/genes12030381

Carré A, Gallina S, Santoni S, Vernet P, Godé C, Castric V, Saumitou-Laprade P (2021) Genetic mapping of sex and self-incompatibility determinants in the androdioecious *plant Phillyrea angustifolia*. bioRxiv, 2021.04.15.439943, ver. 7 peer-reviewed and

recommended by Peer community in Genomics. https://doi.org/10.1101/2021.04.15.439943

Caruso, M., P. Merelo, G. Distefano, S. La Malfa, A. R. Lo Piero *et al.*, 2012 Comparative transcriptome analysis of stylar canal cells identifies novel candidate genes implicated in the self-incompatibility response of Citrus clementina. BMC Plant Biology 12: 20. **https://dx.doi.org/10.1186/1471-2229-12-20**

Castric, V., and X. Vekemans, 2004 Plant self-incompatibility in natural populations: a critical assesment of recent theoretical and empirical advances. Molecular Ecology 13: 2873-2889. **https://dx.doi.org/10.1111/j.1365-294X.2004.02267.x**

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko and J. H. Postlethwait, 2011 Stacks: building and genotyping Loci de novo from short-read sequences. G3 (Bethesda, Md.) 1: 171-182. **https://dx.doi.org/10.1534/g3.111.000240**

Chantha, S.-C., A. C. Herman, V. Castric, X. Vekemans, W. Marande *et al.*, 2017 The unusual S locus of Leavenworthia is composed of two sets of paralogous loci. New Phytologist 216: 1247-1255. **https://dx.doi.org/https://doi.org/10.1111/nph.14764**

Chantha, S.-C., A. C. Herman, A. E. Platts, X. Vekemans and D. J. Schoen, 2013 Secondary Evolution of a Self-Incompatibility Locus in the Brassicaceae Genus Leavenworthia. PLOS Biology 11: e1001560. **https://dx.doi.org/10.1371/journal.pbio.1001560**

Charlesworth, B., 1978 The population genetics of anisogamy. J Theor Biol 73: 347-357. **https://dx.doi.org/https://doi.org/10.1016/0022-5193(78)90195-9**

Charlesworth, B., and D. Charlesworth, 1978 A Model for the Evolution of Dioecy and Gynodioecy. The American Naturalist 112: 975-997. **https://dx.doi.org/10.1086/283342**

Charlesworth, D., 2002 Plant sex determination and sex chromosomes. Heredity 88: 94-101. **https://dx.doi.org/10.1038/sj.hdy.6800016**

Charlesworth, D., 2006 Evolution of Plant Breeding Systems. Curr Biol 16: R726-R735. **https://doi.org/10.1016/j.cub.2006.07.068**

Charlesworth, D., 2015a Plant contributions to our understanding of sex chromosome evolution. New Phytol 208: 52-65. **https://dx.doi.org/10.1111/nph.13497**

Charlesworth, D., 2015b The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. Evolutionary applications 9: 74-90. **https://dx.doi.org/10.1111/eva.12291**

Charlesworth, D., and B. Charlesworth, 1979 A Model for the Evolution of Distyly. The American Naturalist 114: 467-498.

Chen, Y., A. T. Lun and G. K. Smyth, 2016 From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Res 5: 1438. **https://dx.doi.org/10.12688/f1000research.8987.2**

Cocker, J. M., J. Wright, J. Li, D. Swarbreck, S. Dyer *et al.*, 2018 Primula vulgaris (primrose) genome assembly, annotation and gene expression, with comparative genomics on the heterostyly supergene. Scientific reports 8: 17942-17942. **https://dx.doi.org/10.1038/s41598-018-36304-4**

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon *et al.*, 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676. **https://dx.doi.org/10.1093/bioinformatics/bti610**

Cossard, G. G., M. A. Toups and J. R. Pannell, 2019 Sexual dimorphism and rapid turnover in gene expression in pre-reproductive seedlings of a dioecious herb. Annals of botany 123: 1119-1131. https://dx.doi.org/10.1093/aob/mcy183

Cruz, F., I. Julca, J. Gómez-Garrido, D. Loska, M. Marcet-Houben *et al.*, 2016 Genome sequence of the olive tree, Olea europaea. GigaScience 5: 29. https://dx.doi.org/10.1186/s13742-016-0134-5

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156-2158. https://dx.doi.org/10.1093/bioinformatics/btr330

De-Kayne, R., and P. G. D. Feulner, 2018 A European Whitefish Linkage Map and Its Implications for Understanding Genome-Wide Synteny Between Salmonids Following Whole Genome Duplication. G3 (Bethesda, Md.) 8: 3745-3755. https://dx.doi.org/10.1534/g3.118.200552

De Cauwer, I., P. Vernet, S. Billiard, C. Godé, A. Bourceaux *et al.*, 2021 Widespread coexistence of self-compatible and self-incompatible phenotypes in a diallelic self-incompatibility system in Ligustrum vulgare (Oleaceae). Heredity 127: 384-392. https://dx.doi.org/10.1038/s41437-021-00463-4

De Nettancourt, D., 1977 *Incompatibility in Angiosperms*. Springer-Verlag, Berlin, Heidelberg et New York.

De Summa, S., G. Malerba, R. Pinto, A. Mori, V. Mijatovic *et al.*, 2017 GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics 18: 119. https://dx.doi.org/10.1186/s12859-017-1537-8

de Graaf, B. H. J., S. Vatovec, J. A. Juárez-Díaz, L. Chai, K. Kooblall *et al.*, 2012 The Papaver Self-Incompatibility Pollen S-Determinant, PrpS, Functions in Arabidopsis thaliana. Curr Biol 22: 154-159. http://dx.doi.org/10.1016/j.cub.2011.12.006

Dellaporta, S. L., and A. Calderon-Urrea, 1993 Sex determination in flowering plants. The Plant Cell 5: 1241-1251. https://dx.doi.org/10.1105/tpc.5.10.1241

Delph, L. F., 2009 Sex allocation: evolution to and from dioecy. Curr Biol 19: R249-251. https://dx.doi.org/10.1016/j.cub.2009.01.048

Diggle, P. K., V. S. Di Stilio, A. R. Gschwend, E. M. Golenberg, R. C. Moore *et al.*, 2011 Multiple developmental processes underlie sex differentiation in angiosperms. Trends in Genetics 27: 368-376. https://doi.org/10.1016/j.tig.2011.05.003

Dommée, B., A. Geslot, J. D. Thomson, M. Reille and N. Denelle, 1999 Androdioecy in the entomophilous tree Fraxinus ornus (Oleaceae). New Phytologist 143: 419-426. https://dx.doi.org/doi:10.1046/j.1469-8137.1999.00442.x

Dommée, B., J. D. Thompson and F. Cristini, 1992 Distylie chez Jasminum fruticans L.: hypothèse de la pollinisation optimale basée sur les variations de l'écologie intraflorale. Bulletin de la Société Botanique de France. Lettres Botaniques 139: 223-234. https://dx.doi.org/10.1080/01811797.1992.10824960

Dufay, M., P. Champelovier, J. Käfer, J. P. Henry, S. Mousset *et al.*, 2014 An angiosperm-wide analysis of the gynodioecy–dioecy pathway. Annals of Botany 114: 539-548. https://dx.doi.org/10.1093/aob/mcu134

Dulberger, R., 1992 Floral Polymorphisms and Their Functional Significance in the Heterostylous Syndrome, pp. 41-84 in *Evolution and Function of Heterostyly*, edited by

S. C. H. Barrett. Springer Berlin Heidelberg, Berlin, Heidelberg. https://dx.doi.org/10.1007/978-3-642-86656-2_3

Dupin, J., P. Raimondeau, C. Hong-Wa, S. Manzi, M. Gaudeul *et al.*, 2020 Resolving the Phylogeny of the Olive Family (Oleaceae): Confronting Information from Organellar and Nuclear Genomes. Genes (Basel) 11: 1508. https://dx.doi.org/10.3390/genes11121508

Durand, E., R. Méheust, M. Soucaze, P. M. Goubet, S. Gallina *et al.*, 2014 Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. Science 346: 1200-1205. https://dx.doi.org/10.1126/science.1259442

Dussert, Y., L. Legrand, I. D. Mazet, C. Couture, M.-C. Piron *et al.*, 2020 Identification of the First Oomycete Mating-type Locus Sequence in the Grapevine Downy Mildew Pathogen, Plasmopara viticola. Curr Biol 30: 3897-3907.e3894. https://dx.doi.org/10.1016/j.cub.2020.07.057

Evangelistella, C., A. Valentini, R. Ludovisi, A. Firrincieli, F. Fabbrini *et al.*, 2017 De novo assembly, functional annotation, and analysis of the giant reed (Arundo donax L.) leaf transcriptome provide tools for the development of a biofuel feedstock. Biotechnology for Biofuels 10: 138. https://dx.doi.org/10.1186/s13068-017-0828-7

Fernandes Cardoso, J. C., M. Viana, R. Matias, M. Furtado, A. Caetano *et al.*, 2018 Towards a unified terminology for angiosperm reproductive system. Acta Botanica Brasilica 32: 329-348. https://dx.doi.org/10.1590/0102-33062018abb0124

Foote, H. C., J. P. Ride, V. E. Franklin-Tong, E. A. Walker, M. J. Lawrence *et al.*, 1994 Cloning and expression of a distinctive class of self-incompatibility (S) gene from Papaver rhoeas L. Proc Natl Acad Sci U S A 91: 2265-2269.

Fujii, S., K.-i. Kubo and S. Takayama, 2016 Non-self- and self-recognition models in plant self-incompatibility. Nature Plants 2: 1-9. https://dx.doi.org/10.1038/NPLANTS.2016.130

Gaude, T., and D. Cabrillac, 2001 Self-incompatibility in flowering plants: The Brassica model. Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie 324: 537-542. http://dx.doi.org/10.1016/S0764-4469(01)01323-3

Gaude, T., S. Glémin, D. Cabrillac and A. Mignot, 2001 L'auto-incompatibilité chez les plantes à fleurs. Médecine/sciences 17: I-XIV.

Genete, M., V. Castric and X. Vekemans, 2020 Genotyping and De Novo Discovery of Allelic Variants at the Brassicaceae Self-Incompatibility Locus from Short-Read Sequencing Data. Mol. Biol. Evol. 37: 1193-1201. https://dx.doi.org/10.1093/molbev/msz258

Geng, S., P. De Hoff and J. G. Umen, 2014 Evolution of sexes from an ancestral mating-type specification pathway. PLoS biology 12: e1001904-e1001904. https://dx.doi.org/10.1371/journal.pbio.1001904

Gervais, C. E., V. Castric, A. Ressayre and S. Billiard, 2011 Origin and diversification dynamics of self-incompatibility haplotypes. Genetics 188: 625-636. https://dx.doi.org/10.1534/genetics.111.127399

Gilbert, D., 15 Dec 2013 EvidentialGene: tr2aacds, mRNA Transcript Assembly Software, pp.

Goubet, P. M., H. Bergès, A. Bellec, E. Prat, N. Helmstetter *et al.*, 2012 Contrasted Patterns of Molecular Evolution in Dominant and Recessive Self-Incompatibility Haplotypes in Arabidopsis. PLOS Genetics 8: e1002495. https://dx.doi.org/10.1371/journal.pgen.1002495

Gouzy, J., S. Carrere and T. Schiex, 2009 FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics (Oxford, England) 25: 670-671. https://dx.doi.org/10.1093/bioinformatics/btp024

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology 29: 644-652. https://dx.doi.org/10.1038/nbt.1883

Guo, Y.-L., X. Zhao, C. Lanz and D. Weigel, 2011 Evolution of the S-locus region in Arabidopsis relatives. Plant physiology 157: 937-946. https://dx.doi.org/10.1104/pp.111.174912

Harkess, A., K. Huang, R. van der Hulst, B. Tissen, J. L. Caplan *et al.*, 2020 Sex Determination by Two Y-Linked Genes in Garden Asparagus[OPEN]. The Plant Cell 32: 1790-1796. https://dx.doi.org/10.1105/tpc.19.00859

Harrison, P. W., A. E. Wright, F. Zimmer, R. Dean, S. H. Montgomery *et al.*, 2015 Sexual selection drives evolution and rapid turnover of male gene expression. Proceedings of the National Academy of Sciences 112: 4393. https://dx.doi.org/10.1073/pnas.1501339112

Hayman, D. L., 1956 The Genetical Control of Incompatibility in Phalaris Coerulescens Desf. Australian Journal of Biological Sciences 9: 321-331.

Hodgkin, J., 1990 Sex determination compared in Drosophila and Caenorhabditis. Nature 344: 721-728. https://dx.doi.org/10.1038/344721a0

Hooper, D. M., S. C. Griffith and T. D. Price, 2019 Sex chromosome inversions enforce reproductive isolation across an avian hybrid zone. Molecular Ecology 28: 1246-1262. https://doi.org/10.1111/mec.14874

Husse, L., S. Billiard, J. Lepart, P. Vernet and P. Saumitou-Laprade, 2013 A one-locus model of androdioecy with two homomorphic self-incompatibility groups: expected vs. observed male frequencies. J Evol Biol 26: 1269-1280. https://dx.doi.org/10.1111/jeb.12124

Iwano, M., K. Ito, H. Shimosato-Asano, K.-S. Lai and S. Takayama, 2014 Self-Incompatibility in the Brassicaceae, pp. 245-254 in *Sexual Reproduction in Animals and Plants*, edited by H. Sawada, N. Inoue and M. Iwano. Springer Japan, Tokyo.

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

Jay, P., M. Chouteau, A. Whibley, H. Bastide, H. Parrinello *et al.*, 2021 Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. Nature Genetics 53: 288-293. https://dx.doi.org/10.1038/s41588-020-00771-1

Jiménez-Ruiz, J., J. A. Ramírez-Tejero, N. Fernández-Pozo, M. d. l. O. Leyva-Pérez, H. Yan *et al.*, 2020 Transposon activation is a major driver in the genome evolution of cultivated olive trees (Olea europaea L.). The Plant Genome 13: e20010. https://dx.doi.org/https://doi.org/10.1002/tpg2.20010

Johnson, L., 1957 A review of the family Oleaceae. , pp. 395–418 in *Contributions from the New South Wales National Herbarium 2*.

Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet *et al.*, 2016 HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl Plant Sci 4: apps.1600016. https://dx.doi.org/10.3732/apps.1600016

Käfer, J., G. A. B. Marais and J. R. Pannell, 2017 On the rarity of dioecy in flowering plants. Molecular Ecology 26: 1225-1241. https://dx.doi.org/https://doi.org/10.1111/mec.14020

Kappel, C., C. N. Huu and M. Lenhard, 2017 A short story gets longer: recent insights into the molecular basis of heterostyly. J Exp Bot 68: 5719-5730. https://dx.doi.org/10.1093/jxb/erx387

Keller, B., J. D. Thomson and E. Conti, 2014 Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin's primroses: evidence from experimental studies. Functional Ecology 28: 1413-1425. https://doi.org/10.1111/1365-2435.12274

Kelly, L. J., W. J. Plumb, D. W. Carey, M. E. Mason, E. D. Cooper *et al.*, 2020 Convergent molecular evolution among ash species resistant to the emerald ash borer. Nature Ecology & Evolution 4: 1116-1128. https://dx.doi.org/10.1038/s41559-020-1209-3

Kent, W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664. https://dx.doi.org/10.1101/gr.229202

Kodera, C., J. Just, M. Da Rocha, A. Larrieu, L. Riglet *et al.*, 2021 The molecular signatures of compatible and incompatible pollination in Arabidopsis. BMC Genomics 22: 268. https://dx.doi.org/10.1186/s12864-021-07503-7

Kozielska, M., F. J. Weissing, L. W. Beukeboom and I. Pen, 2010 Segregation distortion and the evolution of sex-determining mechanisms. Heredity 104: 100-112. https://dx.doi.org/10.1038/hdy.2009.104

Krzywinski, M. I., J. E. Schein, I. Birol, J. Connors, R. Gascoyne *et al.*, 2009 Circos: An information aesthetic for comparative genomics. Genome Res. https://dx.doi.org/10.1101/gr.092759.109

Kubo, K.-i., T. Paape, M. Hatakeyama, T. Entani, A. Takara *et al.*, 2015 Gene duplication and genetic exchange drive the evolution of S-RNase-based self-incompatibility in *Petunia*. Nature Plants 1: 1-9.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359. https://dx.doi.org/10.1038/nmeth.1923

Lebel-Hardenack, S., and S. Grant, 1997 Genetics of sex determination in flowering plants. Trends in Plant Science 2: 130-136. https://dx.doi.org/10.1016/S1360-1385(97)01012-1

Lepart, J., and B. Dommée, 1992 Is Phillyrea angustifolia L. (Oleaceae) an androdioecious species? Botanical Journal of the Linnean Society 108: 375-387. https://dx.doi.org/10.1111/j.1095-8339.1992.tb00252.x

Lewis, D., 1941 Male sterility un populations of hermaphrodite plants, the equilibrium between females and hermaphrodites to be expected with different types of inheritance. New Phytologist 40: 56-63. https://doi.org/10.1111/j.1469-8137.1941.tb07028.x

Lewis, D., and L. K. Crowe, 1958 Unilateral interspecific incompatibility in flowering plants. Heredity 12: 233-256.

Lewis, D., and D. A. Jones, 1992, pp. 129-150 in *Evolution and Function of Heterostyly*, edited by S. C. H. Barrett. Springer Berlin Heidelberg, Berlin, Heidelberg. https://dx.doi.org/10.1007/978-3-642-86656-2_1

Li, B., and C. N. Dewey, 2011 RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323-323. https://dx.doi.org/10.1186/1471-2105-12-323

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760. https://dx.doi.org/10.1093/bioinformatics/btp324

Li, J., J. M. Cocker, J. Wright, M. A. Webster, M. McMullan *et al.*, 2016 Genetic architecture and evolution of the S locus supergene in Primula vulgaris. Nature plants 2: 16188. https://dx.doi.org/10.1038/nplants.2016.188

Li, W., and A. Godzik, 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659. https://dx.doi.org/10.1093/bioinformatics/btl158

Lin, Z., D. J. Eaves, E. Sanchez-Moran, F. C. H. Franklin and V. E. Franklin-Tong, 2015 The Papaver rhoeas S determinants confer self-incompatibility to Arabidopsis thaliana in planta. Science 350: 684-687.

Llaurens, V., S. Billiard, V. Castric and X. Vekemans, 2009 Evolution of Dominance in Sporophytic Self-Incompatibility Systems: I. Genetic Load and Coevolution of Levels of Dominance in Pollen and Pistil. Evolution 63: 2427-2437.

Lloyd, D. G., 1974 Theoretical sex ratios of dioecious and gynodioecious angiosperms. Heredity 32: 11-34. https://dx.doi.org/10.1038/hdy.1974.2

Lloyd, D. G., 1975 The maintenance of gynodioecy and androdioecy in angiosperms. Genetica 45: 325-339. https://dx.doi.org/10.1007/BF01508307

Lloyd, D. G., 1980 Sexual strategies in plants III. A quantitative method for describing the gender of plants. New Zealand Journal of Botany 18: 103-108. https://dx.doi.org/10.1080/0028825X.1980.10427235

Lundqvist, A., 1956 Self-incompatibility in rye. Hereditas 42: 293-348.

Ma, Y., Q. Li, G. Hu and Y. Qin, 2017 Comparative transcriptional survey between self-incompatibility and self-compatibility in Citrus reticulata Blanco. Gene 609: 52-61. https://dx.doi.org/10.1016/j.gene.2017.01.033

Manzanares, C., S. Barth, D. Thorogood, S. L. Byrne, S. Yates *et al.*, 2016 A Gene Encoding a DUF247 Domain Protein Cosegregates with the S Self-Incompatibility Locus in Perennial Ryegrass. Mol. Biol. Evol. 33: 870-884. https://dx.doi.org/10.1093/molbev/msv335

Mariotti, R., S. Pandolfi, I. De Cauwer, P. Saumitou-Laprade, P. Vernet *et al.*, 2020 Diallelic self-incompatibility is the main determinant of fertilization patterns in olive orchards. Evolutionary Applications n/a. https://doi.org/10.1111/eva.13175

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

Mayer, S. S., and D. Charlesworth, 1991 Cryptic dioecy in flowering plants. Trends Ecol Evol 6: 320-325. https://dx.doi.org/10.1016/0169-5347(91)90039-z

Ming, R., A. Bendahmane and S. S. Renner, 2011 Sex Chromosomes in Land Plants. Annu. Rev. Plant. Biol 62: 485-514. https://dx.doi.org/10.1146/annurev-arplant-042110-103914

Mondragón-Palomino, M., A. John-Arputharaj, M. Pallmann and T. Dresselhaus, 2017 Similarities between Reproductive and Immune Pistil Transcriptomes of Arabidopsis Species. Plant Physiol 174: 1559-1575. https://dx.doi.org/10.1104/pp.17.00390

Müller, N. A., B. Kersten, A. P. Leite Montalvão, N. Mähler, C. Bernhardsson *et al.*, 2020 A single gene underlies the dynamic evolution of poplar sex determination. Nature Plants 6: 630-637. https://dx.doi.org/10.1038/s41477-020-0672-9

Muyle, A., J. Kafer, N. Zemp, S. Mousset, F. Picard *et al.*, 2016 SEX-DETector: A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms. Genome biology and evolution 8: 2530-2543. https://dx.doi.org/10.1093/gbe/evw172

Muyle, A., R. Shearn and G. A. Marais, 2017 The Evolution of Sex Chromosomes and Dosage Compensation in Plants. Genome biology and evolution 9: 627-645. https://dx.doi.org/10.1093/gbe/evw282

Nasrallah, M. E., P. Liu and J. B. Nasrallah, 2002 Generation of self-incompatible *Arabidopsis thaliana* by transfer of two S locus genes from *A. lyrata*. Science 297: 247-249. https://dx.doi.org/10.1126/science.1072205

Olofsson, J. K., I. Cantera, C. Van de Paer, C. Hong-Wa, L. Zedane *et al.*, 2019 Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. Mol Ecol Resour 19: 877-892. https://doi.org/10.1111/1755-0998.13016

Pailler, T., L. Humeau, J. Figier and J. D. Thompson, 1998 Reproductive trait variation in the functionally dioecious and morphologically heterostylous island endemic *Chassalia corallioides*(Rubiaceae). Biological Journal of the Linnean Society 64: 297-313. https://doi.org/10.1006/bijl.1998.0219

Palmer, D. H., T. F. Rogers, R. Dean and A. E. Wright, 2019 How to identify sex chromosomes and their turnover. Molecular ecology 28: 4709-4724. https://dx.doi.org/10.1111/mec.15245

Pannell, J., 1997a Mixed genetic and environmental sex determination in an androdioecious population of Mercurialis annua. Heredity 78: 50-56. https://dx.doi.org/10.1038/hdy.1997.6

Pannell, J., 1997b Widespread functional androdioecy in Mercurialis annua L. (Euphorbiaceae). Biological Journal of the Linnean Society 61: 95-116. https://dx.doi.org/doi:10.1111/j.1095-8312.1997.tb01779.x

Pannell, J. R., S. M. Eppley, M. E. Dorken and R. Berjano, 2014 Regional variation in sex ratios and sex allocation in androdioecious Mercurialis annua. J Evol Biol 27: 1467-1477. https://dx.doi.org/10.1111/jeb.12352

Pannell, J. R., and G. Korbecka, 2010 Mating-System Evolution: Rise of the Irresistible Males. Curr Biol 20: R482-R484. https://doi.org/10.1016/j.cub.2010.04.033

Pannell, J. R., and M. Verdú, 2006 The evolution of gender specialization from dimorphic hermaphroditism: paths from heterodichogamy to gynodioecy and androdioecy. Evolution 60: 660-673. https://dx.doi.org/10.1554/05-481.1

Pannell, J. R., and M. Voillemot, 2015 Plant mating systems: female sterility in the driver's seat. Curr Biol 25: R511-514. https://dx.doi.org/10.1016/j.cub.2015.04.044

Parker, J. S., and M. S. Clark, 1991 Dosage sex-chromosome systems in plants. Plant Science 80: 79-92. https://doi.org/10.1016/0168-9452(91)90274-C

Pei, X., Z. Jing, Z. Tang and Y. Zhu, 2017 Comparative transcriptome analysis provides insight into differentially expressed genes related to cytoplasmic male sterility in broccoli (Brassica oleracea var. italica). Scientia Horticulturae 217: 234-242. https://doi.org/10.1016/j.scienta.2017.01.041

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher and H. E. Hoekstra, 2012 Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one 7: e37135. https://dx.doi.org/10.1371/journal.pone.0037135

Philbrick, C. T., and L. H. Rieseberg, 1994 Pollen Production in the Androdioecious Datisca glomerata (Datiscaceae): Implications for Breeding System Equilibrium. Plant Species Biology 9: 43-46. https://dx.doi.org/doi:10.1111/j.1442-1984.1994.tb00081.x

Rastas, P., 2017 Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics 33: 3726-3732. https://dx.doi.org/10.1093/bioinformatics/btx494

Reboud, X., and C. Zeyl, 1994 Organelle inheritance in plants. Heredity 72: 132-140. https://dx.doi.org/10.1038/hdy.1994.19

Renner, S. S., and R. E. Ricklefs, 1995 Dioecy and its correlates in the flowering plants. American Journal of Botany 82: 596-606. https://doi.org/10.1002/j.1537-2197.1995.tb11504.x

Rice, P., I. Longden and A. Bleasby, 2000 EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276-277. https://dx.doi.org/10.1016/S0168-9525(00)02024-2

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law *et al.*, 2015 limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acid Res 43: e47-e47. https://dx.doi.org/10.1093/nar/gkv007

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field *et al.*, 2010 De novo assembly and analysis of RNA-seq data. Nat Methods 7: 909-912. https://dx.doi.org/10.1038/nmeth.1517

Robinson, M. D., D. J. McCarthy and G. K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140. https://dx.doi.org/10.1093/bioinformatics/btp616

Rochette, N. C., and J. M. Catchen, 2017 Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc 12: 2640-2659. https://dx.doi.org/10.1038/nprot.2017.123

Sakai, A. K., and S. G. Weller, 1999 Gender and Sexual Dimorphism in Flowering Plants: A review of Terminology, Biogeographic Patterns, Ecological Correlates, and Phylogenetic Approaches, pp. 1-31 in *Gender and Sexual Dimorphism in Flowering Plants*, edited by M. A. Geber, T. E. Dawson and L. F. Delph. Springer Berlin Heidelberg, Berlin, Heidelberg. https://dx.doi.org/10.1007/978-3-662-03908-3_1

Saumitou-Laprade, P., P. Vernet, A. Dowkiw, S. Bertrand, S. Billiard *et al.*, 2018 Polygamy or subdioecy? The impact of diallelic self-incompatibility on the sexual system in Fraxinus excelsior (Oleaceae). Proc Biol Sci 285: 20180004. https://dx.doi.org/10.1098/rspb.2018.0004

Saumitou-Laprade, P., P. Vernet, C. Vassiliadis, Y. Hoareau, G. de Magny *et al.*, 2010 A self-incompatibility system explains high male frequencies in an androdioecious plant. Science 327: 1648-1650. https://dx.doi.org/10.1126/science.1186687

Saumitou-Laprade, P., P. Vernet, X. Vekemans, S. Billiard, S. Gallina *et al.*, 2017 Elucidation of the genetic architecture of self-incompatibility in olive: Evolutionary consequences and perspectives for orchard management. Evolutionary Applications: 1-14. https://dx.doi.org/10.1111/eva.12457

Sauquet, H., M. von Balthazar, S. Magallón, J. A. Doyle, P. K. Endress *et al.*, 2017 The ancestral flower of angiosperms and its early diversification. Nature Communications 8: 16047. https://dx.doi.org/10.1038/ncomms16047

Scharmann, M., A. G. Rebelo and J. R. Pannell, 2021 High rates of evolution preceded shifts to sex-biased gene expression in *Leucadendron*, the most sexually dimorphic angiosperms. bioRxiv: 2021.2001.2012.426328. https://dx.doi.org/10.1101/2021.01.12.426328

Schmieder, R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863-864. https://dx.doi.org/10.1093/bioinformatics/btr026

Schwander, T., R. Libbrecht and L. Keller, 2014 Supergenes and complex phenotypes. Curr Biol 24: R288-294. https://dx.doi.org/10.1016/j.cub.2014.01.056

Sébastian, C., 1956 *Étude du genre Phillyrea Tournefort*. Société des sciences naturelles et physiques du Maroc.

She, H., Z. Liu, Z. Xu, H. Zhang, F. Cheng *et al.*, 2020 The female (XX) and male (YY) genomes provide insights into the sex determination mechanism in spinach. bioRxiv: 2020.2011.2023.393710. https://dx.doi.org/10.1101/2020.11.23.393710

Shen, J. J., T. Y. Wang and W. Yang, 2017 Regulatory and evolutionary signatures of sex-biased genes on both the X chromosome and the autosomes. Biology of sex differences 8: 35. https://dx.doi.org/10.1186/s13293-017-0156-4

Shore, J. S., H. J. Hamam, P. D. J. Chafe, J. D. J. Labonne, P. M. Henning *et al.*, 2019 The long and short of the S-locus in Turnera (Passifloraceae). New Phytologist 224: 1316-1329. https://doi.org/10.1111/nph.15970

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus *et al.*, 2011 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, pp. 539 in *Mol Syst Biol*. https://dx.doi.org/10.1038/msb.2011.75

Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212. https://dx.doi.org/10.1093/bioinformatics/btv351

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier *et al.*, 2003 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423: 825-837. https://dx.doi.org/10.1038/nature01722

Smith-Unna, R., C. Boursnell, R. Patro, J. M. Hibberd and S. Kelly, 2016 TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26: 1134-1144. https://dx.doi.org/10.1101/gr.196469.115

Sollars, E. S. A., A. L. Harper, L. J. Kelly, C. M. Sambles, R. H. Ramirez-Gonzalez *et al.*, 2017 Genome sequence and genetic diversity of European ash trees. Nature 541: 212-216. https://dx.doi.org/10.1038/nature20786

Su, S., C. W. Law, C. Ah-Cann, M.-L. Asselin-Labat, M. E. Blewitt *et al.*, 2017 Glimma: interactive graphics for gene expression analysis. Bioinformatics 33: 2050-2052. https://dx.doi.org/10.1093/bioinformatics/btx094

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

Taylor, H., 1945 Cyto-taxonomy and phylogeny of the oleaceae. Brittonia 5: 337-367. https://dx.doi.org/10.2307/2804889

Torres, M. F., L. S. Mathew, I. Ahmed, I. K. Al-Azwani, R. Krueger *et al.*, 2018 Genus-wide sequencing supports a two-locus model for sex-determination in Phoenix. Nature Communications 9: 3969. https://dx.doi.org/10.1038/s41467-018-06375-y

Tsagkogeorga, G., V. Cahais and N. Galtier, 2012 The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate Ciona intestinalis. Genome biology and evolution 4: 740-749. https://dx.doi.org/10.1093/gbe/evs054

Unver, T., Z. Wu, L. Sterck, M. Turktas, R. Lohaus *et al.*, 2017 Genome of wild olive and the evolution of oil biosynthesis. Proc. Natl. Acad. Sci. U.S.A. 114: E9413-e9422. https://dx.doi.org/10.1073/pnas.1708621114

Ushijima, K., R. Nakano, M. Bando, Y. Shigezane, K. Ikeda *et al.*, 2012 Isolation of the floral morph-related genes in heterostylous flax (Linum grandiflorum): the genetic

polymorphism and the transcriptional and post-transcriptional regulations of the S locus. The Plant Journal 69: 317-331. https://doi.org/10.1111/j.1365-313X.2011.04792.x

Van de Paer, C., P. Saumitou-Laprade, P. Vernet and S. Billiard, 2015 The joint evolution and maintenance of self-incompatibility with gynodioecy or androdioecy. J Theor Biol 371: 90-101. https://dx.doi.org/10.1016/j.jtbi.2015.02.003

Vassiliadis, C., J. Lepart, P. Saumitou-Laprade and P. Vernet, 2000 Self-Incompatibility and Male Fertilization Success in Phillyrea angustifolia (Oleaceae). Int J Plant Sci 161: 393-402.

Vassiliadis, C., P. Saumitou-Laprade, J. Lepart and F. Viard, 2002 High male reproductive success of hermaphrodites in the androdioecious Phillyrea angustifolia. Evolution 56: 1362-1373.

Vekemans, X., V. Castric, H. Hipperson, N. A. Müller, H. Westerdahl *et al.*, 2021 Whole-genome sequencing and genome regions of special interest: Lessons from major histocompatibility complex, sex determination, and plant self-incompatibility. Molecular Ecology n/a. https://doi.org/10.1111/mec.16020

Vernet, P., P. Lepercq, S. Billiard, A. Bourceaux, J. Lepart *et al.*, 2016 Evidence for the long-term maintenance of a rare self-incompatibility system in Oleaceae. New Phytologist 210: 1408-1417. https://dx.doi.org/10.1111/nph.13872

Wallander, E., 2001 Evolution of wind-pollination in
Fraxinus (Oleaceae): an ecophylogentic approach., pp. Göteborg University, PhD thesis.

Wallander, E., and V. A. Albert, 2000 Phylogeny and classification of Oleaceae based on rps16 and trnL-F sequence data. American Journal of Botany 87: 1827-1841.

Wang, S., and M. Gribskov, 2017 Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics 33: 327-333. https://dx.doi.org/10.1093/bioinformatics/btw625

Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. Huber *et al.*, 2009 gplots: Various R programming tools for plotting data. R package version 2: 1.

Wei, Q., I. K. Khan, Z. Ding, S. Yerneni and D. Kihara, 2017 NaviGO: interactive tool for visualization and functional similarity and coherence analysis with gene ontology. BMC Bioinformatics 18: 177. https://dx.doi.org/10.1186/s12859-017-1600-5

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-Statistics for the Analysis of Population Structure. Evolution 38: 1358-1370. https://dx.doi.org/10.2307/2408641

Wheeler, M. J., B. H. de Graaf, N. Hadjiosif, R. M. Perry, N. S. Poulter *et al.*, 2009 Identification of the pollen self-incompatibility determinant in Papaver rhoeas. Nature 459: 992-995. https://dx.doi.org/10.1038/nature08027

Wolf, D. E., J. A. Satkoski, K. White and L. H. Rieseberg, 2001 Sex determination in the androdioecious plant Datisca glomerata and its dioecious sister species D. cannabina. Genetics 159: 1243-1257.

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

Wybouw, B., and B. De Rybel, 2019 Cytokinin – A Developing Story. Trends in Plant Science 24: 177-185. https://doi.org/10.1016/j.tplants.2018.10.012

Yamato, K. T., K. Ishizaki, M. Fujisawa, S. Okada, S. Nakayama *et al.*, 2007 Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. Proc Natl Acad Sci U S A 104: 6472-6477. https://dx.doi.org/10.1073/pnas.0609054104

Yang, B., D. Thorogood, I. Armstead and S. Barth, 2008 How far are we from unravelling self-incompatibility in grasses? New Phytologist 178: 740-753.

Yasui, Y., M. Mori, J. Aii, T. Abe, D. Matsumoto *et al.*, 2012 S-LOCUS EARLY FLOWERING 3 is exclusively present in the genomes of short-styled buckwheat plants that exhibit heteromorphic self-incompatibility. PloS one 7: e31264-e31264. https://dx.doi.org/10.1371/journal.pone.0031264

Zemp, N., R. Tavares, A. Muyle, D. Charlesworth, G. A. B. Marais *et al.*, 2016 Evolution of sex-biased gene expression in a dioecious plant. Nature Plants 2: 16168. https://dx.doi.org/10.1038/nplants.2016.168

Zhang, C. C., L. Y. Wang, K. Wei, L. Y. Wu, H. L. Li *et al.*, 2016 Transcriptome analysis reveals self-incompatibility in the tea plant (Camellia sinensis) might be under gametophytic control. BMC Genomics 17: 359. https://dx.doi.org/10.1186/s12864-016-2703-5

Zhang, L.-B., M. P. Simmons, A. Kocyan and S. S. Renner, 2006 Phylogeny of the Cucurbitales based on DNA sequences of nine loci from three genomes: Implications for morphological and sexual system evolution. Mol Phylogenet Evol 39: 305-322. https://doi.org/10.1016/j.ympev.2005.10.002

Zhou, Q., J. Jia, X. Huang, X. Yan, L. Cheng *et al.*, 2014 The large-scale investigation of gene expression in Leymus chinensis stigmas provides a valuable resource for understanding the mechanisms of poaceae self-incompatibility. BMC Genomics 15: 399. https://dx.doi.org/10.1186/1471-2164-15-399

Zhou, R., D. Macaya-Sanz, C. H. Carlson, J. Schmutz, J. W. Jenkins *et al.*, 2020 A willow sex chromosome reveals convergent evolution of complex palindromic repeats. Genome Biol 21: 38. https://dx.doi.org/10.1186/s13059-020-1952-4

Zhou, X. J., Y. Y. Wang, Y. N. Xu, R. S. Yan, P. Zhao *et al.*, 2015 De Novo Characterization of Flower Bud Transcriptomes and the Development of EST-SSR Markers for the Endangered Tree Tapiscia sinensis. Int J Mol Sci 16: 12855-12870. https://dx.doi.org/10.3390/ijms160612855

Annexe

Figure S1. Phillyrea angustifolia sex-averaged linkage map showing the grouping and position of 15812 SNPs.

The length of each of the 23 linkage groups is indicated by the vertical scale in cM. The markers strictly linked to sex and self-incompatibility (SI) phenotypes are shown in red.

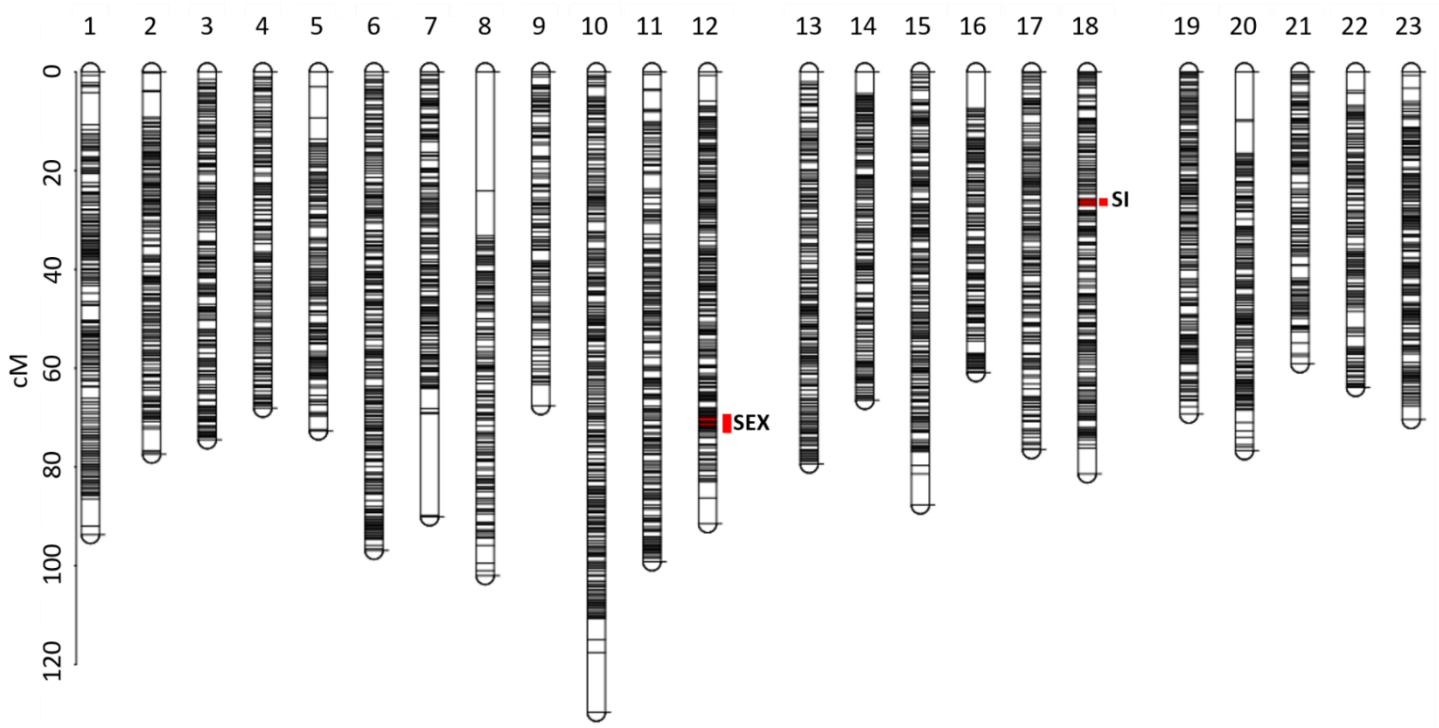


Figure S2. Comparison between the maternal and paternal genetic maps.

The vertical and horizontal lines on LG12 and LG18 indicate the position of markers strictly associated with sex and SI phenotypes.

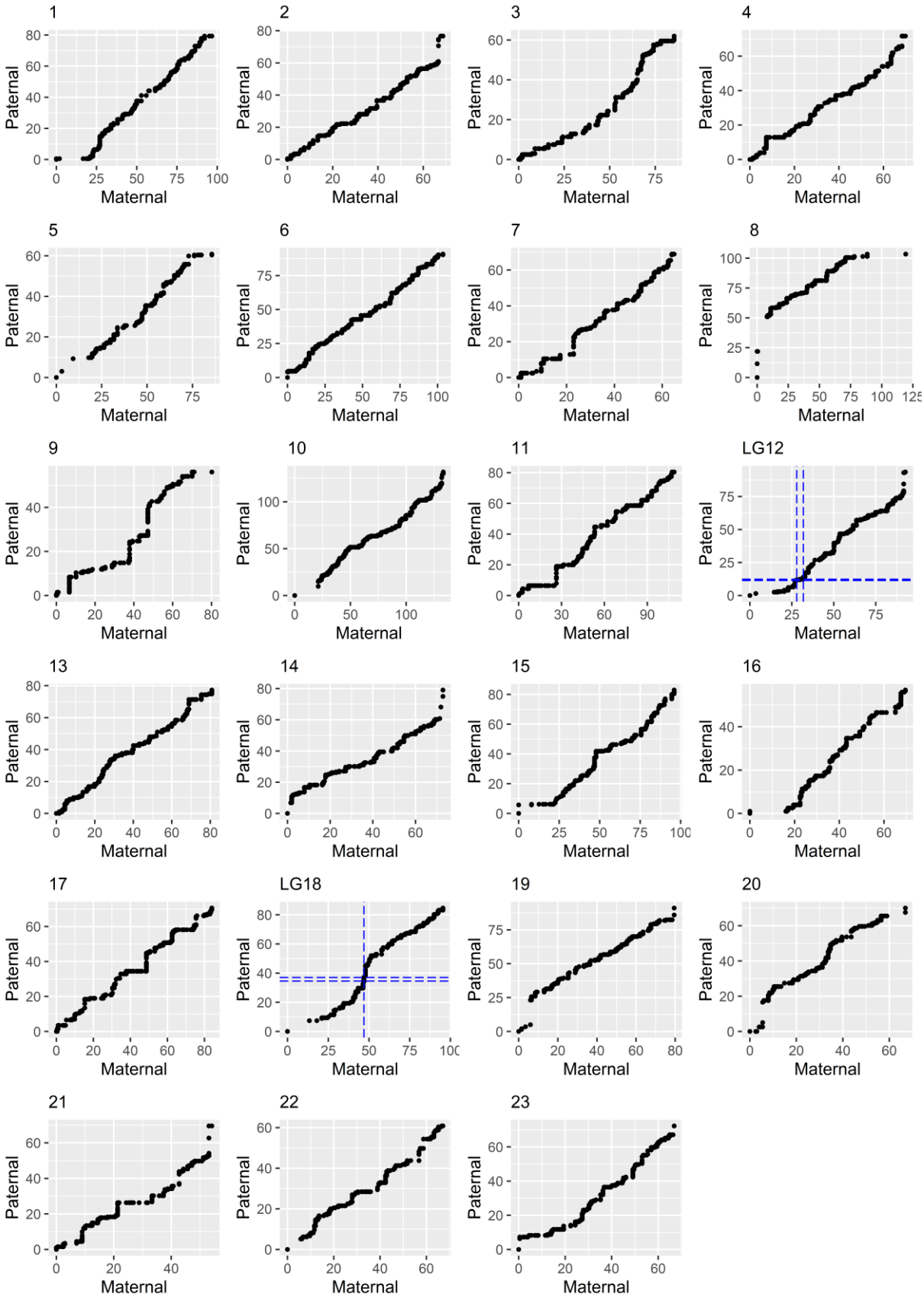


Table S1. List of the 82 gene annotations in the chromosomal interval of the olive tree genome bounded by *P. angustifolia* loci strictly associated with sex.

This list is completed by the 57 gene annotations in the five scaffolds containing loci strongly or loosely associated with sex in *P. angustifolia*. The description field is taken from Unver *et al.* (2020) and is based on automated blast annotation. NA corresponds to predicted gene models with no hit.

genome location	Sequence ID	Description	GO annotation
chr12	Oeu003179.1	ribonuclease h	
chr12	Oeu003182.1	50S ribosomal protein L15, chloroplastic	F:GO:0003735:structural constituent of ribosome ; P:GO:0006412:translation; C:GO:0015934:large ribosomal subunit
chr12	Oeu003184.1	caffeoylshikimate esterase	
chr12	Oeu003185.3	uncharacterized protein LOC111406129	
chr12	Oeu003188.1	uncharacterized protein LOC111406130	C:GO:0016020:membrane ; F:GO:0016757:glycosyltransferase activity
chr12	Oeu003330.1	uncharacterized protein LOC111406115 isoform X1	P:GO:0015031:protein transport
chr12	Oeu003331.1	probable NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 5, mitochondrial	P:GO:0022904:respiratory electron transport chain
chr12	Oeu003332.1	late embryogenesis abundant protein At1g64065-like	
chr12	Oeu003334.1	reticulon-like protein B5	
chr12	Oeu003335.2	WD repeat-containing protein 91 homolog	F:GO:0005515:protein binding
chr12	Oeu003336.1	zinc finger MYM-type protein 1-like	
chr12	Oeu003337.1	probable WRKY transcription factor 19	
chr12	Oeu003338.1	putative SNAP25 homologous protein SNAP30	
chr12	Oeu003339.1	Hypothetical predicted protein	
chr12	Oeu003341.1	uncharacterized protein LOC111406154 isoform X2	
chr12	Oeu003342.1	ABC transporter C family member 10-like	F:GO:0005524:ATP binding; C:GO:0016021:integral component of membrane; F:GO:0042626:ATPase-coupled transmembrane transporter activity; P:GO:0055085:transmembrane transport
chr12	Oeu003343.1	uncharacterized protein LOC111394170	P:GO:0000723:telomere maintenance; F:GO:0003678:DNA helicase activity; P:GO:0006281:DNA repair
chr12	Oeu003344.1	probable leucine-rich repeat receptor kinase At1g68400	F:GO:0004672:protein kinase activity; F:GO:0005524:ATP binding; P:GO:0006468:protein phosphorylation
chr12	Oeu003345.3	probable LRR receptor-like serine/threonine-protein kinase At1g51880	F:GO:0004672:protein kinase activity; F:GO:0005524:ATP binding; P:GO:0006468:protein phosphorylation; F:GO:0046872:metal ion binding
chr12	Oeu003346.1	probable inactive receptor kinase At4g23740	F:GO:0004672:protein kinase activity; F:GO:0005524:ATP binding; P:GO:0006468:protein phosphorylation
chr12	Oeu003347.1	pollen receptor-like kinase 1	F:GO:0004672:protein kinase activity; F:GO:0005524:ATP binding; P:GO:0006468:protein phosphorylation
chr12	Oeu003348.1	probable inactive receptor kinase At4g23740	F:GO:0004672:protein kinase activity; F:GO:0005524:ATP binding; P:GO:0006468:protein phosphorylation
chr12	Oeu010693.1	phytolongin Phyl2.1-like	C:GO:0016021:integral component of membrane
chr12	Oeu010694.1	protein RKD3	F:GO:0003700:DNA-binding transcription factor activity
chr12	Oeu010695.1	nephrocystin-3	F:GO:0005515:protein binding

chr12	Oeu010696.3	hypothetical protein H5410_017114	P:GO:0006606:protein import into nucleus; F:GO:0031267:small GTPase binding
chr12	Oeu010697.1	uncharacterized protein LOC111406111	
chr12	Oeu010698.1	Hypothetical predicted protein	
chr12	Oeu010700.1	thylakoid lumenal 17.4 kDa protein, chloroplastic-like isoform X1	
chr12	Oeu010701.1	Hypothetical predicted protein	
chr12	Oeu010702.1	WD repeat-containing protein YMR102C-like isoform X1	F:GO:0005515:protein binding
chr12	Oeu010703.1	pentatricopeptide repeat-containing protein At2g13600-like	F:GO:0005515:protein binding
chr12	Oeu010704.1	protein yippee-like At4g27745	
chr12	Oeu010705.1	oligopeptide transporter 6-like	P:GO:0055085:transmembrane transport
chr12	Oeu010706.1	molybdate-anion transporter-like	F:GO:0015098:molybdate ion transmembrane transporter activity; P:GO:0015689:molybdate ion transport; C:GO:0016021:integral component of membrane
chr12	Oeu010708.1	receptor-like protein 12	F:GO:0005515:protein binding
chr12	Oeu010709.1	pentatricopeptide repeat-containing protein At3g18020	F:GO:0005515:protein binding
chr12	Oeu010710.1	protein PHOSPHATE STARVATION RESPONSE 3-like	
chr12	Oeu010711.1	protein PHOSPHATE STARVATION RESPONSE 1-like	
chr12	Oeu010714.1	uncharacterized protein LOC111377053	P:GO:0006508:proteolysis; F:GO:0008234:cysteine-type peptidase activity
chr12	Oeu010715.1	Hypothetical predicted protein	
chr12	Oeu010716.1	uncharacterized protein LOC111407041	
chr12	Oeu010717.1	uncharacterized protein LOC111395991	
chr12	Oeu015503.1	nudix hydrolase 25-like	F:GO:0016787:hydrolase activity
chr12	Oeu015505.1	uncharacterized protein LOC111406139 isoform X2	
chr12	Oeu015506.1	uncharacterized protein LOC111406809	
chr12	Oeu015507.1	protein C2-DOMAIN ABA-RELATED 7-like isoform X1	
chr12	Oeu015509.1	uncharacterized protein LOC111390401	
chr12	Oeu015510.1	probable CCR4-associated factor 1 homolog 11	F:GO:0003676; F:GO:0004535; C:GO:0030014
chr12	Oeu015511.1	uncharacterized protein LOC111406141	
chr12	Oeu015512.1	inositol-tetrakisphosphate 1-kinase 1-like	F:GO:0000287:magnesium ion binding; F:GO:0005524:ATP binding; P:GO:0032957:inositol trisphosphate metabolic process; F:GO:0047325:inositol tetrakisphosphate 1-kinase activity; F:GO:0052725:inositol-1,3,4-trisphosphate 6-kinase; F:GO:0052726:inositol-1,3,4-trisphosphate 5-kinase
chr12	Oeu032517.1	5 -3 exoribonuclease 3	
chr12	Oeu032519.1	uncharacterized protein LOC111406806	F:GO:0003677:DNA binding; F:GO:0046983:protein dimerization activity
chr12	Oeu032520.1	Hypothetical predicted protein	
chr12	Oeu032521.1	uncharacterized protein LOC111406134	F:GO:0003677:DNA binding; F:GO:0046983:protein dimerization activity
chr12	Oeu032522.1	ATP-dependent DNA helicase 2 subunit KU80-like	

chr12	Oeu032523.1	uncharacterized protein LOC111373134	
chr12	Oeu037452.1	---NA---	
chr12	Oeu037453.1	---NA---	
chr12	Oeu037454.1	nucleolar complex protein 2 homolog	
chr12	Oeu037965.1	uncharacterized protein LOC111406151 isoform X2	
chr12	Oeu037967.1	GRF1-interacting factor 2-like	F:GO:0003713:transcription coactivator activity
chr12	Oeu037968.1	Hypothetical predicted protein	
chr12	Oeu037969.1	uncharacterized protein LOC111406816	C:GO:0000127:transcription factor TFIIIC complex; F:GO:0004402:histone acetyltransferase activity; F:GO:0005515:protein binding; P:GO:0006384:transcription initiation from RNA polymerase III promoter
chr12	Oeu037970.1	uncharacterized protein LOC111392799	
chr12	Oeu037972.1	uncharacterized protein LOC111404605	
chr12	Oeu037973.2	polyadenylate-binding protein-interacting protein 12-like	F:GO:0003676:nucleic acid binding
chr12	Oeu037975.2	uncharacterized protein LOC111406147	
chr12	Oeu037977.2	uncharacterized protein LOC111406815	P:GO:0006629:lipid metabolic process; F:GO:0008970:phospholipase A1 activity
chr12	Oeu037979.1	nascent polypeptide-associated complex subunit alpha-like protein 2	C:GO:0005854:nascent polypeptide-associated complex
chr12	Oeu037980.1	peroxidase 5-like	F:GO:0004601:peroxidase activity; P:GO:0006979:response to oxidative stress; F:GO:0020037:heme binding
chr12	Oeu037981.1	uncharacterized protein LOC111373264	
chr12	Oeu037982.1	heavy metal-associated isoprenylated plant 9	F:GO:0046872:metal ion binding
chr12	Oeu037983.1	60S ribosomal protein L9	F:GO:0003735:structural constituent of ribosome; C:GO:0005840:ribosome; P:GO:0006412:translation; F:GO:0019843:rRNA binding
chr12	Oeu037984.1	uncharacterized protein LOC105159007	C:GO:0005576:extracellular region; P:GO:0060320:rejection of self pollen
chr12	Oeu050214.1	transcription factor MYB12-like	
chr12	Oeu057647.1	hypothetical protein H0E87_010871	
chr12	Oeu057648.1	pentatricopeptide repeat-containing protein At2g03380, mitochondrial-like isoform X1	F:GO:0005515:protein binding
chr12	Oeu057649.1	uncharacterized protein LOC111406154 isoform X1	
chr12	Oeu057651.1	uncharacterized protein LOC111406155	

chr12	Oeu057652.1	eukaryotic translation initiation factor 5A-2-like	F:GO:0003746:translation elongation factor activity; F:GO:0043022:ribosome binding; P:GO:0045901:positive regulation of translational elongation; P:GO:0045905:positive regulation of translational termination
scaffold1196	Oeu005322.1	ADP-ribosylation factor 1-like 2 isoform X1	F:GO:0003924:GTPase activity; F:GO:0005525:GTP binding
scaffold1196	Oeu005323.1	---NA---	
scaffold1196	Oeu005324.1	serine/threonine-protein phosphatase 4 regulatory subunit 2 isoform X1	F:GO:0019888:protein phosphatase regulator activity; C:GO:0030289:protein phosphatase 4 complex
scaffold1196	Oeu005325.1	importin-5-like	P:GO:0006606:protein import into nucleus
scaffold1196	Oeu005327.1	homeobox-leucine zipper protein ATHB-17-like	F:GO:0003677:DNA binding
scaffold1196	Oeu005328.1	uncharacterized protein LOC111385861	
scaffold1264	Oeu007174.1	receptor kinase At4g00960	
scaffold1264	Oeu007175.1	---NA---	
scaffold1264	Oeu007176.1	Hypothetical predicted protein	
scaffold1264	Oeu007178.1	phosphoglucomutase, cytoplasmic	P:GO:0005975:carbohydrate metabolic process; F:GO:0016868:intramolecular transferase activity, phosphotransferases
scaffold1264	Oeu007179.1	glucan endo-1,3-beta-glucosidase 3-like isoform X1	F:GO:0004553:hydrolase activity, hydrolyzing O-glycosyl compounds; P:GO:0005975:carbohydrate metabolic process
scaffold1264	Oeu007181.1	probable peroxygenase 4	
scaffold1264	Oeu007182.1	ubiquitin-conjugating enzyme E2 variant 1A-like	
scaffold1264	Oeu007185.1	uncharacterized protein LOC111394498	
scaffold1264	Oeu007186.1	E3 ubiquitin- ligase RHA2A-like	
scaffold1264	Oeu007187.1	zinc finger CCCH domain-containing protein 15-like	F:GO:0046872:metal ion binding
scaffold1264	Oeu007190.2	lipid phosphate phosphatase 2-like	P:GO:0006644:phospholipid metabolic process; F:GO:0042577:lipid phosphatase activity
scaffold1264	Oeu007191.1	lipid phosphate phosphatase 2-like	P:GO:0006644:phospholipid metabolic process; F:GO:0042577:lipid phosphatase activity
scaffold1264	Oeu007192.1	uncharacterized protein LOC111374612	F:GO:0005515:protein binding
scaffold393	Oeu041861.2	tRNA(adenine(34)) deaminase, chloroplastic-like	
scaffold393	Oeu041862.2	phosphate transporter PHO1 homolog 1-like isoform X1	C:GO:0016021:integral component of membrane
scaffold393	Oeu041863.1	uncharacterized protein LOC111367484	
scaffold393	Oeu041866.1	uncharacterized protein LOC111392799	

scaffold393	Oeu041867.1	ATPase subunit 4	C:GO:0000276:mitochondrial proton-transporting ATP synthase complex, coupling factor F(o); F:GO:0015078:proton transmembrane transporter activity; P:GO:0015986:ATP synthesis coupled proton transport
scaffold393	Oeu041868.1	---NA---	
scaffold393	Oeu041869.1	---NA---	
scaffold393	Oeu041870.1	protein BOBBER 2-like	
scaffold393	Oeu041871.1	axoneme-associated protein mst101(2)-like	
scaffold393	Oeu041872.1	---NA---	
scaffold393	Oeu041874.1	protein decapping 5-like	
scaffold393	Oeu041875.1	protein CROWDED NUCLEI 2-like	C:GO:0005634:nucleus; P:GO:0006997:nucleus organization
scaffold393	Oeu041876.1	Hypothetical predicted protein	
scaffold393	Oeu041877.1	Hypothetical predicted protein	
scaffold393	Oeu041878.1	Hypothetical predicted protein	
scaffold393	Oeu041879.2	SUPPRESSOR OF GAMMA RESPONSE 1-like	F:GO:0003677:DNA binding; F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription, DNA-templated
scaffold393	Oeu041880.1	---NA---	
scaffold393	Oeu041881.1	uncharacterized protein LOC111386005	
scaffold393	Oeu041882.1	trafficking particle complex subunit 11	
scaffold393	Oeu041883.1	transcription factor HHO3-like	F:GO:0003677:DNA binding; F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription, DNA-templated
scaffold393	Oeu041885.1	transcription factor HHO3-like	F:GO:0003677:DNA binding; F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription, DNA-templated
scaffold393	Oeu041887.1	Hypothetical predicted protein	
scaffold393	Oeu041888.1	---NA---	
scaffold393	Oeu041891.1	GDT1-like protein 4 isoform X1	
scaffold393	Oeu041892.2	CTL-like protein DDB_G0274487 isoform X1	F:GO:0022857:transmembrane transporter activity; P:GO:0055085:transmembrane transport
scaffold393	Oeu041893.1	uncharacterized protein LOC111387182	
scaffold393	Oeu041894.1	serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform-like	F:GO:0005515:protein binding
scaffold393	Oeu041896.1	heavy metal-associated isoprenylated plant protein 6-like	
scaffold393	Oeu041897.1	heat stress transcription factor C-1-like isoform X1	F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription, DNA-templated; F:GO:0043565:sequence-specific DNA binding
scaffold393	Oeu041898.1	heat stress transcription factor C-1	
scaffold393	Oeu041901.1	transcription factor MYB106-like	
scaffold393	Oeu041902.1	---NA---	
scaffold969	Oeu064202.3	hypothetical protein DKX38_015652	F:GO:0005515:protein binding
scaffold969	Oeu064204.1	---NA---	

scaffold969	Oeu064205.1	vacuolar-sorting protein BRO1-like	F:GO:0005515:protein binding
scaffold969	Oeu064206.1	ATP synthase subunit beta, mitochondrial-like	C:GO:0000275:mitochondrial proton-transporting ATP synthase complex, catalytic sector F(1); F:GO:0005524:ATP binding; P:GO:0015986:ATP synthesis coupled proton transport; F:GO:0016887:ATP hydrolysis activity; F:GO:0046933:proton-transporting ATP synthase activity, rotational mechanism
scaffold969	Oeu064207.2	origin of replication complex subunit 4 isoform X4	C:GO:0000808:origin recognition complex; F:GO:0003677:DNA binding; F:GO:0005524:ATP binding; C:GO:0005634:nucleus; P:GO:0006260:DNA replication; F:GO:0016887:ATP hydrolysis activity
scaffold969	Oeu064208.1	nuclear pore complex protein NUP155 isoform X1	C:GO:0005643:nuclear pore; P:GO:0006913:nucleocytoplasmic transport; F:GO:0017056:structural constituent of nuclear pore

Table S2. List of the 32 gene annotations in the chromosomal interval of the olive tree genome bounded by *P. angustifolia* loci strictly associated with SI.

This list is completed by the 111 gene annotations in the eight scaffolds containing loci strongly or loosely associated with SI in *P. angustifolia*. The description field is taken from Unver *et al.* (2020) and is based on automated blast annotation. NA corresponds to predicted gene models with no hit.

Genome location	Sequence ID	Description	GO annotation
chr18	Oeu052727.1	uncharacterized protein LOC111404303	
chr18	Oeu037727.1	GATA transcription factor 5-like	P:GO:0006355:regulation of transcription, DNA-templated; F:GO:0008270:zinc ion binding; F:GO:0043565:sequence-specific DNA binding
chr18	Oeu016310.2	uncharacterized protein LOC111372111 isoform X1	
chr18	Oeu016307.4	Hypothetical predicted protein	
chr18	Oeu016306.2	tubby-like F-box protein 5 isoform X1	F:GO:0005515:protein binding
chr18	Oeu037732.1	uncharacterized protein LOC111367390	
chr18	Oeu037717.1	protein SHI RELATED SEQUENCE 1-like isoform X1	
chr18	Oeu037714.1	---NA---	
chr18	Oeu052725.1	CHAPERONE-LIKE PROTEIN OF POR1, chloroplastic	
chr18	Oeu037735.1	zinc finger MYM-type protein 1-like	
chr18	Oeu037740.1	uncharacterized protein LOC111373987	
chr18	Oeu052728.1	uncharacterized protein LOC111372566	
chr18	Oeu037721.1	hypothetical protein CRG98_024810	F:GO:0008168:methyltransferase activity; C:GO:0016021:integral component of membrane; P:GO:0032259:methylation
chr18	Oeu037725.1	uncharacterized protein LOC111367386	
chr18	Oeu037716.1	protein FAR1-RELATED SEQUENCE 4-like	
chr18	Oeu037719.1	---NA---	
chr18	Oeu037736.3	transmembrane and coiled-coil domain-containing protein 4	
chr18	Oeu037737.1	---NA---	
chr18	Oeu037738.1	uncharacterized protein LOC111394886	
chr18	Oeu016311.1	uncharacterized protein LOC111384618	
chr18	Oeu037724.1	uncharacterized protein LOC111367386	
chr18	Oeu037733.1	uncharacterized protein LOC111378408	
chr18	Oeu037720.1	uncharacterized protein LOC111368283	
chr18	Oeu037730.1	---NA---	
chr18	Oeu037734.1	---NA---	

chr18	Oeu016314.1	microtubule-associated TORTIFOLIA1	C:GO:0005874:microtubule; F:GO:0008017:microtubule binding; C:GO:0045298:tubulin complex
chr18	Oeu037731.1	cytochrome P450 84A1	F:GO:0004497:monooxygenase activity; F:GO:0005506:iron ion binding; F:GO:0016705:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; F:GO:0020037:heme binding
chr18	Oeu037741.1	---NA---	
chr18	Oeu037739.1	---NA---	F:GO:0003676:nucleic acid binding; F:GO:0008270:zinc ion binding
chr18	Oeu052726.1	---NA---	F:GO:0003677:DNA binding; F:GO:0005515:protein binding; F:GO:0046872:metal ion binding
chr18	Oeu016308.1	---NA---	F:GO:0004842:ubiquitin-protein transferase activity; P:GO:0016567:protein ubiquitination
chr18	Oeu016312.1	uncharacterized protein LOC111373662	
scaffold1199	Oeu005362.1	uncharacterized protein LOC111366962	
scaffold1199	Oeu005363.1	Hypothetical predicted protein	
scaffold1199	Oeu005364.1	protein indeterminate-domain 1-like isoform X1	
scaffold1199	Oeu005367.1	deSI-like protein At4g17486	F:GO:0008233:peptidase activity
scaffold1199	Oeu005368.1	peroxisomal membrane protein 11A	C:GO:0005779:integral component of peroxisomal membrane; P:GO:0016559:peroxisome fission
scaffold1199	Oeu005369.2	protein BRASSINAZOLE-RESISTANT 1-like	F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006351:transcription, DNA-templated; P:GO:0009742:brassinosteroid mediated signaling pathway
scaffold1199	Oeu005370.1	NEDD8-conjugating enzyme Ubc12-like	F:GO:0005524:ATP binding; F:GO:0016740:transferase activity
scaffold1199	Oeu005371.1	30S ribosomal protein S13, chloroplastic-like	F:GO:0003723:RNA binding; F:GO:0003735:structural constituent of ribosome; C:GO:0005840:ribosome; P:GO:0006412:translation
scaffold1199	Oeu005372.1	terminal ear1 homolog	F:GO:0003676:nucleic acid binding
scaffold1200	Oeu005581.1	transmembrane protein 87A-like	C:GO:0016021:integral component of membrane
scaffold1200	Oeu005582.1	exocyst complex component EXO70A1-like	C:GO:0000145:exocyst; P:GO:0006887:exocytosis
scaffold1200	Oeu005583.1	protein TIFY 11B-like	
scaffold1200	Oeu005584.1	putative B3 domain-containing protein Os03g0621600	
scaffold1200	Oeu005585.1	zinc finger CCCH domain-containing 48-like isoform X2	F:GO:0005515:protein binding; P:GO:0006364:rRNA processing; F:GO:0034511:U3 snoRNA binding; F:GO:0046872:metal ion binding
scaffold1200	Oeu005587.1	hypothetical protein CK203_000394	
scaffold1200	Oeu005588.1	zinc finger CCCH domain-containing protein 62-like	F:GO:0003723:RNA binding; P:GO:0045892:negative regulation of transcription, DNA-templated; F:GO:0046872:metal ion binding
scaffold1200	Oeu005589.1	25.3 kDa vesicle transport protein-like isoform X1	F:GO:0005484:SNAP receptor activity; P:GO:0006888:endoplasmic reticulum to Golgi vesicle-mediated transport; P:GO:0006890:retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum; C:GO:0016021:integral component of membrane
scaffold1200	Oeu005590.1	Hypothetical predicted protein	
scaffold1200	Oeu005591.1	ABC transporter B family member 19-like	F:GO:0005524:ATP binding; C:GO:0016021:integral component of membrane; P:GO:0055085:transmembrane transport; F:GO:0140359:ABC-type transporter activity
scaffold1287	Oeu007808.1	---NA---	
scaffold1287	Oeu007810.1	Hypothetical predicted protein	

scaffold1287	Oeu007811.1	Hypothetical predicted protein	
scaffold1287	Oeu007813.1	---NA---	
scaffold1287	Oeu007814.1	probable inactive poly [ADP-ribose] polymerase SRO3	F:GO:0003950:NAD+ ADP-ribosyltransferase activity
scaffold1287	Oeu007815.1	KH domain-containing protein HEN4-like isoform X1	F:GO:0003723:RNA binding
scaffold1579	Oeu014557.1	zinc finger BED domain-containing protein RICESLEEPER 2-like	
scaffold1579	Oeu014558.1	nitrate regulatory gene2 protein	
scaffold1579	Oeu014559.1	nitrate regulatory gene2 protein	
scaffold1579	Oeu014560.1	protein POLLENLESS 3-LIKE 2	F:GO:0005515:protein binding
scaffold1579	Oeu014561.2	histidine kinase 1 isoform X2	F:GO:0000155:phosphorelay sensor kinase activity; P:GO:0000160:phosphorelay signal transduction system; P:GO:0016310:phosphorylation
scaffold1579	Oeu014562.1	protein EARLY-RESPONSIVE TO DEHYDRATION 7, chloroplastic-like	
scaffold1579	Oeu014563.1	Hypothetical predicted protein	
scaffold1579	Oeu014567.1	alpha-mannosidase-like	F:GO:0004559:alpha-mannosidase activity; P:GO:0006013:mannose metabolic process; F:GO:0030246:carbohydrate binding
scaffold213	Oeu024812.1	uncharacterized protein LOC111367376 isoform X1	P:GO:0007064:mitotic sister chromatid cohesion
scaffold213	Oeu024815.1	zinc finger BED domain-containing protein RICESLEEPER 1-like	F:GO:0003677:DNA binding; F:GO:0046983:protein dimerization activity
scaffold213	Oeu024816.1	Hypothetical predicted protein	
scaffold213	Oeu024817.1	uncharacterized protein LOC111380358	
scaffold213	Oeu024818.1	blue copper -like	F:GO:0009055:electron transfer activity
scaffold213	Oeu024819.1	Hypothetical predicted protein	
scaffold213	Oeu024820.1	Hypothetical predicted protein	
scaffold213	Oeu024822.1	uncharacterized protein LOC120106791	
scaffold213	Oeu024824.1	ribosomal protein S2	F:GO:0003735:structural constituent of ribosome; P:GO:0006412:translation; C:GO:0009507:chloroplast; C:GO:0015935:small ribosomal subunit
scaffold213	Oeu024825.1	VQ motif-containing protein 9	P:GO:1901001:negative regulation of response to salt stress
scaffold213	Oeu024826.1	Hypothetical predicted protein	
scaffold213	Oeu024827.1	uncharacterized protein LOC111380349 isoform X4	
scaffold213	Oeu024828.1	AP-4 complex subunit mu-like	P:GO:0006886:intracellular protein transport; P:GO:0016192:vesicle-mediated transport; C:GO:0030131:clathrin adaptor complex
scaffold213	Oeu024829.1	uncharacterized protein LOC111406687	
scaffold213	Oeu024830.1	---NA---	
scaffold213	Oeu024832.1	Hypothetical predicted protein	
scaffold269	Oeu032218.1	protein TsetseEP-like	

scaffold269	Oeu032219.1	uncharacterized protein LOC111383442	
scaffold269	Oeu032220.1	DNA (cytosine-5)-methyltransferase 1B-like	F:GO:0003682:chromatin binding; F:GO:0003886:DNA (cytosine-5-)-methyltransferase activity; C:GO:0005634:nucleus; P:GO:0090116:C-5 methylation of cytosine
scaffold269	Oeu032221.1	---NA---	
scaffold269	Oeu032222.1	---NA---	
scaffold269	Oeu032223.3	uncharacterized protein LOC111383435	
scaffold269	Oeu032224.2	membrin-11-like	
scaffold269	Oeu032225.1	hypothetical protein F0562_029638	
scaffold269	Oeu032226.1	caltractin-like isoform X1	F:GO:0005509:calcium ion binding
scaffold269	Oeu032227.1	uncharacterized protein LOC111398231	
scaffold269	Oeu032228.1	uncharacterized protein LOC111394895	F:GO:0008270:zinc ion binding
scaffold269	Oeu032229.1	zinc-finger homeodomain protein 5-like	
scaffold269	Oeu032230.1	protein SUPPRESSOR OF GENE SILENCING 3	P:GO:0031047:gene silencing by RNA; P:GO:0051607:defense response to virus
scaffold269	Oeu032231.1	uncharacterized protein LOC111398231	
scaffold269	Oeu032236.1	protein SUPPRESSOR OF GENE SILENCING 3-like	P:GO:0031047:gene silencing by RNA; P:GO:0051607:defense response to virus
scaffold327	Oeu037602.1	60S ribosomal L34	F:GO:0003735:structural constituent of ribosome; C:GO:0005840:ribosome; P:GO:0006412:translation
scaffold327	Oeu037603.1	uncharacterized protein LOC111373208	
scaffold327	Oeu037604.1	uncharacterized protein LOC111373818	
scaffold327	Oeu037606.1	Hypothetical predicted protein	
scaffold327	Oeu037607.1	probable carbohydrate esterase At4g34215	
scaffold327	Oeu037608.1	probable carbohydrate esterase At4g34215	
scaffold327	Oeu037609.1	Hypothetical predicted protein	
scaffold327	Oeu037610.2	BUD13 homolog	
scaffold327	Oeu037611.1	uncharacterized protein LOC111399181	
scaffold327	Oeu037612.1	mannose-P-dolichol utilization defect 1 homolog 2-like	
scaffold327	Oeu037613.1	uncharacterized protein LOC111394498	
scaffold327	Oeu037614.1	uncharacterized protein LOC111385761	
scaffold327	Oeu037615.1	low-temperature-induced cysteine proteinase-like	P:GO:0006508:proteolysis; F:GO:0008234:cysteine-type peptidase activity
scaffold327	Oeu037616.1	Hypothetical predicted protein	
scaffold327	Oeu037617.1	---NA---	
scaffold327	Oeu037618.1	Hypothetical predicted protein	
scaffold327	Oeu037619.1	probable inactive receptor kinase At5g67200	F:GO:0004672:protein kinase activity; F:GO:0005515:protein binding; F:GO:0005524:ATP binding; P:GO:0006468:protein phosphorylation

scaffold327	Oeu037620.1	uncharacterized protein LOC111385749	
scaffold327	Oeu037621.1	kinesin-like protein KIN-4A isoform X1	F:GO:0003777:microtubule motor activity; F:GO:0005524:ATP binding; P:GO:0007018:microtubule-based movement; F:GO:0008017:microtubule binding
scaffold327	Oeu037623.1	ethylene-responsive transcription factor ERF008-like	F:GO:0003677:DNA binding; F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription, DNA-templated
scaffold327	Oeu037624.1	Hypothetical predicted protein	
scaffold502	Oeu047735.1	purple acid phosphatase 2-like	F:GO:0003993:acid phosphatase activity; F:GO:0046872:metal ion binding
scaffold502	Oeu047736.1	Hypothetical predicted protein	
scaffold502	Oeu047738.1	WALLS ARE THIN 1	C:GO:0016021:integral component of membrane; F:GO:0022857:transmembrane transporter activity
scaffold502	Oeu047739.1	unnamed protein product, partial	F:GO:0005198:structural molecule activity; F:GO:0005515:protein binding; P:GO:0006886:intracellular protein transport; P:GO:0016192:vesicle-mediated transport; C:GO:0030126:COPI vesicle coat
scaffold502	Oeu047747.1	B-box zinc finger 21-like	F:GO:0008270:zinc ion binding
scaffold502	Oeu047748.1	uncharacterized TPR repeat-containing protein At1g05150-like	F:GO:0005509:calcium ion binding; F:GO:0005515:protein binding
scaffold502	Oeu047750.1	probable 1-acylglycerol-3-phosphate O-acyltransferase	F:GO:0003824:catalytic activity
scaffold502	Oeu047751.2	hypothetical protein GIB67_019500	F:GO:0015078:proton transmembrane transporter activity; C:GO:0016021:integral component of membrane; C:GO:0033179:proton-transporting V-type ATPase, V0 domain; P:GO:1902600:proton transmembrane transport
scaffold502	Oeu047754.1	importin beta-like SAD2	
scaffold502	Oeu047755.1	cytokinin dehydrogenase 5-like	P:GO:0009690:cytokinin metabolic process; F:GO:0019139:cytokinin dehydrogenase activity; F:GO:0050660:flavin adenine dinucleotide binding
scaffold502	Oeu047756.1	Hypothetical predicted protein	
scaffold502	Oeu047757.1	Hypothetical predicted protein	
scaffold502	Oeu047759.1	probable ubiquitin-conjugating enzyme E2 18	
scaffold502	Oeu047760.1	bZIP transcription factor 11-like	F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription, DNA-templated
scaffold502	Oeu047761.1	bZIP transcription factor 11-like	F:GO:0003700:DNA-binding transcription factor activity; P:GO:0006355:regulation of transcription, DNA-templated
scaffold502	Oeu047764.1	Hypothetical predicted protein	
scaffold502	Oeu047765.2	bifunctional nuclease 2-like	F:GO:0004518:nuclease activity
scaffold502	Oeu047766.1	---NA---	
scaffold502	Oeu047767.2	uncharacterized protein LOC111400469	F:GO:0003676:nucleic acid binding; F:GO:0008270:zinc ion binding
scaffold502	Oeu047769.1	50S ribosomal protein L31, chloroplastic-like	F:GO:0003735:structural constituent of ribosome; C:GO:0005840:ribosome; P:GO:0006412:translation
scaffold502	Oeu047770.2	zinc finger MYM-type protein 1-like	F:GO:0046983:protein dimerization activity
scaffold502	Oeu047772.1	putative pectinesterase 11	F:GO:0030599:pectinesterase activity; P:GO:0042545:cell wall modification
scaffold502	Oeu047773.1	uncharacterized protein LOC111409242 isoform X1	
scaffold502	Oeu047774.1	cell division homolog 2-1, chloroplastic-like	F:GO:0003924:GTPase activity; F:GO:0005525:GTP binding

scaffold502	Oeu047775.1	uncharacterized protein LOC111391588
scaffold502	Oeu047778.1	Erythronate-4-phosphate dehydrogenase family

Annexe 2.1 Extraction d'ADN végétal en 96 puits

Protocole simplifié pour 2 plaques

Au KingFisher bille Chemagic 5 lavages à l'éthanol

Objectif :

Extraction d'ADN total issue de végétaux à partir de plantes plus difficile (vigne, olivier…)

On obtient de l'ADN végétal, en petite quantité (2 à 5µg) et de haute qualité

Extraction semi-automatiser

Référence bibliographique :

D'après le protocole mis au point par Sylvain Santoni (INRA Montpellier-Unité de Génétique et Amélioration des plantes).

Mode opératoire :

Avant de commencer :

Vérifier que vous avez :

►90ml de tampon d'extraction (rajouter extemporement 900mg de NaBisulfite) préchauffée à 50°C. Attention : *le tampon contient du SDS, si celui si à précipiter, remettre à chauffer la solution pour le redissoudre avant utilisation.*

►1,2ml de RNase (10mg/ml) conservé à -20°C

►30ml d'Acétate de K (conservé à 4°C)

►3.5ml de solution de CTAB 12.5%

►80ml d'AMMCGE

►65ml d'ethanol 96

►120ml d'AMMLav/E

►120ml de wash1

►240ml de wash2

►240ml d'ethanol 75

►20ml de TE 1X

►30ml de chemagic (bien vortexer avant d'utiliser !!)

Mettre le bain marie à 65°C

Préparer un bloc de glace

Mettre la centrifugation à 4°C

Sortir 5 plaques deepwell 96 spécifique KingFisher et 2 plaques standard 96 spécifique KingFisher et 1 tipcomb.

Sortir les 2 pipettes Eppendorf, 3 boites de pointes eppendorf et 9 réservoirs. Et la multipette stream

Attention : mettre en charge les pipettes, quelques heures avant la manip.

Pendant toute l'extraction, pour la pipette Eppendorf se mettre en **Tips long** dans option puis ajustage.

1- Broyage et extraction :

On pèse 10 à 15mg de feuille sèche par individu (tubes 1.2ml dans les blocs bleus).

Ajouter les billes de tungstène (1 bille par tube) en utilisant le distributeur de billes (X96)de chez Qiagen.

Broyer deux fois pendant **40 secondes à fréquence 25Hz** (retourner les blocs entre les 2 broyages).

Centrifuger **2min à 5600g** (6000rpm) pour faire tomber les éventuelles traces de poudre présente sur les bouchons.

Pour 2 plaques mélanger 90ml de tampon d'extraction et 900µl de RNase (10mg/ml)

Pipette Eppendorf: disp 400µl 3X speed ↑ à 5 speed ↓à 5

Distribuer 400µl du mélange par tube.

Rajouter à la multipette 1ml, 16µl de CTAB 12.5%

Retourner doucement la plaque pour mettre le culot en suspension et taper de chaque côté de la boite d'un coup vif sur la table (pour décoller les culots). Puis mélanger en secouant vivement pendant 20sec.

Centrifuger 30sec à 1500g.

Incuber 30min au bain marie à 65°C et sous agitation douce (60rpm).

Remarque : *Pour éviter que les bouchons s'ouvrent pendant l'incubation, laisser les tubes dans la boite, mettre la nacelle de l'ancienne centri dessus avec un poids dedans. Toutes les 10min retourner les plaques pour remettre le culot en suspension.*

2- Déproteinisation et filtration :

Pipette Eppendorf: disp 150µl 6X speed ↑ à 5 speed ↓à 5
Ajouter 150µl d'acétate de K (5M/3M) froid (dans la glace).

Mélanger en retournant doucement la plaque 2 ou 3 fois.

Incuber 5min dans la glace. *Pour cela sortir les tubes des racks pour qu'ils soient bien en contact avec la glace.*

Centrifuger 20 min à 5600g (6000rpm) à 4°C.

3- Préparation des plaques pour le kingfisher :

Wash 1 : dans une plaque deepwell 96 (KF), distribuer 600µl de Wash 1

Pipette Eppendorf : disp 600µl 2X speed ↑ à 5 speed ↓à 5 option ajustage: réglage usine
Wash 2 : dans une plaque deepwell 96 (KF), distribuer 600µl de Wash 2

Pipette Eppendorf : disp 600µl 2X speed ↑ à 5 speed ↓à 5 option ajustage: réglage usine
Ethanol 75% : dans une plaque deepwell 96 (KF), distribuer 600µl d'éthanol 75%

Pipette Eppendorf : disp 600µl 2X speed ↑ à 5 speed ↓à 5 option ajustage: ethanol 75%
AMMLAV/E : dans une plaque deepwell 96 (KF), distribuer 600µl de AMMLav/E

Pipette Eppendorf : disp 600µl 2X speed ↑ à 5 speed ↓à 5 option ajustage: ethanol 75%
Ethanol 75% : dans une plaque deepwell 96 (KF), distribuer 600µl d'éthanol 75

Pipette Eppendorf : disp 600µl 2X speed ↑ à 5 speed ↓à 5 option ajustage: ethanol 75%
Elution : dans une plaque standard 96 (KF), distribuer 100µl de TE

Pipette Eppendorf : disp 100µl 12X speed ↑ à 5 speed ↓à 5 option ajustage: réglage usine
Tip-combs : dans une plaque standard 96 (KF), mettre les tip-combs

Lyse :

Dans une plaque deepwell 96 (KF), distribuer 385µl de AMMCGE

Pipette Eppendorf : disp 385µl 3X speed ↑ à 5 speed ↓à 5 option ajustage: ethanol 75%
Rajouter en distribution 300µl d'ethanol 96.
Pipette Eppendorf : disp 300µl 4X speed ↑ à 5 speed ↓à 5 option ajustage: ethanol 75%
Ajouter à la multipette 15µl de Chemagic (attention bien vortexer avant pour avoir une solution homogène !).
Puis, **transférer 300µl** du surnageant dans la DPW96.

Pipette Eppendorf : pip 300µl speed ↑ à 2 speed ↓à 5 option ajustage: réglage usine

Mélanger par pipettage 2 ou 3 fois.

Mettre un alu adhesif.

A cette étape on peut laisser l'échantillon à 4°C pendant quelques heures (repas, nuit)

4- Chargement du kingfisher et lancement du programme:

Lancer le programme : Chemagic-DNAplant-eth5lav.

Appuyer sur ok et suivre les instructions du robot pour charger les plaques dans le bon ordre.

Le programme se lance.

Attendre 30min et décharger le robot.

Faire la même chose pour la 2ème plaque.

Remarque : *On peut lancer pendant ce temps 2 autres plaques en décaler si on a beaucoup d'échantillons.*

5- Stockage de l'ADN :

Mettre un alu adhésif sur la plaque et le stocker à 4°c en attendant de faire les vérifications.

Puis pour un stockage à long terme, mettre la plaque sur le portoir magnétique Macherey-Nagel, prélever à la multicanaux 100*l les échantillons et les transférer dans une boite 96 1.2ml (boite pour élution). Puis stocker à -20°C

6- Vérification de l'extraction :

Gel d'agarose 1% en plaque 96 puits.

Déposer 5µl d'ADN + 3µl de bleu 6X.

Remarque : *On peut quantifier quelques échantillons au qubit avec le kit BR (voir protocole du Qubit).*

7- Nettoyage et élimination des déchets :

Nettoyer les nacelles de la centrifugeuse à l'eau claire et sécher la centrifugeuse.

Nettoyer la paillasse et changer les papiers filtres.

Nettoyer les billes : récupérer toutes les billes et les mettre dans un mélange eau/RBS pendant 15 min. Puis mettre dans l'acide chloridrique (0.4N) pendant 1 h. Rincer, sécher, compter et ranger les billes.

Jeter les effluents dans le bidon de récupération spécifique prévus à cet usage.

Nettoyage spécifique des plaques et combs tips, **voir protocole spécifique**. Attention un certain nombre de plastique sont réutiliser. Merci de suivre les instructions.

Compléments :

1- Préparation des solutions :

Les solutions stocks suivante sont déjà disponible au laboratoire :Tris 1M ph=8,0 EDTA 0.5M,Acetate de potassium 5M, NaCl 5M, SDS 20%, TE 1X, Ethanol 96% (armoire à solvants) et Acide acétique glacial 37% (armoire à acides).

Elles servent de base aux autres solutions.

A- Tampon d'extraction (Tris ph=8 200mM, EDTA 50mM, NaCl 500mM, SDS 1.25% PVP 40 1% Nabisulfite 1g/100ml) :

Solutions/Poudres Stocks	Pour 100ml	Pour 200ml	Pour 500ml	**Pour 1l**
Tris 1M pH=8	20ml	20ml	50ml	**100ml**
EDTA 0.5M	10ml	20ml	50ml	**100ml**
NaCL 5M	10ml	20ml	50ml	**100ml**
SDS 20%	6.25ml	12.5ml	31.25ml	**62.5ml**
PVP 40000	1g	2g	5g	**10g**
Na bisulfite[1]	1g	2g	5g	**10g**
H2O	QSP 100ml	QSP 200ml	QSP 500ml	**QSP 1l**

[1] Ajouter le NaBisulfite au moment de l'utilisation.

B- Acétate de K (3M K et 5M Ac):

Pour 100ml, mélanger 60ml d'Acetate de K 5M, 11.5ml d'acide acétique glacial (37%) et 28.5ml d'H2O

C- Solution aqueuse à 12.5% de CTAB:

Peser 1.25g de poudre pour 10ml d'eau. Attention mettre sous agitateur chauffant car long à dissoudre.

D- AMMCG (Chlorure de guanidium) :

Préparer la solution de Chlorure de Guanidium 7.8M (Sigma G-3272). Peser 74.5g de chlorure de guanidium dans un bécher. Dans un autre bécher mettre 40ml d'H2O et mettre sous agitation (agitateur chauffant). Verser par petite quantité la poudre de chlorure de guanidium et attendre qu'elle soit dissoute avant d'en rajouter. Quand tout est dissous, compléter à 100ml dans une éprouvette.

Attention : ne se garde pas bien, la mettre rapidement en solution avec de l'éthanol (=AMMCGE) pour la conserver ou en préparer de petite quantité.

E- AMMCGE (Chlorure de guanidium 7.8M/Ethanol) :

Mélanger 1/3 de solution de chlorure de Guanidium 7.8M et 2/3 d'éthanol à 96%. Soit pour 2 plaques : 60ml de chlorure de guanidium et 120ml d'ethanol à 96%.

F- AMMLav/E (Ethanol et sels d'acétate de potassium) :

Préparer une solution aqueuse AMMlav (Acétate de K 160mM, Tris HCl pH=8 22.5mM, EDTA 0.1mM)

Solutions	Pour 100ml	Pour 400ml
Tris 1M pH=8	2.25ml	9ml
EDTA 0.5M	20µl	80µl
Acetate de K 5M	3.2ml	12.8ml
ddH_2O	QSP 100ml	QSP 400ml

Ensuite mélanger 100ml de AMMLav aqueuse et 170ml d'éthanol 96%.

G- Ethanol 75% :

En respectant les proportions de la table de Gay-Lussac mélangé 100ml d'ethanol 96% et 31ml d'eau distillée.

H- Perchlorate de Na 2M

Peser 14,65g dans 60ml d'eau distillée.

Attention c'est un CMR, mettre gants, blouses, masques pour le préparer. Préparer la quantité exacte à utiliser. Ne pas stocker de solution mère.

I- Wash 1 : Ne pas conserver plus d'une semaine

Solution finale	Produits/Solution initiale	Pour 60ml (100 ech)	Pour 125ml (200 ech)	Pour 250ml (400 ech)
Ethanol 30%	Ethanol 96%	18ml	37.5ml	75ml
Acetate de Na 0.15M	Acetate de Na anhydre	1.44g	3g	6g

Perchlorate de Na 1M	Perchlorate de Na 2M	30ml	62.5ml	125ml
Chlorure de Guanidium 1M	AMMCG 7.8M	7.8ml	16.25ml	32.5ml
Triton X100 1%	Triton X100	0.6ml	1.25ml	2.5ml
H2O		QSP 60ml	QSP 125ml	QSP 500ml

J- Wash 2 :Ne pas conserver plus d'une semaine

Solution finale	Produits/Solution initiale	Pour 60ml (100 ech	Pour 125ml (200 ech)	Pour 250ml (400 ech)
Ethanol 30%	Ethanol 96%	18ml	37.5ml	75ml
Perchlorate de Na 1M	Perchlorate de Na 2M	30ml	62.5ml	125ml
Chlorure de Guanidium 1M	AMMCG 7.8M	7.8ml	16.25ml	32.5ml
Triton X100 1%	Triton X100	0.3ml	0.625ml	1.25ml
H2O	H2O	QSP 60ml	QSP 125ml	QSP 500ml

K- Tampon d'élution TE :

TE 1X : Tris 10mM pH=8, EDTA 1mM

Annexe 2.2 Protocol préparation de banque avec le kit NEXT FLEX RAPID DNA SEQ KIT version 2.0

1. **Dosage des ADNs au Qubit kit DNA BR :**

1-1 Préparer de la Working Solution pour X échantillons + 2 tubes (les 2 standards)

Pour 1 tube, on a besoin de 199µl de tampon BR (bouteille dans placard) + 1µl de fluo BR (tube de solution rose dans boite grise dans le placard).
Soit pour N tubes : 199µl X N tampon BR + 1µl X N fluo BR. On peut faire le mélange dans un tube eppendorf 5ml.

1-2 Préparer les standards et les échantillons :

Sortir les tubes 0.5ml spécifique Qubit (dans le placard), écrire le nom de l'échantillon sur la paroi du tube pas sur le bouchon.

- Pour les standards : sortir les tubes du frigo (boite plastique blanche kit BR, tube à bouchon jaune et rouge).
  10µl de standard 1 + 190µl de Working Solution

  10µl de standard 1 + 190µl de Working Solution
- Pour les échantillons :
  2µl d'échantillon + 198µl Working Solution.
- On peut mettre d'abord à la pipette 10µl les ADNs et les standards. Puis on distribue à la multipette 1ml la Working Solution.

1-3 Vortexer ensuite tous les tubes 2 à 3 sec puis incuber 2min à température ambiante et à l'obscurité (dans le placard).

1-4 Ensuite on fait la lecture sur le Qubit en suivant les instructions de l'appareil. Se mettre juste au départ en DNA double brin puis kit BR.

2. **Fragmentation des ADNs :**

2-1 Fragmentation sur le RotorGene

- On a besoin de 100ng dans 64µl. Dans les tubes 0.5ml spécifique au bioruptor (sachet à côté de l'appareil) mettre la quantité d'ADN nécessaire pour avoir 100ng et le volume de Tris-HCl 10mM (ph8.0-8.5) nécessaire pour un volume de 64µl.
- Allumer 1/2h avant le RotorGene et le système de refroidissement (sous la paillasse).
- Centrifuger les tubes 30sec à 11000rpm (il faut éliminer toute les bulles)
- Mettre les tubes sur le rotor du RotorGene. Si on a moins de 12 tubes on complète avec des tubes vides. Attendre 5min pour qu'ils soient à 4°C

- Vérifier le programme : time ON 30sec, Time OFF 30sec, cycle numb : 20.
- Ensuite on lance le programme. (environ 20min)
- Remarque : si les échantillons d'ADNs pas assez pure la fragmentation peut ne pas marcher, faire un 3X bead-based cleanup avec bille AMPure (voir protocole).

2-2 Vérification sur Puce DNA HS sur le bioanalyser

Voir protocole pour la préparation de la puce.

On dépose 1µl de fragmentation. On peut faire migrer 11 échantillons à la fois.

Environ 30-40min de manip.

Si la fragmentation est bonne, on peut passer à la suite ou la congeler si on l'utilise le lendemain ou dans les 36h00. Ne pas attendre trop longtemps avant de faire la suite.

3-End Repair et A tailling :

- Sortir le tube « Nextflex End Repair & Adenylation Buffer Mix 2.0 » 1/2h avant pour le faire décongeler sur glace. (Tube transparent dans la boite de kit avec scotch rose, congélo tiroir du bas)
- Vortexer le tube 5 à 10sec puis mettre sur glace
- Préparation mix Repair & Adenylation dans une chainette PCR:

Pour 1 échantillon

ADN fragmenté	32µl
Nextflex end repair & adenylation Buffer mix 2.0	15µl
Nextflex end repair & adenylation enzyme mix 2.0	3µl
Vol total	50µl

Quand on a plus que 8 échantillons, on peut faire un mix avec le buffer et l'enzyme (mix N échantillons + 0.5)

- Vortexer la chainette et centrifuger
- Lancer le programme PCR :

30 min	20°C
30 min	65°C
End	4°C

Aller dans la machine PCR chez Chris/PerkinElmer/endrepairtail

- Faire immédiatement la ligation des adaptateurs (étape 4).

4- Ligation des adaptateurs :

Attention, c'est là qu'on met les barcode adaptateur (Nextflex Unique DualIndex barcodes) qui vont servir pour l'identification des individus après le séquençage illumina.

- Prendre dans le congélo les tubes d'index pré-aliquoté en chainette, les mettre à décongeler sur glace et centrifuger la chainette. (tube rose dans le kit au congélateur).
- Mettre à décongeler la « nextflex ligase buffer mix 2.0 » à température ambiante.
- Préparer le mix de ligation :

Pour 1 échantillon

ADN fragmenté et end repair	50µl
Nextflex ligase buffer kit mix	44.5µl
Nextflex barcode adaptater	2.5µl
Nextflex ligase enzyme	3µl
Volume total	100µl

On fait le mélange dans la chainette PCR où sont pré-aliquoter les barcode adaptateur.

On peut faire un pré-mix avec la ligase et le buffer (en préparer pour N échantillons +0.5 tubes).

Attention l'enzyme est très visqueuse, mélanger à la pipette une 10aine de fois, puis vortexer et centrifuger.

- Lancer le programme PCR :

15 min	20°C

Aller dans la machine PCR chez Chris/PerkinElmer/end repair tail

- Faire immédiatement l'étape suivante (post ligation clean-up)

5- Post-ligation cleanup

- Avant la manip :

Les billes AMPure doivent être sorties 20min avant leur utilisation et remise en suspension.

Préparer l'éthanol 80% juste avant (dans un falcon 20ml éthanol absolue + 5.718ml d'eau.

Sortir le tampon de resuspension. (dans le congélo ou le frigo)

- Dans des tubes 1.5ml lowbind mettre 35µl de billes AMpure +65µl d'eau (eau comprise dans le kit).
  Vortexer les billes avant de les distribuer.
  Si gouttes sur paroi, centrifuger 1 sec.
- Rajouter les 100µl de l'adaptateur ligation reaction product.
- Mélanger par vortex et/ou pipetage plusieurs fois
- Incuber les tubes à température ambiante 5 min pour lier l'ADN aux billes.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement à la P200, enlever et jeter le surnageant.
- Enlever du portoir et ajouter 200µl d'éthanol 80%, vortexer et attendre 30sec.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement à la P200, enlever et jeter l'éthanol.
- Enlever du portoir et ajouter 200µl d'éthanol 80%, vortexer et attendre 30sec.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement enlever et jeter l'éthanol. Essayer d'enlever toute trace d'éthanol résiduel sans toucher aux billes.
- Centrifuger les tubes pour culotter le reste d'éthanol.
- Placer les tubes sur le portoir et enlever à la P10 le reste d'éthanol.
- Sécher les billes à température ambiante pendant 3 min. Attention un séchage supérieur peut diminuer les rendements. (les billes ont un aspect humide et bombé, elles sont sèches quand aspect lisse).
- Enlever les tubes du portoir magnétique.
- Resuspendre les billes dans 25µl de solution de resuspension. Vortexer.
- Incuber les tubes à température ambiante pendant 2min pour éluer l'ADN hors des billes.
- Placer les tubes sur le portoir magnétique pour capturer les billes. Incuber jusqu'à ce que le liquide soit clair.
- Transférer 23µl de surnageant pour faire l'amplification des librairies en tube 0.2ml.

6-Amplification des librairies

- Sortir les tubes (vert) et les mettre à décongeler sur glace pendant qu'on fait la purification.
- Dans une chainette mettre par échantillon :

Produits	**Volume**
Adaptater-ligated library	23µl
Nextflex PCR master mix 2.0	25µl
Nextflex primer mix 2.0	2µl
Volume total	**50µl**

On peut faire un mix avec le MasterMix et les primers.

- Mélanger et centrifuger brièvement
- Amplifier en utilisant le protocole de cycle suivant :

Etape	Temp	Durée	Cycles
Dénaturation initiale	98°C	30 sec	1

Dénaturation	98°C	15 sec	5
Annealing*	60°C	30 sec	
Extension	72°C	30 sec	
Extension finale	72°C	2 min	1
HOLD	4°C	infini	1

- Nombre minimal pour avoir une amplification optimale (entre 5-8 cycles)
- Aller chez christelle/perkinElmer/amplifbanque

7-Post-amplification Cleanup

- Dans des tubes 1.5ml lowbind mettre les 45µl de billes AMpure. Vortexer les billes avant de les distribuer.Si gouttes sur paroi, centrifuger 1 sec.
- Rajouter les 50µl de produit PCR.
- Mélanger par vortex et/ou pipetage plusieurs fois
- Incuber les tubes à température ambiante 5 min pour lier l'ADN aux billes.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement à la P200, enlever et jeter le surnageant.
- Enlever du portoir et ajouter 200µl d'éthanol 80%, vortexer et attendre 30sec. Il faut que le culot se resuspende…
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement à la P200, enlever et jeter l'éthanol.
- Enlever du portoir et ajouter 200µl d'éthanol 80%, vortexer et attendre 30sec.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement enlever et jeter l'éthanol. Essayer d'enlever toute trace d'éthanol résiduel sans toucher aux billes.
- Centrifuger les tubes pour culotter le reste d'éthanol.
- Placer les tubes sur le portoir et enlever à la P10 le reste d'éthanol.
- Sécher les billes à température ambiante pendant 3 min. Attention un séchage supérieur peut diminuer les rendements. (les billes ont un aspect humide et bombé, elles sont sèches quand aspect lisse).
- Enlever les tubes du portoir magnétique.
- Resuspendre les billes dans 33µl de solution de resuspension. Vortexer.
- Incuber les tubes à température ambiante pendant 2min pour éluer l'ADN hors des billes.
- Placer les tubes sur le portoir magnétique pour capturer les billes. Incuber jusqu'à ce que le liquide soit clair.
- Transférer 30µl de surnageant.

Procéder directement à la double sélection de taille.

8-Selection de taille :

Pas recommander par Perkin mais les profils de banques sont bien meilleurs avec.

8-1 Réaliser la première size cut (0.7X) pour exclure les molécules plus grandes que 450pb.

- Dans un tube lowbinding 1.5ml :

Produit	Volume
ADN dont on veut sélectionner la taille	30µl
AMPure Beads	21µl
Volume total	51µl

- Mélanger doucement en vortexant et/ou en pipetant plusieurs fois.
- Incuber les tubes à température ambiante pendant 5 min pour lier aux billes les fragments plus grands que 450pb.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement transférer 50µl du surnageant contenant les molécules plus petites que 450pb dans un nouveau tube 1.5ml. Attention à ne pas transférer de billes avec le surnageant. Jeter le tube contenant les billes.

8-2 Réaliser la deuxième sélection de taille (0.9X) pour retenir les fragments de taille inférieure à 250pb.

- Dans un tube 1.5ml :

Produit	Volume
ADN dont on veut sélectionner la taille	50µl
AMPure Beads	6.25µl
Volume total	56.25µl

- Mélanger par vortex et/ou pipetage plusieurs fois.
- Incuber la plaque ou les tubes à température ambiante 5 min pour lier l'ADN aux billes.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement enlever et jeter le surnageant.
- Enlever du portoir et ajouter 200µl d'éthanol 80%, vortexer et attendre 30sec.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement enlever et jeter l'éthanol
- Enlever du portoir et ajouter 200µl d'éthanol 80%, vortexer et attendre 30sec.
- Placer les tubes sur le portoir magnétique. Incuber jusqu'à ce que le liquide soit clair.
- Doucement enlever et jeter l'éthanol. Essayer d'enlever toute trace d'éthanol résiduel sans toucher aux billes.
- Centrifuger les tubes pour culotter le reste d'éthanol.
- Placer les tubes sur le portoir et enlever à la P10 le reste d'éthanol.
- Sécher les billes à température ambiante pendant 3min (les billes ont un aspect humide et bombé, elles sont sèches quand elles ont un aspect lisse).
- Enlever les tubes du portoir magnétique.

- Resuspendre les billes dans 25µl de tampon de resuspension. Vortexer.
- Incuber les tubes à température ambiante pendant 2min pour éluer l'ADN hors des billes.
- Placer les tubes sur le portoir magnétique pour capturer les billes. Incuber jusqu'à ce que le liquide soit clair.
- Récupérer les 22µl de surnageant.

9-Vérification sur puce Agilent HS et quantification sur Qubit:

9-1 Quantification au Qubit kit HS

A- Préparer de la Working Solution pour X échantillons + 2 tubes (les 2 standards)

Pour 1 tube, on a besoin de 199µl de tampon HS (bouteille dans placard) + 1µl de fluo HS (tube de solution rose dans boite grise dans le placard).

Soit pour N tubes : 199µl X N tampon HS + 1µl X N fluo HS. On peut faire le mélange dans un tube eppendorf 5ml.

B- Préparer les standards et les échantillons :

Sortir les tubes 0.5ml spécifique Qubit (dans le placard), écrire le nom de l'échantillon sur la paroi du tube pas sur le bouchon.

- Pour les standards : sortir les tubes du frigo (boite plastique blanche **kit HS**, tube à bouchon jaune et rouge).
  10µl de standard 1 + 190µl de Working Solution

  10µl de standard 1 + 190µl de Working Solution
- Pour les échantillons :
  1µl échantillon + 199µl Working Solution.
- On peut mettre d'abord à la pipette 10µl les ADns et les standards. Puis on distribue à la multipette 1ml la Working Solution.

1-2 Vortexer ensuite tous les tubes 2 à 3 sec puis incuber 2min à température ambiante et à l'obscurité (dans le placard).

1-3 Ensuite on fait la lecture sur le Qubit en suivant les instructions de l'appareil. Se mettre juste au départ en DNA double brin puis kit HS.

9-2 Puce DNA HS sur le Bioanalyser Agilent :

Voir protocole pour la préparation de la puce.

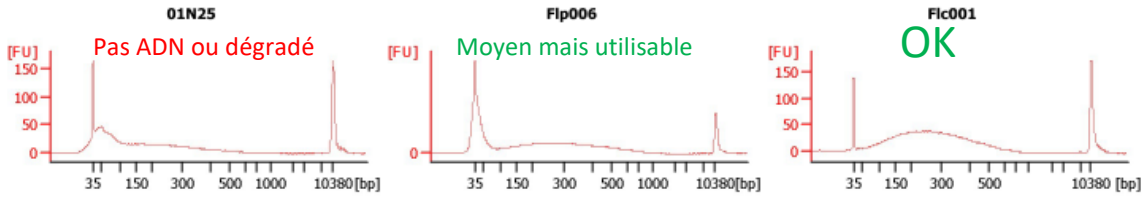
On dépose 1µl de fragmentation. On peut faire migrer 11 échantillons à la fois.

Environ 30-40min de manip.

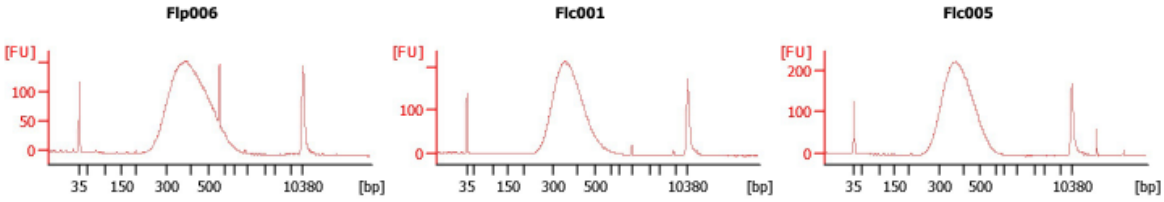
Stocker l'ADN jusqu'à la capture à -80°C sauf si on fait dans les 24h la capture (-20°C suffit).

Compléments :

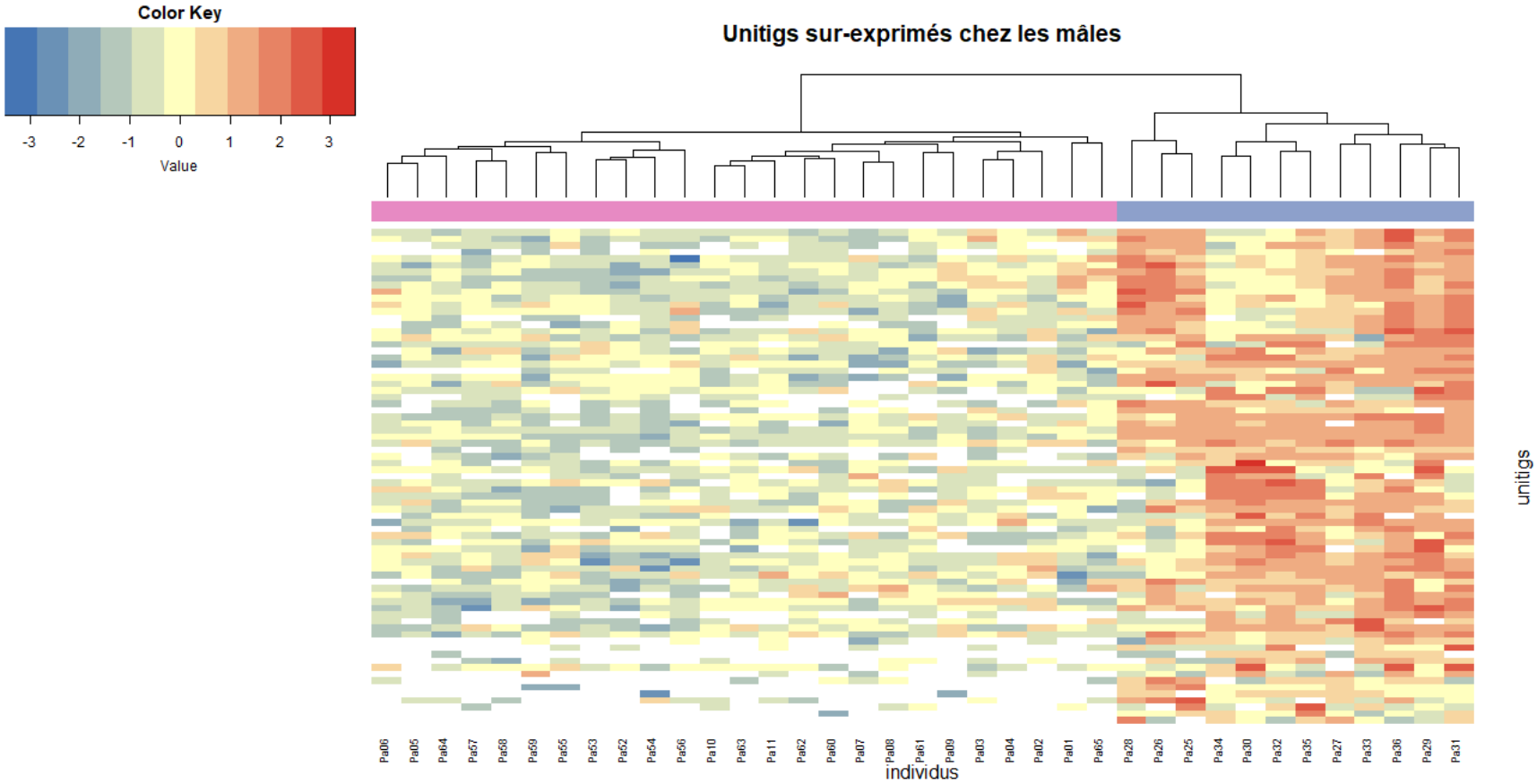
Profils de fragmentation



profils de banque

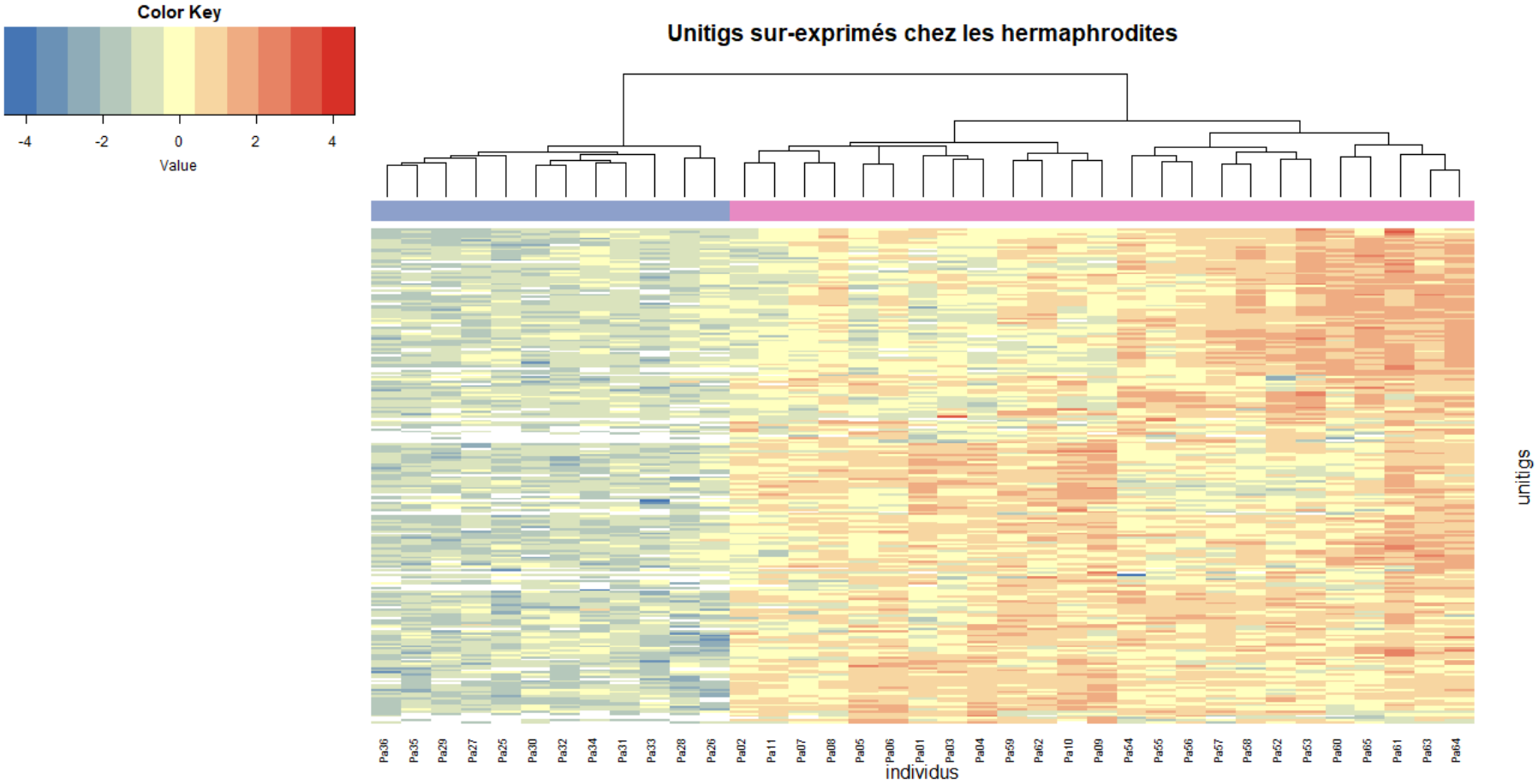


Annexe 2.3 : heatmap des unitigs différentiellement sur-exprimés entre les individus mâles et hermaphrodites utilisés comme cible pour l'expérience de capture



Chaque colonne représente un individu et chaque ligne un unitigs. Les couleurs représentent la divergence d'expression d'un gène particulier dans un échantillon particulier, par rapport à la valeur moyenne de ce gène sur tous les échantillons, centré réduit sur 0 en unités d'écarts types (bleu peu exprimé, jaune expression moyenne, rouge fortement exprimé). En bleu, le regroupement des individus mâles et en rose, les individus hermaphrodites

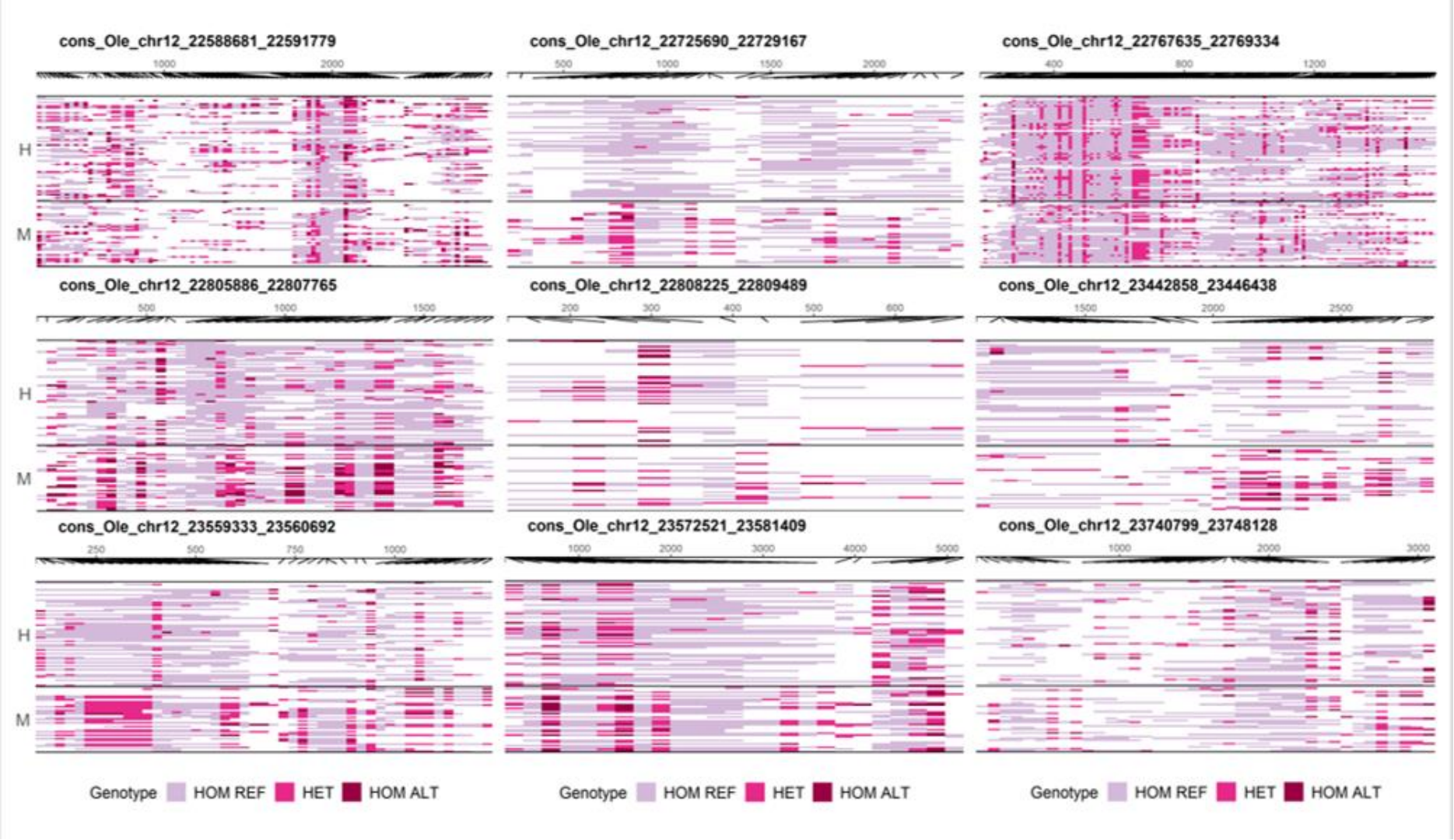
Annexe 2.4 : heatmap des unitigs différentiellement sur-exprimés entre les individus hermaphrodites et mâles utilisés comme cible pour l'expérience de capture

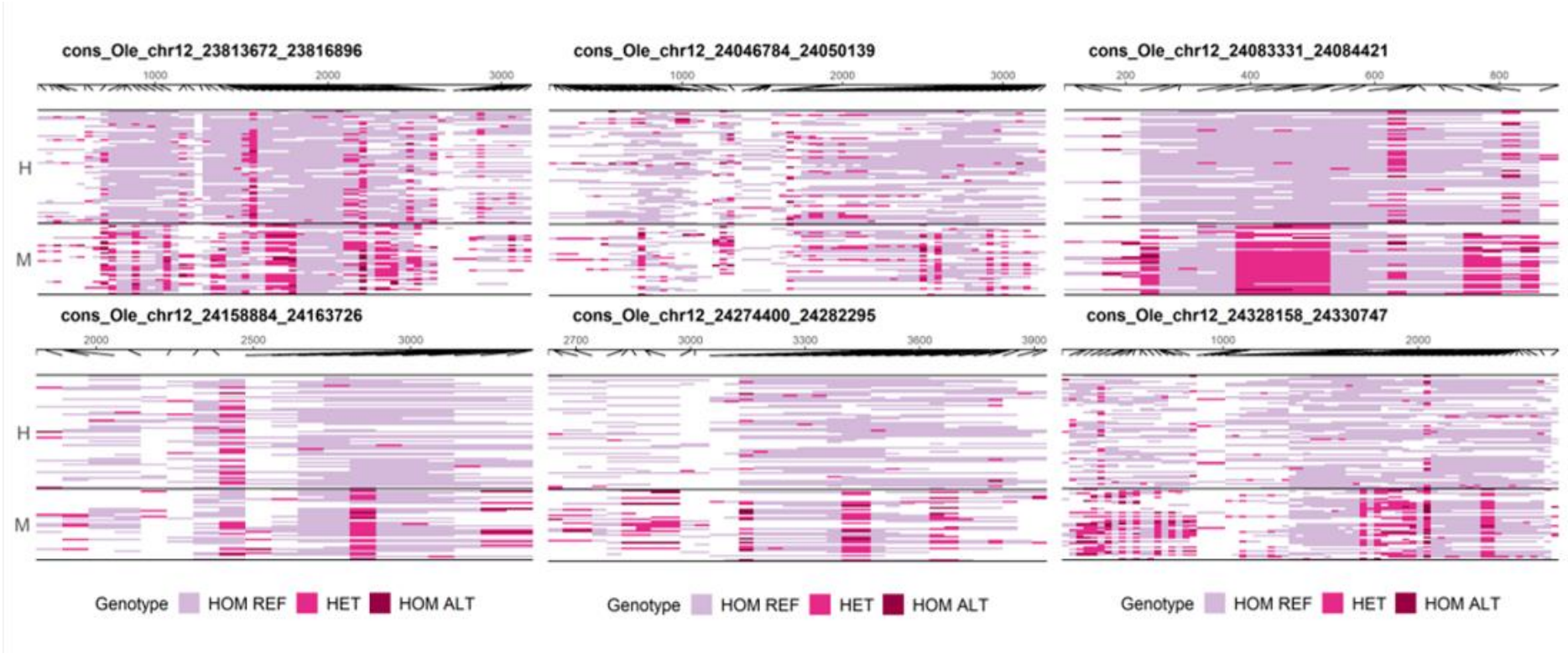


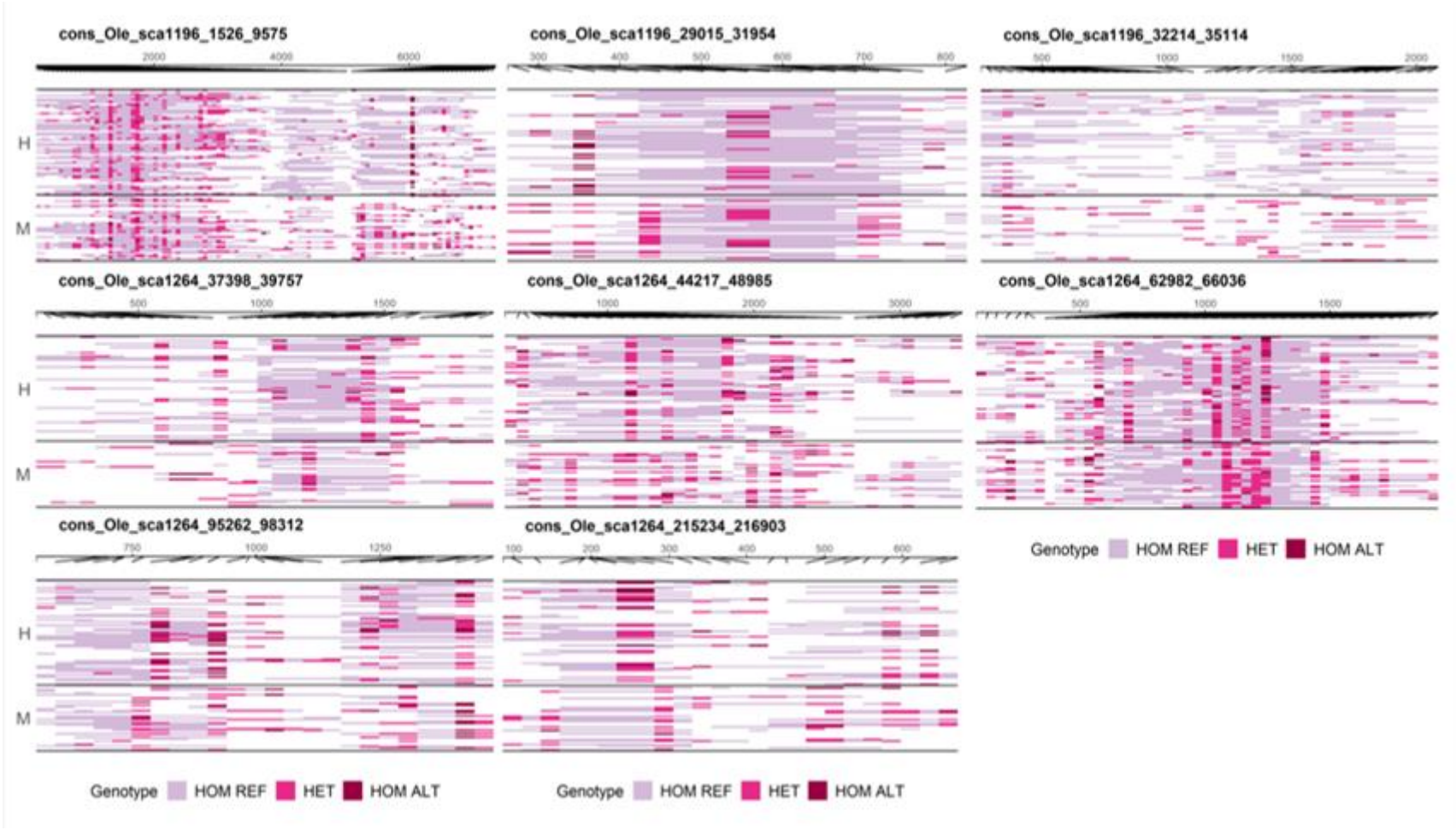
Chaque colonne représente un individu et chaque ligne un unitigs. Les couleurs représentent la divergence d'expression d'un gène particulier dans un échantillon particulier, par rapport à la valeur moyenne de ce gène sur tous les échantillons, centré réduit sur 0 en unités d'écarts types (bleu peu exprimé, jaune expression moyenne, rouge fortement exprimé). En bleu, le regroupement des individus mâles et en rose, les individus hermaphrodites

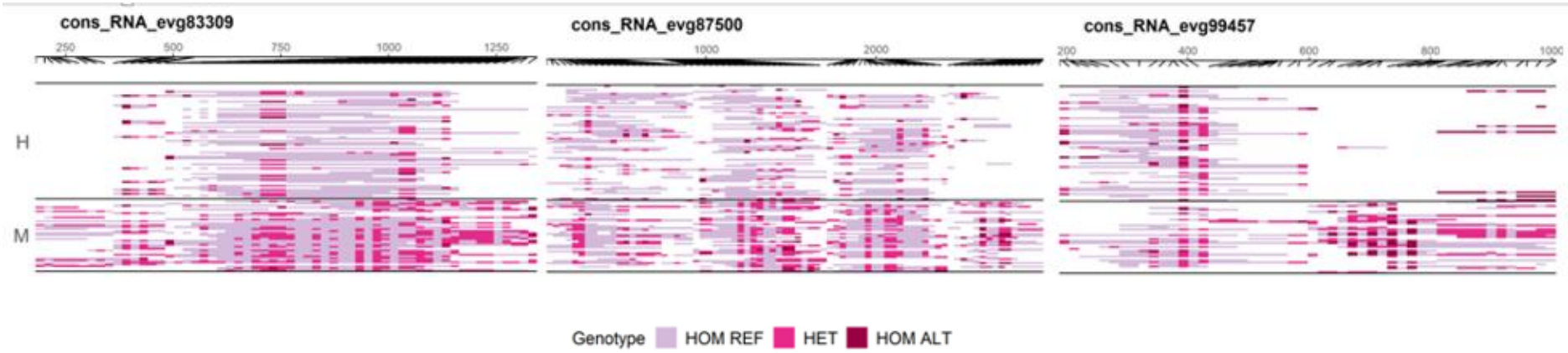
Annexe 2.5 : Visualisation graphique des génotypes pour les 26 séquences d'intérêts génétiquement liées au sexe

On y observe l'état allélique de chacun des SNPs en fonction de leur position sur la séquence (axe horizontal). Chaque individu est représenté par une ligne, et les individus sont regroupés par phénotype (axe vertical)

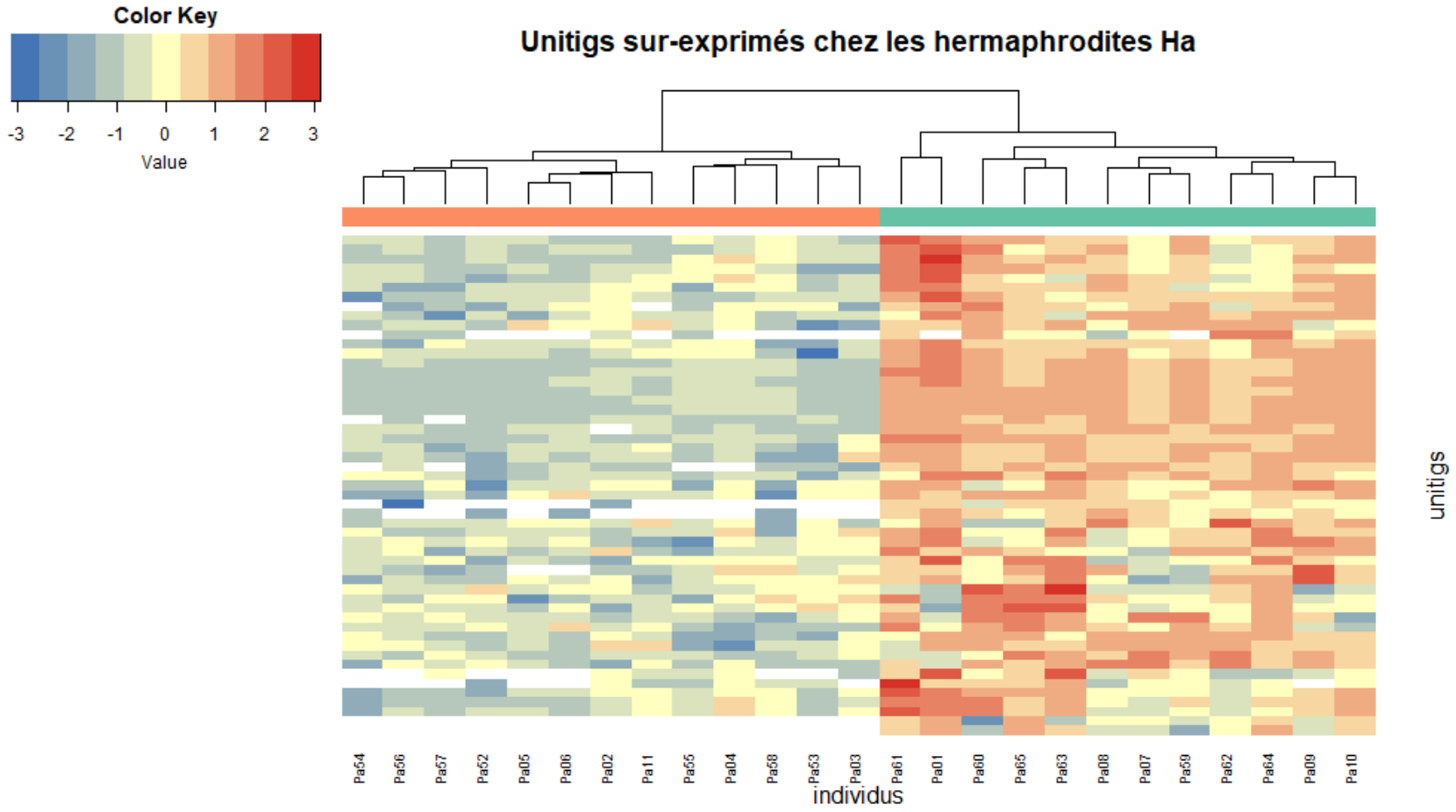



cons_Ole_chr12_23813672_23816896
1000
2000
3000
H
M
cons_Ole_chr12_24046784_24050139
1000
2000
3000
cons_Ole_chr12_24083331_24084421
200
400
600
800
cons_Ole_chr12_24158884_24163726
2000
2500
3000
H
M
cons_Ole_chr12_24274400_24282295
2700
3000
3300
3600
3900
cons_Ole_chr12_24328158_24330747
1000
2000
Genotype HOM REF HET HOM ALT
Genotype HOM REF HET HOM ALT
Genotype HOM REF HET HOM ALT


cons_Ole_sca1196_1526_9575
cons_Ole_sca1196_29015_31954
cons_Ole_sca1196_32214_35114
cons_Ole_sca1264_37398_39757
cons_Ole_sca1264_44217_48985
cons_Ole_sca1264_62982_66036
cons_Ole_sca1264_95262_98312
cons_Ole_sca1264_215234_216903
H
M
Genotype HOM REF HET HOM ALT

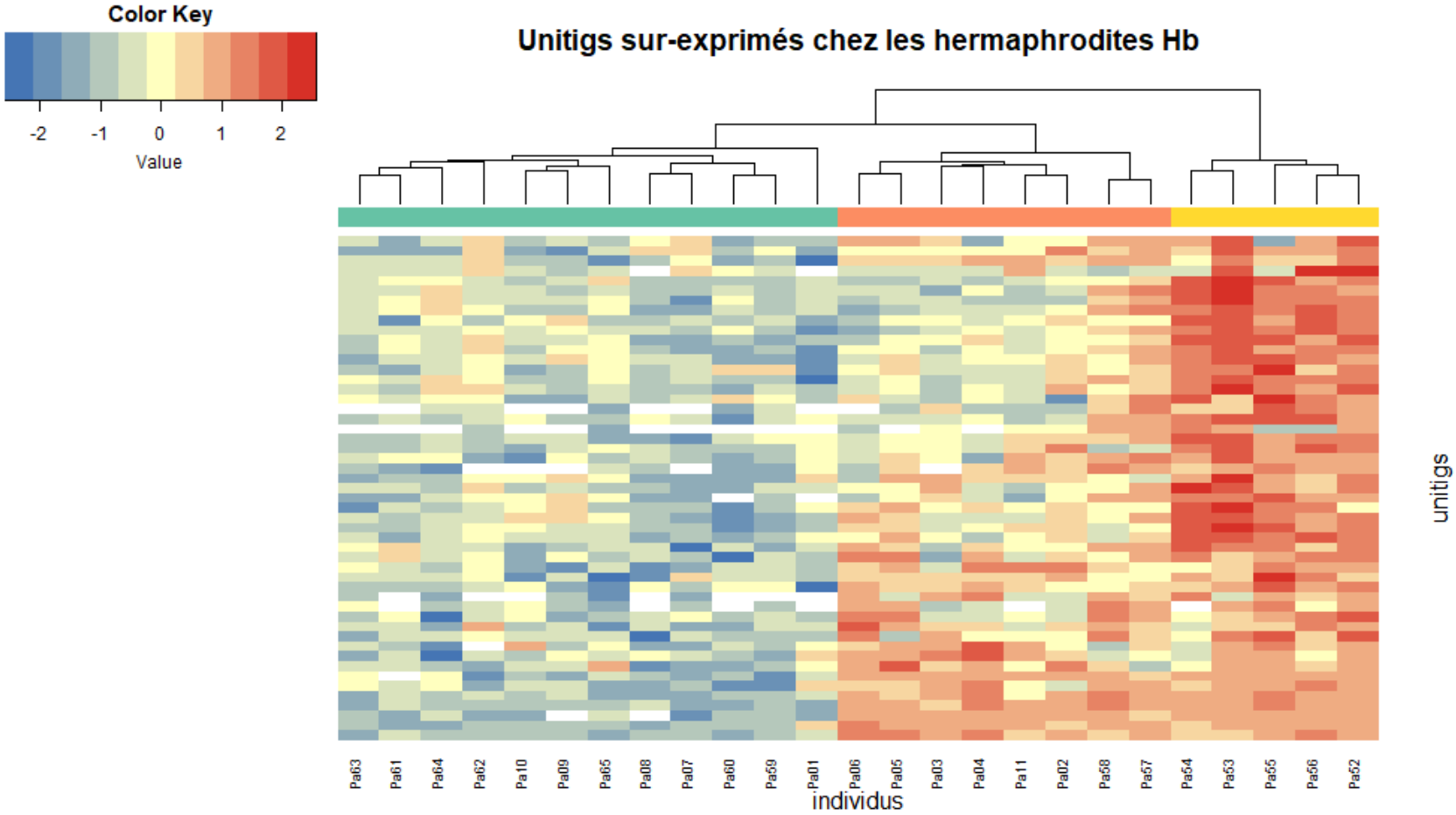

cons_RNA_evg83309
250
500
750
1000
1250
cons_RNA_evg87500
1000
2000
cons_RNA_evg99457
200
400
600
800
1000
H
M
Genotype
HOM REF
HET
HOM ALT

Annexe 3.1 : heatmap des unitigs différentiellement sur-exprimés entre les individus Ha et Hb utilisés comme cible pour l'expérience de capture



Chaque colonne représente un individu et chaque ligne un unitigs. Les couleurs représentent la divergence d'expression d'un gène particulier dans un échantillon particulier, par rapport à la valeur moyenne de ce gène sur tous les échantillons, centré réduit sur 0 en unités d'écarts types (bleu peu exprimé, jaune expression moyenne, rouge fortement exprimé). En vert, le regroupement des individus Ha et orange des individus Hb

Annexe 3.2 : heatmap des unitigs différentiellement sur-exprimés entre les individus Hb et Ha utilisés comme cible pour l'expérience de capture



Chaque colonne représente un individu et chaque ligne un unitigs. Les couleurs représentent la divergence d'expression d'un gène particulier dans un échantillon particulier, par rapport à la valeur moyenne de ce gène sur tous les échantillons, centré réduit sur 0 en unités d'écarts types (bleu peu exprimé, jaune expression moyenne, rouge fortement exprimé). En vert, le regroupement des individus Ha et les individus Hb se séparent en deux sous-groupes en orange et jaune

Résumé

Chez les angiospermes les systèmes d'accouplement et sexués présentent une diversité spectaculaire, mais sont souvent mal connus. Élucider leur déterminisme génétique et identifier les facteurs moléculaires contrôlant leur évolution représentent des enjeux majeurs en biologie évolutive et en génomique. Dans la famille des Oleaeceae, l'espèce *Phillyrea angustifolia* représente un organisme modèle passionnant, car l'espèce présente deux systèmes de compatibilité sexuelle distincts et rares : androdioécie (coexistence d'individus mâles et hermaphrodites) et un système homomorphe d'auto-incompatibilité ne comportant que deux spécificités alléliques qui sépare les hermaphrodites en deux groupes tandis que les mâles sont compatibles avec les deux groupes.

Au cours de cette thèse, j'ai développé plusieurs approches afin d'identifier et de caractériser les régions génomiques associées aux déterminismes du sexe et de l'auto-incompatibilité chez *P. angustifolia*. Dans un premier temps j'ai utilisé une approche de cartographie génétique haute-densité, basée sur un croisement F1, ce qui m'a permis de confirmer que les locus du sexe et de l'auto-incompatibilité sont situés dans deux groupes de liaison distincts (LG12 et LG18) et correspondent à des systèmes de type XY. La carte de liaison de *P. angustifolia* montre une synténie robuste avec le génome de l'olivier dans son ensemble et plusieurs marqueurs strictement associés au phénotype d'auto-incompatibilité chez *P. angustifolia* colocalisent avec la région chromosomique de 741 kb récemment identifiée comme étant liée au phénotype d'auto-incompatibilité chez l'olivier. J'ai ensuite développé une approche transcriptomique qui m'a permis d'identifier plusieurs transcrits dont l'expression est associée soit au phénotype sexuel soit au groupe d'incompatibilité, qui sont des candidats robustes pour identifier les déterminants moléculaires responsables et/ou participant à la détermination des deux phénotypes étudiés. Enfin, j'ai utilisé les résultats de la cartographie génétique et de l'analyse transcriptomique, pour établir une liste de 560 séquences cibles que nous avons capturées par hybridation ciblée chez *P. angustifolia*. Cette dernière approche effectuée sur des individus issus du croisement de cartographie et de populations naturelles, ce qui nous a permis de réduire à 69 728 pb la taille de la région liée au sexe initialement mise en évidence dans la cartographie génétique et d'identifier 26 gènes candidats. La mise en œuvre de cette approche pour l'analyse génétique des phénotypes d'auto-incompatibilité s'est révélée plus difficile, possiblement en raison de remaniements chromosomiques plus importants. Nous avons cependant pu exclure la liaison génétique avec les phénotypes d'incompatibilités des séquences candidates issues de l'analyse transcriptomique. Certains transcrits ayant une expression spécifique à chacun des groupes de compatibilité restent cependant des candidats intéressants pour comprendre la cascade moléculaire intervenant dans le déterminisme du phénotype d'incompatibilité.

Globalement, ce travail de thèse constitue une avancée dans la perspective de l'identification des déterminants moléculaires de ces deux phénotypes sexuels, ce qui constituera une étape importante dans la compréhension de leur évolution conjointe au sein de la famille des Oleaeceae.