



**HAL**  
open science

# Analyses empiriques des étapes précoces et tardives de l'évolution des chromosomes sexuels chez les plantes grâce à *Silene acaulis*, *Cannabis sativa* et *Humulus lupulus*

Djivan Prentout

► **To cite this version:**

Djivan Prentout. Analyses empiriques des étapes précoces et tardives de l'évolution des chromosomes sexuels chez les plantes grâce à *Silene acaulis*, *Cannabis sativa* et *Humulus lupulus*. Génomique, Transcriptomique et Protéomique [q-bio.GN]. Université de Lyon, 2021. Français. NNT : 2021LYSE1278 . tel-03708847

**HAL Id: tel-03708847**

**<https://theses.hal.science/tel-03708847>**

Submitted on 29 Jun 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N°d'ordre NNT :  
2021LYSE1278



**THESE de DOCTORAT DE L'UNIVERSITE DE LYON**  
opérée au sein de  
**l'Université Claude Bernard Lyon 1**

**Ecole Doctorale 341**  
Ecosystèmes Evolution Modélisation Microbiologie

**Spécialité de doctorat :**  
Génomique évolutive

Soutenue publiquement le 03/12/2021, par :

**Djivan Prentout**

---

**Analyses empiriques des étapes précoces et tardives de l'évolution des chromosomes sexuels chez les plantes grâce à *Silene acaulis*, *Cannabis sativa* et *Humulus lupulus***

---

Devant le jury composé de :

**Dr. Frédérique ABERLENC**, IR, IRD, DIADE  
**Dr. Tatiana GIRAUD**, DR, CNRS, ESE

Rapporteure  
Rapporteure

**Dr. Christophe DOUADY**, Pr., UCBL Lyon 1  
**Dr. Christelle FRAISSE**, CR, CNRS, EEP

Examinateur  
Examinatrice

**Dr. Gabriel MARAIS**, DR, CNRS, LBBE-CIBIO  
**Dr. Jos KÄFER**, CR, CNRS, LBBE

Directeur de thèse  
Co-directeur de thèse

# **Université Claude Bernard – LYON 1**

Président de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-Président du Conseil d'Administration	M. Didier REVEL
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. Philippe CHEVALLIER
Vice-Président de la Commission de Recherche	M. Petru MIRONESCU
Directeur Général des Services	M. Pierre ROLLAND

## **COMPOSANTES SANTE**

Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté de Médecine Lyon-Est	Doyen : M. Gilles RODE
Institut des Sciences et Techniques de la Réadaptation (ISTR)	Directeur : M. Xavier PERROT
Institut des Sciences Pharmaceutiques et Biologiques (ISBP)	Directrice : Mme Christine VINCIGUERRA

## **COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE**

Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur M. Marc BUFFAT
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : Gérard PIGNAULT
Institut de Science Financière et d'Assurances (ISFA)	Directeur : M. Nicolas LEBOISNE
Institut National du Professorat et de l'Education	Administrateur Provisoire : M. Pierre CHAREYRON
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL
Polytechnique Lyon	Directeur : Emmanuel PERRIN
UFR Biosciences	Administratrice provisoire : Mme Kathrin GIESELER
UFR des Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
UFR Faculté des Sciences	Directeur : M. Bruno ANDRIOLETTI

## Résumé

L'évolution d'une paire de chromosomes sexuels à longtermes été décrite par une trajectoire qui serait universelle à tous les systèmes, néanmoins, des études ont récemment nuancé ce «modèle» unique. D'après ce modèle, il y aurait dans un premier temps l'émergence d'une région non-recombinante (XY ou ZW), puis, une expansion de celle-ci. Simultanément, l'absence de recombinaison induit ce que l'on appelle la dégénérescence du chromosome Y (ou W). La dégénérescence est supposée augmenter et, après un certain temps évolutif, devrait conduire à un système dans lequel le chromosome Y (ou W) serait plus petit que le chromosome X (ou Z), voire disparaît. Cependant, seulement une trentaine de paires de chromosomes sexuels de plantes ont été étudiées empiriquement, parmi plus de 15 000 espèces dioïques (*i.e.* plantes à sexes séparés). Il en résulte que certaines étapes du systèmes sont moins supportées que d'autres. Plus précisément, la formation de la région non-recombinante a essentiellement été étudiée de manière théorique, tandis qu'une forte dégénérescence avec un chromosome Y (ou W) plus petit que le chromosome X (ou Z) n'a été décrite que chez les animaux.

Afin de mieux décrire la première étape du modèle, l'émergence de la région non recombinante, le premier axe de cette thèse représente une étude de *Silene acaulis* ssp *exscapa*, la seule sous-espèce dioïque du complexe *Silene acaulis*. En effet, ceci laisse supposer que ce système sexuel est un caractère dérivé, donc probablement récent. Le mécanisme du déterminisme du sexe n'étant pas connu, j'ai voulu savoir si une région non-recombinante typique d'une paire de chromosomes sexuels est présente chez cette sous-espèce. Pour cela, j'ai utilisé un outil récemment publié basé sur l'analyse de fréquences génotypiques et phénotypiques de mâles et de femelles au sein d'une population. Deux jeux de données RNA-seq provenant de deux populations ont permis d'identifier 27 gènes potentiellement XY, et suggèrent que la paire de chromosomes sexuels serait récente. Des analyses complémentaires restent nécessaires pour confirmer ces résultats.

Deuxièmement, afin de tester l'existence d'une paire ancienne de chromosomes sexuels avec une forte dégénérescence chez les plantes, le deuxième axe de cette thèse est une étude de deux espèces dioïques de la famille des Cannabaceae, *Cannabis sativa* et *Humulus lupulus*. En effet, l'ancêtre commun de ces deux espèces, qui divergent depuis plusieurs dizaines de millions d'années, était probablement dioïque. De plus, des analyses cytologiques ont identifié des paires de chromosomes sexuels qui pourraient être anciennes. Pour caractériser l'âge et le niveau de dégénérescence de ces paires de chromosomes sexuels, des données RNA-seq d'un croisement ont été générées pour chacune des deux espèces. Un outil probabiliste analysant les ségrégations alléliques au sein d'un croisement a permis d'identifier la première paire de chromosomes sexuels homologue entre deux genres chez les plantes. De plus, ces chromosomes sexuels sont parmi les plus vieux et les plus dégénérés actuellement décrits chez les plantes. Par ailleurs, la détection de séquences Y-spécifiques permettrait d'améliorer la culture de ces deux espèces. En effet, seules les femelles ont un intérêt agronomique et le faible dimorphisme sexuel limite une détection précoce du sexe. Au cours de cette thèse, j'ai développé des amorces PCR qui montrent des résultats prometteurs.

Plus généralement, ces résultats apportent de nouvelles informations concernant les étapes les moins bien décrites de l'évolution des chromosomes sexuels chez les plantes. Premièrement, nous montrons qu'une paire de chromosomes sexuels a probablement émergé récemment dans une espèce, et confirmons l'intérêt de continuer à l'étudier. Deuxièmement, nous confirmons que des chromosomes sexuels vieux et fortement dégénérés existent chez les plantes.

Mots-clés : Dioécie, Chromosomes Sexuels, Dégénérescence du Y, Arrêt de recombinaison, *Silene acaulis*, *Cannabis sativa*, *Humulus lupulus*

# Abstract

**Title : Analyses of early and late steps of sex chromosome evolution in plants with *Silene acaulis*, *Cannabis sativa* and *Humulus lupulus***

The evolution of a pair of sex chromosomes has long been described as universal to all systems, nonetheless, recent studies have moderated this "model". First, a non-recombining region (XY or ZW) should emerge. Then, this region is predicted to expand, and at the same time, the local loss of recombination between the sex chromosomes should induce a degeneration of the Y (or W) chromosome. This degeneration is then expected to progress, and after a certain amount of time, the Y (or W) chromosome should become smaller than the X (or Z) chromosome, or even disappear. A main issue of this model is that it is based on the genomic study of only about thirty plant sex chromosome systems, while there are more than 15,000 dioecious species (*i.e.*, plants with separate sexes). As a consequence, while some of these steps are well characterized, others are clearly lacking empirical support. In particular, the formation of a non-recombining region has only been studied theoretically and a strongly degenerated system with a Y (or W) chromosome smaller than the X (or Z) has only been genomically studied in animals.

In order to better understand the first step of this model, *i.e.* the emergence of a non-recombining region, the first part of this thesis focuses on *Silene acaulis* ssp *exscapa*, the only subspecies in the *Silene acaulis* complex that is dioecious. This pattern suggests that dioecy is a recently derived character in this subspecies. Since the sex determination mechanism was unknown in this subspecies, I sought to test whether a non-recombining region typical of a pair of sex chromosomes exists. In order to do so, RNA-seq data was generated from two different populations, to which I applied a new probabilistic tool based on the analysis of genotyping and phenotyping frequencies. This approach allowed me to identify 27 potential XY genes and to propose that they belong to a young sex chromosome system. However, complementary analyses will be required to validate this hypothesis.

In order to test whether plants can also present old and strongly degenerated non-recombining regions of sex chromosomes, the second part of this thesis focusses on two dioecious species from the Cannabaceae family, *Cannabis sativa* et *Humulus lupulus*. Indeed, these two species diverged 20-25 million years ago and their more recent common ancestor is supposed to be dioecious. Cytological analyses have furthermore shown that sex chromosomes in these species may be old. To estimate the level of degeneration and the age of this sex chromosome system, I analyzed RNA-seq data generated from one pedigree in each species with a probabilistic tool based on allele segregation patterns. I was able to identify the first sex chromosome system homologous between two genera in plants. This is among the oldest and most degenerated sex chromosome system described in plants so far. Otherwise, detecting Y-specific sequences in these species could improve their crops since only the females are useful and the sexual dimorphism is weak. I designed PCR-primers that showed promising results.

Overall, these results shed light on some of the less well-studied steps of the sex chromosome evolution in plants. Indeed, on one hand, we point to a potentially recently emerged system that deserves further attention; on the other hand, we confirm that old and strongly degenerated sex chromosomes also exist in plants.

Keywords : Dioecy, Sex chromosomes, Y degeneration, Arrest of recombination, *Silene acaulis*, *Cannabis sativa*, *Humulus lupulus*

# Remerciements

Au cours de ces trois dernières années, j'ai régulièrement constaté la chance que j'avais de réaliser une thèse de doctorat dans les conditions qui m'ont été offertes. La liste des personnes ayant participé à rendre cette expérience aussi agréable est sûrement trop longue pour être présentée exhaustivement dans ces remerciements, mais je vais tenter de n'oublier personne. Si jamais quelques noms venaient à manquer, je vous prie de ne pas trop m'en vouloir.

Ne perdons pas de temps, et commençons par l'essentiel. Jos et Gabriel, il n'est pas évident de trouver les mots justes pour vous remercier en quelques lignes seulement. Je vous suis très reconnaissant de l'encadrement de qualité dont j'ai bénéficié. Au-delà de vos connaissances et de votre expertise de la méthode scientifique qui, au passage, ont été une aubaine pour ma curiosité, vous m'avez accordé beaucoup de disponibilité et de confiance. Je souhaitais le souligner et vous en remercier sincèrement. La mise en évidence des difficultés est, à mon sens, essentielle dans le processus d'apprentissage, et vous avez su aborder mes lacunes avec une bienveillance telle que je ne pouvais que vouloir les surmonter. Je tiens aussi à vous remercier pour cela. Nos routes commençant à se séparer, il me faut maintenant trouver des personnes aussi riches scientifiquement et humainement que vous deux. J'espère que vous n'avez pas mis la barre trop haute!

L'environnement de travail dont j'ai bénéficié au LBBE a évidemment joué un rôle majeur dans le plaisir que j'ai pris à faire mon doctorat. Nombreuses sont les personnes à remercier au sein du laboratoire.

Commençons par les amis et amies eux aussi en doctorat, dont je ne pourrai citer tous les noms au risque de finir avec des remerciements plus longs que mon introduction... Les nombreuses discussions scientifiques que nous avons eu, dans des cadres plus ou moins informels (durant les pauses, le séminaire non-permanent, ou parfois autour d'un verre) ont été essentielles pour alimenter ma passion pour la science. Il me faut aussi vous remercier pour le plaisir que ce fut d'allier professionnalisme et camaraderie (j'entends par là les heures passées au Condorcet ou autres lieux de production scientifique sans précédent). Vous m'avez beaucoup appris, merci!!

Je tiens aussi à remercier les différents pôles du laboratoire. Mon expérience au LBBE est relativement récente, mais il me semble tout de même avoir compris que ce laboratoire ne pourrait pas fonctionner sans vous. Merci au pôle administratif pour tout le travail que vous faites. Mes travaux m'ont principalement amené à travailler avec le pôle informatique. Merci particulièrement à Stéphane et à Bruno (mais aussi Adil, Simon et Philippe), cette thèse n'aurait pas pu être réalisée sans vous. Je tiens aussi à vous remercier pour la sympathie dont vous faites toujours preuve (malgré les innombrables fois où je vous ai sollicités...). Enfin, Hélène, ce fut un plaisir de travailler à tes côtés, merci pour ta collaboration (on va la monter notre start-up cannabis!).

Merci aussi aux permanents du laboratoire. Les Journal Club, séminaires internes et autres moments de discussions scientifiques ont été très importants pour moi. C'est une vraie chance d'avoir pu échanger aussi régulièrement avec des personnes expertes dans leur domaine, surtout lorsque ceci est fait avec la simplicité dont vous faites preuve. Merci en particulier à tous les membres de l'équipe sexe et évolution.

Je voudrais aussi remercier les membres du Laboratoire d'Écologie, Évolution et Paléontologie qui m'ont accueilli pendant une partie de ma troisième année de thèse. Ce fut un plaisir de découvrir votre laboratoire. Merci notamment à Pascal.

La science étant bien plus efficace, et agréable, lorsqu'elle est collaborative, il est important de noter que les travaux que j'ai réalisés pendant cette thèse n'auraient pas été possibles sans les nombreuses personnes avec qui j'ai eu la chance de travailler. Merci à tous mes collaborateurs et collaboratrices.

Au-delà des commentaires de fond, mes «qualités» littéraires ont nécessité plusieurs relectures de ce manuscrit, merci aux nombreuses personnes ayant accepté de se faire mal aux yeux.

Aline, Charlie, Jacquie, Pascal, et Tristan, je tiens à vous remercier pour les discussions et le recul que vous nous avez apporté lors des différents comités de pilotage. Je tiens aussi à remercier les personnes qui ont accepté d'évaluer mon travail. Christelle, Christophe, Frédérique et Tatiana merci beaucoup de prendre le temps de lire et de corriger ce manuscrit.

Il n'aurait probablement pas été possible de faire cette thèse sans l'environnement social et familial dans lequel j'évolue. Merci à vous qui participez, ou avez participé, quotidiennement, ou ponctuellement, à mon bien-être. La grande majorité d'entre vous n'étant pas scientifique (de profession), vous m'apportez beaucoup en me faisant découvrir vos passions (que je partage pour certaines) et en vous ouvrant aux miennes. Une pensée singulière s'adresse à mes parents.

Pour finir, cette chance d'avoir pu travailler dans un domaine qui me passionne je la dois aussi à toutes les personnes qui ont permis à notre société d'avoir financé mes études, dont ce doctorat.

# Table des matières

<b>1 Introduction</b>	<b>11</b>
1.1 Systèmes sexuels chez les plantes . . . . .	12
1.2 Évolution de la dioécie . . . . .	13
1.3 Déterminisme . . . . .	17
1.4 Évolution des chromosomes sexuels . . . . .	17
1.5 Les chromosomes sexuels actuellement décrits chez les plantes .	24
1.6 Méthodes disponibles pour étudier les chromosomes sexuels . .	25
1.7 Objectifs de la thèse . . . . .	29
<b>2 Chapter 2 : Are sex chromosomes emerging in <i>Silene acaulis</i> ssp <i>exscapa</i> ?</b>	<b>49</b>
2.1 Abstract . . . . .	51
2.2 Introduction . . . . .	51
2.3 Materials and methods . . . . .	53
2.4 Results . . . . .	54
2.5 Discussion . . . . .	58
2.6 References . . . . .	60
2.7 Supplementary Material . . . . .	64
<b>3 Chapter 3 : An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of <i>Cannabis sativa</i></b>	<b>67</b>
3.1 Abstract . . . . .	69
3.2 Introduction . . . . .	69
3.4 Results . . . . .	70
3.5 Discussion . . . . .	72



3.3 Materials and methods . . . . .	74
3.6 References . . . . .	75
3.7 Supplementary Material . . . . .	78
<b>4 Chapter 4 : Development of genetic markers for sexing <i>Cannabis sativa</i> seedlings</b>	<b>89</b>
4.1 Abstract . . . . .	91
4.2 Introduction . . . . .	91
4.3 Materials and methods . . . . .	92
4.4 Results . . . . .	92
4.5 Discussion . . . . .	93
4.6 References . . . . .	93
<b>5 Chapter 5 : Plant genera <i>Cannabis</i> and <i>Humulus</i> share the same pair of well-differentiated sex chromosomes</b>	<b>95</b>
5.1 Abstract . . . . .	97
5.2 Introduction . . . . .	97
5.3 Materials and methods . . . . .	98
5.4 Results . . . . .	101
5.5 Discussion . . . . .	104
5.6 References . . . . .	107
5.7 Supplementary Material . . . . .	110
<b>6 Discussion</b>	<b>123</b>
6.1 Résumé des principaux résultats de thèse . . . . .	124
6.2 Perspectives . . . . .	124
6.3 Apports de mes travaux pour la compréhension de l'évolution des chromosomes sexuels . . . . .	128
6.4 Identifier des chromosomes sexuels chez des espèces non-modèles	129
6.5 Conclusion . . . . .	131

6.6 Références . . . . .	132
--------------------------	-----

Cette thèse est une thèse sur articles. L'introduction et la discussion ont été rédigées en français et constituent les chapitres 1 et 6 de cette thèse. Les chapitres 2, 3, 4 et 5 sont des articles scientifiques, parfois publiés, et sont donc rédigés en anglais.



# 1

## Introduction

## Contexte

Cette thèse a pour objectif d'apporter des éléments de compréhension sur l'évolution des chromosomes sexuels chez les plantes. En effet, bien que l'idée que toutes les paires de chromosomes sexuels suivraient une trajectoire évolutive unique soit de moins en moins admise, les modèles expliquant l'évolution des chromosomes sexuels chez les plantes sont basés sur un nombre limité d'études empiriques. Par ailleurs, certaines étapes de l'évolution d'une paire de chromosomes sexuels sont particulièrement moins bien décrites. C'est notamment le cas de la formation et de l'évolution très tardive d'une paire de chromosomes sexuels chez les plantes. Avant de présenter le projet de thèse, je vais commencer par introduire ce qu'est la dioécie (système sexuel avec des individus mâles et des individus femelles), puis faire un état de la connaissance actuelle de l'évolution des chromosomes sexuels chez les plantes.

# 1 Systèmes sexuels chez les plantes

Bien que les raisons expliquant le succès d'une reproduction sexuée ne soient toujours pas bien comprises, plus de 99.9% des espèces eucaryotes y ont recours (voir Barton et Charlesworth, 1998; Otto et Lenormand, 2002; Otto, 2009). Chez les plantes, en particulier, les systèmes sexuels sont très divers, bien plus que chez les animaux (Barrett, 2002). Ces différents systèmes sexuels reposent sur le polymorphisme des fleurs existant dans l'espèce ainsi que sur le polymorphisme floral entre individus de l'espèce. On peut identifier trois types de fleurs : les fleurs mâles, les fleurs femelles, et les fleurs bisexuelles (aussi appelées fleurs parfaites). Le polymorphisme floral entre individus est un peu plus complexe, avec des individus pouvant porter un ou plusieurs types de fleurs. Dans la Table 1, je liste 8 de ces systèmes sexuels, dont les trois principaux : (1) L'hermaphrodisme (monoclinie et distilie), qui est le système majoritaire, dans lequel les individus ont des fleurs bisexuelles, (2) La monoécie, lorsque les individus ont des fleurs mâles et des fleurs femelles, (3) Et la dioécie, lorsque certains individus portent des fleurs mâles et certains individus des fleurs femelles (Bawa et Beach, 1981; Barrett, 2002). D'autres systèmes sexuels incluant 3 morphes individuels existent, comme la trioécie (individus mâles, femelles, et hermaphrodites), mais ne sont pas présentés dans ce tableau.

TABLE 1 – Tableau récapitulatif des différents systèmes sexuels chez les plantes. Ce tableau renseigne le polymorphisme entre individus et entre fleurs au sein d’un même système sexuel, ainsi que la fréquence de chaque système sexuel parmi les angiospermes (ce tableau est obtenu d’après Käfer, Marais et Pannell, 2017).

Système sexuel	Polymorphisme inter-individuel du système sexuel	Polymorphisme floral du système sexuel	Fréquence
Monoclinie	Aucun	Aucun	~85%
Distilie	Aucun	- Longues étamines et courts styles - Longs styles et courtes étamines	<1%
Gynomonoécie	Aucun	- Fleurs parfaites - Fleurs femelles	Na
Gynodioécie	- Individus hermaphrodites - Individus femelles	- Fleurs parfaites - Fleurs femelles	<1%
Monoécie	Aucun	- Fleurs mâles - Fleurs femelles	6-7%
Dioécie	- Individus mâles - Individus femelles	- Fleurs mâles - Fleurs femelles	5-6%
Andromonoécie	Aucun	- Fleurs parfaites - Fleurs mâles	Na Na
Androdioécie	- Individus hermaphrodites - Individus mâles	- Fleurs parfaites - Fleurs mâles	«1%

Comme indiqué dans la Table 1, les fréquences des différents systèmes sexuels sont fortement variables chez les plantes. Pour les 3 principaux systèmes sexuels chez les angiospermes, nous comptons 77-85% d’hermaphrodites, 5-6% de dioïques et légèrement plus de monoïques (Renner, 2014). Les 15 600 espèces d’angiospermes dioïques sont distribuées dans 38% des familles et 7% des genres des plantes à fleurs (Renner, 2014). D’après la distribution des espèces dioïques chez les angiospermes, il a été estimé qu’entre 871 et 5000 transitions vers la dioécie auraient eu lieu dans ce clade (Renner, 2014).

## 2 Évolution de la dioécie

### 2.1 Allo-fécondation et dioécie

Les individus hermaphrodites ou monoïques développent à la fois des appareils reproducteurs mâles et des appareils reproducteurs femelles. Ces individus ont donc tout le matériel biologique nécessaire à la reproduction avec soi-même, c’est-à-dire, l’autofécondation. Bien que l’autofécondation ait pour avantage de garantir la reproduction, elle peut induire de la dépression de consanguinité qui peut être fortement délétère (Charlesworth et Charlesworth, 1987). L’allo-fécondation, le fait de se reproduire avec un autre individu, permet de réduire

cette dépression de consanguinité. Ainsi, plusieurs mécanismes obligeant l'allo-fécondation ont été sélectionnés chez les plantes (résumé dans Carey, Yu et Harkess, 2021). Par exemple, nous pouvons noter l'auto-incompatibilité, la séparation temporelle du développement des appareils reproducteurs mâles et femelles ou encore la distanciation spatiale de ces appareils sur un même individu (résumé dans Carey, Yu et Harkess, 2021). Une autre stratégie est de séparer les individus avec des fleurs mâles et les individus avec des fleurs femelles, c'est à dire, la dioécie.

La dioécie a longtemps été décrite comme un cul-de-sac évolutif (Heilbuth, 2000; Vamosi et Vamosi, 2004), cependant, depuis quelques années des travaux empiriques contrastent cette hypothèse (Käfer et al., 2014; Sabath et al., 2016). Käfer et al. (2014) ont montré que la dioécie augmente la diversification alors que Sabath et al. (2016) ont montré que la dioécie augmente la diversification dans certains clades mais la réduit dans d'autres. Outre les conséquences de la dioécie sur la diversification, les mécanismes conduisant à son évolution ne sont toujours pas bien compris non plus. Il a été proposé que la dioécie peut évoluer chez une espèce initialement hermaphrodite ou monoïque (Barrett, 2002).

## 2.2 De l'hermaphrodisme à la dioécie

Charlesworth et Charlesworth (1978) ont développé un modèle théorique pour expliquer la transition vers la dioécie depuis un état hermaphrodite. Cette transition se ferait par l'accumulation de deux mutations nucléaires. La première mutation inhibe le développement de l'appareil reproducteur mâle chez les hermaphrodites, ce qui conduit au développement de femelles (*i.e.* mutation de stérilité mâle, Figure 1 **A**). La population est, à ce stade, composée de femelles et d'hermaphrodites, donc est gynodioïque. La seconde mutation inhibe le développement de l'appareil reproducteur femelle et conduit donc au développement de mâles (*i.e.* mutation de stérilité femelle, Figure 1 **A**). Si les deux mutations apparaissent sur des chromosomes homologues, un arrêt de recombinaison de la région incluant ces deux mutations est sélectionné car il empêche un haplotype de porter les deux mutations de stérilité. Un chromosome portera la mutation stérilisant les mâles alors que le second portera celle stérilisant les femelles. Si la mutation de stérilité femelle est dominante, alors nous parlons de systèmes XY, les mâles sont hétérogamétiques (XY) et les femelles homogamétiques (XX). Si la mutation de stérilité mâle est dominante, ce sont cette fois-ci les femelles qui sont hétérogamétiques (ZW) et les mâles homogamétiques (ZZ), nous parlons de système ZW. Suite à cette seconde mutation puis à l'arrêt de recombinaison, la population est composée de femelles, d'hermaphrodites et de mâles. On parle de population trioïque.

Un autre modèle décrivant la transition entre l'hermaphrodisme et la dioécie est celui de la stérilité cytoplasmique chez les mâles (Schultz, 1994; Maurice et al., 1994; résumé dans Fruhard et Marais, 2021). Une mutation inhibant le développement des fonctions mâles a lieu dans l'ADN cytoplasmique et conduit au développement d'individus femelles (*i.e.* CMS pour "Cytoplasmic Male Sterility" Figure 1 **B**). La transmission du cytoplasme étant maternelle,

la proportion de femelle augmentera rapidement dans la population après l'apparition de cette mutation. Une mutation permettant la restauration de la fertilité mâle sera fortement sélectionnée si elle apparaît dans la population. En effet, la proportion de femelle peut devenir très importante dans la population, ce qui augmente fortement les pressions de sélection pour restaurer la production de gamètes mâles (Schultz, 1994; Maurice et al., 1994; résumé dans Fruchard et Marais, 2021). Les individus portant le facteur de restauration de fertilité mâle seront hermaphrodites, les autres seront des femelles. Tout comme pour le modèle de Charlesworth et Charlesworth (1978), la population est à ce stade gynodioïque. Nous pouvons noter ici qu'un déterminisme nucléo-cytoplasmique du sexe peut être un mécanisme permettant le maintien de la gynodioécie (voir Maurice et al., 1994). Ensuite, une mutation de stérilité femelle peut se produire chez un individu portant le gène restaurateur de la fertilité mâle, développant ainsi un individu mâle (*i.e.* mutation de stérilité femelle, Figure 1 **B**). À nouveau, un arrêt de recombinaison va être sélectionné afin de lier le gène de restauration de la fertilité mâle avec celui de la stérilité femelle. Comme pour le modèle précédent, la population est trioïque.

Il y a deux avantages principaux pour les individus uni-sexués dans une population trioïque. Premièrement, la production de gamètes mâles d'un individu mâle est supposée être plus importante et de meilleure qualité que celle d'un individu hermaphrodite car l'ensemble des ressources dédiées à la reproduction sont allouées à la production de gamètes mâles uniquement. Il en est de même pour les gamètes femelles chez des individus femelles. Les gamètes des individus uni-sexués sont donc plus compétitives que celles d'individus hermaphrodites. Deuxièmement, l'allo-fécondation obligatoire des individus uni-sexués réduit la dépression de consanguinité dans leurs descendance (voir sous-section précédente). Ainsi, plus le taux d'auto-fécondation est important dans la population, plus les descendants d'individus uni-sexués sont avantagés par le brassage génétique de l'allo-fécondation. En fonction de l'intensité de ces deux avantages, les mâles et les femelles peuvent être sélectionnés. Les individus hermaphrodites sont donc éliminés et la population deviendra dioïque.

Les deux modèles présentés précédemment impliquent une transition par la gynodioécie car la mutation de stérilité mâle apparaît avant la mutation de stérilité femelle. Si la mutation de stérilité femelle apparaît en premier, la transition vers la dioécie se fera via l'androdioécie. Cependant, une mutation de stérilité femelle ne pourra pas être transmise par le cytoplasme puisque ce dernier est transmis par voie maternelle. Par ailleurs, une transition par l'androdioécie est considérée comme peu probable car l'avantage des mâles sur les hermaphrodites dans une population sans femelles implique une augmentation très importante de la production de pollen (voir Charlesworth et Charlesworth, 1978). Il est donc attendu que les mâles soient sélectionnés plutôt après l'apparition de femelles dans la population (voir Charlesworth et Charlesworth, 1978; Barrett, 2002; Fruchard et Marais, 2021).

Par ailleurs, Charlesworth et Charlesworth (1978) montrent qu'il est plus probable que la mutation de stérilité mâle soit récessive, et donc la mutation de stérilité femelle dominante, ce qui conduit à un système XY. Dans le modèle impliquant une stérilité mâle cytoplasmique, la



mutation de restauration de fertilité mâle et de celle de stérilité femelle sont aussi attendues pour être dominante (Maurice et al., 1994). Ceci explique probablement la sur-représentation des systèmes XY chez les plantes (voir Table 2 dans la section "Les chromosomes sexuels actuellement décrits chez les plantes").

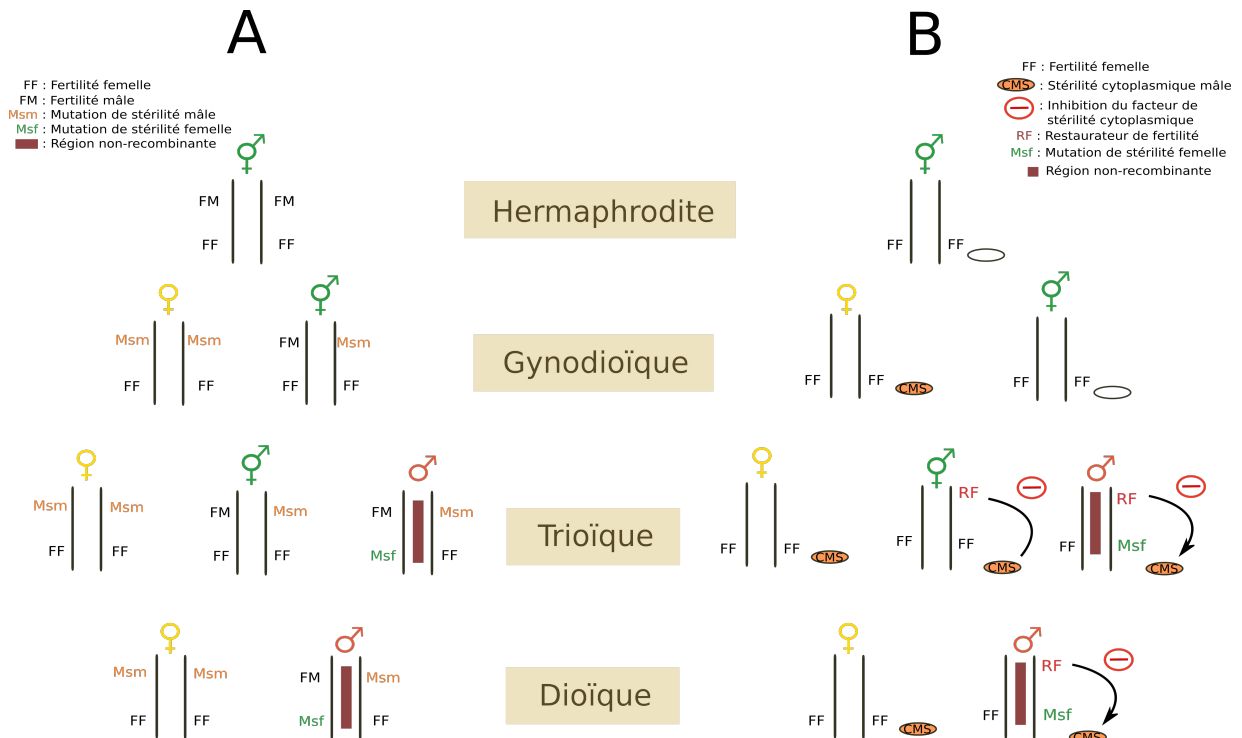


FIGURE 1 – Transition de l’hermaphroditisme vers la dioécie par la voie gynodioïque (d’après Fruchard et Marais, 2021). (A) modèle de Charlesworth et Charlesworth (1978) et (B) modèle de la stérilité cytoplasmique mâle (Schultz, 1994; Maurice et al., 1994). Les exemples de cette figure conduisent à l’émergence d’une paire de chromosomes sexuels XY.

## 2.3 De la monoécie à la dioécie

Renner et Ricklefs (1995) et Renner (2014) ont montré une association phylogénétique entre la monoécie et la dioécie plus importante qu’entre la dioécie et la gynodioécie. Cette association laisse supposer que la transition vers la dioécie se produit majoritairement depuis la monoécie, mais cette question reste ouverte (Renner et Ricklefs, 1995; Renner, 2014; Dufay et al., 2014; Renner et Müller, 2021). Les deux modèles génétiques présentés précédemment n’incluent pas de transition par la monoécie. Un individu monoïque produit à la fois des fleurs mâles et des fleurs femelles, le ratio entre les deux pouvant varier. La transition vers la dioécie pourrait donc résulter de ratios extrêmes conduisant à la production d’un seul type de fleur par individu (Golenberg et West, 2013). Aujourd’hui, aucun modèle mathématiquement formalisé ne permet d’expliquer génétiquement cette transition et les analyses empiriques sont encore trop peu nombreuses pour formuler des hypothèses (résumé dans Fruchard et Marais, 2021).

## 3 Déterminisme du sexe

Le modèle de Charlesworth et Charlesworth (1978) et celui incluant une mutation cytoplasmique impliquent deux gènes. Ils sont donc appelés modèles à deux facteurs. Des paires de chromosomes sexuels avec deux gènes du déterminisme ont été identifiées chez plusieurs espèces d'angiospermes (*e.g. Asparagus officinalis*, *Actinidia chinensis* var *chinensis*) (Akagi et al., 2019; Harkess et al., 2020). Cependant, des données empiriques ont montré que le sexe chez certaines plantes était déterminé par un seul gène (résumé dans Ponnikas et al., 2018; Carey, Yu et Harkess, 2021; Renner et Müller, 2021). Ce modèle, appelé le modèle à un facteur, n'est donc pas concordant avec les modèles expliqués précédemment. Plusieurs paires de chromosomes sexuels soutenant ce modèle ont été identifiées chez les plantes (*e.g.* chez *Populus* et *Salix*, *Diospyros lotus*, suggéré chez *Spinacia oleracea*) (Akagi et al., 2014; West et Golenberg, 2018; Müller et al., 2020)). Ce qui est intéressant, c'est que ces deux types de déterminisme du sexe peuvent conduire à l'émergence de chromosomes sexuels.

Il est important de noter que le déterminisme peut être environnemental, voire être un mécanisme multifactoriel avec une interaction de l'environnement et de la génétique (voir Golenberg et West, 2013; Pannell, 2017). Cependant, aucun mécanisme environnemental du déterminisme du sexe n'a été mis en évidence chez les plantes jusqu'à présent (Pannell, 2017). Il est donc supposé que la plupart des espèces dioïques ont un déterminisme génétique, et peut-être des chromosomes sexuels.

## 4 Évolution des chromosomes sexuels

### 4.1 Formation des chromosomes sexuels

Comme expliqué précédemment, il existe deux types de chromosomes sexuels. Les chromosomes sexuels XY (hétérogamétie chez les mâles) ou ZW (hétérogamétie chez les femelles). Dans le reste de cette introduction, je parlerais de X et de Y bien que ce qui est dit pour le Y soit transposable au W et ce qui est dit pour le X le soit pour le Z.

Il n'est pas évident de savoir à partir de quel moment nous pouvons parler de chromosomes sexuels (Charlesworth, 2021). Nous avons vu avec les modèles présentés précédemment qu'un arrêt de recombinaison qui empêche la liaison de la mutation stérilité mâle et de la mutation de stérilité femelle sur le même chromosome serait sélectionné. Dans la plupart des systèmes étudiés, il a été observé que cette région non-recombinante s'agrandit (détaillé plus tard). Généralement, lorsque seuls les gènes responsables du déterminisme du sexe ont arrêté de recombiner, on parle de proto-X et proto-Y (Ming, Bendahmane et Renner, 2011). La notion de chromosomes sexuels ne sera utilisée que pour décrire des chromosomes sur lesquels la région non-recombinante s'est déjà étendue au-delà des gènes du déterminisme du sexe.

Ainsi, ce que j'appelle la formation d'une paire de chromosomes sexuels correspond en réalité

à la formation du proto-X et du proto-Y suite au premier arrêt de recombinaison. Plusieurs formations indépendantes de chromosomes sexuels ont été rapportées chez les eucaryotes (Bachtrog et al., 2011, 2014; Wright et al., 2016). En revanche, les mécanismes conduisant à un arrêt de recombinaison restent peu compris aujourd'hui (Wright et al., 2016). Les régions non-recombinantes bien caractérisées existent, généralement, depuis plusieurs millions d'années et ne permettent plus d'isoler les mutations ayant conduit à leur formation (voir la section "Les chromosomes sexuels actuellement décrits chez les plantes", et Charlesworth (2019)).

Les inversions génomiques ont longtemps été proposées comme le mécanisme sélectionné qui permet un arrêt de recombinaison rapide sur une large région du chromosome (Charlesworth, Charlesworth et Marais, 2005). Bien que certaines données soutiennent un arrêt de recombinaison dû à une inversion chromosomique, ce n'est pas le cas pour tous les systèmes (Branco et al., 2017; résumé dans Wright et al., 2016). Un autre mécanisme a été proposé suite à des observations faites chez *Salix viminalis* (Almeida et al., 2020). Il s'agirait d'une accumulation d'éléments transposables qui permettrait un arrêt de recombinaison du fait d'une augmentation de la divergence entre les deux néo-chromosomes sexuels (*i.e.* Chalopin et al., 2015; Almeida et al., 2020. Ici, nous pouvons noter qu'il n'est pas évident de savoir si l'arrêt de recombinaison précède ou succède l'insertion d'éléments transposables dans les données empiriques.

Par ailleurs, il est assez bien décrit que le taux de recombinaison n'est pas homogène sur l'ensemble du génome (Gaut et al., 2007; Stapley et al., 2017; Zelkowski et al., 2019), et qu'il peut aussi varier en fonction du sexe (Lenormand et Dutheil, 2005). L'hétérochiasmie (taux de recombinaison différents entre mâles et femelles) et l'achiasmie (absence de recombinaison dans un sexe) ont été observées chez plusieurs espèces (Lenormand et Dutheil, 2005). Les régions faiblement recombinantes dans un sexe, voire dans les deux, pourraient donc être des localisations génomiques favorables à la formation d'une région non-recombinante de chromosomes sexuels (*e.g.* chez les grenouilles du genre *Hyla* et chez les plantes du genre *Rumex*) (Dufresnes et al., 2021; Rifkin et al., 2021).

Actuellement, les théories sur la formation des chromosomes sexuels sont principalement sélectives (résumé dans Jeffries et al., 2021). Récemment, un modèle montrant une évolution neutre d'une région non-recombinante de chromosomes sexuels a été publié (Jeffries et al., 2021). Dans ce modèle, une augmentation de divergence entre deux régions homologues réduit le taux de recombinaison entre ces deux régions. Donc, si le taux de mutation génère de la divergence qui n'a pas le temps d'être éliminée par la recombinaison, la divergence entre deux régions homologues peut devenir suffisamment importante pour supprimer complètement la recombinaison. Cependant, ce modèle n'a pas encore été testé avec des données empiriques. Il est aussi possible que la région du déterminisme du sexe change de chromosome, on parle alors de «turnover» de chromosomes sexuels (résumé dans Vicoso, 2019). Ce turnover peut se produire via une translocation du locus du déterminisme du sexe, ou par l'émergence d'un nouveau locus de déterminisme du sexe sur un autre chromosome. Suite à ce turnover, l'an-

cienne paire de chromosomes sexuels redevient autosomale (voir Renner et Müller, 2021). Ces mécanismes peuvent être fréquents chez les animaux (*e.g.* chez des poissons, amphibiens, reptiles), et ont été observés chez plusieurs plantes (Vicoso, 2019; Renner et Müller, 2021). Ceci pourrait expliquer, en partie, que de larges régions non-recombinantes ont rarement été observées chez les plantes (Charlesworth, 2019))(voir section "Les chromosomes sexuels actuellement décrits chez les plantes").

## 4.2 Évolution de la région non-recombinante

Une fois la région non-recombinante formée, elle peut s'agrandir avec le temps (Charlesworth et Charlesworth, 1978; Charlesworth, Charlesworth et Marais, 2005; Ming, Bendahmane et Renner, 2011). Il a longtemps été proposé que des gènes à antagonisme sexuel (avantageux pour un sexe mais délétères pour l'autre) soient sélectionnés aux abords des gènes du déterminisme du sexe (résumé dans Charlesworth, Charlesworth et Marais, 2005; Wright et al., 2017). Une extension de la région non-recombinante serait sélectionnée afin de conserver les gènes avantageux pour les mâles, mais délétères pour les femelles, dans la région Y-spécifique (dans le cas d'un système XY). Cette hypothèse a été validée chez un nombre limité d'espèces. Parmi celles-ci, on trouve l'exemple du chromosome Y des guppys, qui code la couleur chez les mâles (Wright et al., 2017, 2019). Récemment, Shearn et al. (2020) ont montré que l'hypothèse de l'expansion de la région non-recombinante par l'accumulation de mutations à antagonisme sexuel était validée chez des primates (les strepsirrhiniens). En revanche, cette hypothèse reste non validée pour plusieurs systèmes de chromosomes sexuels (Ironsides, 2010; Cavoto et al., 2018; Perrin, 2021). Perrin (2021) n'a pas mis en évidence de lien entre sélection antagoniste et évolution de la région non-recombinante chez deux familles d'amphibiens. Des travaux théoriques supportent l'hypothèse que la recombinaison serait toujours bénéfique pour les mâles même avec une forte sélection sexuelle antagoniste (Cavoto et al., 2018). L'intensité de la sélection sexuelle antagoniste modifie le taux de recombinaison optimal pour les mâles et peut donc le réduire fortement. En revanche, d'après leur modèle, cette sélection antagoniste ne serait pas assez intense pour sélectionner un arrêt total de recombinaison (Cavoto et al., 2018).

Une autre hypothèse propose que l'arrêt de recombinaison soit sélectionné afin d'empêcher l'homozygotie de mutations délétères récessives (voir Charlesworth et Wall, 1999). Récemment, des travaux théoriques ont montré qu'un arrêt de recombinaison dans une région juxtaposant le locus du sexe peut être sélectionné si la région possède moins de mutations délétères récessives que le reste de la population (Jay, Tezenas et Giraud, 2021; Lenormand et Roze, 2021). Plus précisément, cet arrêt de recombinaison doit avoir lieu chez un individu qui possède moins de mutations récessives dans la région inversée que la moyenne dans la population (voir Jay, Tezenas et Giraud, 2021). Lenormand et Roze (2021) suggèrent que la fixation de l'inversion dans la population n'est possible que si l'expression des gènes présents dans cette inversion est modifiée par des éléments régulateurs sur des autosomes. D'après ce modèle, la

modification de l'expression permettrait de compenser l'effet de l'accumulation de mutations délétères (Lenormand et Roze, 2021).

L'absence de recombinaison entre la copie X et Y ne sera pas sans conséquence, notamment pour le chromosome Y. Ce dernier sera isolé chez les mâles alors que les chromosomes X recombineront chez les femelles (Charlesworth et Charlesworth, 2000; Charlesworth, Charlesworth et Marais, 2005; Bachtrog, 2006). Sans recombinaison, les différents gènes du chromosome Y seront liés. Cette liaison entre gènes entraîne des interférences dans les pressions de sélection auxquelles ils sont soumis (Hill et Robertson, 1966). Globalement, ces interférences réduisent les pressions de sélection au sein de la région non-recombinante, ce qui favorise la fixation de mutations délétères et l'accumulation d'éléments transposables (Orr et Kim, 1998; Charlesworth, Langley et Sniegowski, 1997; Charlesworth et Charlesworth, 2000). Différents processus expliquent cette réduction de l'efficacité de sélection (Charlesworth et Charlesworth, 2000; résumé dans Bachtrog, 2006) :

(1) la sélection d'arrière-plan : pour qu'une mutation avantageuse soit sélectionnée, il faut donc que son avantage évolutif soit supérieur au désavantage évolutif des allèles délétères qui lui sont liés. Des mutations avantageuses seront donc contre sélectionnées.

(2) l'effet auto-stop : lorsque les loci sont liés, la fixation d'un allèle entraîne la fixation des allèles qui lui sont liés. Un allèle fortement avantageux peut donc conduire à la fixation d'allèles faiblement délétères qui lui sont liés.

(3) le cliquet de Muller : considérons plusieurs allèles avec différents niveaux de dégénérescence au sein d'une population. En l'absence de recombinaison, la perte de l'allèle le moins dégénéré (par dérive génétique par exemple) augmentera le niveau de dégénérescence minimal de la population. Avec le temps, des copies de plus en plus dégénérées ségrégeront dans la population, sans qu'il soit possible de revenir en arrière (d'où le nom de cliquet).

(4) réduction de la taille efficace : dans la région non-recombinante, les tailles efficaces des chromosomes sexuels sont réduites à  $\frac{1}{4}$  de celles des autosomes pour le Y, et  $\frac{3}{4}$  de celles des autosomes pour le X. Cette réduction de la taille efficace augmente la dérive génétique, qui à son tour augmente la probabilité de fixation de mutations faiblement délétères par hasard. L'effet du cliquet de Muller est donc amplifié (par dérive génétique). Par ailleurs, cette réduction de la taille efficace réduit la probabilité de fixation d'une mutation avantageuse.

L'accumulation d'éléments transposables et de mutations délétères a des conséquences négatives sur les séquences protéiques et les niveaux d'expression des gènes. Cette détérioration de la qualité des gènes du chromosome Y est appelée la dégénérescence du chromosome Y. Ce phénomène a été observé chez un nombre important d'espèces (Bachtrog, 2013). En revanche, nous ne savons toujours pas quelle est la part de chacun des processus détaillés ci-dessus dans la dégénérescence du chromosome Y (Charlesworth et Charlesworth, 2000; Bachtrog, 2006).

### 4.3 Hétéromorphie chromosomique

La divergence entre les chromosomes X et Y va donc augmenter avec l'âge du système, tout comme la dégénérescence du chromosome Y (Charlesworth, Charlesworth et Marais, 2005). Après plusieurs millions de générations, le contenu des chromosomes sexuels peut devenir divergent au point d'observer des chromosomes de tailles différentes. On parle d'hétéromorphie chromosomique. Le chromosome Y peut devenir plus grand que le chromosome X, ce qui est observé chez quelques plantes (*e.g. Silene latifolia, Coccinia grandis*) (résumé dans Ming, Bendahmane et Renner, 2011). Cet agrandissement du chromosome Y est dû à une accumulation d'éléments transposables dans la région non-recombinante (Ming, Bendahmane et Renner, 2011). Le Y peut aussi devenir plus petit que le X à cause de la perte de matériel génétique (Charlesworth, Charlesworth et Marais, 2005; Bachtrog, 2006; Ming, Bendahmane et Renner, 2011). Comme on peut le voir dans la Figure 2, le modèle actuel suppose une hétéromorphie chromosomique pour les systèmes qui divergent depuis longtemps (Charlesworth, Charlesworth et Marais, 2005; Bachtrog, 2006; Ming, Bendahmane et Renner, 2011). Une homomorphie chromosomique est donc attendue pour les systèmes les plus jeunes.

Il est important de noter que la définition d'hétéromorphie chromosomique ne fait pas consensus dans la communauté (voir Wright et al., 2016). Dans certains cas, l'hétéromorphie concerne uniquement la taille des chromosomes et la notion de taille est elle-même ambiguë puisqu'en fonction de l'état de la chromatine deux chromosomes avec le même nombre de bases peuvent montrer des tailles différentes en analyse cytologique (voir Charlesworth, 2016). Dans d'autres cas, c'est le contenu des chromosomes sexuels qui peut définir leur homomorphie (discuté dans Ming, Bendahmane et Renner, 2011). Ainsi, des chromosomes de tailles similaires mais avec une grande région non-recombinante dont les contenus diffèrent entre X et Y pourraient être appelés hétéromorphes. Si l'on considère uniquement la taille, le seuil permettant de définir une paire de chromosomes sexuels comme hétéromorphique est assez arbitraire. Il n'existe pas de convention sur ce seuil pour le moment. Par exemple, chez le *Cannabis sativa*, des auteurs considèrent la paire de chromosomes sexuels comme hétéromorphique (Razumova et al., 2016) alors que d'autres non (Ming, Bendahmane et Renner, 2011). Cette classification peut être plus continue, et des chromosomes dont les tailles divergent de moins de 10%, par exemple, peuvent être considérés comme «quasi homomorphiques» (terme employé, par exemple, dans Odierna et al., 1993; Priore et Pigozzi, 2017).

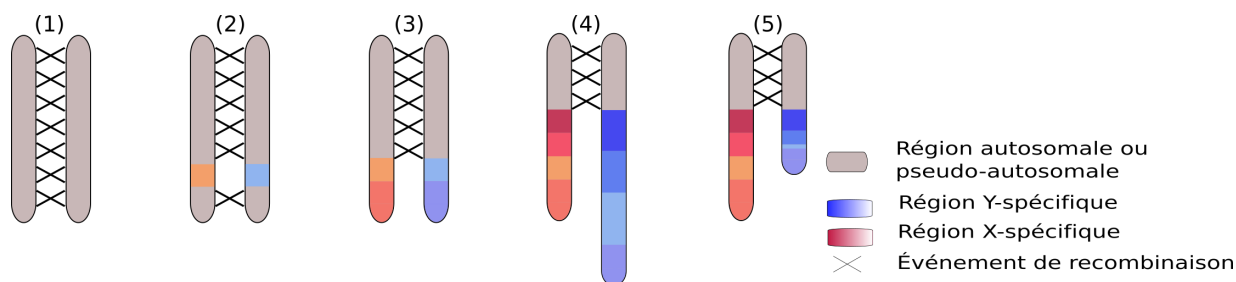


FIGURE 2 – Schéma du modèle d'évolution des chromosomes sexuels chez les plantes. (1) Paire d'autosomes avec plusieurs événements de recombinaison le long de la paire ; (2) Formation de la région non-recombinante ; (3) Expansion de la région non-recombinante avec la formation de strates évolutives ; (4) L'accumulation d'éléments transposables rend le chromosome Y plus grand que le X ; (5) Réduction du chromosome Y à cause de la perte de matériel génétique. Les différentes intensités de bleu et de rouge représentent les strates évolutives (voir la sous-section «strates évolutives» pour plus de détails).

## 4.4 Réduction de l'expression et perte de gènes

L'accumulation de mutations délétères et l'insertion d'éléments transposables ont des conséquences sur le niveau d'expression des gènes ainsi que leurs séquences codantes (résumé dans Bachtrog, 2013; Muyle, Shearn et Marais, 2017). Si ces événements se produisent dans les régions régulatrices des gènes, ils vont probablement réduire voire inhiber leur transcription. Par ailleurs, une accumulation de mutations dans la séquence codante modifiera la fonction du gène ou pourra conduire à l'inhibition de son expression (Zhou et Bachtrog, 2012; Lenormand et al., 2020). Ce processus peut être graduel avec une réduction partielle puis totale de l'expression du gène, mais il est aussi tout à fait possible qu'une mutation fortement délétère inhibe immédiatement l'expression du gène. Certains gènes auront donc une expression plus faible sur le X que sur le Y alors que d'autres gènes seront exprimés uniquement par le chromosome X. On appelle perte de gènes la fraction de gènes qui n'est plus exprimée par le chromosome Y. Les gènes qui ne sont plus exprimés sur le chromosome Y sont appelés X-hémizygotés. Une fois que les gènes ne sont plus exprimés, il n'y a plus de pressions de sélection pour les maintenir dans le génome. Des délétions peuvent donc les amener à disparaître totalement de la région non-recombinante du chromosome Y.

## 4.5 Compensation de dosage

La réduction, et parfois l'arrêt, de l'expression des gènes sur le chromosome Y entraîne un déséquilibre entre l'expression de ces gènes chez les mâles et chez les femelles (résumés dans Mank, 2009, 2013). Ce déséquilibre aura des conséquences négatives chez les mâles pour les gènes dont le niveau d'expression est important. Ainsi, un mécanisme de sur-expression de la copie X permet de compenser la réduction de l'expression sur le Y. C'est ce que l'on appelle la compensation de dosage (résumé dans Mank, 2009, 2013). La compensation de dosage a premièrement été étudiée chez les mammifères thériens, des drosophiles et *Caenorhabditis elegans*. Chez ces espèces, trois mécanismes différents ont été mis en évidence (résumé dans

Mank, 2009, 2013). Comme le résume Mank (2013), l'expression sur le X est doublée dans les deux sexes chez les mammifères thériens, mais l'expression d'un des deux X est inhibée chez les femelles. Le niveau d'expression est donc le même dans les deux sexes. Chez des drosophiles, seul le X chez les mâles voit son expression doublée. Enfin, chez *C. elegans*, les chromosomes X sont sur-exprimés dans les deux sexes, mais un second mécanisme réduit simultanément l'expression de certains gènes X chez les hermaphrodites (résumé dans Mank, 2013).

D'autre part, l'étude de la compensation de dosage chez d'autres espèces a mis en évidence des mécanismes relativement locaux de compensation de dosage (résumé dans Mank, 2013). Dans ces cas, seuls certains gènes sont compensés. De plus, il a été montré que la compensation de dosage est faible chez les oiseaux, bien que leurs chromosomes sexuels soient très anciens (Ellegren et al., 2007; Itoh et al., 2007; Julien et al., 2012; Sigeman et al., 2018). L'étude de la compensation de dosage chez les plantes est plus récente, mais ce mécanisme a été identifié chez plusieurs espèces (résumé dans Muyle, Shearn et Marais, 2017). Pour le moment, aucun mécanisme de compensation complète de dosage n'a été identifié chez les plantes (Muyle et al., 2018; Fruchard et al., 2020).

## 4.6 Les strates évolutives

Les premières analyses détaillées des chromosomes sexuels de l'homme ont montré que l'âge de l'arrêt de recombinaison n'est pas constant sur l'ensemble de la région non-recombinante (Lahn et Page, 1999). Ce résultat indique que différents événements d'arrêt de recombinaison se produisent lors de l'évolution d'une paire de chromosomes sexuels (Charlesworth, Charlesworth et Marais, 2005). Une région relativement large (généralement plusieurs Mb) issue d'un événement unique d'arrêt de recombinaison est appelée strate évolutive (Lahn et Page, 1999). Il est important de noter qu'il n'y a pas de consensus sur la taille minimale de cette région pour la qualifier de strate évolutive, et qu'une évolution continue (*i.e.* qui n'implique pas de strates) de la région non-recombinante est possible (résumé dans Wright et al., 2016). Le mécanisme initialement proposé pour la formation de strates évolutives est une inversion chromosomique (Lahn et Page, 1999). Ensuite, des translocations depuis des autosomes ont été identifiées dans le chromosome Y de l'homme (Waters et al., 2001). Cependant, des données empiriques soutiennent la présence de strates sans inversion, ni translocation (Bergero et al., 2008; Wang et al., 2012). Chez *Carica papaya*, par exemple, une région colinéaire entre le X et le Y écarte l'hypothèse d'une inversion pour expliquer l'arrêt de recombinaison. Pour plusieurs paires de chromosomes sexuels, les mécanismes responsables de l'expansion de la région non-recombinante restent inconnus (résumé dans Bergero et Charlesworth, 2009). De plus, à la suite d'un arrêt de recombinaison, les pressions de sélection pour empêcher une inversion dans la région non-recombinante sont faibles, il est donc possible que l'inversion chromosomique se produise après l'arrêt de recombinaison (Ross et al., 2005; Lemaitre et al., 2009; Wright et al., 2016).



La détection de strates évolutives à partir d'alignements de la région X avec la région Y est souvent compliquée à cause des nombreux réarrangements dans les régions non-recombinantes (Pandey, Wilson Sayres et Azad, 2013). Des analyses de divergence synonyme le long des chromosomes sexuels permettent de limiter le problème et d'identifier des strates évolutives (Lahn et Page, 1999; Nicolas et al., 2005; Wang et al., 2012). Des analyses phylogénétiques sont parfois utilisées pour valider l'existence de strates évolutives (Wilson & Makova, 2009). Chez les plantes, des strates évolutives ont été identifiées seulement chez *C. papaya* et *Silene latifolia* (Nicolas et al., 2005; Wang et al., 2012; Papadopulos et al., 2015).

## 5 Les chromosomes sexuels actuellement décrits chez les plantes

Seulement une trentaine de paires de chromosomes sexuels ont été décrites chez les plantes (Ming, Bendahmane et Renner, 2011; Renner et Müller, 2021; Charlesworth, 2021). La table 2 renseigne la taille de la région non-recombinante, son contenu en gènes ainsi que l'estimation de son âge pour la majorité des systèmes actuellement décrits chez les plantes. De plus, il faut noter que tous ces systèmes n'ont pas été étudiés avec les mêmes méthodes (cf. section sur les méthodes). Le nombre d'études publiées pour chaque espèce est aussi variable. Les informations présentées ici sont donc susceptibles de varier dans les prochaines années et les comparaisons inter-spécifiques doivent être interprétées prudemment.

Cependant, nous pouvons voir que l'ensemble des systèmes présentés dans ce tableau sont âgés d'au moins un million d'années. Ce tableau montre aussi qu'une corrélation entre l'âge et la taille de la région non-recombinante, comme présentée dans la Figure 2, ne semble pas universelle. Par exemple, *Vitis vinifera sylvestris* et *Diospyros lotus* ont des systèmes vieux de plusieurs dizaines de millions d'années mais des tailles de régions non-recombinantes relativement petites (<2 Mb). *Silene latifolia* possède une très grande région non-recombinante (>100 Mb) mais son système est bien plus jeune que celui des deux espèces précédentes (11 Ma). En revanche, certaines espèces montrent une évolution de leur paire de chromosomes sexuels concordante avec l'attendu du modèle. C'est par exemple le cas de *Fragaria chiloensis* qui possède une région non-recombinante de plusieurs centaines de kilobases et vieille d'environ 1 Ma. Ce tableau montre donc que le modèle présenté dans la Figure 2 ne permet pas de décrire toutes les paires de chromosomes sexuels.

De plus, certaines étapes du modèle manquent particulièrement de données empiriques. Premièrement, aucune espèce avec une paire de chromosomes sexuels très jeune (<<1 Ma) n'a été décrite pour le moment. Deuxièmement, aucun système évoluant depuis plusieurs dizaines de millions d'années avec une région non-recombinante très large (>50 Mb) n'a été décrit chez les plantes. Le manque d'identification de systèmes anciens et fortement dégénérés questionne donc l'existence de tels systèmes chez les plantes. Cette dernière étape du modèle d'évolution des chromosomes sexuels chez les plantes est inspirée d'observations faites chez

les animaux.

TABLE 2 – Tableau récapitulatif des chromosomes sexuels décrits chez les plantes. Ce tableau compile pour chaque espèce, le type de système chromosomique (XY ou ZW), la taille de la région non-recombinante, le nombre de gamétochromosomes identifiés et l'âge du système. Les espèces sont triées en fonction de la taille de leur région non-recombinante. Les \* indiquent que l'estimation de l'âge a été faite avec des données fossiles et non pas moléculaires. La majorité des données ont été récupérées à partir des revues suivantes : Charlesworth, 2021 ; Renner & Muller, 2021, Carey et al, 2021. Des données manquantes dans ces revues ont été ajoutées (Wu & Moore, 2015 ; Veltsos et al, 2018 ; Tennessen et al, 2018 ; Krasovec et al, 2018 ; Fruchard et al, 2020). Les Na signifient que l'information n'est pas disponible.

Espèces	Type de système	Taille région Non-recombinante (Mb)	Nombre de gamétochromosomes	Âge (Ma)	Références
<i>Morella rubra</i>	ZW	0,06	7	Na	Carey, Yu et Harkess (2021)
<i>Populus alba</i>	ZW	0,10	~20	4	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Vitis vinifera sylvestris</i>	XY	0,15	20	20-100	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Fragaria chiloensis</i>	ZW	0,28	~70	1	Tennessen et al. (2018); Carey, Yu et Harkess (2021) Renner et Müller (2021)
<i>Populus deltoides</i>	XY	0,30	Na	Na	Charlesworth (2021); Carey, Yu et Harkess (2021)
<i>Actinidia chinensis chinensis</i>	XY	0,80	30	20-100	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Asparagus officinalis</i>	XY	1,00	13	1-6*	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Diospyros lotus</i>	XY	1,30	32	20-28*	Charlesworth (2021); Carey, Yu et Harkess (2021) Renner et Müller (2021)
<i>Populus tremula</i>	XY	1,50	26	2-4*	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Amaranthus palmerii</i>	XY	1,30-2,00	121	Na	Carey, Yu et Harkess (2021)
<i>Salix nigra</i>	XY	2,10	Na	Na	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Salix viminalis</i>	ZW	3,10	>100	8.6	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Amborella trichopoda</i>	ZW	4,00	150	9.5-14.5	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Amaranthus tuberculatus</i>	XY	4,60	147	Na	Carey, Yu et Harkess (2021)
<i>Salix purpurea</i>	ZW	8,10	>400	8.6	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Carica papaya</i>	XY	4,00	~100	7	Wu et Moore (2015); Carey, Yu et Harkess (2021) Renner et Müller (2021); Charlesworth (2021)
<i>Spinacia oleracea</i>	XY	10,00	210	<5.7	Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Phoenix dactylifera</i>	XY	13,00	~100	~50*	Charlesworth (2021); Carey, Yu et Harkess (2021) Renner et Müller (2021)
<i>Mercurialis annua</i>	XY	>14,50	>400	1.5	Veltsos et al. (2018); Charlesworth (2021) Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Silene latifolia</i>	XY	>100,00	>1000	11	Tennessen et al. (2018); Charlesworth (2021) Carey, Yu et Harkess (2021); Renner et Müller (2021)
<i>Solanum appendiculatum</i>	XY	Na	>20	<4	Carey, Yu et Harkess (2021)
<i>Coccinia grandis</i>	XY	Na	>1300	8.7-34.7	Charlesworth (2021); Carey, Yu et Harkess (2021) Renner et Müller (2021)

## 6 Méthodes disponibles pour étudier les chromosomes sexuels

### 6.1 Détection des séquences de la région non-recombinante

Il existe différentes approches pour identifier une région non-recombinante (résumées dans Muyle, Shearn et Marais, 2017; Palmer et al., 2019) :

### 6.1.1 Ratio de couverture d’alignement mâles et femelles

Une première approche consiste à aligner des séquences d’ADN d’individus homogamétiques sur une référence construite à partir d’un individu hétérogamétique. Les contigs avec une couverture de mapping faible, voire nulle, correspondent alors à des séquences Y. En revanche, cette méthode ne permet pas d’identifier les séquences X (*e.g.* Darolti et al., 2019). Il est aussi possible d’aligner des séquences d’ADN d’individus mâles et femelles sur une référence construite à partir d’un individu homogamétique. Une méthode appelée Chromosome Quotient (Hall et al., 2013) permet de contrôler le bruit lié aux biais d’alignement. Cette méthode permet d’identifier la région non-recombinante sur le chromosome X mais ne permet pas d’identifier directement les séquences Y-spécifiques. Il est cependant possible de récupérer les lectures d’individus hétérogamétiques n’ayant pas été alignées puis de les assembler pour reconstruire les séquences Y-spécifiques.

Enfin, il est possible d’utiliser une approche dite des *k-mers* (voir Carvalho et Clark, 2013). Elle consiste à fragmenter les génomes mâles et femelles en fragments de longueurs *k* puis d’identifier les fragments dont les ratios de couverture *femelles/mâles* diffèrent de 1. Les *k-mers* spécifiques du chromosome Y auront un ratio de couverture *femelles/mâles* de 0, les *k-mers* des séquences X auront un ratio de couverture *femelles/mâles* de 2.

Ces approches sont plus puissantes lorsque la divergence entre les séquences X et Y est importante. Ainsi, elles ne sont pas les plus efficaces pour étudier des systèmes jeunes (Palmer et al., 2019).

### 6.1.2 La densité en SNPs

Cette approche consiste à estimer la densité en SNPs chez les mâles et les femelles (voir Wright et al., 2017; Darolti et al., 2019). Elle a pour avantage de fonctionner même si la divergence entre mâles et femelles est récente. Dans le cas d’un système XY, la densité en SNPs sera plus importante chez les mâles car plus de sites sont hétérozygotes. En revanche, pour des gènes qui ont arrêté de recombiner il y a longtemps et dont la copie Y n’est plus présente sur le chromosome, la densité en SNPs sera plus importante chez les femelles car les mâles seront hémizygotes pour ces gènes. Avec cette approche, il est donc possible de distinguer les régions fortement dégénérées (pour lesquelles les gènes ne sont plus présents sur le chromosome Y) des régions avec une faible divergence.

### 6.1.3 La ségrégation allélique

La ségrégation allélique permet d’identifier au sein d’un croisement les allèles qui sont transmis uniquement d’un parent à un type de descendant (Chibalina et Filatov, 2011; Bergero et Charlesworth, 2011; Muyle et al., 2016). Pour un système XY, les allèles transmis uniquement du parent mâle aux descendants mâles seront des allèles Y. Une méthode pro-

babilliste, SEX-DETECTOR (Muyle et al., 2016), analyse les ségrégations alléliques au sein d'un croisement pour identifier les séquences d'une région non-recombinante de chromosomes sexuels. SEX-DETECTOR permet aussi de tester la présence de chromosomes sexuels et, le cas échéant, quel système de chromosomes sexuels est le plus probable (XY ou ZW). L'analyse des ségrégations alléliques a pour avantage d'être efficace sur de jeunes chromosomes sexuels, mais de moins bien fonctionner pour des systèmes avec beaucoup de divergence XY. Cette méthode nécessite aussi d'être capable de faire un croisement, ce qui n'est pas le cas pour toutes les espèces. Enfin, un polymorphisme trop faible entre le parent mâle et le parent femelle limitera la détection de séquences liées au sexe, particulièrement pour les séquences X-hémizygotés.

#### 6.1.4 Méthode GWAS

Outre l'étude des chromosomes sexuels, il existe plusieurs méthodes pour identifier des associations génétiques. La méthode Genome Wide Association Study (GWAS) en fait partie. Pour cela, on considère le sexe comme une variable binaire et la méthode GWAS identifie les positions du génome associées à un sexe uniquement. Cette méthode est efficace si elle est utilisée avec un assemblage du génome, idéalement au niveau chromosomique. En revanche, plusieurs études ont mis en avant des problèmes de faux-positifs (résumé dans Palmer et al., 2019).

#### 6.1.5 Utiliser des modèles de génétique des populations

Dans une population naturelle, les fréquences alléliques et génotypiques ne sont pas supposées varier entre les sexes, sauf pour les séquences dans la région non-recombinante. En identifiant les séquences dont ces fréquences s'écartent des équilibres attendus avec les équations de génétique des populations, il est possible de trouver celles présentes sur les chromosomes sexuels (Käfer et al., 2021). Par exemple, un gène pour lequel un allèle est trouvé uniquement chez les mâles est probablement un allèle du chromosome Y. Un outil probabiliste a récemment été publié, SDpop (Käfer et al., 2021). Tout comme SEX-DETECTOR, cet outil permet à la fois d'identifier des séquences liées au sexe, mais aussi de tester quel modèle de chromosomes sexuels est le plus probable (Käfer et al., 2021).

#### 6.1.6 Construire une carte de liaison

Au lieu de chercher des traces génomiques typiques d'une absence de recombinaison il est possible de reconstruire des cartes de liaisons qui permettent d'identifier les régions qui ne recombinent plus dans un sexe (*e.g.* Bergero et al., 2019). Cette méthode peut être fastidieuse si le taux de recombinaison est faible et que la région non-recombinante est petite

(peut nécessiter plusieurs croisements avec plusieurs centaines de descendants pour chaque croisement). De plus, cette méthode ne fonctionnera pas s'il y a de l'achiasmie (recombinaison dans un seul sexe) ou une très forte hétérochiasmie (taux de recombinaison différents entre mâles et femelles).

## 6.2 Estimation de l'âge du système

Il existe deux approches principales pour estimer l'âge d'une paire de chromosomes sexuels. La première s'appuie sur la phylogénie (résumé dans Bergero et Charlesworth, 2009; Charlesworth, 2021). En fonction des données disponibles pour les espèces proches, il est possible d'inférer des états ancestraux et donc de déterminer les deux nœuds qui se situent avant et après la formation de la région non-recombinante. En estimant l'âge de ces nœuds il est possible d'approcher l'âge de la région non-recombinante. La connaissance d'un seul des deux nœuds permettra tout de même d'estimer l'âge minimal ou maximal de la région non-recombinante.

La seconde approche utilise des estimations de la divergence synonyme ( $dS$ ) entre les copies X et Y. Comme le résume Charlesworth (2021), à l'aide d'une horloge moléculaire, du temps de génération si l'horloge moléculaire le nécessite et du  $dS$ , il est possible d'inférer depuis combien de temps les séquences X et Y ont arrêté de recombiner. Cependant, le temps de génération est susceptible d'évoluer chez une espèce. Les estimations faites aujourd'hui peuvent ne pas correspondre au temps de génération de la plante y a plusieurs millions d'années. Certaines horloges moléculaires utilisent un taux de mutation par génération, leurs estimations seront donc erronées si le temps de génération est mal estimé.

De plus, sachant que la région non-recombinante est rarement issue d'un unique événement d'arrêt de recombinaison, il semble peu judicieux d'utiliser l'ensemble des valeurs de  $dS$ . Une solution est de faire la moyenne, ou la médiane, des valeurs de  $dS$  les plus élevées (les 5% les plus élevées, par exemple)(*e.g.* dans Crowson, Barrett et Wright, 2017; Veltsos et al., 2018). En revanche, cette approche comporte plusieurs biais. Premièrement, il est nécessaire de calculer la divergence synonyme sur les premiers gènes qui ont arrêté de recombiner. Or, ces mêmes gènes seront probablement les plus dégénérés donc potentiellement absents du chromosome Y. Si les premiers gènes non-recombinants ont été perdus du chromosome Y, alors l'âge sera sous-estimé. Deuxièmement, le taux de mutations n'étant pas forcément constant le long du génome (Smith, Arndt et Eyre-Walker, 2018), un gène avec un fort  $dS$  peut ne pas être ancien mais simplement dans un point chaud de mutation. Enfin, les  $dS$  étant estimés par gènes, donc sur des séquences courtes, l'estimation de l'âge sera bruitée.

Combiner des estimations moléculaires et phylogénétiques basées sur des données fossiles permettrait de limiter les erreurs dans les estimations. Cependant, les deux types de données sont rarement disponibles ensemble.

Les estimations absolues de l'âge des chromosomes sexuels sont donc à prendre avec précaution, car elles restent approximatives (voir Table 2). De plus, en fonction des méthodes

utilisées il peut être inapproprié de comparer les âges de différentes espèces.

## 7 Objectifs de la thèse

### 7.1 Brève introduction des objectifs

Comme mentionné au début de l'introduction, les objectifs de cette thèse sont d'étudier avec des données empiriques les étapes les moins bien comprises du modèle d'évolution des chromosomes sexuels chez les plantes.

Le premier axe de cette thèse a consisté à étudier la formation des chromosomes sexuels. Pour cela j'ai étudié une sous-espèce dioïque du complexe *Silene acaulis*, à savoir, *S. acaulis* ssp *exscapa*. *S. acaulis* est un modèle rare pour l'étude de l'évolution des systèmes sexuels, puisque des sous-espèces gynodioïques, trioïques et dioïques ont été identifiées. La transition vers la dioécie est probablement très récente chez *S. acaulis* ssp *exscapa*. Si une région non-recombinante existe chez cette sous-espèce, une analyse comparative avec la région homologue des autres sous-espèces permettrait de mieux comprendre la formation des chromosomes sexuels. Ainsi, la question à laquelle j'ai tenté de répondre dans le premier axe de la thèse est de savoir si des chromosomes sexuels existent chez *S. acaulis* ssp *exscapa*.

Le deuxième axe de la thèse concerne l'évolution tardive d'une paire de chromosomes sexuels chez les plantes. Deux plantes de la famille des Cannabaceae, *Cannabis sativa* et *Humulus lupulus*, ont été étudiées pour cet axe. Des analyses cytologiques et phylogénétiques soutenaient l'hypothèse que les chromosomes sexuels de ces espèces étaient anciens (détaillé ci-dessous). Valider cette hypothèse avec une approche génomique permettrait de savoir si des chromosomes sexuels de plusieurs dizaines de millions d'années et fortement dégénérés existent chez les plantes. Outre l'intérêt fondamental d'étudier les chromosomes sexuels de ces espèces, il existe aussi des intérêts économiques à identifier les séquences spécifiques de leur chromosome Y. Les chapitres 3 et 5 de cette thèse sont les analyses des chromosomes sexuels de *C. sativa* et *H. lupulus*, respectivement. Par ailleurs, nous avons aussi développé des marqueurs génétiques spécifiques du chromosome Y de *C. sativa* qui sont toujours en phase de validation expérimentale au laboratoire.

### 7.2 Axe 1 : La transition vers la dioécie chez une silène, *Silene acaulis*.

Le genre *Silene*, de la famille des Caryophyllaceae, est un clade modèle de l'étude de l'évolution des systèmes sexuels chez les plantes, et plus généralement de l'écologie et l'évolution (Bernasconi et al., 2009). Ce genre est composé d'environ 700 espèces parmi lesquelles 98 ont un système sexuel bien décrit (Casimiro-Soriguer, Buide et Narbona, 2015). L'hermaphrodisme est le système sexuel majoritaire puisqu'il est présent chez 58.2% des 98 espèces

décrites (Casimiro-Soriguer, Buide et Narbona, 2015). Parmi les espèces dont le système sexuel a été décrit, la dioécie est le second système sexuel le plus représenté, avec 14.3% des 98 espèces (Casimiro-Soriguer, Buide et Narbona, 2015). Ensuite viennent la gynodioécie et la gynodioécie-gynomonoécie qui représentent respectivement 13.3% et 12.1% des silènes dont le système sexuel est connu (Casimiro-Soriguer, Buide et Narbona, 2015). Le pourcentage d'espèces dioïques de cette étude suggère une sur-représentation de la dioécie dans ce genre comparé à la moyenne chez les angiospermes (Renner, 2014). Il est tout de même important de noter que plus de 80% des espèces de silènes ne sont pas intégrées dans cette étude, et que ces pourcentages pourraient donc être relativement écartés de la réalité. De plus, la gynodioécie est plus difficile à détecter que la dioécie. Il y a donc probablement une surestimation de la dioécie comparée à la gynodioécie.

Par ailleurs, l'absence de monoécie au sein des silènes suggère que la transition vers la dioécie a eu lieu via la gynodioécie (Casimiro-Soriguer, Buide et Narbona, 2015; Fruchard et Marais, 2021). Desfeux et al. (1996) ont estimé que la dioécie a émergé au moins deux fois au sein du genre *Silene*. Dans cette étude, il est suggéré que la dioécie chez *S. acaulis* et les espèces dioïques de la section *Otites* pourrait être issue d'un même événement de transition vers la dioécie. Cependant, peu d'espèces proches de *S. acaulis* et de la section *Otites* sont présentes dans cette étude. Slancarova et al. (2013) ont montré que la gynodioécie est probablement le système sexuel ancestral de la section *Otites*. De plus, comme le montre la Figure 3, il est plus parcimonieux de suggérer que la dioécie chez *S. acaulis* et les espèces dioïques de la section *Otites* soit issue de deux événements indépendants.

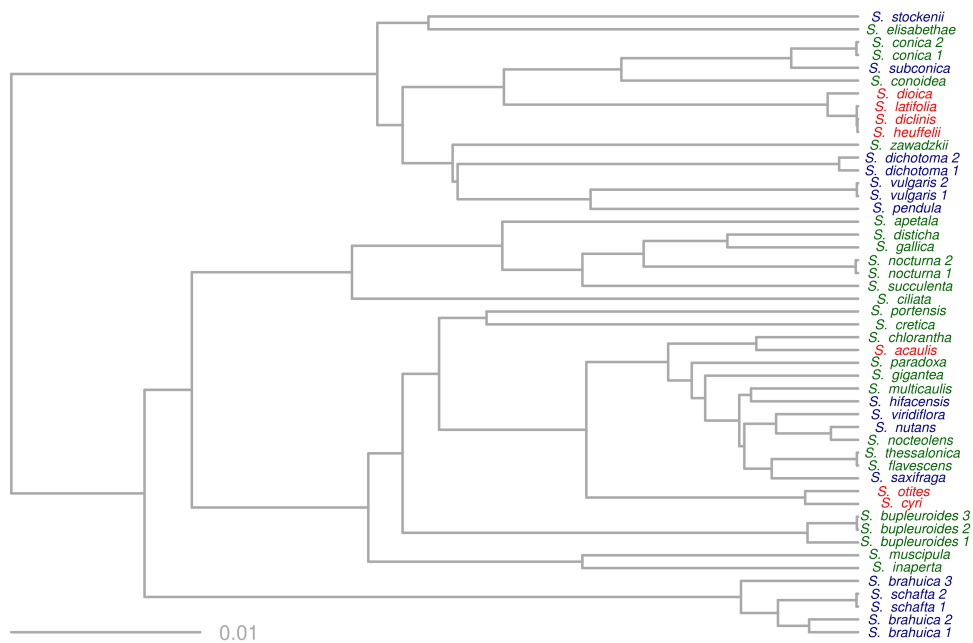


FIGURE 3 – Phylogénie des silènes avec les systèmes sexuels des espèces. Les systèmes sexuels des espèces ont été récupérés depuis Casimiro-Soriguer, Buide et Narbona (2015). En vert sont représentées les espèces hermaphroditiques ; en bleu les espèces gynodioïques et en rouge les espèces dioïques. La phylogénie de ces espèces a été obtenue grâce à un sous-échantillonnage des espèces présentes dans Casimiro-Soriguer, Buide et Narbona (2015) dans la phylogénie de Jafari et al. (2020). L'échelle pour la longueur des branches est indiquée en bas à gauche de la figure.

Ces événements indépendants de transition vers la dioécie font du genre *Silene* un genre modèle pour l'étude de l'évolution des chromosomes sexuels chez les plantes. En effet, mis à part pour *S. acaulis* ssp *exscapa*, des chromosomes sexuels ont été trouvés chez toutes les espèces dioïques du genre *Silene* qui ont été étudiées (Mrackova et al., 2008; Balounova et al., 2019; Martin et al., 2019; Muyle et al., 2021). Cinq espèces dioïques ont été décrites dans la section Melandrium, et leurs paires de chromosomes sexuels sont homologues, donc toutes héritées d'un ancêtre commun (Marais et al., 2011). L'histoire des chromosomes sexuels au sein de la section Otites semble plus compliquée. *Silene otites* possède un système ZW alors que *Silene pseudotites* possède un système XY (Martin et al., 2019). Par ailleurs, *Silene colpophylla* possède aussi un système XY, et il est probable que le locus du sexe de *S. pseudotites* soit issu d'une introgression du locus du sexe de *S. colpophylla* (Balounova et al., 2019). Récemment, une paire de chromosomes sexuels a même été identifiée dans une espèce gynodioïque de la section Otites, *Silene sibirica* (B. Janousek, résultats non publiés). Par ailleurs, *Silene latifolia* est une espèce modèle de l'évolution des chromosomes sexuels chez les plantes (Bernasconi et al., 2009; Kejnovsky et Vyskot, 2010; Krasovec et al., 2018). Comme mentionné précédemment, il existe un polymorphisme des systèmes sexuels chez *S. acaulis* (Maurice et al., 1998; Gussarova et al., 2015). Le premier axe de cette thèse a consisté



à tester la présence, ou non, d'une région non-recombinante des populations dioïques de *S. acaulis*. Si elle existe, cette région est supposée être jeune, donc avec peu de divergence entre les gamétologues. Par ailleurs, pour un système très récent, nous nous attendons à identifier peu de gènes non-recombinants. L'identification des gènes du déterminisme du sexe sera donc plus simple que chez une espèce comme *S. latifolia*, pour laquelle plus de 1000 gènes sont dans la région non-recombinante.

Cependant, aucun génome de référence assemblé au niveau chromosomique n'est disponible chez les silènes. Ceci écarte les approches de ratio de couverture *mâles/femelles*. Par ailleurs, nous ne maîtrisons pas bien la culture de cette plante sous serre et le temps de génération peut-être de plusieurs années (Morris et Doak, 1998). Il n'est donc pas envisageable de faire un croisement rapidement. Une approche SEX-DETECTOR et la construction d'une carte de liaison ne sont donc pas non plus envisageables. Pour tenter d'identifier des gènes dans une région non-recombinante, j'ai utilisé SDpop, une approche basée sur la génétique des populations, avec deux jeux de données provenant de deux populations différentes.

## 7.3 Axe 2 : Les chromosomes sexuels chez les Cannabaceae : étude des stades avancés de l'évolution des chromosomes sexuels et intérêts pour l'agronomie

### 7.3.1 Les intérêts fondamentaux des chromosomes sexuels de *Cannabis sativa* et *Humulus lupulus*

*Cannabis sativa* et *Humulus lupulus* sont deux espèces dioïques de la famille des Cannabaceae. Les analyses phylogénétiques estiment que la divergence entre les genres *Humulus* et *Cannabis* a commencé il y a entre 21 et 25 millions d'années (Divashuk et al., 2014; Jin et al., 2020). Par ailleurs, des analyses cytologiques ont identifié une paire de chromosomes sexuels XY chez *C. sativa* et chez *H. lupulus* (Shephard et al., 2000; Divashuk et al., 2011, 2014). Chez *H. lupulus*, les chromosomes sexuels sont hétéromorphes et le chromosome Y est plus petit que le chromosome X. De plus, ces analyses ont montré que les régions non-recombinantes sont grandes (plus de la moitié du chromosome X) chez les deux espèces (Divashuk et al., 2011, 2014). Bien que les gènes du déterminisme du sexe ne soient pas connus chez ces espèces, il a été proposé que le sexe soit déterminé par le ratio du nombre de chromosomes X / nombre d'autosomes (ratio  $X/A$ ) chez *H. lupulus* (Neve, 1961; Shephard et al., 2000). Des expérimentations faites dans les années 1940 chez *C. sativa* ont montré que des individus XXXY étaient principalement des femelles, ce qui laissait penser que le sexe est déterminé par un ratio  $X/A$  (résumé dans Westergaard (1958)). Plus récemment, des études ont plutôt soutenu l'hypothèse d'un Y actif (Sakamoto et al., 1995; Matsunaga et Kawano, 2001; Ming, Bendahmane et Renner, 2011).

La famille des Cannabaceae comporte 10 genres et plus de 110 espèces (Zhang et al., 2018).

Yang et al. (2013) ont étudié l'évolution de différents traits d'histoire de vie au sein de cette famille, notamment les différents systèmes sexuels (Figure 4). Parmi les 42 feuilles utilisées pour leur phylogénie, 16 (38%) sont monoïques strictes, 6 (14%) sont dioïques strictes, 4 (10%) sont andromonoïques, 11 (26%) sont monoïques et dioïques, et enfin 5 (12%) sont hermaphrodites Yang et al. (2013). Les auteurs concluent que la monoécie est l'état ancestral de la famille des Cannabaceae le plus probable. Nous pouvons noter que la dioécie est sur-représentée dans cette famille comparée à la moyenne chez les angiospermes. Plus récemment, Zhang et al. (2018) ont estimé que la dioécie était le système sexuel de l'ancêtre commun des Cannabaceae, des Urticaceae, des Artocarpeae et des Moraceae. Cependant l'article ne précise pas clairement si l'ancêtre commun des Cannabaceae était dioïque. De plus, bien que tous les systèmes sexuels ne soient pas connus au sein des Cannabaceae, il y a probablement une surestimation de la fréquence des Cannabaceae dioïques dans leur échantillonnage (2/5 soit 40%) (Zhang et al., 2018). Il n'y a donc toujours pas de consensus sur le système sexuel de l'ancêtre commun des Cannabaceae mais il n'est pas impossible que ce dernier fut dioïque (Yang et al., 2013; Jin et al., 2020). En revanche, il est probable que l'ancêtre commun des genres *Humulus* et *Cannabis* était dioïque (Yang et al., 2013).

Il est important de noter que (Yang et al., 2013) considèrent les genres *Humulus* et *Cannabis* comme étant monoïques et dioïques, alors que la plupart des auteurs définissent *C. sativa* et *H. lupulus* comme des espèces dioïques (Sakamoto et al., 1995; Ming, Bendahmane et Renner, 2011; Renner, 2014; Divashuk et al., 2014). La limite entre les deux définitions est probablement assez fine. Chez *H. lupulus*, il a été reporté que des individus monoïques peuvent émerger (Skof et al., 2012) et plusieurs variétés monoïques de *Cannabis sativa* sont cultivées (Razumova et al., 2016). La monoécie chez *C. sativa* a été sélectionnée lors de la domestication de cette plante (Small, 2015).

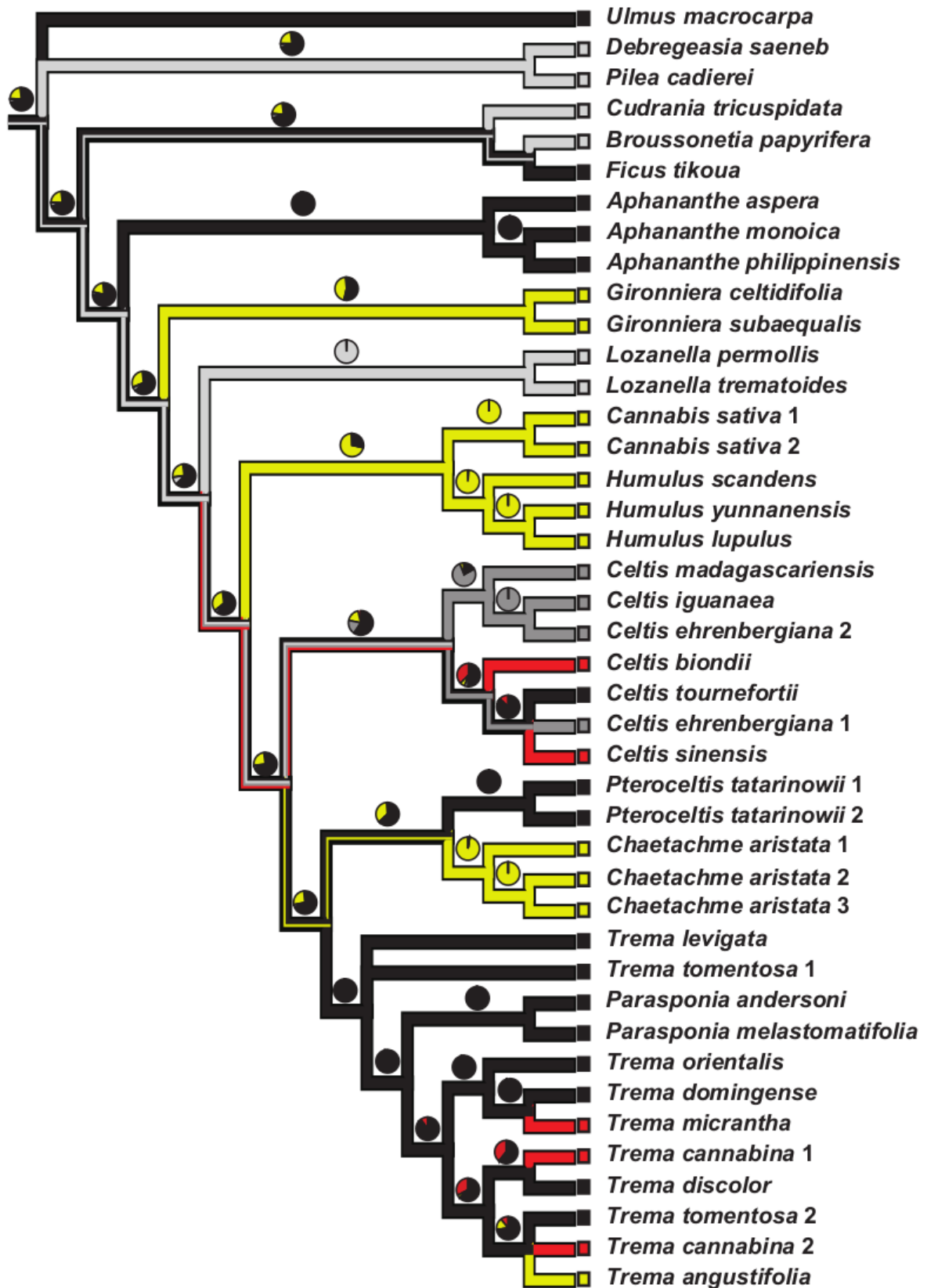


FIGURE 4 – Figure produite par Yang et al. (2013). Reconstruction des systèmes sexuels le long de la phylogénie des Cannabaceae. En rouge, les espèces hermaphrodites ; en jaune, les espèces monoïques et dioïques ; en gris foncé, les andromonoïques ; en gris clair, les dioïques strictes ; en noir, les monoïques strictes. Les probabilités postérieures pour les nœuds le long de la phylogénie sont représentées avec des diagrammes.

La dioécie est donc probablement ancienne chez *C. sativa* et *H. lupulus*. Ayant toutes les deux une paire de chromosomes sexuels XY, ces espèces sont de bons modèles pour tester si des chromosomes sexuels de plusieurs dizaines de millions d'années avec de grandes régions non-recombinantes fortement dégénérées existent chez les plantes. Comme mentionné précédemment, il était déjà admis avant le début de cette thèse que les régions non-recombinantes étaient grandes chez ces deux espèces. En revanche, il n'y avait pas de données génomiques, donc leurs âges et leurs niveaux de dégénérescence (*i.e.* perte de gènes, réduction de l'expression Y et compensation de dosage) n'étaient pas connus. Ces espèces présentent aussi l'opportunité de tester si des chromosomes sexuels qui divergent depuis plus de 20 millions d'années peuvent être homologues chez les plantes. De tels chromosomes sexuels ont uniquement été observés chez les animaux jusqu'à présent (*i.e.* chez les oiseaux et les mammifères)(Ohno, 1969; Fridolfsson et al., 1998; Cortez et al., 2014).

Pour répondre à ces questions la méthode SEX-DETECTOR, basée sur l'analyse de ségrégations alléliques, a été utilisée. Au début des analyses, aucun génome n'avait été assemblé au niveau chromosomique, ce qui limitait les approches avec le ratio de couvertures *mâles/femelles* le long du génome, par exemple. De plus, ce sont des plantes dont les techniques de cultures sont très bien maîtrisées, ce qui permet de faire des croisements facilement. Enfin, pour étudier la réduction de l'expression Y et la compensation de dosage, les données RNA-seq sont essentielles. L'approche SEX-DETECTOR avec des données transcriptomiques représentait l'effort économique, expérimental et bio-informatique (pas besoin d'assembler de génome) le plus rentable pour tenter de répondre à nos questions. Au cours de cette étude, plusieurs génomes de *C. sativa* assemblés au niveau chromosomique ont été publiés, ce qui a permis d'approfondir nos analyses.

### **7.3.2 *Cannabis sativa* et *Humulus lupulus* sont des plantes économiquement importantes**

Outre les intérêts fondamentaux de travailler sur les chromosomes sexuels de *C. sativa* et de *H. lupulus*, l'identification de régions Y-spécifiques chez ces espèces représente un intérêt économique. *C. sativa* est une des premières plantes cultivées (Pisanti et Bifulco, 2019; Crocq, 2020) et ses usages sont aujourd'hui multiples (résumé dans Schilling et al., 2020). Les premiers sont à des fins médicinales ou récréatives avec les variétés de marijuana (Schilling et al., 2020). Les seconds reposent davantage sur les fibres et les huiles des variétés de chanvre. Les fibres de chanvre sont aujourd'hui principalement utilisées comme bio-matériaux de construction ou pour la conception de tissus et de papier (Schilling et al., 2020). L'huile est obtenue à partir des graines et principalement utilisée comme bio-carburant, source de nutrition ou encore à des fins cosmétiques. Enfin, le chanvre suscite un grand intérêt pour la lutte contre le changement climatique puisqu'il peut aussi être utilisé pour capturer du carbone (Schilling et al., 2020). La distinction entre le chanvre et la marijuana se fait sur la quantité de cannabinoïdes produite. Les cannabinoïdes sont les molécules psychoactives

utilisées à des fins médicinales ou récréatives, dont les plus abondantes sont le tétrahydrocannabinol (THC) et le cannabidiol (CBD). Au cours du dernier siècle la prohibition de la marijuana a été très forte dans la grande majorité des pays. Depuis maintenant une douzaine d'années un nombre croissant de pays légalisent la marijuana à des fins médicinales et/ou récréatives (Offord, 2018; Yeager, 2018). Cette légalisation s'est accompagnée du développement de produits dérivés (huiles, confiseries, alimentation, ...) et d'une industrialisation de la culture de marijuana. La culture du *Cannabis* connaît donc une explosion mondiale depuis quelques années. Aux États-Unis, des estimations annoncent que l'industrie de la marijuana dépassera les 100 milliards de dollars en 2022 (référence : marijuana business daily<sup>1</sup>). Ceci entraîne le développement de biotechnologies pour optimiser la culture de la marijuana.

Le houblon est aussi une plante à forte valeur économique. Il est l'un des ingrédients clefs de la production de bière, un marché estimé à plus de 100 milliards de dollars en 2017 (référence : National Beer Sales & Production Data<sup>2</sup>). La fabrication de la bière a évolué ces dernières années et le houblon n'a plus comme unique fonction d'accroître la conservation et l'amertume de la bière. Il est aujourd'hui utilisé comme ingrédient aromatique dans un nombre croissant de bières, appelées les «craft beers». Depuis plusieurs années, la production de bière utilisant beaucoup de houblon a significativement augmenté (King et Pavlovič, 2017), on parle même de «Craft Beer revolution». Ceci a pour conséquence d'augmenter radicalement la production de houblon (+30% entre 2007 et 2017 ; référence : The Barth report<sup>3</sup>) et de favoriser l'émergence de nouveaux croisements avec des arômes différents (Lafontaine et Shellhammer, 2019).

Pour la marijuana comme pour le houblon, seules les femelles sont utiles. Chez le cannabis, les cannabinoïdes sont majoritairement produits dans les trichomes (fines excroissances concentrées sur les fleurs) des femelles non pollinisées (résumé dans McKernan et al., 2020). Chez le houblon, la lupuline est aussi synthétisée essentiellement dans les inflorescences femelles, appelées cônes (OKADA et ITO, 2001). Comme pour les cannabinoïdes, la synthèse de lupuline est interrompue si la femelle est pollinisée (Thomas et Neve, 1976). Il est donc préférable de ne pas avoir de mâles dans les cultures de marijuana et de houblon. La problématique est que le dimorphisme sexuel est faible chez ces deux espèces, il est donc difficile d'identifier les mâles et les femelles avant la période de floraison (McAdam et al., 2014; Campbell, Peach et Wizenberg, 2021).

Les stratégies de culture sont différentes entre les deux espèces. Chez le houblon, les producteurs pratiquent la multiplication végétative d'individus femelles. Ceci permet de garantir une descendance femelle. Des croisements sont tout de même effectués pour l'amélioration variétale. Il serait intéressant de pouvoir identifier le sexe le plus précocement possible pour ne pas être tributaire de la floraison, qui peut mettre plusieurs années à se faire chez le houblon (Patzak et al., 2002). Chez la marijuana, les producteurs utilisent les graines pour la production. Il n'y a donc pas de contrôle sur le sexe. Plusieurs technologies ont été mises en

---

1. <https://mjbizdaily.com/marijuana-industry-expected-to-add-92-billion-to-us-economy-in-2021/>

2. <https://www.brewersassociation.org/statistics-and-data/national-beer-stats/>

3. [https://www.barthhaas.com/fileadmin/user\\_upload/news/2019-07-23/barthreport20182019en.pdf](https://www.barthhaas.com/fileadmin/user_upload/news/2019-07-23/barthreport20182019en.pdf)

place pour optimiser la proportion de femelle dans les cultures (Résumées dans McKernan et al., 2020). L'utilisation de phytohormone permet par exemple de changer le sexe des femelles afin d'obtenir des mâles. Ceci permet de produire 100% de pollen avec un chromosome X. Les graines issues de ces grains de pollen seront toutes XX, donc théoriquement femelles (Mohan Ram et Sett, 1982). La limite de cette approche est l'augmentation de la proportion d'hermaphrodite dans la descendance (McKernan et al., 2020).

Il serait donc intéressant d'identifier le sexe des individus de façon précoce grâce à des marqueurs génétiques. Des marqueurs ont été développés chez *C. sativa* mais ils correspondent à des séquences répétées et pourraient ne pas être universels. En identifiant une séquence codante Y suffisamment divergente de la séquence homologue X nous pourrions développer des amorces PCR qui seraient amplifiées seulement chez les mâles, et dont la fiabilité pourrait être supérieure à celle des marqueurs existants.

## Références

- Akagi, T., I. M. Henry, R. Tao, et L. Comai. 2014, A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science* **346** :646–650.
- Akagi, T., S. M. Pilkington, E. Varkonyi-Gasic, et al. 2019, Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nature Plants* **5** :801–809.
- Almeida, P., E. Proux-Wera, A. Churcher, et al. 2020, Genome assembly of the basket willow, *Salix viminalis*, reveals earliest stages of sex chromosome expansion. *BMC Biology* **18** :78.
- Bachtrog, D. 2006, A dynamic view of sex chromosome evolution. *Current Opinion in Genetics & Development* **16** :578–585.
- . 2013, Y-chromosome evolution : emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics* **14** :113–124.
- Bachtrog, D., M. Kirkpatrick, J. E. Mank, S. F. McDaniel, J. C. Pires, W. Rice, et N. Valenzuela. 2011, Are all sex chromosomes created equal? *Trends in Genetics* **27** :350–357.
- Bachtrog, D., J. E. Mank, C. L. Peichel, et al. 2014, Sex Determination : Why So Many Ways of Doing It? *PLOS Biology* **12** :e1001899.
- Balounova, V., R. Gogela, R. Cegan, et al. 2019, Evolution of sex determination and heterogamety changes in section *Otites* of the genus *Silene*. *Scientific Reports* **9** :1045.
- Barrett, S. C. H. 2002, The evolution of plant sexual diversity. *Nature Reviews. Genetics* **3** :274–284.
- Barton, N. H. et B. Charlesworth. 1998, Why Sex and Recombination? *Science* **281** :1986–1990.

- Bawa, K. S. et J. H. Beach. 1981, Evolution of Sexual Systems in Flowering Plants. *Annals of the Missouri Botanical Garden* **68** :254–274.
- Bergero, R. et D. Charlesworth. 2009, The evolution of restricted recombination in sex chromosomes. *Trends in Ecology & Evolution* **24** :94–102.
- . 2011, Preservation of the Y Transcriptome in a 10-Million-Year-Old Plant Sex Chromosome System. *Current Biology* **21** :1470–1474.
- Bergero, R., D. Charlesworth, D. A. Filatov, et R. C. Moore. 2008, Defining Regions and Rearrangements of the *Silene latifolia* Y Chromosome. *Genetics* **178** :2045–2053.
- Bergero, R., J. Gardner, B. Bader, L. Yong, et D. Charlesworth. 2019, Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. *Proceedings of the National Academy of Sciences* **116** :6924–6931.
- Bernasconi, G., J. Antonovics, A. Biere, et al. 2009, *Silene* as a model system in ecology and evolution. *Heredity* **103** :5–14.
- Branco, S., H. Badouin, R. C. R. d. l. Vega, et al. 2017, Evolutionary strata on young mating-type chromosomes despite the lack of sexual antagonism. *Proceedings of the National Academy of Sciences* **114** :7067–7072.
- Campbell, L. G., K. Peach, et S. B. Wizenberg. 2021, Dioecious hemp (*Cannabis sativa* L.) plants do not express significant sexually dimorphic morphology in the seedling stage. *Scientific Reports* **11** :16825.
- Carey, S., Q. Yu, et A. Harkess. 2021, The Diversity of Plant Sex Chromosomes Highlighted through Advances in Genome Sequencing. *Genes* **12** :381.
- Carvalho, A. B. et A. G. Clark. 2013, Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Research* **23** :1894–1907.
- Casimiro-Soriguer, I., M. L. Buide, et E. Narbona. 2015, Diversity of sexual systems within different lineages of the genus *Silene*. *AoB PLANTS* **7**.
- Cavoto, E., S. Neuenschwander, J. Goudet, et N. Perrin. 2018, Sex-antagonistic genes, XY recombination and feminized Y chromosomes. *Journal of Evolutionary Biology* **31** :416–427.
- Chalopin, D., J.-N. Volff, D. Galiana, J. L. Anderson, et M. Schartl. 2015, Transposable elements and early evolution of sex chromosomes in fish. *Chromosome Research* **23** :545–560.
- Charlesworth, B. et D. Charlesworth. 1978, A Model for the Evolution of Dioecy and Gynodioecy. *The American Naturalist* **112** :975–997.
- . 2000, The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences* **355** :1563–1572.

- Charlesworth, B., C. Langley, et P. Sniegowski. 1997, Transposable element distributions in *Drosophila*. *Genetics* **147** :1993.
- Charlesworth, B. et J. D. Wall. 1999, Inbreeding, heterozygote advantage and the evolution of neo-X and neo-Y sex chromosomes. *Proceedings of the Royal Society of London. Series B : Biological Sciences* **266** :51–56.
- Charlesworth, D. 2016, Plant Sex Chromosomes. *Annual Review of Plant Biology* **67** :397–420.
- . 2019, Young sex chromosomes in plants and animals. *New Phytologist* **224** :1095–1107.
- . 2021, When and how do sex-linked regions become sex chromosomes? *Evolution* **75** :569–581.
- Charlesworth, D. et B. Charlesworth. 1987, Inbreeding Depression and Its Evolutionary Consequences. *Annual Review of Ecology and Systematics* **18** :237–268.
- Charlesworth, D., B. Charlesworth, et G. Marais. 2005, Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95** :118–128.
- Chibalina, M. V. et D. A. Filatov. 2011, Plant Y Chromosome Degeneration Is Retarded by Haploid Purifying Selection. *Current Biology* **21** :1475–1479.
- Cortez, D., R. Marin, D. Toledo-Flores, L. Froidevaux, A. Liechti, P. D. Waters, F. Grützner, et H. Kaessmann. 2014, Origins and functional evolution of Y chromosomes across mammals. *Nature* **508** :488–493.
- Crocq, M.-A. 2020, History of cannabis and the endocannabinoid system. *Dialogues in Clinical Neuroscience* **22** :223–228.
- Crowson, D., S. C. Barrett, et S. I. Wright. 2017, Purifying and Positive Selection Influence Patterns of Gene Loss and Gene Expression in the Evolution of a Plant Sex Chromosome System. *Molecular Biology and Evolution* **34** :1140–1154.
- Darolti, I., A. E. Wright, B. A. Sandkam, et al. 2019, Extreme heterogeneity in sex chromosome differentiation and dosage compensation in livebearers. *Proceedings of the National Academy of Sciences* **116** :19031–19036.
- Desfeux, C., S. Maurice, J.-p. Henry, B. Lejeune, et P.-h. Gouyon. 1996, Evolution of reproductive systems in the genus *Silene*. *Proceedings of the Royal Society of London. Series B : Biological Sciences* **263** :409–414.
- Divashuk, M. G., O. S. Alexandrov, P. Y. Kroupin, et G. I. Karlov. 2011, Molecular Cytogenetic Mapping of *Humulus lupulus* Sex Chromosomes. *Cytogenetic and Genome Research* **134** :213–219.



- Divashuk, M. G., O. S. Alexandrov, O. V. Razumova, I. V. Kirov, et G. I. Karlov. 2014, Molecular Cytogenetic Characterization of the Dioecious *Cannabis sativa* with an XY Chromosome Sex Determination System. *PLOS ONE* **9** :e85118.
- Dufay, M., P. Champelovier, J. Käfer, J. P. Henry, S. Mousset, et G. A. B. Marais. 2014, An angiosperm-wide analysis of the gynodioecy–dioecy pathway. *Annals of Botany* **114** :539–548.
- Dufresnes, C., A. Brelsford, F. Baier, et N. Perrin. 2021, When Sex Chromosomes Recombine Only in the Heterogametic Sex : Heterochiasmy and Heterogamety in *Hyla* Tree Frogs. *Molecular Biology and Evolution* **38** :192–200.
- Ellegren, H., L. Hultin-Rosenberg, B. Brunström, L. Dencker, K. Kultima, et B. Scholz. 2007, Faced with inequality : chicken do not have a general dosage compensation of sex-linked genes. *BMC Biology* **5** :40.
- Fridolfsson, A.-K., H. Cheng, N. G. Copeland, N. A. Jenkins, H.-C. Liu, T. Raudsepp, T. Woodage, B. Chowdhary, J. Halverson, et H. Ellegren. 1998, Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proceedings of the National Academy of Sciences* **95** :8147–8152.
- Fruchard, C., H. Badouin, D. Latrassé, R. S. Devani, A. Muyle, B. Rhoné, S. S. Renner, A. K. Banerjee, A. Bendahmane, et G. A. B. Marais. 2020, Evidence for Dosage Compensation in *Coccinia grandis*, a Plant with a Highly Heteromorphic XY System. *Genes* **11** :787.
- Fruchard, C. et G. A. B. Marais. 2021, The Evolution of Sex Determination in Plants. *In* L. Nuño de la Rosa et G. B. Müller, eds., *Evolutionary Developmental Biology : A Reference Guide*, 683–696, Springer International Publishing, Cham.
- Gaut, B. S., S. I. Wright, C. Rizzon, J. Dvorak, et L. K. Anderson. 2007, Recombination : an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics* **8** :77–84.
- Golenberg, E. M. et N. W. West. 2013, Hormonal interactions and gene regulation can link monoecy and environmental plasticity to the evolution of dioecy in plants. *American Journal of Botany* **100** :1022–1037.
- Gussarova, G., G. A. Allen, Y. Mikhaylova, L. J. McCormick, V. Mirré, K. L. Marr, R. J. Hebda, et C. Brochmann. 2015, Vicariance, long-distance dispersal, and regional extinction–recolonization dynamics explain the disjunct circumpolar distribution of the arctic-alpine plant *Silene acaulis*. *American Journal of Botany* **102** :1703–1720.
- Hall, A. B., Y. Qi, V. Timoshevskiy, M. V. Sharakhova, I. V. Sharakhov, et Z. Tu. 2013, Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics* **14** :273.

- Harkess, A., K. Huang, R. van der Hulst, B. Tissen, J. L. Caplan, A. Koppula, M. Batish, B. C. Meyers, et J. Leebens-Mack. 2020, Sex Determination by Two Y-Linked Genes in Garden Asparagus[OPEN]. *The Plant Cell* **32** :1790–1796.
- Heilbut, J. C. 2000, Lower Species Richness in Dioecious Clades. *The American Naturalist* **156** :221–241.
- Hill, W. G. et A. Robertson. 1966, The effect of linkage on limits to artificial selection. *Genetics Research* **8** :269–294.
- Ironside, J. E. 2010, No amicable divorce? Challenging the notion that sexual antagonism drives sex chromosome evolution. *BioEssays* **32** :718–726.
- Itoh, Y., E. Melamed, X. Yang, et al. 2007, Dosage compensation is less effective in birds than in mammals. *Journal of Biology* **6** :2.
- Jafari, F., S. Zarre, A. Gholipour, F. Eggen, R. K. Rabeler, et B. Oxelman. 2020, A new taxonomic backbone for the infrageneric classification of the species-rich genus *Silene* (Caryophyllaceae). *TAXON* **69** :337–368.
- Jay, P., E. Tezenas, et T. Giraud. 2021, A deleterious mutation-sheltering theory for the evolution of sex chromosomes and supergenes. Tech. rep.
- Jeffries, D. L., J. F. Gerchen, M. Scharmann, et J. R. Pannell. 2021, A neutral model for the loss of recombination on sex chromosomes. *Philosophical Transactions of the Royal Society B : Biological Sciences* **376** :20200096.
- Jin, J.-J., M.-Q. Yang, P. W. Fritsch, R. van Velzen, D.-Z. Li, et T.-S. Yi. 2020, Born migrators : Historical biogeography of the cosmopolitan family Cannabaceae. *Journal of Systematics and Evolution* **58** :461–473.
- Julien, P., D. Brawand, M. Soumillon, A. Necșulea, A. Liechti, F. Schütz, T. Daish, F. Grützner, et H. Kaessmann. 2012, Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS biology* **10** :e1001328.
- Kejnovsky, E. et B. Vyskot. 2010, *Silene latifolia* : The Classical Model to Study Heteromorphic Sex Chromosomes. *Cytogenetic and Genome Research* **129** :250–262.
- King, M. A. et M. Pavlovič. 2017, Analysis of hop use in craft breweries in slovenia. *Journal of Agriculture Food and Development* **3** :21–26.
- Krasovec, M., M. Chester, K. Ridout, et D. A. Filatov. 2018, The Mutation Rate and the Age of the Sex Chromosomes in *Silene latifolia*. *Current Biology* **28** :1832–1838.e4.
- Käfer, J., H. J. de Boer, S. Mousset, A. Kool, M. Dufay, et G. a. B. Marais. 2014, Dioecy is associated with higher diversification rates in flowering plants. *Journal of Evolutionary Biology* **27** :1478–1490.

- Käfer, J., N. Lartillot, G. A. B. Marais, et F. Picard. 2021, Detecting sex-linked genes using genotyped individuals sampled in natural populations. *Genetics* **218**.
- Käfer, J., G. A. B. Marais, et J. R. Pannell. 2017, On the rarity of dioecy in flowering plants. *Molecular Ecology* **26** :1225–1241.
- Lafontaine, S. R. et T. H. Shellhammer. 2019, How hoppy beer production has redefined hop quality and a discussion of agricultural and processing strategies to promote it. *MBAA TQ* **56** :1–12.
- Lahn, B. T. et D. C. Page. 1999, Four Evolutionary Strata on the Human X Chromosome. *Science* **286** :964–967.
- Lemaitre, C., M. D. V. Braga, C. Gautier, M.-F. Sagot, E. Tannier, et G. A. B. Marais. 2009, Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in Human Sex Chromosomes. *Genome Biology and Evolution* **1** :56–66.
- Lenormand, T. et J. Dutheil. 2005, Recombination Difference between Sexes : A Role for Haploid Selection. *PLOS Biology* **3** :e63.
- Lenormand, T., F. Fyon, E. Sun, et D. Roze. 2020, Sex chromosome degeneration by regulatory evolution. *Current Biology* **30** :3001–3006.
- Lenormand, T. et D. Roze. 2021, Y recombination arrest and degeneration in the absence of sexual dimorphism. Tech. rep.
- Mank, J. E. 2009, The W, X, Y and Z of sex-chromosome dosage compensation. *Trends in Genetics* **25** :226–233.
- . 2013, Sex chromosome dosage compensation : definitely not for everyone. *Trends in Genetics* **29** :677–683.
- Marais, G. A. B., A. Forrest, E. Kamau, J. Käfer, V. Daubin, et D. Charlesworth. 2011, Multiple Nuclear Gene Phylogenetic Analysis of the Evolution of Dioecy and Sex Chromosomes in the Genus *Silene*. *PLOS ONE* **6** :e21915.
- Martin, H., F. Carpentier, S. Gallina, C. Godé, E. Schmitt, A. Muyle, G. A. B. Marais, et P. Touzet. 2019, Evolution of Young Sex Chromosomes in Two Dioecious Sister Plant Species with Distinct Sex Determination Systems. *Genome Biology and Evolution* **11** :350–361.
- Matsunaga, S. et S. Kawano. 2001, Sex Determination by Sex Chromosomes in Dioecious Plants. *Plant Biology* **3** :481–488.
- Maurice, S., E. Belhassen, D. Couvet, et P.-H. Gouyon. 1994, Evolution of dioecy : can nuclear–cytoplasmic interactions select for maleness ? *Heredity* **73** :346–354.

- Maurice, S., C. Desfeux, A. Mignot, et J.-P. Henry. 1998, Is *Silene acaulis* (Caryophyllaceae) a trioecious species? Reproductive biology of two subspecies. *Canadian Journal of Botany* **76** :478–485.
- McAdam, E. L., R. E. Vaillancourt, A. Koutoulis, et S. P. Whittock. 2014, Quantitative genetic parameters for yield, plant growth and cone chemical traits in hop (*Humulus lupulus*L.). *BMC Genetics* **15** :22.
- McKernan, K. J., Y. Helbert, L. T. Kane, et al. 2020, Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes. Tech. rep.
- Ming, R., A. Bendahmane, et S. S. Renner. 2011, Sex Chromosomes in Land Plants. *Annual Review of Plant Biology* **62** :485–514.
- Mohan Ram, H. Y. et R. Sett. 1982, Induction of fertile male flowers in genetically female *Cannabis sativa* plants by silver nitrate and silver thiosulphate anionic complex. TAG. Theoretical and applied genetics. *Theoretische und angewandte Genetik* **62** :369–375.
- Morris, W. F. et D. F. Doak. 1998, Life history of the long-lived gynodioecious cushion plant *Silene acaulis* (Caryophyllaceae), inferred from size-based population projection matrices. *American Journal of Botany* **85** :784–793.
- Mrackova, M., M. Nicolas, R. Hobza, I. Negrutiu, F. Monéger, A. Widmer, B. Vyskot, et B. Janousek. 2008, Independent Origin of Sex Chromosomes in Two Species of the Genus *Silene*. *Genetics* **179** :1129–1133.
- Muyle, A., J. Käfer, N. Zemp, S. Mousset, F. Picard, et G. A. Marais. 2016, SEX-DETECTOR : A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms. *Genome Biology and Evolution* **8** :2530–2543.
- Muyle, A., H. Martin, N. Zemp, et al. 2021, Dioecy Is Associated with High Genetic Diversity and Adaptation Rates in the Plant Genus *Silene*. *Molecular Biology and Evolution* **38** :805–818.
- Muyle, A., R. Shearn, et G. A. Marais. 2017, The Evolution of Sex Chromosomes and Dosage Compensation in Plants. *Genome Biology and Evolution* **9** :627–645.
- Muyle, A., N. Zemp, C. Fruchard, et al. 2018, Genomic imprinting mediates dosage compensation in a young plant xy system. *Nature plants* **4** :677–680.
- Müller, N. A., B. Kersten, A. P. Leite Montalvão, et al. 2020, A single gene underlies the dynamic evolution of poplar sex determination. *Nature Plants* **6** :630–637.
- Neve, R. 1961, Sex determination in the cultivated hop, *Humulus lupulus*. Ph.D. thesis, University of London (Wye College).

- Nicolas, M., G. Marais, V. Hykelova, B. Janousek, V. Laporte, B. Vyskot, D. Mouchiroud, I. Negrutiu, D. Charlesworth, et F. Monéger. 2005, A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *PLoS biology* **3** :e4.
- Odierna, G., T. Caprigilone, L. A. Kupriyanova, et E. Olmo. 1993, Further data on sex chromosomes of Lacertidae and a hypothesis on their evolutionary trend. *Amphibia-Reptilia* **14** :1–11.
- Offord, C. 2018, Uk to legalize medicinal cannabis. *The Scientist* .
- Ohno, S. 1969, Evolution of Sex Chromosomes in Mammals. *Annual Review of Genetics* **3** :495–524.
- OKADA, Y. et K. ITO. 2001, Cloning and Analysis of Valerophenone Synthase Gene Expressed Specifically in Lupulin Gland of Hop (*Humulus lupulus* L.). *Bioscience, Biotechnology, and Biochemistry* **65** :150–155.
- Orr, H. A. et Y. Kim. 1998, An Adaptive Hypothesis for the Evolution of the Y Chromosome. *Genetics* **150** :1693–1698.
- Otto, S. 2009, The Evolutionary Enigma of Sex. *The American Naturalist* **174** :S1–S14.
- Otto, S. P. et T. Lenormand. 2002, Resolving the paradox of sex and recombination. *Nature Reviews Genetics* **3** :252–261.
- Palmer, D. H., T. F. Rogers, R. Dean, et A. E. Wright. 2019, How to identify sex chromosomes and their turnover. *Molecular Ecology* **28** :4709–4724.
- Pandey, R. S., M. A. Wilson Sayres, et R. K. Azad. 2013, Detecting Evolutionary Strata on the Human X Chromosome in the Absence of Gametologous Y-Linked Sequences. *Genome Biology and Evolution* **5** :1863–1871.
- Pannell, J. R. 2017, Plant Sex Determination. *Current Biology* **27** :R191–R197.
- Papadopulos, A. S. T., M. Chester, K. Ridout, et D. A. Filatov. 2015, Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proceedings of the National Academy of Sciences* **112** :13021–13026.
- Patzak, J., V. Nesvadba, P. Vejl, et S. Skupinova. 2002, Identification of sex in F1 progenies of hop (*Humulus lupulus*) by molecular marker. *Rostlinna Vyroba - UZPI (Czech Republic)* .
- Perrin, N. 2021, Sex-chromosome evolution in frogs : what role for sex-antagonistic genes? *Philosophical Transactions of the Royal Society B : Biological Sciences* **376** :20200094.
- Pisanti, S. et M. Bifulco. 2019, Medical Cannabis : A plurimillennial history of an evergreen. *Journal of Cellular Physiology* **234** :8342–8351.

- Ponnikas, S., H. Sigeman, J. K. Abbott, et B. Hansson. 2018, Why Do Sex Chromosomes Stop Recombining? *Trends in genetics* : TIG **34** :492–503.
- Priore, L. d. et M. I. Pigozzi. 2017, Broad-scale recombination pattern in the primitive bird *Rhea americana* (Ratites, Palaeognathae). *PLOS ONE* **12** :e0187549.
- Razumova, O. V., O. S. Alexandrov, M. G. Divashuk, T. I. Sukhorada, et G. I. Karlov. 2016, Molecular cytogenetic analysis of monoecious hemp (*Cannabis sativa* L.) cultivars reveals its karyotype variations and sex chromosomes constitution. *Protoplasma* **253** :895–901.
- Renner, S. S. 2014, The relative and absolute frequencies of angiosperm sexual systems : Dioecy, monoecy, gynodioecy, and an updated online database. *American Journal of Botany* **101** :1588–1596.
- Renner, S. S. et N. A. Müller. 2021, Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. *Nature Plants* **7** :392–402.
- Renner, S. S. et R. E. Ricklefs. 1995, Dioecy and its correlates in the flowering plants. *American Journal of Botany* **82** :596–606.
- Rifkin, J. L., F. E. G. Beaudry, Z. Humphries, B. I. Choudhury, S. C. H. Barrett, et S. I. Wright. 2021, Widespread Recombination Suppression Facilitates Plant Sex Chromosome Evolution. *Molecular Biology and Evolution* **38** :1018–1030.
- Ross, M. T., D. V. Grafham, A. J. Coffey, et al. 2005, The dna sequence of the human x chromosome. *Nature* **434** :325–337.
- Sabath, N., E. E. Goldberg, L. Glick, M. Einhorn, T.-L. Ashman, R. Ming, S. P. Otto, J. C. Vamosi, et I. Mayrose. 2016, Dioecy does not consistently accelerate or slow lineage diversification across multiple genera of angiosperms. *New Phytologist* **209** :1290–1300.
- Sakamoto, K., K. Shimomura, Y. Komeda, H. Kamada, et S. Satoh. 1995, A Male-Associated DNA Sequence in a Dioecious Plant, *Cannabis sativa* L. *Plant and Cell Physiology* **36** :1549–1554.
- Schilling, S., C. A. Dowling, J. Shi, L. Ryan, D. Hunt, E. O'Reilly, A. S. Perry, O. Kinnane, P. F. McCabe, et R. Melzer. 2020, The cream of the crop : Biology, breeding and applications of *cannabis sativa*. *Authorea Preprints* .
- Schultz, S. T. 1994, Nucleo-Cytoplasmic Male Sterility and Alternative Routes to Dioecy. *Evolution* **48** :1933–1945.
- Shearn, R., A. E. Wright, S. Mousset, C. Régis, S. Penel, J.-F. Lemaitre, G. Douay, B. Crouau-Roy, E. Lecompte, et G. A. Marais. 2020, Evolutionary stasis of the pseudoautosomal boundary in strepsirrhine primates. *eLife* **9** :e63650.
- Shephard, H. L., J. S. Parker, P. Darby, et C. C. Ainsworth. 2000, Sexual development and sex chromosomes in hop. *New Phytologist* **148** :397–411.

- Sigeman, H., S. Ponnikas, E. Videvall, H. Zhang, P. Chauhan, S. Naurin, et B. Hansson. 2018, Insights into avian incomplete dosage compensation : sex-biased gene expression coevolves with sex chromosome degeneration in the common whitethroat. *Genes* **9** :373.
- Skof, S., A. Cerenak, J. Jakse, B. Bohanec, et B. Javornik. 2012, Ploidy and sex expression in monoecious hop (*Humulus lupulus*). *Botany* **90** :617–626.
- Slancarova, V., J. Zdanska, B. Janousek, et al. 2013, Evolution of Sex Determination Systems with Heterogametic Males and Females in *Silene*. *Evolution* **67** :3669–3677.
- Small, E. 2015, Evolution and Classification of *Cannabis sativa* (Marijuana, Hemp) in Relation to Human Utilization. *The Botanical Review* **81** :189–294.
- Smith, T. C., P. F. Arndt, et A. Eyre-Walker. 2018, Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS genetics* **14** :e1007254.
- Stapley, J., P. G. D. Feulner, S. E. Johnston, A. W. Santure, et C. M. Smadja. 2017, Variation in recombination frequency and distribution across eukaryotes : patterns and processes. *Philosophical Transactions of the Royal Society B : Biological Sciences* **372** :20160455.
- Tenessen, J. A., N. Wei, S. C. K. Straub, R. Govindarajulu, A. Liston, et T.-L. Ashman. 2018, Repeated translocation of a gene cassette drives sex-chromosome turnover in strawberries. *PLOS Biology* **16** :e2006062.
- Thomas, G. G. et R. A. Neve. 1976, Studies on the Effect of Pollination on the Yield and Resin Content of Hops (*humulus Lupulus* L.). *Journal of the Institute of Brewing* **82** :41–45.
- Vamosi, J. C. et S. M. Vamosi. 2004, The Role of Diversification in Causing the Correlates of Dioecy. *Evolution* **58** :723–731.
- Veltsos, P., G. Cossard, E. Beaudoin, G. Beydon, D. Savova Bianchi, C. Roux, S. C. González-Martínez, et J. R. Pannell. 2018, Size and Content of the Sex-Determining Region of the Y Chromosome in Dioecious *Mercurialis annua*, a Plant with Homomorphic Sex Chromosomes. *Genes* **9** :277.
- Vicoso, B. 2019, Molecular and evolutionary dynamics of animal sex-chromosome turnover. *Nature Ecology & Evolution* **3** :1632–1641.
- Wang, J., J.-K. Na, Q. Yu, et al. 2012, Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proceedings of the National Academy of Sciences* **109** :13710–13715.
- Waters, P. D., B. Duffy, C. J. Frost, M. L. Delbridge, et J. a. M. Graves. 2001, The human Y chromosome derives largely from a single autosomal region added to the sex chromosomes 80–130 million years ago. *Cytogenetic and Genome Research* **92** :74–79.

- West, N. W. et E. M. Golenberg. 2018, Gender-specific expression of GIBBERELLIC ACID INSENSITIVE is critical for unisexual organ initiation in dioecious *Spinacia oleracea*. *New Phytologist* **217** :1322–1334.
- Westergaard, M. 1958, The Mechanism of Sex Determination in Dioecious Flowering Plants. *In* M. Demerec, ed., *Advances in Genetics*, vol. 9, 217–281, Academic Press.
- Wright, A. E., I. Darolti, N. I. Bloch, V. Oostra, B. Sandkam, S. D. Buechel, N. Kolm, F. Breden, B. Vicoso, et J. E. Mank. 2017, Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. *Nature Communications* **8** :14251.
- Wright, A. E., I. Darolti, N. I. Bloch, V. Oostra, B. A. Sandkam, S. D. Buechel, N. Kolm, F. Breden, B. Vicoso, et J. E. Mank. 2019, On the power to detect rare recombination events. *Proceedings of the National Academy of Sciences* **116** :12607–12608.
- Wright, A. E., R. Dean, F. Zimmer, et J. E. Mank. 2016, How to make a sex chromosome. *Nature Communications* **7** :12087.
- Wu, M. et R. C. Moore. 2015, The Evolutionary Tempo of Sex Chromosome Degradation in *Carica papaya*. *Journal of Molecular Evolution* **80** :265–277.
- Yang, M.-Q., R. van Velzen, F. T. Bakker, A. Sattarian, D.-Z. Li, et T.-S. Yi. 2013, Molecular phylogenetics and character evolution of Cannabaceae. *TAXON* **62** :473–485.
- Yeager, A. 2018, Canada could come to the fore in cannabis research. *The Scientist* .
- Zelkowski, M., M. A. Olson, M. Wang, et W. Pawlowski. 2019, Diversity and Determinants of Meiotic Recombination Landscapes. *Trends in Genetics* **35** :359–370.
- Zhang, H., J. Jin, M. J. Moore, T. Yi, et D. Li. 2018, Plastome characteristics of Cannabaceae. *Plant Diversity* **40** :127–137.
- Zhou, Q. et D. Bachtrog. 2012, Chromosome-wide gene silencing initiates y degeneration in *Drosophila*. *Current Biology* **22** :522–525.





# 2

## Chapter 2 : Are sex chromosomes emerging in *Silene acaulis* ssp *exscapa* ?

En 2014, Paul Jay a effectué son stage de Master 1 sous l'encadrement de Aline Muyle et Gabriel Marais. L'objectif de ce stage était d'identifier des gènes liés au sexe chez *Silene acaulis* ssp *exscapa* à partir de données RNA-seq issues d'une population séquencée au cours de l'été 2013 en utilisant une méthode empirique développée précédemment (Muyle et al. 2012). À la fin de ce stage, les résultats suggéraient la présence d'un système XY chez *S. acaulis* ssp *exscapa* mais demandaient à être confirmés.

Depuis la fin de ce stage, SDpop a été développé, ce qui offre un cadre statistique plus robuste que la méthode employée dans le stage de Paul Jay. Nous avons donc décidé de tester à nouveau la présence de chromosomes sexuels chez *S. acaulis* ssp *exscapa* avec SDpop en ajoutant un second jeu de données provenant d'une autre population.

Ces travaux ont été compilés sous la forme d'un article scientifique qui constitue le deuxième chapitre de ma thèse. Les résultats obtenus ne permettent pas de répondre avec certitude à la question pour le moment et n'ont donc pas encore été soumis à un journal scientifique. Ce chapitre n'a donc pas été évalué par les pairs.

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., & Marais, G. A. (2012). Rapid de novo evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *PLoS biology*, 10(4), e1001308

# Are sex chromosomes emerging in *Silene acaulis* ssp *exscapa*?

Prentout D<sup>1</sup>, Jay P<sup>1,2</sup>, Muyle A<sup>1</sup>, Marais G.A.B<sup>1,3,\*</sup>, and Käfer J<sup>1,\*</sup>

<sup>1</sup>Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Université de Lyon, Université Lyon 1, CNRS, Villeurbanne F-69622, France

<sup>2</sup>Current adress: Ecologie Systématique Evolution, Bâtiment 360, CNRS, AgroParisTech, Université Paris-Saclay, 91400 Orsay, France

<sup>3</sup>Current adress: CIBIO, Biopolis, Campus de Vairao, Univ. Porto, Portugal

\*Equal contributions

## Abstract

The *Silene* genus is a model to study the evolution of sexual systems in plants since several independent transitions to dioecy have been reported. Among this genus, *S. acaulis* displays a unique sexual system polymorphism with gynodioecious, trioecious and dioecious subspecies. This polymorphism suggests a transition to dioecy through the gynodioecious pathway that may be very recent in this species. The sex determination mechanism remains unknown in *S. acaulis* but a pair of sex chromosomes has been identified in all other studied *Silene* dioecious species. Thus, if the determination of the sex is also genetic in *S. acaulis*, this model could be insightful for understanding the first steps of sex chromosome evolution. Here, we tested whether the dioecious subspecies of *S. acaulis*, *S. acaulis* ssp *exscapa*, has sex chromosomes or not. We used a new probabilistic tool, called SDpop, that analyses allele and genotype frequencies in males and females in population-based data. This likelihood-based method allows to test the presence of sex chromosomes and to identify sex-linked genes. We combined this method with a comparison of male versus female heterozygosities. We studied two datasets from two French Alps dioecious populations and we found 27 genes that are potentially XY in at least one population. Further analyses are required to clarify whether a non-recombining region exists in this subspecies but our preliminary results suggest that a sex chromosome system may have emerged recently in *S. acaulis*.

**Keywords.** Sex chromosomes, Dioecy, *Silene*, *Silene acaulis*, arrest of recombination.

## Introduction

Among >260,000 angiosperm species, about 15,000 (5-6%) are dioecious, with separate sexes (Renner, 2014). Despite the low proportion of dioecious species, dioecy is widespread among angiosperms. Indeed, this sexual system is present in 7% of angiosperm genera and in 43% of angiosperm families (Renner, 2014). Unlike in animals, the mechanisms of sex determination remain poorly understood in plants (Ming et al., 2011; Pannell, 2017; Fruchard and Marais, 2017). So far, about 30 sex chromosome systems have been studied with genomic data in plants, and only a handful of master sex-determining genes have been identified (reviewed in Carey et al., 2021; Renner and Müller, 2021).

Why sex chromosomes initially stop recombining is not well known (reviewed in Charlesworth, 2021). The arrest of recombination leads to the formation of a Y (or W) specific region and an X (or Z) specific region (Charlesworth et al., 2005). In plants, the formation of this non-recombining region should occur during a transition from hermaphroditism (or monoecy) to dioecy (reviewed in Fruchard and Marais, 2017). The transition to dioecy from hermaphroditism may arise through a gynodioecious stage (*i.e.* population with hermaphrodites and females individuals) (Lewis, 1942; Charlesworth and Charlesworth, 1978; Lloyd, 1980).

Two main models explaining the emergence of sex chromosomes while a transition through gynodioecy

are present in the literature (Charlesworth and Charlesworth, 1978; Schultz, 1994; Maurice et al., 1994; summarized in Fruchard and Marais, 2017). In both models, a male sterility mutation and a female sterility mutation occurred one after another. Charlesworth and Charlesworth (1978) proposed that both mutations arise in the nuclear genome. More precisely, one mutation occurs on a chromosome and the other mutation on the homologous chromosome. Then, an arrest of recombination is selected to avoid the linkage of these mutations on the same haplotype. In the other model, the male sterility mutation occurred in cytoplasmic genome (cytoplasmic male sterility - CMS) (Schultz, 1994; Maurice et al., 1994). Then, a mutation restoring male fertility is expected before the female sterility mutation. These two last mutations occurred in a region of a chromosome and an arrest of recombination is selected to linked them. However, none of these models describing the genetic architecture of the transition to dioecy has been tested with empirical data. Indeed, all sex chromosomes described so far have regions that are already non-recombining, and often since several My (Ming et al., 2011; Charlesworth, 2019, 2021; Renner and Müller, 2021). A ZW system with a 13kb female specific region have been identified in *Fragaria vesca ssp vesca* (Tennessen et al., 2018). Dioecy evolved recently in *Solanum appendiculatum*, and a male heterogamety have been newly identified (Wu et al., 2021). In these species, further analyses seem necessary to conclude on the mechanisms of the formation of the non-recombining region. *Asparagus officinalis* sex chromosome analysis provided informations for the genetic transition to dioecy with the identification of two genes likely responsible of the sex determination in a small non-recombining region (Harkess et al., 2017). However, the mechanisms responsible of the arrest of recombination in this species remain unknown. Finally, recent work on *Spinacia oleracea* showed that sex chromosomes should be less than 1My old (Okazaki et al., 2019), however the non-recombining region already carries more than 200 genes and is rich in transposable elements (Yu et al., 2021). Here again, this non-recombining region evolved for too long ago to help to describe the formation of sex chromosomes. *Silene acaulis* is a long-lived species (may exceed 500 years), present in a large part of the northern hemisphere, that lives in arctic-alpine environments (Morris and Doak, 1998; Gussarova et al., 2015) and has recently been classified within the Siphonomorpha section of *Silene* genus (Naciri et al., 2017; Jafari et al., 2020). This species is particular since several subspecies with different sexual system have been described worldwide (Maurice et al., 1998; Gussarova et al., 2015). All subspecies found in North America are gynodioecious (summarized in Gussarova et al., 2015). The subspecies *Silene acaulis ssp exscapa* mostly described in French Alps is dioecious (Maurice et al., 1998; Gussarova et al., 2015), although

a study reported gynodioecious populations of this subspecies in North America (Hermanutz and Innes, 1994). Otherwise, Maurice et al. (1998) suggested that *Silene acaulis ssp cenisia* may be a true trioecious subspecies. Desfeux et al. (1996) reported that dioecy evolved at least twice in *Silene* genus. However, the sample size was small and following works showed that two independent transitions likely to have happened for Otites section and *S. acaulis*. Indeed, species close to *S. acaulis* are mostly gynodioecious and gynodioecy seems to be the ancestral sexual system of the Otites section (Slancarova et al., 2013). These transitions to dioecy make this genus a model for studying the evolution of sexual systems and sex chromosomes in plants. Moreover, the importance of gynodioecy in this genus suggests that dioecy evolved through gynodioecy (Fruchard and Marais, 2017). In *S. acaulis* complex, most populations are gynodioecious and closely relative species are also gynodioecious. This suggests that gynodioecy was the ancestral sexual system of *S. acaulis* and, thus, dioecy should be recent in *S. acaulis ssp exscapa*. The mechanism regulating the dioecy should have evolved recently in this subspecies. The sex-ratio reported for *S. acaulis ssp exscapa* in French Alps is generally 1:1, which is expected for dioecious species with a genetic sex determination (Maurice et al., 1998). All dioecious species that belong to Otites and Melandrium sections have sex chromosomes (Marais et al., 2011; Martin et al., 2019; Balounova et al., 2019). This suggests that sex may be determined by a pair of sex chromosomes in *S. acaulis ssp exscapa*. Otherwise, nuclear-cytoplasmic sex determination have been identified in several *Silene* species (Hermanutz and Innes, 1994; Garraud et al., 2011; Sloan et al., 2012), including gynodioecious populations of *S. acaulis* (Hermanutz and Innes, 1994). Thus, the emergence of a pair of sex chromosomes with a CMS factor is conceivable in *S. acaulis ssp exscapa*. This subspecies, therefore, could be an insightful model to understand the genetic transition from hermaphroditism to dioecy through the gynodioecious pathway, thus the formation of sex chromosomes (Charlesworth and Charlesworth, 1978; Schultz, 1994; Fruchard and Marais, 2017). If a non-recombining region exists in the dioecious subspecies, it might be really recent since it has been estimated that the maximal divergence between different *S. acaulis* subspecies is lower than one My (Gussarova et al., 2015). However, it is worth noting that divergence may be more important since dioecious populations present in French alps are not studied in Gussarova et al. (2015) The recently published probabilistic tool SDpop uses allele and genotype frequencies of males and females from natural populations to identify sex-linked sequences (Käfer et al., 2021b). SDpop is a probabilistic framework that uses population-based data to assign a probability for each gene of being sex-linked. Simulations and tests with five to ten individuals of each sex showed that

SDpop should be able to identify the sex chromosome system (XY or ZW) and sex-linked sequences (Käfer et al., 2021a,b).

Here, we aimed to test whether a non-recombining region on a pair of sex chromosomes exists in *S. acaulis* ssp *exscapa*. A very young sex-linked region may allow the identification of first event leading to an arrest of recombination. We sequenced males and females transcriptomes of two distinct dioecious populations from the French Alps to identify sex-linked genes. Using SDpop and a heterozygosity analysis we identified a list of 27 genes that may be in, or close to, the non-recombining region of a pair of sex chromosomes.

## Materials and methods

### Sampling

Two populations have been sampled, as indicated in Table 1. The first one in 2013 at the “Col du Lautaret” (“massif des Écrins”, between 2000m and 2100m above sea level, population studied in Maurice et al. (1998)). The second one in 2018 at “Rocher Blanc” (“massif d’Allevard”, between 2800m and 2900m above sea level). The distance as the crow flies separating both sampling locations is about 30km.

According to tissues available on the individuals during sampling, flowers or flower buds were picked. The number of sampled individuals for each sex and the collected tissues are compiled in Table 1. Both flowers and flower buds were sampled for two individuals from the population sampled in 2018 in order to have sufficient quantity of RNA (Table 1).

Table 1: Sampling informations for the two populations.

Sampling date	July 2013	August 10th 2018
# of males (total)	15	8
# of males (flower buds)	14	7
# of males (flower)	1	1
# of males (both tissues)	0	0
# of females (total)	15	11
# of females (flower buds)	3	8
# of females (flower)	12	1
# of females (both tissues)	0	2

### RNA extraction and sequencing

#### Data sampled in 2013

All the RNA preparation was done at the AGAP laboratory (<http://umr-agap.cirad.fr/>). The total RNA was extracted through the Spectrum Plant Total RNA kit (Sigma, Inc., USA) following the manufacturer’s protocol and treated with a DNase. Libraries were prepared

with the TruSeq RNA sample Preparation v2 kit (Illumina Inc., USA). Each 2 nM cDNA library was sequenced using a paired-end protocol on a HiSeq2000 sequencer, producing 100 bp reads. Demultiplexing was performed using CASAVA 1.8.1 (Illumina) to produce paired sequence files containing reads for each sample in Illumina FASTQ format.

#### Data sampled in 2018

The RNA preparation was done by the AGAP laboratory for this sampling as well. For all individuals, a part of the tissues was sent freshly, and another was dried on the field for a zeolite protocol. Fresh samples were frozen with the flash freezing methods (Freshfreeze). Then, the RNA was extracted following the SIGMA protocol. Since the RNA qualities were similar between both protocols, we sequenced all fresh samples and three zeolites samples (two males and one female). For the sequencing, we used the HiSeq3000-HWI-J00115 technology, producing 150bp paired reads.

#### De novo transcriptome assemblies and published transcriptome assemblies

A total of ten assemblies was used for this analysis, of which five were *de novo* assembled. Among the five *de novo* assemblies, four were performed by the use of DRAP pipeline with default parameters and RNA-seq data from the population sampled in 2018 (Cabau et al., 2017). We firstly assembled 13 transcriptomes (five with male RNA-seq data and eight with female RNA-seq data). DRAP allows to merge and compact different assemblies into a “meta-assembly”. Thus, we made the meta-assembly (1) of a male sampled twice (Zeolite + Freshfreeze), (2) of a female sampled twice (Zeolite + Freshfreeze), (3) of five males and five females, and (4) of eight females.

We also assembled a transcriptome with data sampled in 2013. All reads from three females and five males flower bud samples were assembled with Trinity with default paired-end options (Haas et al., 2013). Poly-A tails longer than five bp were trimmed with PRINSEQ (Schmieder and Edwards, 2011) and the ribosomal RNA through Ribopicker (Schmieder et al., 2012) with the SILVA Small Subunit RNA database (SSR version 21/09/2013) and the SILVA Large Subunit non-redundant truncated database (LSR version 07/10/2013) (Quast et al., 2013). ORFs were predicted with a Trinity tool, Transdecoder (Haas et al., 2013), and CAP3 was used to group together the similar contigs due to polymorphism (Huang and Madan, 1999) (parameter -p 90). This stage was important in order to have X and Y or Z and W sequences assembled in a single contig.

Otherwise, we retrieved five assemblies already assem-

bled in the literature: *Silene otites* transcriptome assembly (Martin et al., 2019); *Silene pseudotites* transcriptome assembly (Martin et al., 2019), *Silene schafta* transcriptome assembly (Bertrand et al., 2018), *Beta vulgaris* transcriptome assembly (Dohm et al., 2014) and *Chenopodium quinoa* transcriptome assembly (Zou et al., 2017).

## Mapping & Genotyping

We aligned the RNA-seq with BWA (version: 0.7.15-r1140; (Li and Durbin, 2009)). Five mismatches were authorized for the mapping on the four assemblies obtained with DRAP, 15 mismatches for the mapping on the 3 other *Silene* species assemblies and the assembly with data sampled in 2013, and 25 for mapping on *Beta vulgaris* and *Chenopodium quinoa* assemblies. Then, with samtools we only conserved mapped reads and sorted the output mapping files for the genotyping (version 1.9; (Li et al., 2009)). The genotyping was done with reads2snp (version: 2.0; (Gayral et al., 2013)) and only positions supported by at least 10 reads were conserved. The allelic expression biases were taken into account, and paralogous positions were kept.

## Detection of sex linkage

To determine the sex chromosome system and detect sex-linked contigs, we used a probabilistic framework, called SDpop. Four sex chromosome models have been implemented in SDpop (1) XY sex chromosome system; (2) ZW sex chromosome system; (3) absence of sex chromosomes; (4) both type of sex chromosomes. The Bayesian Information Criterion (BIC) is computed for each model, which allows to determine which sex chromosome system is the most likely.

Sdpop estimates a probability for each contigs of being (1) autosomal; (2) haploid; (3) paralogous; (4) hemizygous; or (5) gametologous (Käfer et al., 2021b). For each segregation type, different equilibrium between allele frequencies and genotype frequencies are expected. For example, for autosomal genes (or SNPs), a Hardy-Weinberg equilibrium without differences between males and females is expected. We fixed a no prior probability threshold at 0.8 to determine the segregation of the contig as recommended for a SDpop analysis (Käfer et al., 2021b). Contigs were unclassified if none of the classification probabilities was greater than 0.8

We analyzed both populations independently, as suggested in Käfer et al. (2021b). For each mapping on the ten assemblies, we run (1) the model without sex chromosomes, (2) the XY sex chromosome model and (3) the ZW sex chromosome model. For the population sequenced in 2018, we run SDpop on genotyping data from a subset of females. Indeed, to have the same number of males and females in the analysis we randomly selected 8 females.

The heterozygosity for each contig of each individual was computed with a home-made program that computes the proportion of heterozygous sites among all genotyped sites. Then, we computed a Wilcoxon test (two.sided test) to compare male heterozygosity versus female heterozygosity for each gene. Given the large number of genes tested ( $> 17,000$ ), a  $p$ -value fixed at 0.05 should provide several hundreds of false positives under the null hypothesis ( $0.05 \times 17,000 = 850$ ). Thus, we selected genes with a  $p$ -value lower than 0.005 and reported the exact  $p$ -values. We didn't make a multiple test correction (*e.g.* Bonferroni) to keep a quantitative approach instead of a binary qualitative approach with an arbitrary threshold.

Statistics and graphical representation were done with R (version 4.10; R core Team, 2021).

## Results

### Assemblies statistics

Table S1 gathers the number of contigs in each assembly and the proportion of RNA-seq that mapped on them for both populations. Table S2 shows some assembly statistics for references assembled *de novo* with DRAP (Cabau et al., 2017). As we can see in Table S2, all *de novo* assemblies provided similar results for the BUSCO scores. However, the transcriptome assembly from a single male has the highest transrate score.

### SDpop outputs

We compared the Bayesian Information Criterion (BIC) of the three SDpop models, as indicated in Table 2 and in Table 3, for population sampled in 2013 and 2018, respectively. Among a total of 20 cases (ten assemblies and two populations), 17 indicate that an XY sex chromosome system is more likely than a ZW system or an absence of sex chromosomes. The three others cases indicate that a ZW sex chromosome system is the most likely.

Table 2: Bayesian Information Criterion (BIC) with data sampled in 2013 for 3 models of SDpop. In bold, the lowest BIC values, thus, the highest likelihood scores.

Reference transcriptome	2013		
	No sex chromosomes	XY sex chromosomes	ZW sex chromosomes
Meta-assembly (5 males 5 females)	37520799.010145	<b>37520047.429903</b>	37520134.075563
Meta-assembly (8 females)	37553314.308531	<b>37552618.942646</b>	37552706.569836
Meta-assembly (1 male)	21337641.875933	<b>21336607.667168</b>	21337634.834565
Meta-assembly (1 female)	21246580.553666	<b>21246182.092990</b>	21246600.285932
Assembly with 2013 data	36329830.714485	<b>36325225.956627</b>	36329155.586754
<i>Silene otites</i>	30585496.673568	<b>30579553.981373</b>	30584906.395290
<i>Silene pseudotites</i>	30193164.270942	<b>30188589.541943</b>	30192639.162986
<i>Silene schafta</i>	16489364.428070	<b>16479965.815826</b>	16488688.679298
<i>Beta vulgaris</i>	3683690.070789	<b>3675887.854363</b>	3683382.416713
<i>Chenopodium quinoa</i>	4667737.722044	<b>4658504.849482</b>	4667200.976898

Table 3: Bayesian Information Criterion (BIC) with data sampled in 2018 for 3 models of SDpop. In bold, the lowest BIC values, thus, the highest likelihood scores.

Reference transcriptome	2018		
	No sex chromosomes	XY sex chromosomes	ZW sex chromosomes
Meta-assembly (5 males 5 females)	21745230.392795	21745272.067670	<b>21744516.027270</b>
Meta-assembly (8 females)	32219211.873911	32219158.789205	<b>32217263.606859</b>
Meta-assembly (1 male)	14210350.142996	<b>14210007.860333</b>	14210206.222230
Meta-assembly (1 female)	14193776.109634	14193815.807236	<b>14193599.097217</b>
Assembly with 2013 data	27650769.672496	<b>27645454.952283</b>	27650808.415108
<i>Silene otites</i>	21304841.200189	<b>21301414.595068</b>	21304448.861405
<i>Silene pseudotites</i>	20739016.026881	<b>20736965.080254</b>	20738967.438153
<i>Silene schafta</i>	8002648.845329	<b>7998138.224306</b>	8002296.557424
<i>Beta vulgaris</i>	443309.844118	<b>442349.467910</b>	443278.078326
<i>Chenopodium quinoa</i>	550507.916204	<b>550347.558061</b>	550481.307867

Table 4: Numbers of contigs and SNPs for which SDpop estimated of probability greater to 0.8 of being XY or ZW. These results were obtained with data from the both populations.

Reference transcriptome	# XY contigs		# XY SNPs		# ZW contigs		# ZW SNPs	
	2013	2018	2013	2018	2013	2018	2013	2018
Meta-assembly (5 males 5 females)	12	0	85	0	14	0	0	0
Meta-assembly (8 females)	21	13	66	11	9	5	2	0
Meta-assembly (1 male)	10	6	204	54	11	15	9	0
Meta-assembly (1 female)	9	8	76	0	0	8	0	7
Assembly with 2013 data	70	81	801	667	47	37	111	0
<i>Silene otites</i>	33	30	933	479	30	26	95	43
<i>Silene pseudotites</i>	24	32	755	360	21	23	83	13
<i>Silene schafta</i>	28	20	1292	608	11	7	124	66
<i>Beta vulgaris</i>	37	8	1275	116	10	3	57	12
<i>Chenopodium quinoa</i>	57	11	1539	26	27	12	84	12

We also looked at the proportions of each segregation type, as reported in Table S3 and Table S4 for populations sampled in 2013 and 2018, respectively. The Tables S3 and S4 show very low proportions of sex-linked sequences across all runs. The proportions of hemizygous sequences are mostly equal to zero and range between 0.1% and 0.6% in six runs (with a ZW system, every time). An absence of gametologous sequences is reported for 3 runs only, and the proportion of XY se-

quences is generally greater than the proportion of ZW sequences (except in two cases).

Otherwise, the proportion of paralogous sequences seems positively correlated to the phylogenetic distance between *S. acaulis* and the species used as reference for the mapping. Indeed, *Silene otites* and *Silene pseudotites* are closer to *S. acaulis* than *Silene schafta*, and the proportion of paralogous sequences is lower with a mapping on references from the two former species than



from *S. schafta*. For non-Silene species (*i.e.* *Beta vulgaris* and *Chenopodium quinoa*), the proportion of paralogous sequences ranged between 49.4% and 63.9%.

We classified a contig (or a SNP) as gametologous if its probability associated to the segregation type is greater than 0.8. The results for XY and ZW systems for both populations are compiled in Table 4. Here again, there seems to be a positive correlation between the number of SNP detected as gametologous and the phylogenetic distance between *S. acaulis* and the species used as the reference. SDpop detected more gametologous contigs and gametologous SNPs with a mapping on a reference from a distant species than references from *S. acaulis*. This could be explained by independent assemblies of X (or Z) and Y (or W) sequences with RNA-seq data from *S. acaulis*. However, this should not be the case in references assembled with data from homogametic sex or non-dioecious individuals. Otherwise, Table 4 shows that for several runs with a mapping on a *S. acaulis* reference, SDpop detected contigs with a probability of being ZW greater than 0.8 but didn't identify SNP in this condition. This suggests these ZW sequences may be false positives.

## Choice of the assembly

For the rest of the results section, we decided to report the results obtained with a mapping on *S. schafta* transcriptome (a gynodioecious species). The choice of this assembly is based on several criteria (1) number of contigs in the assembly, (2) quality of the RNA-seq mapping, (3) proportion of sex-linked contigs estimated by SDpop, (4) absolute number of XY (or ZW) and X (or Z)-hemizygous contigs. The mapping on *S. schafta* transcriptome assembly fulfills all these criteria in a balanced way (Table S1 - Table S4). For example, despite a high number of SNPs classified as XY with a mapping on *Chenopodium quinoa* and *Beta vulgaris* transcriptome assemblies, the low mapping rate or the proportion of paralogous sequences avoid the use of these assemblies for further analysis. Otherwise, the mapping qualities were better with a mapping on a *de novo* assembly of *S. acaulis* but the number of sex-linked SNPs (and also contigs) was low. In order to test whether the contigs identified as sex-linked by SDpop are false positives we decided to choose an assembly with a number of sex-linked genes as high as possible. Otherwise, an assembly done with heterogametic individual could lead to assemble independently gametologous sequences. To ensure proper mapping on a single sequence it is preferable to use homogametic individuals for reference assemblies, which is unknown in *S. acaulis*. Thus, we preferred the mapping on *S. schafta* transcriptome assembly rather than on a *de novo* assembly of *S. acaulis*.

## Male and female heterozygosities

In addition to SDpop analysis, we computed the heterozygosity for all individuals for each gene from a mapping on *S. schafta* reference transcriptome. We averaged heterozygosities in males and in females for each gene. Then, we used a Wilcoxon paired test to identify genes with the largest heterozygosity differences between males and females. As SDpop classifications are not based on SNPs densities, thus, computing the heterozygosity differences between males and females is an alternative analysis that is independent and complementary to SDpop analysis. Figure 1 show the  $p$ -value of the Wilcoxon test as a function of the probability of being gametologous inferred by SDpop. In an XY system the heterozygosity is expected to be greater in males whereas a greater heterozygosity is expected in females for a ZW system. Figures 1 **A** and 1 **B** show genes with a heterozygosity greater in males as a function of the probability of being XY for populations sampled in 2013 and 2018, respectively. The Figures 1 **C** and 1 **D** show genes with a heterozygosity greater in females as a function of the probability of being ZW for populations sampled in 2013 and 2018, respectively.

We considered a gene as potentially sex-linked if its probability of being gametologous is greater than 0.8 and if the associated  $p$ -value for the Wilcoxon test is lower than 0.005. In both populations, we found 14 potential XY genes among which only one is detected in both populations. A single ZW gene was found in both populations and these two genes are unique to a population. Thus, a total of 27 XY genes (14 genes in each population including one common to the both) and 2 ZW genes (one in each population) have been identified. The Table 5 reports the exact  $p$ -value and XY probability for the 27 potential XY genes. Other parameters obtained from SDpop are also reported in this table: (1) the number of polymorphic sites between X and Y sequences, (2) the number of fixed sites between X and Y sequences, (3) the sequence length, (4) the diversity of X and Y sequences ( $\pi$ ), (5) the divergence between X and Y sequences (Käfer et al., 2021b). We can see that populations not only differ from the set of genes identified as sex-linked, but also from several SDpop parameters. For example, except for two genes, Table 5 shows that a gene with fixed mutations in a population has no fixed mutations in the other. Also, about 15 genes (among the 27 putative sex-linked genes) have an X-Y divergence in a population that is at least twice as high as in the other. Finally, four genes have a probability of being XY greater than 0.8 in a population and lower than 0.01 in the other. However, if we fix the minimal probability of being XY at 0.5 instead of 0.8, three more genes fulfill the conditions in both analysis. It is worth noting that with a probability greater than 0.5 the XY segregation type remains the classification with greatest probability.

ity. The highest X-Y divergence reaches 8.1% and 8.3% for the population sampled in 2013 and in 2018, respectively. As a comparison, the mean heterozygosity is equal to 1.78% and 2.26% in population sampled in 2013 and 2018, respectively. We used these mean heterozygosities to estimate the polymorphism of *S. acaulis* ssp *exscapa* ( $\pi = \frac{1.78+2.26}{2} = 2.02$ ). Considering these highest divergence and the polymorphism in both populations the X and Y divergence time could be around  $6N_e$  generations:

$$D = \pi(4N_e 2t + 1) \quad (1)$$

$$\Leftrightarrow t = 4N_e \frac{D - 1}{2\pi} \quad (2)$$

So,  $t \approx 6N_e$  generations, with  $D \approx 0.08$  and  $\pi \approx 0.02$ . If we fix the minimal probability of being XY at 0.5 instead of 0.8, three more genes fulfill the conditions in both analysis. It is worth noting that with a probability greater than 0.5 the XY segregation type remains the classification with greatest probability. We identified the predicted function of these genes with blast in NCBI database. However, none of these functional annotations correspond to a function previously linked to the determination of the sex ((1) TBC1 domain family member 15; (2) pyruvate dehydrogenase E1alpha subunit; (3) alaline/glyoxylate aminotransferase; (4) ecosyst complex component SEC3A-like).

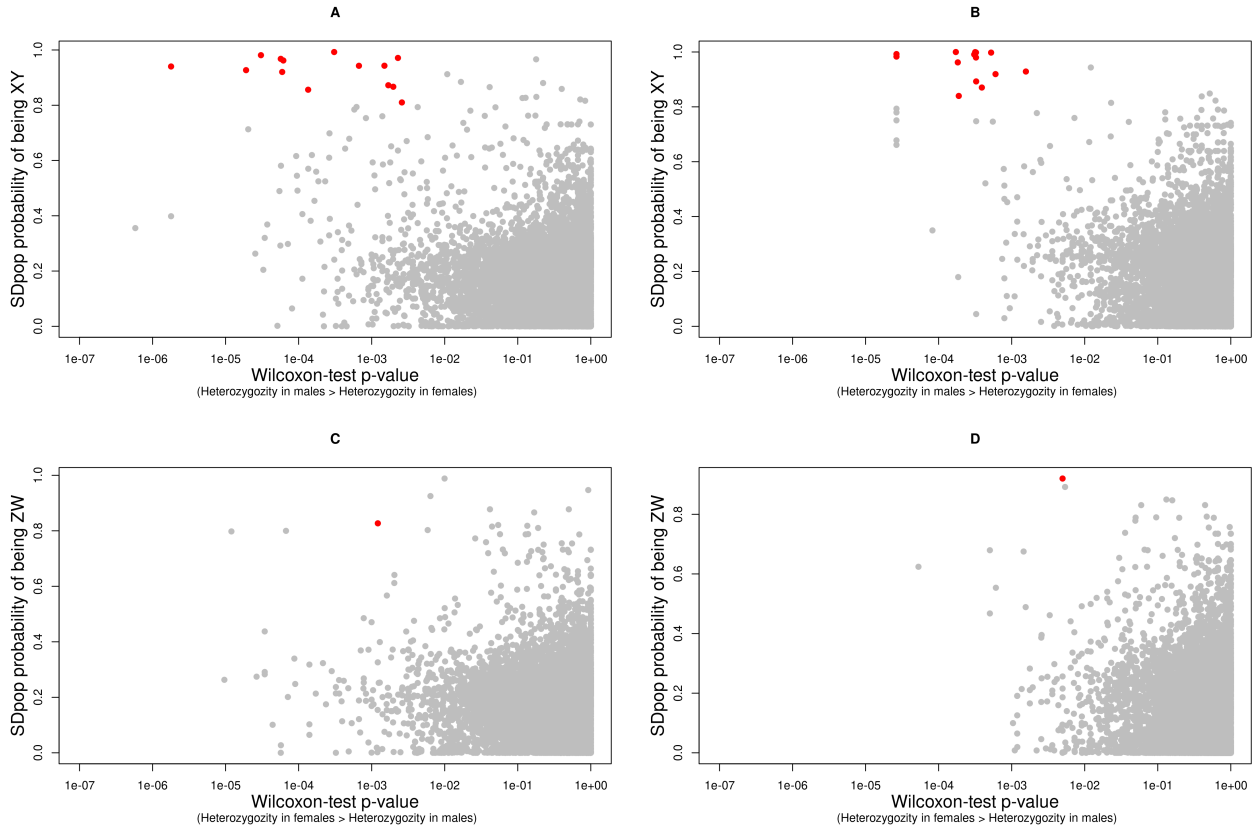


Figure 1: SDpop probabilities and heterozygosity differences for both populations under both SDpop sex chromosome models. The X axis is, at the logarithmic scale, the  $p$ -value of the Wilcoxon test (with a greater heterozygosity in females for ZW model and a greater heterozygosity in males for XY models) and the Y axis represents the SDpop probability of being gametologous. Red dots are genes with a  $p$ -value  $< 0.005$  and a probability  $> 0.8$ . The results presented in this figure were obtained with data from the population sampled in **A**) 2013 and analyzed under XY sex chromosome model, **B**) 2018 and analyzed under XY sex chromosome model, **C**) 2013 and analyzed under ZW sex chromosome model, **D**) 2018 and analyzed under ZW sex chromosome model. The data were mapped on the *Silene schafta* reference transcriptome.

Table 5: Informations of the 27 potential sex-linked genes. In bold, the  $p$ -values  $< 0.005$  and the XY probabilities  $> 0.8$ . Underlined, the XY probabilities  $> 0.5$ .

	$p$ -value		XY probability		# polymorphic sites		# fixed sites		# sites		$\pi$		divergence	
	2013	2018	2013	2018	2013	2018	2013	2018	2013	2018	2013	2018	2013	2018
	>SeqIndex111_X	$3.7 \times 10^{-2}$	<b><math>2.6 \times 10^{-6}</math></b>	$2.1 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	252	201	0	79	1986	1865	0.016082	0.012460	0.015584
>SeqIndex111_Y	$3.7 \times 10^{-2}$	<b><math>2.6 \times 10^{-6}</math></b>	$2.1 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	252	201	0	79	1986	1865	0.006930	0.009279	0.015584	0.069662
>SeqIndex11318_X	NA	<b><math>1.8 \times 10^{-4}</math></b>	NA	<b><math>8.3 \times 10^{-1}</math></b>	70	12	0	2	550	552	0.049332	0.000958	0.050846	0.006892
>SeqIndex11318_Y	NA	<b><math>1.8 \times 10^{-4}</math></b>	NA	<b><math>8.3 \times 10^{-1}</math></b>	70	12	0	2	550	552	0.009839	0.002614	0.050846	0.006892
>SeqIndex11360_X	<b><math>2.6 \times 10^{-3}</math></b>	$1.6 \times 10^{-2}$	<b><math>8.1 \times 10^{-1}</math></b>	$1.3 \times 10^{-1}$	70	58	1	0	486	491	0.034609	0.033550	0.072109	0.040515
>SeqIndex11360_Y	<b><math>2.6 \times 10^{-3}</math></b>	$1.6 \times 10^{-2}$	<b><math>8.1 \times 10^{-1}</math></b>	$1.3 \times 10^{-1}$	70	58	1	0	486	491	0.006436	0.028069	0.072109	0.040515
>SeqIndex11872_X	<b><math>1.7 \times 10^{-3}</math></b>	$7.0 \times 10^{-1}$	<b><math>8.7 \times 10^{-1}</math></b>	$2.5 \times 10^{-1}$	51	58	0	0	475	487	0.004334	0.009576	0.039687	0.012201
>SeqIndex11872_Y	<b><math>1.7 \times 10^{-3}</math></b>	$7.0 \times 10^{-1}$	<b><math>8.7 \times 10^{-1}</math></b>	$2.5 \times 10^{-1}$	51	58	0	0	475	487	0.034642	0.012254	0.039687	0.012201
>SeqIndex11994_X	$1.7 \times 10^{-1}$	<b><math>2.6 \times 10^{-5}</math></b>	<b><math>5.6 \times 10^{-1}</math></b>	<b><math>9.8 \times 10^{-1}</math></b>	231	133	0	44	1939	1931	0.015182	0.007109	0.032248	0.041620
>SeqIndex11994_Y	$1.7 \times 10^{-1}$	<b><math>2.6 \times 10^{-5}</math></b>	<b><math>5.6 \times 10^{-1}</math></b>	<b><math>9.8 \times 10^{-1}</math></b>	231	133	0	44	1939	1931	0.028371	0.009163	0.032248	0.041620
>SeqIndex12137_X	$1.1 \times 10^{-1}$	<b><math>1.8 \times 10^{-4}</math></b>	$2.3 \times 10^{-1}$	<b><math>9.6 \times 10^{-1}</math></b>	116	103	0	38	723	742	0.022033	0.021217	0.044524	0.079631
>SeqIndex12137_Y	$1.1 \times 10^{-1}$	<b><math>1.8 \times 10^{-4}</math></b>	$2.3 \times 10^{-1}$	<b><math>9.6 \times 10^{-1}</math></b>	116	103	0	38	723	742	0.036921	0.007380	0.044524	0.079631
>SeqIndex12541_X	<b><math>5.9 \times 10^{-5}</math></b>	<b><math>1.9 \times 10^{-3}</math></b>	<b><math>9.2 \times 10^{-1}</math></b>	$2.0 \times 10^{-1}$	19	36	4	0	254	252	0.006910	0.032857	0.037765	0.044141
>SeqIndex12541_Y	<b><math>5.9 \times 10^{-5}</math></b>	<b><math>1.9 \times 10^{-3}</math></b>	<b><math>9.2 \times 10^{-1}</math></b>	$2.0 \times 10^{-1}$	19	36	4	0	254	252	0.015561	0.022941	0.037765	0.044141
>SeqIndex13137_X	<b><math>2.0 \times 10^{-3}</math></b>	NA	<b><math>8.7 \times 10^{-1}</math></b>	NA	74	20	15	0	627	625	0.027618	0.012234	0.082938	0.011460
>SeqIndex13137_Y	<b><math>2.0 \times 10^{-3}</math></b>	NA	<b><math>8.7 \times 10^{-1}</math></b>	NA	74	20	15	0	627	625	0.001911	0.000182	0.082938	0.011460
>SeqIndex13151_X	<b><math>1.5 \times 10^{-3}</math></b>	$1.1 \times 10^{-1}$	<b><math>9.4 \times 10^{-1}</math></b>	$2.0 \times 10^{-1}$	78	63	0	0	731	734	0.029385	0.019222	0.070772	0.014037
>SeqIndex13151_Y	<b><math>1.5 \times 10^{-3}</math></b>	$1.1 \times 10^{-1}$	<b><math>9.4 \times 10^{-1}</math></b>	$2.0 \times 10^{-1}$	78	63	0	0	731	734	0.003958	0.002784	0.070772	0.014037
>SeqIndex13850_X	<b><math>2.3 \times 10^{-3}</math></b>	<b><math>2.6 \times 10^{-5}</math></b>	<b><math>9.7 \times 10^{-1}</math></b>	<b><math>6.6 \times 10^{-1}</math></b>	150	149	4	33	1395	1380	0.030232	0.026597	0.071842	0.065562
>SeqIndex13850_Y	<b><math>2.3 \times 10^{-3}</math></b>	<b><math>2.6 \times 10^{-5}</math></b>	<b><math>9.7 \times 10^{-1}</math></b>	<b><math>6.6 \times 10^{-1}</math></b>	150	149	4	33	1395	1380	0.000462	0.021044	0.071842	0.065562
>SeqIndex14059_X	<b><math>1.9 \times 10^{-5}</math></b>	<b><math>3.3 \times 10^{-3}</math></b>	<b><math>9.3 \times 10^{-1}</math></b>	<b><math>6.6 \times 10^{-1}</math></b>	84	71	23	1	843	849	0.019670	0.022533	0.064294	0.044156
>SeqIndex14059_Y	<b><math>1.9 \times 10^{-5}</math></b>	<b><math>3.3 \times 10^{-3}</math></b>	<b><math>9.3 \times 10^{-1}</math></b>	<b><math>6.6 \times 10^{-1}</math></b>	84	71	23	1	843	849	0.008678	0.009579	0.064294	0.044156
>SeqIndex14402_X	<b><math>5.6 \times 10^{-2}</math></b>	<b><math>3.3 \times 10^{-4}</math></b>	$2.0 \times 10^{-1}$	<b><math>8.9 \times 10^{-1}</math></b>	120	102	0	25	1316	1319	0.026613	0.010857	0.030927	0.042530
>SeqIndex14402_Y	<b><math>5.6 \times 10^{-2}</math></b>	<b><math>3.3 \times 10^{-4}</math></b>	$2.0 \times 10^{-1}$	<b><math>8.9 \times 10^{-1}</math></b>	120	102	0	25	1316	1319	0.002941	0.008926	0.030927	0.042530
>SeqIndex14864_X	<b><math>1.2 \times 10^{-3}</math></b>	<b><math>6.0 \times 10^{-4}</math></b>	$1.4 \times 10^{-1}$	<b><math>9.2 \times 10^{-1}</math></b>	32	25	0	5	429	435	0.013772	0.008408	0.014872	0.031440
>SeqIndex14864_Y	<b><math>1.2 \times 10^{-3}</math></b>	<b><math>6.0 \times 10^{-4}</math></b>	$1.4 \times 10^{-1}$	<b><math>9.2 \times 10^{-1}</math></b>	32	25	0	5	429	435	0.008307	0.006002	0.014872	0.031440
>SeqIndex14937_X	$2.7 \times 10^{-1}$	<b><math>5.2 \times 10^{-4}</math></b>	$2.8 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	8	19	0	12	214	208	0.002148	0.006355	0.003044	0.068005
>SeqIndex14937_Y	$2.7 \times 10^{-1}$	<b><math>5.2 \times 10^{-4}</math></b>	$2.8 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	8	19	0	12	214	208	0.003117	0.005954	0.003044	0.068005
>SeqIndex24360_X	$2.8 \times 10^{-1}$	<b><math>3.9 \times 10^{-4}</math></b>	$2.7 \times 10^{-1}$	<b><math>8.7 \times 10^{-1}</math></b>	16	17	0	6	258	272	0.007461	0.005340	0.006016	0.029459
>SeqIndex24360_Y	$2.8 \times 10^{-1}$	<b><math>3.9 \times 10^{-4}</math></b>	$2.7 \times 10^{-1}$	<b><math>8.7 \times 10^{-1}</math></b>	16	17	0	6	258	272	0.003298	0.008766	0.006016	0.029459
>SeqIndex4671_X	<b><math>5.7 \times 10^{-5}</math></b>	$7.1 \times 10^{-1}$	<b><math>9.7 \times 10^{-1}</math></b>	$6.5 \times 10^{-3}$	74	41	15	0	552	553	0.027010	0.015533	0.070380	0.017378
>SeqIndex4671_Y	<b><math>5.7 \times 10^{-5}</math></b>	$7.1 \times 10^{-1}$	<b><math>9.7 \times 10^{-1}</math></b>	$6.5 \times 10^{-3}$	74	41	15	0	552	553	0.005156	0.006784	0.070380	0.017378
>SeqIndex4672_X	<b><math>3.0 \times 10^{-5}</math></b>	$5.0 \times 10^{-2}$	<b><math>9.8 \times 10^{-1}</math></b>	$2.7 \times 10^{-1}$	53	43	16	0	394	386	0.021936	0.026916	0.072143	0.029872
>SeqIndex4672_Y	<b><math>3.0 \times 10^{-5}</math></b>	$5.0 \times 10^{-2}$	<b><math>9.8 \times 10^{-1}</math></b>	$2.7 \times 10^{-1}$	53	43	16	0	394	386	0.010492	0.009631	0.072143	0.029872
>SeqIndex5556_X	$5.0 \times 10^{-1}$	<b><math>1.7 \times 10^{-4}</math></b>	$6.0 \times 10^{-4}$	<b><math>9.9 \times 10^{-1}</math></b>	30	25	0	19	273	274	0.040726	0.002928	0.041803	0.081158
>SeqIndex5556_Y	$5.0 \times 10^{-1}$	<b><math>1.7 \times 10^{-4}</math></b>	$6.0 \times 10^{-4}$	<b><math>9.9 \times 10^{-1}</math></b>	30	25	0	19	273	274	0.015758	0.005298	0.041803	0.081158
>SeqIndex5747_X	<b><math>6.7 \times 10^{-4}</math></b>	<b><math>3.2 \times 10^{-4}</math></b>	<b><math>9.4 \times 10^{-1}</math></b>	<b><math>9.9 \times 10^{-1}</math></b>	69	64	0	42	588	586	0.029616	0.004680	0.069294	0.085349
>SeqIndex5747_Y	<b><math>6.7 \times 10^{-4}</math></b>	<b><math>3.2 \times 10^{-4}</math></b>	<b><math>9.4 \times 10^{-1}</math></b>	<b><math>9.9 \times 10^{-1}</math></b>	69	64	0	42	588	586	0.003044	0.005197	0.069294	0.085349
>SeqIndex6022_X	<b><math>6.1 \times 10^{-5}</math></b>	<b><math>5.5 \times 10^{-4}</math></b>	<b><math>9.6 \times 10^{-1}</math></b>	<b><math>7.5 \times 10^{-1}</math></b>	119	64	38	0	893	895	0.011529	0.003150	0.078263	0.031027
>SeqIndex6022_Y	<b><math>6.1 \times 10^{-5}</math></b>	<b><math>5.5 \times 10^{-4}</math></b>	<b><math>9.6 \times 10^{-1}</math></b>	<b><math>7.5 \times 10^{-1}</math></b>	119	64	38	0	893	895	0.022879	0.023272	0.078263	0.031027
>SeqIndex6591_X	$1.3 \times 10^{-1}$	<b><math>3.2 \times 10^{-4}</math></b>	$2.0 \times 10^{-4}$	<b><math>9.8 \times 10^{-1}</math></b>	128	90	0	31	1047	1051	0.052189	0.007445	0.057894	0.055236
>SeqIndex6591_Y	$1.3 \times 10^{-1}$	<b><math>3.2 \times 10^{-4}</math></b>	$2.0 \times 10^{-4}$	<b><math>9.8 \times 10^{-1}</math></b>	128	90	0	31	1047	1051	0.025744	0.010053	0.057894	0.055236
>SeqIndex6752_X	$1.4 \times 10^{-1}$	<b><math>3.1 \times 10^{-4}</math></b>	$1.5 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	19	18	0	8	163	170	0.034893	0.016514	0.037566	0.068633
>SeqIndex6752_Y	$1.4 \times 10^{-1}$	<b><math>3.1 \times 10^{-4}</math></b>	$1.5 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	19	18	0	8	163	170	0.005104	0.005436	0.037566	0.068633
>SeqIndex7460_X	<b><math>3.1 \times 10^{-4}</math></b>	NA	<b><math>9.9 \times 10^{-1}</math></b>	NA	203	26	14	0	1649	1656	0.005286	0.003392	0.071842	0.002711
>SeqIndex7460_Y	<b><math>3.1 \times 10^{-4}</math></b>	NA	<b><math>9.9 \times 10^{-1}</math></b>	NA	203	26	14	0	1649	1656	0.034853	0.005574	0.071842	0.002711
>SeqIndex7824_X	$2.6 \times 10^{-1}$	<b><math>1.6 \times 10^{-3}</math></b>	$2.2 \times 10^{-3}$	<b><math>9.3 \times 10^{-1}</math></b>	28	20	0	4	341	343	0.018695	0.014224	0.033650	0.034705
>SeqIndex7824_Y	$2.6 \times 10^{-1}$	<b><math>1.6 \times 10^{-3}</math></b>	$2.2 \times 10^{-3}$	<b><math>9.3 \times 10^{-1}</math></b>	28	20	0	4	341	343	0.023891	0.005747	0.033650	0.034705
>SeqIndex8001_X	<b><math>1.3 \times 10^{-4}</math></b>	<b><math>3.2 \times 10^{-1}</math></b>	<b><math>8.6 \times 10^{-1}</math></b>	$1.9 \times 10^{-1}$	76	58	8	0	1021	1012	0.012652	0.016214	0.039941	0.012715
>SeqIndex8001_Y	<b><math>1.3 \times 10^{-4}</math></b>	<b><math>3.2 \times 10^{-1}</math></b>	<b><math>8.6 \times 10^{-1}</math></b>	$1.9 \times 10^{-1}$	76	58	8	0	1021	1012	0.007475	0.000531	0.039941	0.012715
>SeqIndex8753_X	<b><math>3.9 \times 10^{-4}</math></b>	<b><math>3.2 \times 10^{-4}</math></b>	$2.8 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	48	46	5	28	465	466	0.036021	0.014627	0.049505	0.080913
>SeqIndex8753_Y	<b><math>3.9 \times 10^{-4}</math></b>	<b><math>3.2 \times 10^{-4}</math></b>	$2.8 \times 10^{-1}$	<b><math>9.9 \times 10^{-1}</math></b>	48	46	5	28	465	466	0.005374	0.001607	0.049505	0.080913
>SeqIndex9704_X	<b><math>1.8 \times 10^{-6}</math></b>	<b><math>2.5 \times 10^{-3}</math></b>	<b><math>9.4 \times 10^{-1}</math></b>	$1.8 \times 10^{-1}$	65	30	3	0	676	676	0.014389	0.008683	0.048149	0.009691
>SeqIndex9704_Y	<b><math>1.8 \times 10^{-6}</math></b>	<b><math>2.5 \times 10^{-3}</math></b>	<b><math>9.4 \times 10^{-1}</math></b>	$1.8 \times 10^{-1}$	65	30	3	0	676	676	0.018368	0.006583	0.048149	0.009691

## Discussion

The transition to dioecy could be recent in *S. acaulis* ssp *excapa*. Thus, if the determination of the sex is genetic, we expect to identify very young sex chromosomes in this subspecies. For the purpose of testing this, we conducted two different analysis to identify sex-linked genes in *S. acaulis* ssp *excapa*. We combined an SDpop analysis (Käfer et al., 2021b) with a heterozygosity analysis on two datasets from two different populations. From these analysis, we found 27 genes that are potentially XY in at least one of the two populations. Our results may support the hypothesis that the pair of sex chromosomes is young in *S. acaulis* ssp *excapa*. First, we identified only 27 potential sex-linked genes. The low divergence between X and Y chromosomes in young systems should provide a small subset of

sex-linked genes (Charlesworth et al., 2005; Bachtrog,

supported by SDpop and heterozygosity analysis, which suggest that the ZW system is more likely to be wrong than the XY system. It remains intriguing that a single XY gene is identified by both methods in both populations. Moreover, some genes were inferred as XY with a probability greater than 0.8 without fixed XY SNPs. Some genes have XY fixed mutations in a population but not in the other. An hypothesis to explain the absence of fixed SNPs, and thus the weak SDpop support, is the sex reversion. Sex reversions have been reported in several plants (*i.e.* *Amborella trichopoda*) (Freeman et al., 1980; Korpelainen, 1998; Anger et al., 2017; Käfer et al., 2021a) and individuals with a phenotypic sex different from the genotypic sex avoid the detection of fixed SNPs and reduce the sex-linked probability inferred by SDpop. Indeed, if an XY female is present in the sampling, Y-specific SNPs will be identified in females, and thus not be considered as fixed. In reversed females, it is possible that both copies are expressed in most XY genes, but a few XY genes did not express the Y copy. Since we used RNA-seq data, it could explain that among the putative XY genes, some have fixed mutations between X and Y copies but not all.

The formation of a pair of sex chromosomes may be a long evolutionary process and it is not clear what are the expectations for different populations of *S. acaulis* ssp *exscapa*. For example, does each population have its genetic sex determination? Or, is there a core subset of sex-linked genes for all populations, plus a specific subset of genes per population? In guppies, for example, it has been reported that the absence of inversion could explain a remarkable population variation in Y chromosomes diversities (Almeida et al., 2021). Sampling more divergent dioecious populations could determine if the potential XY genes identified in this study are shared with other populations or unique. Adding several populations would be ideal, but even a single one may provide interesting results.

As other tools used for sex-linked genes identification, SDpop becomes less effective with young sex-linked regions and weak divergence (Palmer et al., 2019; Käfer et al., 2021b). In Käfer et al. (2021b), it is reported that precision is high (>95%) for systems older than  $4N_e$  generations and with a non-recombining region larger than 1% of the genome. For systems younger than  $2N_e$  generations, the precision and the power of SDpop decrease (Käfer et al., 2021b). Nevertheless, Table 5 shows that highest divergence between X and Y sequences reached 8%. With an estimated polymorphism close to 1.5%, the divergence time is estimated to be about  $6N_e$ . Such a divergence should allow SDpop to identify sex-linked genes with high precision. It is thus possible that sex chromosomes in *S. acualis* ssp *exscapa* are older than expected. A biological hypothesis may explain this results. Some sex determining loci have

been found in recombining regions but with a low recombination rate (Perrin, 2009; Rodrigues et al., 2018; Rifkin et al., 2021). In these regions, a small part is strongly heterozygous, and the heterozygosity rapidly declines with the distance to the sex-determining locus. Thus, there are only a few number of genes with fixed mutations between X and Y sequences compared to the age of the sex-determining locus. Moreover, this could explain the polymorphism between populations regarding the set of genes that we identified.

Some analysis or tools could help to clarify whether sex chromosomes exist in this species. First, the greenhouse cultivation of *S. acaulis* ssp *exscapa* remains difficult. Increasing our abilities to grow this plant in controlled conditions would enable to make a cross. This would enable SEX-DETECTOR analysis (Muyle et al., 2016). SEX-DETECTOR analyses the alleles segregations within a cross with a small number of offsprings (about five of each sex) to identify genes in a non-recombining region. Since SEX-DETECTOR uses about 10 offsprings from a single generation, few recombination events between homologous chromosomes should be present in the data. Therefore, the size of the non-recombining region would be overestimated, and some genes in the pseudo-autosomal region considered as sex-linked. An overestimation of the genes present in the non-recombining region could help to determine whether such a region exists or not. However, since some sex-linked genes would be partially and not fully sex-linked, a SEX-DETECTOR approach will limit analysis of which genes are firstly recruited in the non-recombining region (*e.g.* which are the sex-determining genes).

As mentioned before, SDpop needs divergent time greater than  $2N_e$  generations to obtain reliable results (Käfer et al., 2021b). It would be useful to go further in the expected results for very young sex chromosomes. To do so, we could simulate several datasets differing in several variables (*e.g.* sex-linked genes number, divergence level). This would allow us to test whether the results we generated fit with certain young and weakly divergent non-recombining region.

Finally, a crucial tool to acquire would be a chromosomal level assembly of *S. acaulis* ssp *exscapa* genome. Determining the position of the potential sex-linked genes could determine whether they are false positives. If they are sparse all along the genome, we could assume that they are false positives. Although the synteny of these genes would not be sufficient to ensure that we identified genes in a non-recombining region it would support this hypothesis. Moreover, if the formation of the non recombining region is induced by an inversion it could be possible to detect this inversion with a high quality chromosomal level assembly of the genome. Otherwise, non-coding parts of the non-recombining region should display greater rate of heterozygosity since the majority of the non-coding regions is expected to

evolve under neutral evolution. We could measure the heterozygosity in a male and a female along the whole genome to identify a non-recombining region. Also, we used a single tissue at a single stage of development, therefore many genes are not present in the RNA-seq. Working with DNA-seq data, or sequencing the transcriptomes of other tissues and stages of development could increase the number of genes analyzed by SD-pop. Finally, DNA-seq data could allow us to test the hypothesis of monoallelic expression in female reversed individuals.

On another note, the environmental determination of the sex is poorly understood in plants (Pannell, 2017). Having a better estimation of the impact of the role of the environment in the determination of the sex in *S. acaulis* would be useful. It could be interesting to study the impact of several environmental conditions on the sex (*i.e.* temperatures, photoperiods, resources), as it have been done in *Carica papaya* or in *Cucumis sativus* (Lin et al., 2016; Lai et al., 2018). Such an experiment requires the capacity of growing *S. acaulis* in a greenhouse, which is not the case so far. Besides what help the estimation of the environmental impact on the sex could bring to our study, a better understanding of the environmental sex determination in plants seems necessary.

To conclude, our work confirms that *S. acaulis* ssp *exscapa* is an interesting model for understanding the genetic transition to dioecy. Further analyses seem necessary to determine whether this subspecies has sex chromosomes or not. However, our results support the hypothesis that sex chromosomes, if they exist, should be young with a small non-recombining region.

## Author contributions

Conceptualization of the study: G.A.B.M. and J.K.; methodology: G.A.B.M. and J.K.; software: D.P., P.J. and A.M.; formal analysis: D.P.; investigation: D.P., J.K., and G.A.B.M.; resources: A.M., and J.K.; writing—original draft: D.P. ; writing—review and editing: D.P.; visualization: D.P.; supervision: G.A.B.M., and J.K.; project administration: G.A.B.M.; funding acquisition: G.A.B.M.

## Acknowledgments

We thank the Jardin Alpin Botanique du Col du Lautaret for assistance with field work and Sylvain Santoni for sequencing and help with the sampling protocol. This work was performed using the computing facilities of the CC LBBE/PRABI; we thank Bruno Spataro and Stéphane Delmotte for cluster maintenance. Virtual machines from the Institut Français de Bioinformatique

also were used to perform this work. This work received financial support ANR grant (ANR-14-CE19-0021-01) to Marais G.A.B.

## References

- Almeida, P., Sandkam, B. A., Morris, J., Darolti, I., Breden, F., and Mank, J. E. (2021). Divergence and Remarkable Diversity of the Y Chromosome in Guppies. *Molecular Biology and Evolution*, 38(2):619–633.
- Anger, N., Fogliani, B., Scutt, C. P., and Gâteblé, G. (2017). Dioecy in *Amborella trichopoda*: evidence for genetically based sex determination and its consequences for inferences of the breeding system in early angiosperms. *Annals of Botany*, 119(4):591–597.
- Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics*, 14(2):113–124.
- Balounova, V., Gogela, R., Cegan, R., Cangren, P., Zluvova, J., Safar, J., Kovacova, V., Bergero, R., Hobza, R., Vyskot, B., Oxelman, B., Charlesworth, D., and Janousek, B. (2019). Evolution of sex determination and heterogamety changes in section *Otites* of the genus *Silene*. *Scientific Reports*, 9(1):1045.
- Bertrand, Y. J. K., Petri, A., Scheen, A.-C., Töpel, M., and Oxelman, B. (2018). De novo transcriptome assembly, annotation, and identification of low-copy number genes in the flowering plant genus *Silene* (Caryophyllaceae). *bioRxiv*, page 290510.
- Cabau, C., Escudié, F., Djari, A., Guiguen, Y., Bobe, J., and Klopp, C. (2017). Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. *PeerJ*, 5:e2988.
- Carey, S., Yu, Q., and Harkess, A. (2021). The Diversity of Plant Sex Chromosomes Highlighted through Advances in Genome Sequencing. *Genes*, 12(3):381.
- Charlesworth, B. and Charlesworth, D. (1978). A Model for the Evolution of Dioecy and Gynodioecy. *The American Naturalist*, 112(988):975–997.
- Charlesworth, D. (2019). Young sex chromosomes in plants and animals. *New Phytologist*, 224(3):1095–1107.
- Charlesworth, D. (2021). When and how do sex-linked regions become sex chromosomes? *Evolution*, 75(3):569–581.
- Charlesworth, D., Charlesworth, B., and Marais, G. (2005). Steps in the evolution of heteromorphic sex chromosomes. *Heredity*, 95(2):118–128.

- Desfeux, C., Maurice, S., Henry, J.-p., Lejeune, B., and Gouyon, P.-h. (1996). Evolution of reproductive systems in the genus *Silene*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263(1369):409–414.
- Dohm, J. C., Minoche, A. E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sörensen, T. R., Stracke, R., Reinhardt, R., Goesmann, A., Kraft, T., Schulz, B., Stadler, P. F., Schmidt, T., Gabaldón, T., Lehrach, H., Weisshaar, B., and Himmelbauer, H. (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, 505(7484):546–549.
- Freeman, D. C., Harper, K. T., and Charnov, E. L. (1980). Sex change in plants: Old and new observations and new hypotheses. *Oecologia*, 47(2):222–232.
- Fruchard, C. and Marais, G. A. B. (2017). The Evolution of Sex Determination in Plants. In Nuno de la Rosa, L. and Müller, G., editors, *Evolutionary Developmental Biology: A Reference Guide*, pages 1–14. Springer International Publishing, Cham.
- Garraud, C., Brachi, B., Dufay, M., Touzet, P., and Shykoff, J. A. (2011). Genetic determination of male sterility in gynodioecious *Silene nutans*. *Heredity*, 106(5):757–764. Bandiera\_abtest: a Cg\_type: Nature Research Journals Number: 5 Primary\_atype: Research Publisher: Nature Publishing Group Subject\_term: Model plants;Plant breeding;Plant genetics Subject\_term\_id: model-plants;plant-breeding;plant-genetics.
- Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., Lourenco, J. M., Alves, P. C., Ballenghien, M., Faivre, N., Belkhir, K., Cahais, V., Loire, E., Bernard, A., and Galtier, N. (2013). Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. *PLOS Genetics*, 9(4):e1003457.
- Gussarova, G., Allen, G. A., Mikhaylova, Y., McCormick, L. J., Mirré, V., Marr, K. L., Hebda, R. J., and Brochmann, C. (2015). Vicariance, long-distance dispersal, and regional extinction–recolonization dynamics explain the disjunct circumpolar distribution of the arctic-alpine plant *Silene acaulis*. *American Journal of Botany*, 102(10):1703–1720.
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512.
- Harkess, A., Zhou, J., Xu, C., Bowers, J. E., Van der Hulst, R., Ayyampalayam, S., Mercati, F., Riccardi, P., McKain, M. R., Kakrana, A., Tang, H., Ray, J., Groenendijk, J., Arikrit, S., Mathioni, S. M., Nakano, M., Shan, H., Telgmann-Rauber, A., Kanno, A., Yue, Z., Chen, H., Li, W., Chen, Y., Xu, X., Zhang, Y., Luo, S., Chen, H., Gao, J., Mao, Z., Pires, J. C., Luo, M., Kudrna, D., Wing, R. A., Meyers, B. C., Yi, K., Kong, H., Lavrijsen, P., Sunseri, F., Falavigna, A., Ye, Y., Leebens-Mack, J. H., and Chen, G. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature Communications*, 8(1):1279.
- Hermanutz, L. A. and Innes, D. J. (1994). Gender variation in *Silene acaulis* (Caryophyllaceae). *Plant Systematics and Evolution*, 191(1):69–81.
- Huang, X. and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9):868–877.
- Jafari, F., Zarre, S., Gholipour, A., Eggens, F., Rabaler, R. K., and Oxelman, B. (2020). A new taxonomic backbone for the infrageneric classification of the species-rich genus *Silene* (Caryophyllaceae). *TAXON*, 69(2):337–368.
- Korpelainen, H. (1998). Labile sex expression in plants. *Biological Reviews*, 73(2):157–180.
- Käfer, J., Bewick, A., Andres-Robin, A., Lapetoule, G., Harkess, A., Caius, J., Fogliani, B., Gâteblé, G., Ralph, P., dePamphilis, C. W., Picard, F., Scutt, C., Marais, G. A. B., and Leebens-Mack, J. (2021a). A derived ZW chromosome system in *Amborella trichopoda*, representing the sister lineage to all other extant flowering plants. *The New Phytologist*.
- Käfer, J., Lartillot, N., Marais, G. A. B., and Picard, F. (2021b). Detecting sex-linked genes using genotyped individuals sampled in natural populations. *Genetics*, 218(2):iyab053.
- Lai, Y.-S., Shen, D., Zhang, W., Zhang, X., Qiu, Y., Wang, H., Dou, X., Li, S., Wu, Y., Song, J., Ji, G., and Li, X. (2018). Temperature and photoperiod changes affect cucumber sex expression by different epigenetic regulations. *BMC Plant Biology*, 18(1):268.
- Lewis, D. (1942). The Evolution of Sex in Flowering Plants. *Biological Reviews*, 17(1):46–67.
- Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- Lin, H., Liao, Z., Zhang, L., and Yu, Q. (2016). Transcriptome analysis of the male-to-hermaphrodite sex reversal induced by low temperature in papaya. *Tree Genetics & Genomes*, 12(5):94.
- Lloyd, D. G. (1980). The Distributions of Gender in Four Angiosperm Species Illustrating Two Evolutionary Pathways to Dioecy. *Evolution*, 34(1):123–134.
- Marais, G. A. B., Forrest, A., Kamau, E., Käfer, J., Daubin, V., and Charlesworth, D. (2011). Multiple Nuclear Gene Phylogenetic Analysis of the Evolution of Dioecy and Sex Chromosomes in the Genus *Silene*. *PLOS ONE*, 6(8):e21915.
- Martin, H., Carpentier, F., Gallina, S., Godé, C., Schmitt, E., Muyle, A., Marais, G. A. B., and Touzet, P. (2019). Evolution of Young Sex Chromosomes in Two Dioecious Sister Plant Species with Distinct Sex Determination Systems. *Genome Biology and Evolution*, 11(2):350–361.
- Maurice, S., Belhassen, E., Couvet, D., and Gouyon, P.-H. (1994). Evolution of dioecy: can nuclear–cytoplasmic interactions select for maleness? *Heredity*, 73(4):346–354.
- Maurice, S., Desfeux, C., Mignot, A., and Henry, J.-P. (1998). Is *Silene acaulis* (Caryophyllaceae) a tri-ocious species? Reproductive biology of two subspecies. *Canadian Journal of Botany*, 76(3):478–485.
- Ming, R., Bendahmane, A., and Renner, S. S. (2011). Sex Chromosomes in Land Plants. *Annual Review of Plant Biology*, 62(1):485–514.
- Morris, W. F. and Doak, D. F. (1998). Life history of the long-lived gynodioecious cushion plant *Silene acaulis* (Caryophyllaceae), inferred from size-based population projection matrices. *American Journal of Botany*, 85(6):784–793.
- Muyle, A., Käfer, J., Zemp, N., Mousset, S., Picard, F., and Marais, G. A. (2016). SEX-DETECTOR: A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms. *Genome Biology and Evolution*, 8(8):2530–2543.
- Naciri, Y., Pasquier, P.-E. D., Lundberg, M., Jeanmonod, D., and Oxelman, B. (2017). A phylogenetic circumscription of *Silene* sect. *Siphonomorpha* (Caryophyllaceae) in the Mediterranean Basin. *TAXON*, 66(1):91–108.
- Okazaki, Y., Takahata, S., Hirakawa, H., Suzuki, Y., and Onodera, Y. (2019). Molecular evidence for recent divergence of X- and Y-linked gene pairs in *Spinacia oleracea* L. *PLOS ONE*, 14(4):e0214949. Publisher: Public Library of Science.
- Palmer, D. H., Rogers, T. F., Dean, R., and Wright, A. E. (2019). How to identify sex chromosomes and their turnover. *Molecular Ecology*, 28(21):4709–4724.
- Pannell, J. R. (2017). Plant Sex Determination. *Current Biology*, 27(5):R191–R197.
- Perrin, N. (2009). Sex Reversal: A Fountain of Youth for Sex Chromosomes? *Evolution*, 63(12):3043–3049.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596.
- Renner, S. S. (2014). The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database. *American Journal of Botany*, 101(10):1588–1596.
- Renner, S. S. and Müller, N. A. (2021). Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. *Nature Plants*, 7(4):392–402.
- Rifkin, J. L., Beaudry, F. E. G., Humphries, Z., Choudhury, B. I., Barrett, S. C. H., and Wright, S. I. (2021). Widespread Recombination Suppression Facilitates Plant Sex Chromosome Evolution. *Molecular Biology and Evolution*, 38(3):1018–1030.
- Rodrigues, N., Studer, T., Dufresnes, C., and Perrin, N. (2018). Sex-Chromosome Recombination in Common Frogs Brings Water to the Fountain-of-Youth. *Molecular Biology and Evolution*, 35(4):942–948.
- Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.
- Schmieder, R., Lim, Y. W., and Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics*, 28(3):433–435.
- Schultz, S. T. (1994). Nucleo-Cytoplasmic Male Sterility and Alternative Routes to Dioecy. *Evolution*, 48(6):1933–1945.
- Slancarova, V., Zdanska, J., Janousek, B., Talianova, M., Zschach, C., Zluvova, J., Siroky, J., Kovacova, V., Blavet, H., Danihelka, J., Oxelman, B., Widmer, A., and Vyskot, B. (2013). Evolution of Sex Determination Systems with Heterogametic Males and Females in *Silene*. *Evolution*, 67(12):3669–3677.

- Sloan, D. B., Müller, K., McCauley, D. E., Taylor, D. R., and Štorchová, H. (2012). Intraspecific variation in mitochondrial genome sequence, structure, and gene content in *Silene vulgaris*, an angiosperm with pervasive cytoplasmic male sterility. *New Phytologist*, 196(4):1228–1239. [\\_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.2012.04340.x](https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.2012.04340.x).
- Tennessen, J. A., Wei, N., Straub, S. C. K., Govindarajulu, R., Liston, A., and Ashman, T.-L. (2018). Repeated translocation of a gene cassette drives sex-chromosome turnover in strawberries. *PLOS Biology*, 16(8):e2006062.
- Veltsos, P., Cossard, G., Beaudoin, E., Beydon, G., Savova Bianchi, D., Roux, C., C. González-Martínez, S., and R. Pannell, J. (2018). Size and Content of the Sex-Determining Region of the Y Chromosome in Dioecious *Mercurialis annua*, a Plant with Homomorphic Sex Chromosomes. *Genes*, 9(6):277.
- Wu, M., Haak, D. C., Anderson, G. J., Hahn, M. W., Moyle, L. C., and Guerrero, R. F. (2021). Inferring the Genetic Basis of Sex Determination from the Genome of a Dioecious Nightshade. *Molecular Biology and Evolution*, 38(7):2946–2957.
- Yu, L., Ma, X., Deng, B., Yue, J., and Ming, R. (2021). Construction of high-density genetic maps defined sex determination region of the Y chromosome in spinach. *Molecular Genetics and Genomics*, 296(1):41–53.
- Zou, C., Chen, A., Xiao, L., Muller, H. M., Ache, P., Haberer, G., Zhang, M., Jia, W., Deng, P., Huang, R., Lang, D., Li, F., Zhan, D., Wu, X., Zhang, H., Bohm, J., Liu, R., Shabala, S., Hedrich, R., Zhu, J.-K., and Zhang, H. (2017). A high-quality genome assembly of quinoa provides insights into the molecular basis of salt bladder-based salinity tolerance and the exceptional nutritional value. *Cell Research*, 27(11):1327–1340.



## Supplementary Material

Table S1: Indicates the number of contigs for each assemblies, as well as the mean percentage of mapped reads for males and females.

Reference transcriptome	# contigs	% mapped (males 2018)	% mapped (females 2018)	% mapped (males 2013)	% mapped (females 2013)
Meta-assembly (5 males 5 females)	51,146	77.3	76.5	84.4	86.2
Meta-assembly (8 females)	55,234	71.1	78.8	85.2	88.8
Meta-assembly (1 male)	28,554	55.8	52.4	61.6	61.7
Meta-assembly (1 female)	29,778	50.0	53.0	58.1	61.5
Assembly with 2013 data	120,129	55.4	54.9	65.1	64.6
<i>Silene otites</i>	71,884	53.2	50.6	60.0	59.6
<i>Silene pseudotites</i>	60,193	53.3	50.7	53.3	59.7
<i>Silene schafta</i>	26,845	36.8	36.9	49.9	45.9
<i>Beta vulgaris</i>	37,802	3.5	3.8	6.3	6.0
<i>Chenopodium quinoa</i>	74,835	3.8	3.9	6.8	6.4

Table S2: Assemblies statistics for transcriptomes assembled with DRAP.

		Meta-assembly (5 males 5 females)	Meta-assembly (8 females)	Meta-assembly (1 male)	Meta-assembly (1 female)
# contigs		51,146	55,234	28,554	29,778
Transrate score		0.17	0.18	0.27	0.21
BUSCO score (eudicotyledon)	Complete	93%	94%	91%	90%
	Single	81%	82%	84%	83%
	Duplicated	12%	12%	7%	7%
	Fragmented	2%	2%	4%	5%

Table S3: Proportions of each segregation type for a mapping of data sampled in 2013 on the ten assemblies.

Reference transcriptome	Model	Autosomal	Haploid	Paralogous	Hemizygous	Gametologous
Meta-assembly (5 males 5 females)	XY	0.964	0.002	0.032	0.001	0.000
	ZW	0.964	0.002	0.032	0.001	0.000
Meta-assembly (8 females)	XY	0.966	0.002	0.031	0.001	0.000
	ZW	0.966	0.002	0.031	0.001	0.000
Meta-assembly (1 male)	XY	0.940	0.000	0.059	0.000	0.001
	ZW	0.941	0.000	0.059	0.000	0.000
Meta-assembly (1 female)	XY	0.942	0.000	0.057	0.000	0.001
	ZW	0.943	0.000	0.057	0.000	0.000
Assembly with 2013 data	XY	0.892	0.000	0.105	0.000	0.003
	ZW	0.893	0.000	0.106	0.000	0.001
<i>Silene otites</i>	XY	0.874	0.000	0.122	0.000	0.004
	ZW	0.877	0.000	0.123	0.000	0.001
<i>Silene pseudotites</i>	XY	0.872	0.000	0.124	0.000	0.004
	ZW	0.875	0.000	0.124	0.000	0.001
<i>Silene schafta</i>	XY	0.783	0.000	0.207	0.000	0.010
	ZW	0.790	0.000	0.207	0.000	0.002
<i>Beta vulgaris</i>	XY	0.335	0.000	0.633	0.000	0.033
	ZW	0.356	0.000	0.639	0.000	0.006
<i>Chenopodium quinoa</i>	XY	0.363	0.000	0.603	0.000	0.034
	ZW	0.384	0.000	0.609	0.000	0.007

Table S4: Proportions of each segregation type for a mapping of data sampled in 2018 on the ten assemblies.

Reference transcriptome	Model	Autosomal	Haploid	Paralogous	Hemizygous	Gametologous
Meta-assembly (5 males 5 females)	XY	0.980	0.002	0.018	0.000	0.000
	ZW	0.977	0.002	0.018	0.004	0.000
Meta-assembly (8 females)	XY	0.981	0.006	0.013	0.000	0.000
	ZW	0.977	0.006	0.013	0.006	0.000
Meta-assembly (1 male)	XY	0.970	0.001	0.029	0.000	0.001
	ZW	0.968	0.001	0.029	0.000	0.000
Meta-assembly (1 female)	XY	0.942	0.000	0.057	0.000	0.001
	ZW	0.943	0.000	0.057	0.002	0.000
Assembly with 2013 data	XY	0.928	0.000	0.070	0.000	0.003
	ZW	0.930	0.000	0.070	0.000	0.000
<i>Silene otites</i>	XY	0.924	0.000	0.074	0.000	0.002
	ZW	0.925	0.000	0.074	0.000	0.001
<i>Silene pseudotites</i>	XY	0.925	0.000	0.073	0.000	0.001
	ZW	0.926	0.000	0.073	0.000	0.000
<i>Silene schafta</i>	XY	0.843	0.000	0.152	0.000	0.005
	ZW	0.846	0.000	0.152	0.000	0.001
<i>Beta vulgaris</i>	XY	0.446	0.000	0.532	0.000	0.023
	ZW	0.461	0.000	0.534	0.000	0.005
<i>Chenopodium quinoa</i>	XY	0.501	0.001	0.0934	0.000	0.005
	ZW	0.501	0.001	0.494	0.000	0.005



# 3

## Chapter 3 : An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of *Cannabis sativa*

Au cours de mes stages de master (M1 et M2) j'ai travaillé sur les chromosomes sexuels de *Cannabis sativa*. Avant mon arrivée dans l'équipe, Jos Käfer et Gabriel Marais avaient déjà monté une collaboration avec une équipe russe pour obtenir un croisement chez cette espèce. En parallèle le Beijing Genomics Institute (BGI) avait lancé un appel d'offre pour financer des séquençages RNA-seq de plantes médicinales avec sa propre technologie de séquençage (Complete Genomics). Le projet d'identification d'une paire de chromosomes sexuels chez *C. sativa* a été retenu par le BGI, ce qui a permis d'obtenir un séquençage RNA-seq du croisement généré par les collaborateurs.

La première année de thèse a, en partie, servi à finir les analyses qui n'avaient pas pu l'être à la fin du Master 2 et à rédiger l'article scientifique sur les chromosomes sexuels de *C. sativa*. Ces travaux ont été publiés dans le journal Genome Research (numéro de février 2020) et cette publication a été mise en avant sur le site de l'université Lyon 1<sup>1</sup>. Par ailleurs, j'ai été invité pour présenter ce travail lors d'une conférence internationale (conférence annuelle de la «Society of Experimental Biology»).

Cet article constitue le troisième chapitre de ma thèse.

---

1. <https://www.univ-lyon1.fr/actualites/actu-recherche/les-chromosomes-sexuels-de-la-plante-cannabis-sativa-identifies>

# An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of *Cannabis sativa*

Djivan Prentout,<sup>1</sup> Olga Razumova,<sup>2,3</sup> Bénédicte Rhoné,<sup>1,4</sup> Hélène Badouin,<sup>1</sup> Hélène Henri,<sup>1</sup> Cong Feng,<sup>5,6</sup> Jos Käfer,<sup>1</sup> Gennady Karlov,<sup>2</sup> and Gabriel A.B. Marais<sup>1</sup>

<sup>1</sup>Laboratoire de Biométrie et Biologie Évolutive UMR 5558, Université Lyon 1, CNRS, F-69622 Villeurbanne, France; <sup>2</sup>Laboratory of Applied Genomics and Crop Breeding, All-Russia Research Institute of Agricultural Biotechnology, Moscow 127550, Russia; <sup>3</sup>N.V. Tsitsin Main Botanical Garden of Russian Academy of Sciences, Moscow 127276, Russia; <sup>4</sup>Institut de Recherche pour le Développement, UMR DIADE, IRD, Université de Montpellier, F-34394 Montpellier, France; <sup>5</sup>Chongqing Medical University, Yuzhong District, Chongqing, 400016, China; <sup>6</sup>BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China

*Cannabis sativa*-derived tetrahydrocannabinol (THC) production is increasing very fast worldwide. *C. sativa* is a dioecious plant with XY Chromosomes, and only females (XX) are useful for THC production. Identifying the sex chromosome sequence would improve early sexing and better management of this crop; however, the *C. sativa* genome projects have failed to do so. Moreover, as dioecy in the Cannabaceae family is ancestral, *C. sativa* sex chromosomes are potentially old and thus very interesting to study, as little is known about old plant sex chromosomes. Here, we RNA-sequenced a *C. sativa* family (two parents and 10 male and female offspring, 576 million reads) and performed a segregation analysis for all *C. sativa* genes using the probabilistic method SEX-DETECTOR. We identified >500 sex-linked genes. Mapping of these sex-linked genes to a *C. sativa* genome assembly identified the largest chromosome pair being the sex chromosomes. We found that the X-specific region (not recombining between X and Y) is large compared to other plant systems. Further analysis of the sex-linked genes revealed that *C. sativa* has a strongly degenerated Y Chromosome and may represent the oldest plant sex chromosome system documented so far. Our study revealed that old plant sex chromosomes can have large, highly divergent nonrecombining regions, yet still be roughly homomorphic.

[Supplemental material is available for this article.]

*Cannabis sativa* is an ancient crop (Schultes et al. 1974) with two main traditional uses: marijuana and hemp (Small 2015). Marijuana, which is used in folk medicine, as a recreational drug, and lately in conventional medicine (Alexander 2016), has a narcotic effect owing to tetrahydrocannabinol (THC) and other cannabinoids produced in high concentration by some *C. sativa* cultivars. Until recently, the use of marijuana was prohibited in almost all countries, but *C. sativa*-derived products with high THC concentrations are now legal, for example, in several US states, Australia, Germany, Peru, and the UK for medicinal purposes (Offord 2018) and also in Uruguay, Canada, and several US states for recreational use (Yeager 2018). In the US, marijuana legal economy amounted to ~\$17 billion in 2016 and may reach as much as \$70 billion/year by 2021 (McVey 2017). However, legalization of marijuana is so recent that very few biotech tools have been developed for high THC-producing *C. sativa* cultivars (Yeager 2018).

THC reaches the highest concentrations in female inflorescences (bracts), so that only female *C. sativa* plants are of economic importance; furthermore, pollinated female plants produce smaller inflorescences and therefore less THC (Small 2015). It is thus important to avoid growing male plants as they are a waste of resources, labor, and space. Interest in hemp is also increasing as it is a crop for the sustainable production of fibers and oils (Andre et al. 2016; Salentijn et al. 2019). Hemp cultivars usually have a low level of THC and can be legally grown in many countries where marijuana is illegal. Features of male and female

hemp plants differ, and early sexing is also useful (Salentijn et al. 2019).

Sexual dimorphism in *C. sativa* is weak as in many dioecious plants (Barrett and Hough 2013), and sex can be determined with certainty only when the plants start flowering (Small 2015). *C. sativa* is a dioecious plant in which sex is determined by an XY Chromosome pair (Divashuk et al. 2014). So far, a few Y-linked genetic markers have been identified and are used to sex *C. sativa* seedlings (e.g., Techen et al. 2010). However, it is not known whether these markers work with all cultivars. The *C. sativa* sex chromosomes sequences would thus be an important genomic resource that could help improve agricultural yields. Currently, the *C. sativa* genome projects (van Bakel et al. 2011; Grassa et al. 2018; Laverty et al. 2019) have failed to identify the sex chromosomes, despite chromosome-level assemblies in the latest projects.

*C. sativa* is one of 15,600 dioecious species of flowering plants (Renner 2014). Dioecy and sex chromosomes have evolved multiple times in plants (Renner 2014), but very few plant systems have been studied in detail (Ming et al. 2011; Charlesworth 2015; Muyle et al. 2017). Historically, sex chromosomes have been classified using results from light microscopy (Ming et al. 2011). The terms homomorphic and heteromorphic refer to these results, with the former being roughly of similar size and the latter clearly different (but see Palmer et al. 2019 for another definition of

**Corresponding author:** gabriel.marais@univ-lyon1.fr

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.251207.119>.

© 2020 Prentout et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

heteromorphy). The extent to which recombination is suppressed between the sex chromosomes largely explains whether a sex chromosome pair becomes heteromorphic or not (Charlesworth et al. 2005; Bergero and Charlesworth 2009). Homomorphic XY tend to have large recombining regions and heteromorphic XY large nonrecombining ones. There is a loose correlation between the level of heteromorphy and age, but some old homomorphic systems have been described in animals and algae (e.g., Touns and Hahn 2010; Vicoso et al. 2013; Ahmed et al. 2014; Yazdi and Ellegren 2014). In plants, a model for the evolution of sex chromosomes heteromorphy with six stages has been proposed (Ming et al. 2011). In the initial stages, the sex chromosomes have small to intermediate nonrecombining regions and are homomorphic. After some time has elapsed since recombination cessation, DNA sequence and gene content can differ substantially in the nonrecombining X and Y regions, even though the sex chromosomes might be homomorphic under light microscopy (Ming et al. 2011; Wang et al. 2012; Veltsos et al. 2019). The well-studied heteromorphic systems in plants are characterized by a Y Chromosome larger than the X due to fast accumulation of repeats on the former, as in *Silene latifolia* or *Coccinia grandis* (Matsunaga et al. 1994; Sousa et al. 2013, 2016; Hobza et al. 2017), or multiple Y Chromosomes due to chromosomal fission-fusion events, as in *Rumex* species (Ming et al. 2011; Hough et al. 2014; Crowson et al. 2017). However, these sex chromosomes systems are still relatively young (less than 15 million years), and the late stages in the current model for the evolution of sex chromosomes heteromorphy in plants have not yet received attention from genomic studies. In particular, it is not clear whether the plant Y Chromosomes can shrink, as found in the ancient heteromorphic animal systems such as those of humans and some *Drosophila* species (Bachtrog 2013).

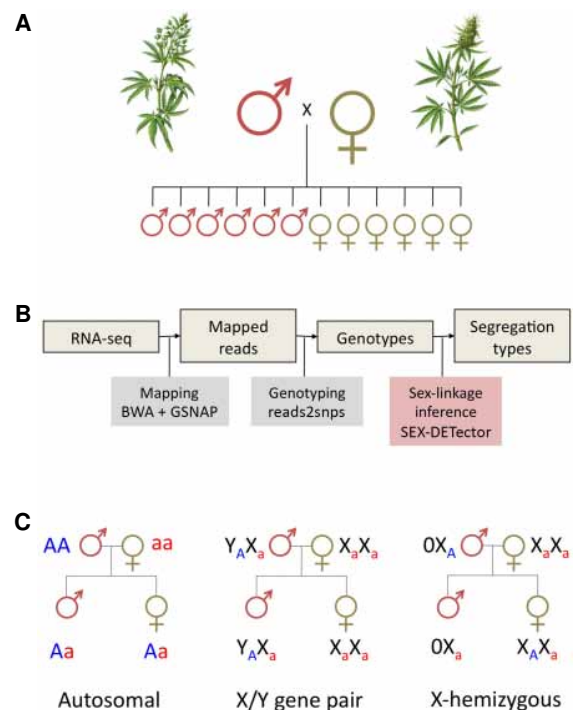
The Cannabaceae and related families (Urticaceae, Moraceae) derive from a dioecious common ancestor (Zhang et al. 2019). Despite being of similar size (Divashuk et al. 2014), the sex chromosomes of *C. sativa* could thus be much older than those of the species studied so far. Here, we used a recently developed statistical tool to identify X- and Y-linked genes, SEX-DETECTOR (Muyle et al. 2016). We applied SEX-DETECTOR to *C. sativa*, inferred sex-linked genes, and used those genes to (1) identify the sex chromosomes of *C. sativa* in an available reference genome assembly, and (2) characterize the *C. sativa* XY system and compare it to other plant systems.

## Results

### Identifying sex-linked genes in *C. sativa*

SEX-DETECTOR requires genotyping data from a cross (two parents and a few offspring individuals) (see Fig. 1). As explained in Muyle et al. (2016), patterns of allele transmission from parents to progeny differ for an autosomal or a sex-linked gene. For example, an allele only transmitted from father to sons is clearly indicative of a Y-linked allele. SEX-DETECTOR relies on a probabilistic model that accounts for typical errors in genotyping data and is used to compute, for each gene, the probability of autosomal and sex-linked segregation types. This key feature of SEX-DETECTOR makes it better at making inferences about segregation type than an empirical approach relying on data filtering to remove genotyping errors would do (better sensitivity, similar specificity).

More than 576 million 50-bp single-end reads of the parents and five male and five female offspring were mapped to the refer-



**Figure 1.** Experimental design and bioinformatic pipeline to identify sex-linked genes. (A, B) The SEX-DETECTOR analysis relies on obtaining genotyping data from a cross (parents + F1 progeny). (C) SEX-DETECTOR infers the segregation type based on how alleles are transmitted from parents to offspring. Three segregation types are included: autosomal (alleles of the parents are transmitted to the progeny the same way in both sexes, in a Mendelian way), XY (one allele of the father—the Y allele—is transmitted exclusively to sons), X-hemizygous (the single allele of the father is transmitted exclusively to daughters; the sons get one allele from the mother only). See Methods for more information. *C. sativa* male and female plants pencil illustration by annarepp/Shutterstock.com.

ence transcriptome of van Bakel et al. (2011), and all individuals were genotyped (see Methods). From these data, 11,515 genes were inferred as autosomal and 565 as sex-linked (i.e., 4.6% of the genes for which SEX-DETECTOR produced an assignment). The latter included 347 XY gene pairs and 218 X-hemizygous genes (i.e., genes lacking Y copies) (see Methods and Table 1).

### Identifying the sex chromosome pair in *C. sativa*

A total of 363 sex-linked genes (out of the 555 that we could map) mapped to Chromosome 1 in the reference genome of Grassa et al. (2018): 166 out of 340 XY gene pairs (48.8%) and 197 out of 215 X-hemizygous genes (91.6%) (Fig. 2). This indicates that Chromosome pair 1 is the sex chromosome pair. The remaining 192 sex-linked genes that could be mapped (i.e., 35% of all sex-linked genes) mapped to other chromosomes. Whether these genes are likely to be false positives or not is discussed below and in Supplemental Text S1. Note that, for the remaining analyses, we calculate statistics on all sex-linked genes as well as on the sex-linked genes from Chromosome 1 only.

Sex chromosomes typically have nonrecombining regions in which the synonymous divergence between the X and Y copies of a sex-linked gene (also called gametologs) can be substantial (Charlesworth 2015; Muyle et al. 2017). Using the sex-linked

**Table 1.** Summary of the results of the SEX-DETECTOR analysis

	Numbers
All genes <sup>a</sup>	30,074
Genes with at least one SNP detected, used for SEX-DETECTOR analysis	28,456
Genes with undetermined segregation type <sup>b</sup>	16,381
Autosomal genes	11,510
All sex-linked genes	565
XY gene pairs	347
X-hemizygous genes	218
Estimated Y gene loss rate	70%

<sup>a</sup>Transcripts from gene annotation of the reference genome (van Bakel et al. 2011).

<sup>b</sup>All posterior probabilities < 0.8, or absence of SNPs without errors.

SNPs inferred by SEX-DETECTOR, we are able to quantify the synonymous divergence ( $d_s$ ) between X and Y copies. The  $d_s$  reaches 0.4 in the two most divergent XY gene pairs (Fig. 3A); furthermore, most XY gene pairs with high X-Y  $d_s$  values mapped to a part of Chromosome 1. Two regions can be distinguished on this chromosome (Fig. 2): region 1 (from 30 to 105 Mb) where the XY gene pairs with the highest  $d_s$  values are found (mean X-Y  $d_s$  = 0.079, top 5% X-Y  $d_s$  = 0.32, top 10% X-Y  $d_s$  = 0.28) and where 58.6% of the sex-linked genes in the region are X-hemizygous, and region 2 (from 1 to 30 Mb) including mainly autosomal genes (791 genes, i.e., 96.1% of the genes in this region), in which the genes inferred as XY gene pairs show little divergence (mean X-Y  $d_s$  = 0.014, top 5% X-Y  $d_s$  = 0.05, top 10% X-Y  $d_s$  = 0.04), and few X-hemizygous genes are present (only 9.3% of the sex-linked genes). These observations suggest region 1 is the X-specific region (not recombining in males) and region 2 the pseudo-autosomal region (= PAR, still recombining in males).

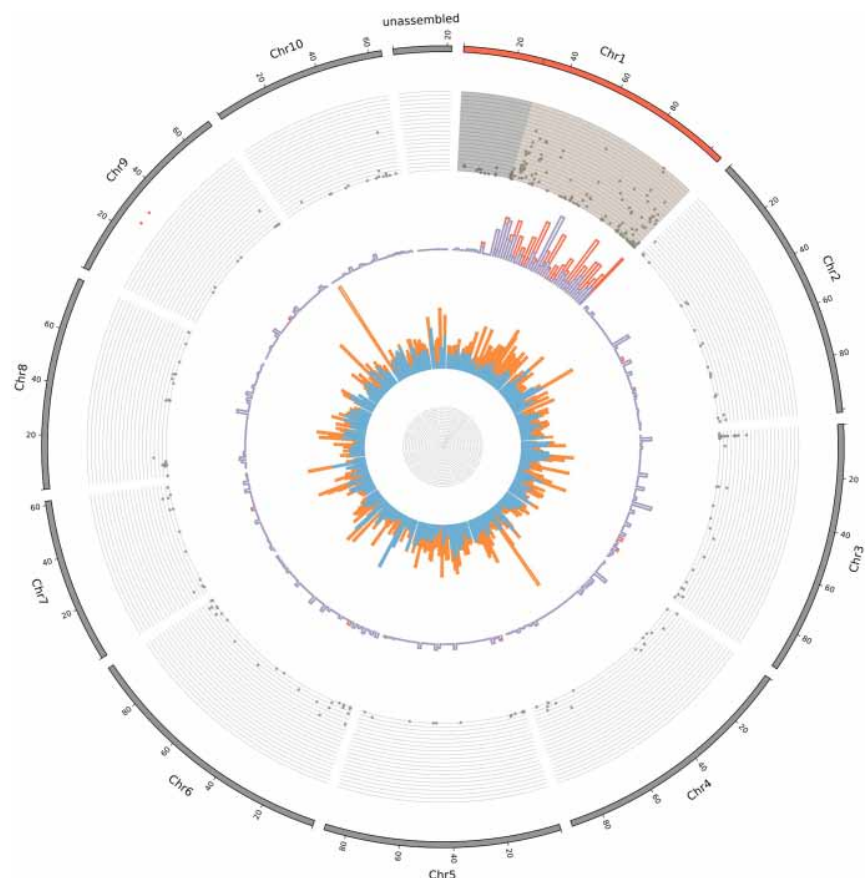
### Age of the *C. sativa* sex chromosome system

We then used the 565 sex-linked genes to study the evolution of sex chromosomes in *C. sativa*. First, we used the  $d_s$  values of the XY gametologs and different molecular clock estimates for plants to infer the age of the sex chromosomes on *C. sativa*. Using the maximum observed  $d_s$  value (0.4), we estimated that recombination suppression between X and Y Chromosomes was initiated 26.7–28.6 million years (myr) ago in *C. sativa*. If we use the  $d_s$  values of the 5% or 10% most divergent gene pairs to be more conservative when estimating the maximum X-Y divergence, we obtain more recent ages for the initial recombination suppression (17.3–20 myr old using the top 5% X-Y  $d_s$  values; 12–18.6 myr old using the top 10% X-Y  $d_s$  values) (see Table 2).

### Degeneration of the Y Chromosome and dosage compensation in *C. sativa*

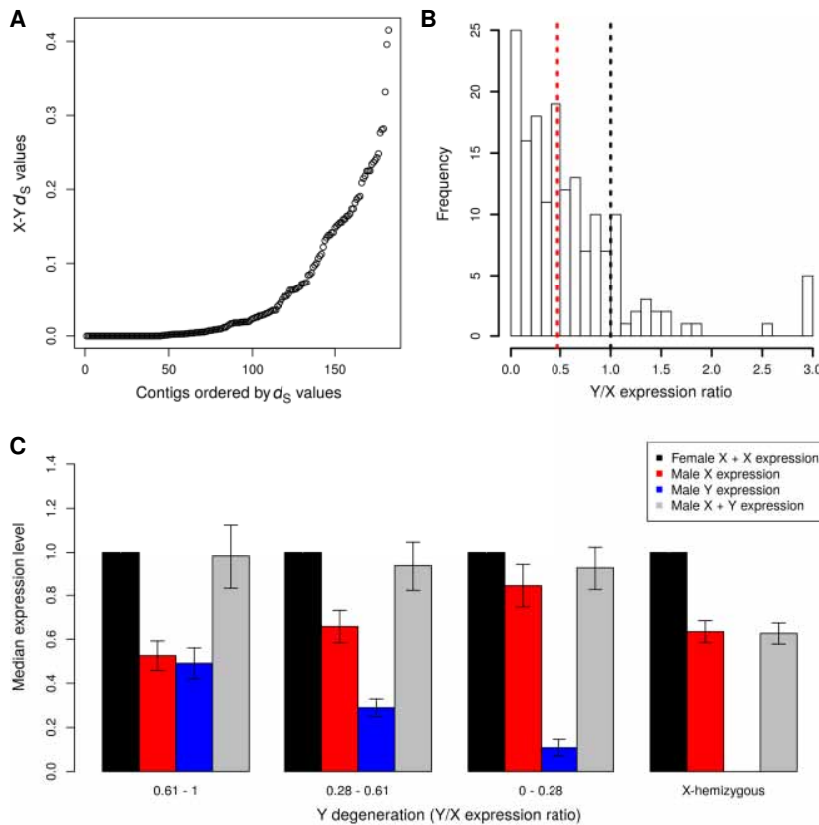
Second, we studied the extent of Y degeneration in *C. sativa* and estimated gene loss using the X-hemizygous genes. This measure of Y gene loss is, of course, a rough estimate as it reflects both true loss and simply the absence of expression of the Y copy in flower buds (Bergero and Charlesworth 2011; Bergero et al. 2015). SNP-based methods, such as ours, underestimate the number of X-hemizygous genes with respect to XY gene pairs, as X-hemizygous genes can only be detected when there is polymorphism in the X. To correct for this, we compared the number of X-hemizygous genes (218) and the XY gene pairs with polymorphism in the X copy (89), and we found that ~70% of the Y-linked genes may have been lost in *C. sativa*. The results were similar when focusing on sex-linked genes found on Chromosome 1 only (72.5%).

To further study Y degeneration, we focused on the expression of the sex-linked genes. Allele-specific expression analysis at the XY gene pairs revealed a median Y/X expression ratio of 0.50 overall (347 genes) and 0.47 for Chromosome 1 genes only (166 genes) (see Fig. 3B), much lower than the expected 1.0 value in the case of equal Y/X expression (i.e., no Y degeneration). We found some evidence for dosage compensation, as in males expression of



**Figure 2.** Distribution of the sex-linked and sex-biased genes onto the *C. sativa* reference genome. From outer to inner rings: (1) Chromosomes from 1 to 10 and unassembled scaffolds of the reference genome (Grassa et al. 2018); (2) X-Y  $d_s$  values (from 0 to 0.4); (3) proportion of XY-linked genes (in blue) and X-hemizygous genes (in red) in 2-Mb windows; (4) proportion of genes with sex-biased expression in 2-Mb windows: male-biased (light blue), female-biased (orange). *THCAD5* and *CBDAC* genes found in Grassa et al. (2018) are indicated by two red dots near the outer ring.





**Figure 3.** Patterns of molecular evolution of *C. sativa* sex chromosomes. (A) X-Y  $d_S$  values for the XY gene pairs. (B) Y/X expression ratio for the XY gene pairs; the black dotted line shows the expected value for no Y degeneration, the red dotted line shows the median observed here (median = 0.47). Both are significantly different (Wilcoxon paired test  $P$ -value  $< 10^{-16}$ ). (C) Dosage compensation in *C. sativa*. The expression levels of the X and Y alleles in males and females are shown for gene categories (from left to right, categories 0.61-1 and 0.28-0.61:  $N = 44$ , category 0-0.28:  $N = 43$ , X-hemizygous:  $N = 184$ ) with different levels of Y degeneration (measured by the Y/X expression ratio). Sex-biased genes (with strong and significant differences in male and female expression) have been removed, as they are not expected to exhibit dosage compensation (see Muyle et al. 2012, 2018). Only sex-linked genes mapping to Chromosome 1 have been included here. Supplemental Figure S1 shows the same analyses with all sex-linked genes.

X was increased when expression of Y was reduced (Fig. 3C). The results were unchanged when using all inferred sex-linked genes or only those found on Chromosome 1 (Fig. 3; Supplemental Fig. S1).

**Genomic distribution of the sex-biased genes in *C. sativa***

Of the genes expressed in flower buds, 15.7% are differentially expressed between male and female individuals (sex-biased genes) (see Table 3; Supplemental Fig. S2). The male-biased genes are significantly more numerous than the female-biased genes (9.06% vs. 6.64%, Fisher’s exact test  $P$ -value  $< 10^{-16}$ ) (see Table 3), a pattern that is common in dioecious plants (Harkess et al. 2015; Zemp et al. 2016; Muyle et al. 2017; Cossard et al. 2019, but see Darolti et al. 2018; Sanderson et al. 2019). Sex-biased genes are distributed all over the *C. sativa* genome (see Fig. 2). The sex-linked genes were significantly enriched among the sex-biased genes (25.8%) compared with the autosomal genes (13.9%; Fisher’s exact test  $P$ -value =  $3.7 \times 10^{-13}$ ), again a very common pattern in dioecious plants (for review, see Muyle et al. 2017; see also Darolti et al. 2018; Sanderson et al. 2019).

**Discussion**

**Chromosome pair 1 is the sex chromosome pair in *C. sativa***

Using SEX-DETECTOR, we have been able to identify a large number of sex-linked genes with a moderate sequencing effort (576 millions of single-end 50-bp reads). While most of these sex-linked genes were found to be located on Chromosome 1 of the assembly of Grassa et al. (2018), some mapped elsewhere on the reference genome. These probably include some false positives, but many are likely to result from assembly errors (see Supplemental Text S1). Nevertheless, we were able to clearly identify a chromosome pair (Chromosome 1 in the assembly of Grassa et al. 2018) as the sex chromosomes of *C. sativa*. We propose that the PAR is ~30 Mb large based on gene content and also taking into account the fact that SEX-DETECTOR tends to overestimate the size of the PAR (discussed in Muyle et al. 2016). Indeed, in a family setting, partially sex-linked pseudo-autosomal genes close to the pseudo-autosomal boundary can be inferred as fully sex-linked by SEX-DETECTOR. However, as these pseudo-autosomal genes still recombine normally, the  $d_S$  values between the X and Y alleles identified by SEX-DETECTOR should not exceed the genome-wide nucleotide polymorphism, which is around 1% in our data. Only from 30 Mb onward, the  $d_S$  values are above this value, leading us to consider the 0–30 Mb region as pseudo-autosomal. However, more data will be needed (e.g., sex-specific genetic maps) to define precisely the limit of the PAR.

**The sex chromosomes are the largest in the *C. sativa* genome**

This pair is the largest pair of the *C. sativa* genome, in agreement with cytogenetic data (Divashuk et al. 2014). This is frequent in plants with heteromorphic sex chromosomes (e.g., *S. latifolia*, *C. grandis*) (for review, see Muyle et al., 2017) but also in species with homomorphic chromosomes, such as papaya, where both

**Table 2.** Estimates of the age of the *C. sativa* sex chromosome system

	Age estimate using all XY gene pairs <sup>a</sup>	Age estimate using XY gene pairs on Chr 1 <sup>a</sup>
Maximum X-Y $d_S$ value	26.7–28.6	26.7–28.6
Top 5% X-Y $d_S$ values	17.3–18.6	18.7–20
Top 10% X-Y $d_S$ values	12–13	17.3–18.6

<sup>a</sup>Estimates obtained using two different molecular clocks (see Methods).

**Table 3.** List of sex-biased genes

Gene categories	Total	Female-biased expression	Male-biased expression	P-value <sup>a</sup>
All genes with sex biased-expression <sup>b</sup>	3483	1473	2010	<10 <sup>-16</sup>
Autosomal genes <sup>c</sup>	1599	725	874	2.1 × 10 <sup>-4</sup>
Sex-linked genes <sup>c</sup>	146	79	67	0.36
XY gene pairs <sup>c</sup>	87	34	53	0.053
X-hemizygous genes <sup>c</sup>	59	45	14	6.5 × 10 <sup>-5</sup>

<sup>a</sup>Exact binomial test, with a theoretical mean equal to 0.5.

<sup>b</sup>Among all 30074 *C. sativa* genes (see Table 1).

<sup>c</sup>Autosomal, sex-linked (XY and X-hemizygous) genes inferred by SEX-DETECTOR (see Table 1).

the X and Y increased in size due to the accumulation of repeats (Gschwend et al. 2012; Wang et al. 2012). We did not observe signs of such a process in *C. sativa*, as the gene density on the X Chromosome is similar to the gene density on the autosomes (32 genes/Mb vs. 33 genes/Mb). It is thus possible that the sex chromosomes are the largest in *C. sativa* simply because the sex-determining genes happened to evolve on the largest pair of chromosomes.

### *C. sativa* sex chromosomes are relatively old

Age estimates of the *C. sativa* sex chromosomes range from ~12 myr to ~29 myr old (Table 2). This is plausible, as dioecy probably is ancestral for the whole Cannabaceae family that diversified ~80 myr ago (Zhang et al. 2019). They may be the oldest sex chromosomes in plants for which the age was inferred from sequence data (Ming et al. 2011; Charlesworth 2015; Muyle et al. 2017). For instance, sex chromosomes are ~11 myr old in *S. latifolia* (Krasovec et al. 2018) and 8–16 myr old in two dioecious *Rumex* species (Crowson et al. 2017). However, only more precise molecular clocks for these plants and age estimates in more plant systems will give a precise picture on where the *C. sativa* sex chromosomes stand in the age distribution of plant sex chromosomes.

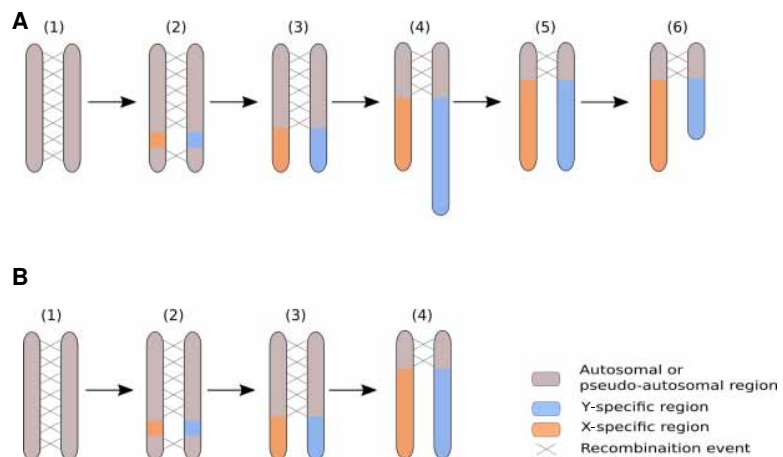
Further evidence that the *C. sativa* sex chromosomes are older than those of *S. latifolia* and *Rumex hastatulus* is the fact that the median Y/X expression ratio is ~0.5, much lower than what has been reported for the other species [~0.8 for *S. latifolia* (Bergero and Charlesworth 2011; Muyle et al. 2012) and ~0.8 for the old sex-linked genes *R. hastatulus* (Hough et al. 2014)]. Moreover, Y gene loss is about 70% in *C. sativa*, which is much higher than other species where Y gene loss has been estimated using the same methodology: ~40% for *S. latifolia* (Muyle et al. 2018; see also

Papadopulos et al. 2015) and 30% in *R. hastatulus* (Hough et al. 2014). In *R. rothschildianus*, gene loss amounts to ~90%, but the degeneration speed, not the age of the system, is believed to explain this observation (Crowson et al. 2017). Thus, the Y Chromosome of *C. sativa* seems more strongly degenerated than the Y Chromosomes of species with strong sex chromosome heteromorphy.

### Implications for the sex chromosome evolution model

Most of the plant sex chromosome systems that have been studied so far with genomic approaches either have small nonrecombining regions and homomorphic sex chromosomes (e.g., *Carica papaya*, *Asparagus officinalis*, *Diospyros lotus*) or have large nonrecombining regions and heteromorphic sex chromosomes, with the Y being larger than the X (e.g., *Silene latifolia*, *Coccinia grandis*). We here found that in a species with homomorphic sex chromosomes, the nonrecombining region is large, as it represents ~70% (75/105 Mb) of the *C. sativa* sex chromosomes (as suggested in Divashuk et al. 2014, based on cytogenetic data).

In the current scenario for the evolution of the sex chromosomes heteromorphy in plants (Ming et al. 2011; Charlesworth 2015; Muyle et al. 2017), it is unclear where these XY Chromosomes fit. Indeed, sex chromosome evolution in plants is thought to start with a small nonrecombining region on the Y Chromosome, which accumulates DNA repeats and tends to grow (Fig. 4). In papaya, the Y nonrecombining region is ~8 Mb large while the X homologous region is ~4 Mb (Wang et al. 2012). In some dioecious plants, DNA repeat accumulation in the Y nonrecombining region has been fast, and Y Chromosomes that are much larger than the X have evolved in *Silene latifolia* (Matsunaga et al. 1994) and *Coccinia grandis* (Sousa et al. 2013).



**Figure 4.** Revisiting the model for the evolution of plant sex chromosomes heteromorphy with *C. sativa*. (A) The current model for the evolution of plant sex chromosomes heteromorphy is as follows: (1) Sex chromosomes originate from autosomes on which sex-determining genes evolve; (2) the region encompassing the sex-determining genes stops recombining; (3) the non-recombining region grows larger due to additional events of recombination suppression; (4) the nonrecombining region of the Y Chromosome accumulates repeats and can become larger than the corresponding region on the X Chromosome; (5–6) the Y Chromosome undergoes large deletions and ultimately becomes smaller than the X Chromosome. Steps 1–4 have been previously documented in plants (e.g., Charlesworth et al. 2005; Ming et al. 2011; Muyle et al. 2017 for review), while steps 5–6 are speculative. Our study is supportive of this scenario if we assume that the *C. sativa* Y Chromosome has been larger in the past. (B) It is possible, however, that the accumulation of repeats has been slow in the Y Chromosome of the *C. sativa* lineage and that X and Y Chromosomes have always been of similar size. Here, step 4 does not imply the elongation of the Y Chromosome.

Either DNA repeat accumulation on the Y has been slow in the *C. sativa* lineage, or the Y used to be larger than it is today and has undergone genomic shrinking, a process that is reminiscent of the evolution of the sex chromosomes heteromorphy in animals (Ming et al. 2011; Bachtrog 2013), where old Y Chromosomes can be tiny compared to their X counterpart (Fig. 4). Distinct assemblies for the X and Y Chromosomes in *C. sativa* and also sequencing of other dioecious Cannabaceae species will help in testing this idea in the future.

## Methods

### Plant material, RNA extraction, and sequencing

One male and one female *C. sativa* plant (“Zenitsa” cultivar) were grown in controlled conditions in a greenhouse. A female was crossed with a male plant (controlled pollination). Seeds from this cross were sown to produce the F1. Flower buds (chosen because they are RNA-rich) of 3–5 d before expected flowering time (~1–3 mm) were sampled (5–7 buds per individual) from the parents and five offspring of each sex, as in Muyle et al. (2012). Total RNA was isolated from young flower buds using the RNeasy Plant Mini (Qiagen) plant isolation kit as recommended by the manufacturer. Isolated RNA was placed in RNastable tubes (Sigma-Aldrich). One library per individual was prepared. RNA-sequencing was conducted using the Complete Genomic (CG) technology, which provides 20 million ~50-bp single-end reads per sample (Liu et al. 2012). Two CG runs were done, and we obtained a mean of 48 million reads per individual (see Supplemental Table S1). Read quality was good (Phred score >35 for all reads), and no trimming was performed.

### Mapping, genotyping, and SEX-DETECTOR analysis

The SEX-DETECTOR analysis requires mapping the reads of the individuals to a reference transcriptome and performing SNP-calling to genotype all individuals for all expressed genes. Ideally, the reference transcriptome is from a female individual so that the X and Y reads map to the same transcript and XY SNPs can be identified by SEX-DETECTOR (Muyle et al. 2016). We extracted the 30,074 transcripts from the annotation of the 2011 complete genome from a Purple Kush female individual (van Bakel et al. 2011). The initial mapping analyses were done using BWA, allowing for five mismatches per read (version 0.7.15-r1140, `bwa aln -n 5` and `bwa samse`) (see Li and Durbin 2010). For comparison, an alternative mapping was performed with Bowtie 2 (version 2.1.0, `bowtie2-build` and `bowtie2 -x`) (see Langmead and Salzberg 2012), which yielded similar results. We used SAMtools (version 1.3.1, `samtools view -t output.fa -F 4 -h` and `samtools sort -m 2G`) (see Li et al. 2009) to remove unmapped reads and to prepare the files for the genotyping.

The genotyping was performed using reads2snp (version 2.0.64, `reads2snp -aeb -min 3 -par 0`) (see Gayral et al. 2013), as recommended by Muyle et al. (2016) (i.e., by accounting for allelic expression biases and without filtering for paralogous SNPs). Only SNPs supported by at least three reads were conserved for subsequent analysis (except in Supplemental Table S2).

We ran SEX-DETECTOR (`-system xy/zw/no_sex_chr -seq -detail -detail-sex-linked -L -SEM -thr 0.8`) (see Muyle et al. 2016) on genotyping data of the 12 individuals. SEX-DETECTOR uses a maximum likelihood approach to estimate the parameters of its model, which include several genotyping error parameters. The posterior probability of being autosomal ( $P_A$ ), XY ( $P_{XY}$ ), or X-hemizygous ( $P_{Xh}$ ) is computed for each SNP and for each transcript (combining the posterior probabilities of all SNPs) (see Muyle et al. 2016). A

transcript was inferred as sex-linked when its posterior probability of being either XY or X-hemizygous was  $\geq 0.8$  (i.e.,  $P_{XY} + P_{Xh} \geq 0.8$ ) and if at least one sex-linked SNP had no genotyping error; autosomal segregation was inferred similarly ( $P_A \geq 0.8$  and at least one autosomal SNP without genotyping error) (see Muyle et al. 2016). The remaining transcripts were considered undetermined and were not used for further analysis unless explicitly mentioned. To identify X-hemizygous genes among the sex-linked genes, we selected (1) the genes that have only X-hemizygous SNPs, of which at least one is without genotyping error, and (2) the genes that have no Y expression and at least one SNP without genotyping error. The second set of genes typically has mainly X-hemizygous SNPs and only a few X/Y SNPs with many Y genotyping errors. After averaging Y expression across all SNPs and individuals of these genes, Y expression is null. Only a few genes were added with step 2.

SEX-DETECTOR runs on the first mapping with BWA (and also Bowtie 2) yielded high Y genotyping error (YGE) parameter values, which could be the result of mapping errors of Y-linked reads (Muyle et al. 2016). The reference transcriptome used for mapping was derived from the genome of a female plant (van Bakel et al. 2011), which may result in a mapping bias against the Y-linked reads. To solve this problem, we used GSNAP (version 2017-11-15, `gsnap -m 5`) (see Wu and Nacu 2010), which can be used to map RNA-seq reads onto a divergent reference. GSNAP was thus used in a SNP-informed mode that adjusts read alignment onto a reference taking into account a user-provided list of SNPs that are not considered mismatches. For this procedure, we first mapped reads with BWA and collected all the SNPs present in SEX-DETECTOR's output, which were provided to GSNAP. We ran four iterations of GSNAP. For each iteration, SEX-DETECTOR detected new sex-linked SNPs, which were added to the list of SNPs provided to GSNAP. As expected, the Y genotyping error parameter value decreased from 0.84 with BWA to 0.07 with the fourth GSNAP iteration (Supplemental Table S2) and the mapping rate from 82.57% to 87% (Supplemental Table S1). All inferred sex-linked genes are available in Supplemental Table S3.

### Circular representations of location of sex-linked genes in the *C. sativa* genome

To map the sex-linked genes to the *C. sativa* genome, we used BLAST (Altschul et al. 1990) to find the best hit of each *C. sativa* transcript in the van Bakel et al. (2011) transcriptome on one of the recent reference genomes (`blastn -max_target_seqs 1 -max_hsps 1`). For this mapping, we used the *C. sativa* reference genome with the best assembly statistics (size = 875 Mb, 10 pseudomolecules, 220 scaffolds, N50 = 91 Mb) (see <https://www.ncbi.nlm.nih.gov/genome/genomes/11681> and Grassa et al. 2018), which was, however, unannotated. We then used Circos (version 0.69-6) (Krzywinski et al. 2009) for visualizing the location of sex-linked genes. We split each chromosome in windows of 2 Mb using BEDTools `makewindows` (version v2.26.0) (Quinlan & Hall 2010). BEDTools `intersect` (version v2.26.0, `-c` option) (Quinlan & Hall 2010) was used for computing proportions of sex-linked genes per window. Proportions of sex-linked genes were computed by dividing the number of XY gene pairs (or X-hemizygous genes) by the number of all genes (sex-linked, autosomal, and undetermined) that blasted in the same window. A similar analysis was done for sex-biased genes. A comparison of the genomes of Grassa et al. (2018) and Lavery et al. (2019) is also shown in the supplemental material (Supplemental Fig. S4). Chromosome 1 in the assembly of Grassa et al. (2018) apparently corresponds to Chromosome 10 in the assembly of Lavery et al. (2019). Note, however, that the assembly of Chromosome 10 in

Laverty et al. (2019) seems to be much less complete than that of Chromosome 1 in Grassa et al. (2018): Chromosome 10 is enriched in sex-linked genes, but many sex-linked genes fall in the unassembled scaffolds.

## Analysis of the sex-linked genes

### Y gene loss

To estimate the rate of gene loss in the Y Chromosome, we compared the number of XY gene pairs and the number of X-hemizygous genes, as in Bergero and Charlesworth (2011). Identifying XY gene pairs relies on fixed XY differences, while identifying X-hemizygous genes relies on X-polymorphism only, which makes detection of X-hemizygous genes less likely (see Bergero and Charlesworth 2011; Muyle et al. 2016). The Y gene loss proportion estimate was thus corrected for this bias as follows:

$$Y \text{ gene loss} = \frac{X\text{-hemizygous gene number}}{(X\text{-hemizygous gene number} + XY \text{ gene pair with X polymorphism number})}$$

### Values of synonymous divergence ( $d_s$ ) and age of the XY system

The X and Y open reading frame sequences were aligned using the translated reference transcripts to get reading-frame informed alignments. X-Y  $d_s$  values were obtained using codeml (PAML version 4.9) (see Yang 2007) in pairwise mode. To estimate the age of the *C. sativa* XY system, we considered maximum X-Y  $d_s$  values and used two different molecular clocks for plants:  $1.5 \times 10^{-8}$  substitutions/site/year (Koch et al. 2000) and  $7 \times 10^{-9}$  mutations/site/generation (Ossowski et al. 2010). We obtained the age of the XY system as follows:

$$\text{age (in years)} = \frac{d_s \text{max}}{\text{rate}},$$

using the molecular clock of Koch et al. (2000), and

$$\text{age (in number of generations)} = \frac{d_s \text{max}}{2\mu},$$

using the molecular clock of Ossowski et al. (2010).

The age in million years from the Ossowski et al. (2010) molecular clock was obtained assuming one generation per year in natural populations of *C. sativa* (which is a tall annual plant).

### Allele-specific expression analyses

We used allele-specific expression estimates at XY gene pairs provided by SEX-DETECTOR (Muyle et al. 2016) for the estimation of the Y/X expression ratio and patterns of dosage compensation (see Fig. 3B–C). These estimates relied on counting reads spanning XY SNPs only and were normalized using the total read number in a library for each individual. These estimates were further normalized by the median autosomal expression for each individual.

### Identifying sex-biased genes

As the differential gene expression analysis methods currently available vary in performance (Schurch et al. 2016; Costa-Silva et al. 2017), we chose to combine several methods. Analyses contrasting the gene expression level between our 12 male and female individuals were thus performed using three R packages: (1) DESeq2 version 1.10.1 (Love et al. 2014); (2) edgeR version 3.26.9 (Robinson et al. 2010), both relying on negative binomial distribution of read count modeling; and (3) *limma*-voom version

3.26.9 (Ritchie et al. 2015), based on log-normal distribution modeling to take into account the sampling variance of small read counts. Very lowly expressed genes were discarded from the analysis, keeping only genes covered by at least 10 reads in a minimum of two replicates. Using a FDR-adjusted *P*-value cut-off of 0.0001, we retained as sex-biased the genes that had significant differences in expression between males and females in at least two of the three methods (Supplemental Fig. S3).

### Statistics

All statistical tests and figures were done using R version 3.2.3 (R Core Team 2016).

### Data access

All RNA-seq data for the *C. sativa* samples generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA549804.

### Acknowledgments

We thank the BGI for free sequencing, thanks to their call for RNA-seq for medicinal plants. We thank Aline Muyle for advice with SEX-DETECTOR and discussions. We thank Dr. Tatyana Sukhorada, P. P. Lukyanenko Krasnodar Research and Development Institute of Agriculture for providing seeds of the *C. sativa* cultivar “Zenitsa.” We thank three anonymous referees and the editor for their useful comments that helped improve this manuscript. This work was performed using the computing facilities of the CC LBBE/PRABI; we thank Bruno Spataro and Stéphane Delmotte for cluster maintenance. This project was supported through Agence Nationale de la Recherche (ANR) grant ANR-14-CE19-0021-01 to G.A.B.M.

**Author contributions:** Conceptualization of the study: G.A.B.M. and G.K.; methodology: G.A.B.M. and G.K.; software: D.P., B.R., and H.B.; formal analysis: D.P.; investigation: D.P., O.R., B.R., H.B., H.H., J.K., G.K., and G.A.B.M.; resources: O.R., C.F., and G.K.; data curation: C.F.; writing—original draft: G.A.B.M., D.P., and J.K.; writing—review and editing: all authors; visualization: D.P.; supervision: G.A.B.M., J.K., and G.K.; project administration: G.A.B.M.; funding acquisition: G.A.B.M. and G.K.

### References

- Ahmed S, Cock JM, Pessia E, Luthringer R, Cormier A, Robuchon M, Sterck L, Peters AF, Dittami SM, Corre E, et al. 2014. A haploid system of sex determination in the brown alga *Ectocarpus* sp. *Curr Biol* **24**: 1945–1957. doi:10.1016/j.cub.2014.07.042
- Alexander SP. 2016. Therapeutic potential of cannabis-related drugs. *Prog Neuropsychopharmacol Biol Psychiatry* **64**: 157–166. doi:10.1016/j.pnpbp.2015.07.001
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410. doi:10.1016/S0022-2836(05)80360-2
- Andre CM, Hausman JF, Guerriero G. 2016. *Cannabis sativa*: the plant of the thousand and one molecules. *Front Plant Sci* **7**: 19. doi:10.3389/fpls.2016.00019
- Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* **14**: 113–124. doi:10.1038/nrg3366
- Barrett SC, Hough J. 2013. Sexual dimorphism in flowering plants. *J Exp Bot* **64**: 67–82. doi:10.1093/jxb/ers308
- Bergero R, Charlesworth D. 2009. The evolution of restricted recombination in sex chromosomes. *Trends Ecol Evol* **24**: 94–102. doi:10.1016/j.tree.2008.09.010

- Bergero R, Charlesworth D. 2011. Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. *Curr Biol* **21**: 1470–1474. doi:10.1016/j.cub.2011.07.032
- Bergero R, Qiu S, Charlesworth D. 2015. Gene loss from a plant sex chromosome system. *Curr Biol* **25**: 1234–1240. doi:10.1016/j.cub.2015.03.015
- Charlesworth D. 2015. Plant contributions to our understanding of sex chromosome evolution. *New Phytol* **208**: 52–65. doi:10.1111/nph.13497
- Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. *Heredity (Edinb)* **95**: 118–128. doi:10.1038/sj.hdy.6800697
- Cossard GG, Toups MA, Pannell JR. 2019. Sexual dimorphism and rapid turnover in gene expression in pre-reproductive seedlings of a dioecious herb. *Ann Bot* **123**: 1119–1131. doi:10.1093/aob/mcy183
- Costa-Silva J, Domingues D, Lopes FM. 2017. RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS One* **12**: e0190152. doi:10.1371/journal.pone.0190152
- Crowson D, Barrett SCH, Wright SI. 2017. Purifying and positive selection influence patterns of gene loss and gene expression in the evolution of a plant sex chromosome system. *Mol Biol Evol* **34**: 1140–1154. doi:10.1093/molbev/msx064
- Darolti I, Wright AE, Pucholt P, Berlin S, Mank JE. 2018. Slow evolution of sex-biased genes in the reproductive tissue of the dioecious plant *Salix viminalis*. *Mol Ecol* **27**: 694–708. doi:10.1111/mec.14466
- Divashuk MG, Alexandrov OS, Razumova OV, Kirov IV, Karlov GI. 2014. Molecular cytogenetic characterization of the dioecious *Cannabis sativa* with an XY chromosome sex determination system. *PLoS One* **9**: e85118. doi:10.1371/journal.pone.0085118
- Gayral P, Melo-Ferreira J, Glémin S, Bierre N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, et al. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genet* **9**: e1003457. doi:10.1371/journal.pgen.1003457
- Grassa CJ, Wenger JP, Dabney C, Poplawski SG, Motley ST, Michael TP, Schwartz C, Weiblen GD. 2018. A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. bioRxiv doi:10.1101/458083
- Gschwend AR, Yu Q, Tong EJ, Zeng F, Han J, VanBuren R, Aryal R, Charlesworth D, Moore PF, Paterson AH, et al. 2012. Rapid divergence and expansion of the X chromosome in papaya. *Proc Natl Acad Sci* **109**: 13716–13721. doi:10.1073/pnas.1121096109
- Harkess A, Mercati F, Shan HY, Sunseri F, Falavigna A, Leebens-Mack J. 2015. Sex-biased gene expression in dioecious garden asparagus (*Asparagus officinalis*). *New Phytol* **207**: 883–892. doi:10.1111/nph.13389
- Hobza R, Cegan R, Jesionek W, Kejnovsky E, Vyskot B, Kubat Z. 2017. Impact of repetitive elements on the Y chromosome formation in plants. *Genes (Basel)* **8**: 302. doi:10.3390/genes8110302
- Hough J, Hollister JD, Wang W, Barrett SC, Wright SI. 2014. Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*. *Proc Natl Acad Sci* **111**: 7713–7718. doi.org/10.1073/pnas.1319227111
- Koch M, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabidopsis* and related genera (Brassicaceae). *Mol Biol Evol* **17**: 1483–1498. doi:10.1093/oxfordjournals.molbev.a026248
- Krasovec M, Chester M, Ridout K, Filatov DA. 2018. The mutation rate and the age of the sex chromosomes in *Silene latifolia*. *Curr Biol* **28**: 1832–1838.e4. doi:10.1016/j.cub.2018.04.069
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645. doi:10.1101/gr.092759.109
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359. doi:10.1038/nmeth.1923
- Laverty KU, Stout JM, Sullivan MJ, Shah H, Gill N, Holbrook L, Deikus G, Sebra R, Hughes TR, Page JE, et al. 2019. A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the *THC/CBD acid synthase* loci. *Genome Res* **29**: 146–156. doi:10.1101/gr.242594.118
- Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**: 589–595. doi:10.1093/bioinformatics/btp698
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. doi:10.1093/bioinformatics/btp352
- Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of next-generation sequencing systems. *Biomed Res Int* **2012**: 251364. doi:10.1155/2012/251364
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550. doi:10.1186/s13059-014-0550-8
- Matsunaga S, Hizume M, Kawano S, Kuroiwa T. 1994. Cytological analyses in *Melandrium album*: genome size, chromosome size and fluorescence *in situ* hybridization. *Cytologia (Tokyo)* **59**: 135–141. doi:10.1508/cytologia.59.135
- McVey E. 2017. U.S. marijuana industry's economic impact to approach \$70B by 2021. In *Marijuana Business Daily*. Published June 12, 2017. https://mjbizdaily.com/chart-u-s-marijuana-industrys-economic-impact-approach-70b-2021/
- Ming R, Bendahmane A, Renner SS. 2011. Sex chromosomes in land plants. *Annu Rev Plant Biol* **62**: 485–514. doi:10.1146/annurev-arplant-042110-103914
- Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais GA. 2012. Rapid de novo evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *PLoS Biol* **10**: e1001308. doi:10.1371/journal.pbio.1001308
- Muyle A, Käfer J, Zemp N, Mousset S, Picard F, Marais GA. 2016. SEX-DETECTOR: a probabilistic approach to study sex chromosomes in non-model organisms. *Genome Biol Evol* **8**: 2530–2543. doi:10.1093/gbe/evw172
- Muyle A, Shearn R, Marais GA. 2017. The evolution of sex chromosomes and dosage compensation in plants. *Genome Biol Evol* **9**: 627–645. doi:10.1093/gbe/evw282
- Muyle A, Zemp N, Fruchard C, Cegan R, Vrana J, Deschamps C, Tavares R, Hobza R, Picard F, Widmer A, et al. 2018. Genomic imprinting mediates dosage compensation in a young plant XY system. *Nat Plants* **4**: 677–680. doi:10.1038/s41477-018-0221-y
- Offord C. 2018. UK to legalize medicinal *Cannabis*. In *The Scientist*. Published July 27, 2018. https://www.the-scientist.com/news-opinion/uk-to-legalize-medicinal-cannabis-64574
- Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**: 92–94. doi:10.1126/science.1180677
- Palmer DH, Rogers TF, Dean R, Wright AE. 2019. How to identify sex chromosomes and their turnover. *Mol Ecol* **28**: 4709–4724. doi:10.1111/mec.15245
- Papadopulos AS, Chester M, Ridout K, Filatov DA. 2015. Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proc Natl Acad Sci* **112**: 13021–13026. doi:10.1073/pnas.1508454112
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. doi:10.1093/bioinformatics/btq033
- R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/
- Renner SS. 2014. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Am J Bot* **101**: 1588–1596. doi:10.3732/ajb.1400196
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**: e47. doi:10.1093/nar/gkv007
- Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140. doi:10.1093/bioinformatics/btp616
- Salentijn EM, Petit J, Trindade LM. 2019. The complex interactions between flowering behavior and fiber quality in hemp. *Front Plant Sci* **10**: doi:10.3389/fpls.2019.00614
- Sanderson BJ, Wang L, Tiffin P, Wu Z, Olson MS. 2019. Sex-biased gene expression in flowers, but not leaves, reveals secondary sexual dimorphism in *Populus balsamifera*. *New Phytol* **221**: 527–539. doi:10.1111/nph.15421
- Schultes RE, Klein WM, Plowman T, Lockwood TE. 1974. *Cannabis*: an example of taxonomic neglect. *Bot Mus Leaf Harv Univ* **23**: 337–367. doi:10.1515/9783110812060.21
- Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**: 839–851. doi:10.1261/rna.053959.115
- Small E. 2015. Evolution and classification of *Cannabis sativa* (marijuana, hemp) in relation to human utilization. *Bot Rev* **81**: 189–294. doi:10.1007/s12229-015-9157-3
- Sousa A, Fuchs J, Renner SS. 2013. Molecular cytogenetics (FISH, GISH) of *Coccinia grandis*: a ca. 3 myr-old species of Cucurbitaceae with the largest Y/autosome divergence in flowering plants. *Cytogenet Genome Res* **139**: 107–118. doi:10.1159/000345370
- Sousa A, Bellot S, Fuchs J, Houben A, Renner SS. 2016. Analysis of transposable elements and organellar DNA in male and female genomes of a species with a huge Y chromosome reveals distinct Y centromeres. *Plant J* **88**: 387–396. doi:10.1111/tj.13254

- Techen N, Chandra S, Lata H, Elshohly MA, Khan IA. 2010. Genetic identification of female *Cannabis sativa* plants at early developmental stage. *Planta Med* **76**: 1938–1939. doi:10.1055/s-0030-1249978
- Toups MA, Hahn MW. 2010. Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. *Genetics* **186**: 763–766. doi:10.1534/genetics.110.118794
- van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, Page JE. 2011. The draft genome and transcriptome of *Cannabis sativa*. *Genome Biol* **12**: R102. doi:10.1186/gb-2011-12-10-r102
- Veltsos P, Ridout KE, Toups MA, González-Martínez SC, Muyle A, Emery O, Rastas P, Hudzieczek V, Hobza R, Vyskot B, et al. 2019. Early sex-chromosome evolution in the diploid dioecious plant *Mercurialis annua*. *Genetics* **212**: 815–835. doi:10.1534/genetics.119.302045
- Vicoso B, Kaiser VB, Bachtrog D. 2013. Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proc Natl Acad Sci* **110**: 6453–6458. doi:10.1073/pnas.1217027110
- Wang J, Na JK, Yu Q, Gschwend AR, Han J, Zeng F, Aryal R, VanBuren R, Murray JE, Zhang W, et al. 2012. Sequencing papaya X and Y<sup>h</sup> chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc Natl Acad Sci* **109**: 13710–13715. doi:10.1073/pnas.1207833109
- Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873–881. doi:10.1093/bioinformatics/btq057
- Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**: 1586–1591. doi:10.1093/molbev/msm088
- Yazdi HP, Ellegren H. 2014. Old but not (so) degenerated—slow evolution of largely homomorphic sex chromosomes in ratites. *Mol Biol Evol* **31**: 1444–1453. doi:10.1093/molbev/msu101
- Yeager A. 2018. Canada could come to the fore in *Cannabis* research. In *The Scientist*. Published July 6, 2018. <https://www.the-scientist.com/news-opinion/canada-could-come-to-the-fore-in-cannabis-research-64455>
- Zemp N, Tavares R, Muyle A, Charlesworth D, Marais GA, Widmer A. 2016. Evolution of sex-biased gene expression in a dioecious plant. *Nat Plants* **2**: 16168. doi:10.1038/nplants.2016.168
- Zhang Q, Onstein RE, Little SA, Sauquet H. 2019. Estimating divergence times and ancestral breeding systems in *Ficus* and *Moraceae*. *Ann Bot* **123**: 191–204. doi:10.1093/aob/mcy159

Received July 31, 2019; accepted in revised form January 24, 2020.

# Supplementary material

## **List of supplementary items**

Text S1: study of the sex-linked genes that do not map onto the *C. sativa* Chromosome 1

Table S1: Sequencing and mapping statistics

Table S2: Mapping iterations and SEX-DETECTOR results.

Figure S1: patterns of molecular evolution of *C. sativa* sex chromosomes.

Figure S2: Log-log plot of male expression versus female expression for all the *C. sativa* genes.

Figure S3: Venn diagram showing the numbers of sex-biased genes found by DESeq2, EdgeR and limma-voom.

Figure S4: correspondence between the assemblies of Grassa et al. (2018) and Laverty et al. (2019).

**Text S1: study of the sex-linked genes that do not map onto the *C. sativa* Chromosome 1**

192 sex-linked genes (i.e 35% of all sex-linked genes) mapped to chromosomes other than Chromosome 1 in the assembly of Grassa et al (2018). Here we tried to understand this observation. A first possibility is that these genes are false positives of the SEX-DETECTOR method. SEX-DETECTOR's rate of false positives was found to be low in previous work (<5%, e.g. Muyle et al. 2016, 2018; Martin et al. 2019). However, we used permissive mapping here and this could increase the rate of false positives compared to previous work. To check this possibility, we used more stringent criteria to assign a gene as sex-linked. If the sex-linked genes mapping outside Chromosome 1 (hereafter called “atypical sex-linked genes”) are false positives, their number should go down, as they most likely do not match the criteria for sex-linkage less perfectly than the real sex-linked genes. However, more stringent criteria resulted in similar numbers and % of atypical sex-linked genes, except for one conservative criterion (see table below).

**Table:** number of sex-linked genes using different criteria. We show the number of sex-linked genes in and out of Chromosome 1 using the criterion indicated in the main text: probability of being XY (P\_XY) + probability of being X-hemizygous (P\_Xh)  $\geq 0.8$  (i.e. SEX-DETECTOR's default criteria) and other more stringent criteria.

	P_XY + P_Xh $\geq 0.8^*$	P_XY or P_Xh $\geq 0.8$	P_XY or P_Xh $\geq 0.9$	P_XY + P_Xh $\geq 0.8$ and no SNP with error
All mapped sex-linked genes	555	447	321	342
Sex-linked genes mapping to Chromosome 1**	363 (65%)	266 (60 %)	216 (67 %)	259 (76%)
Sex-linked genes mapping to other chromosomes**	192 (35%)	181 (40 %)	105 (33 %)	83 (24%)

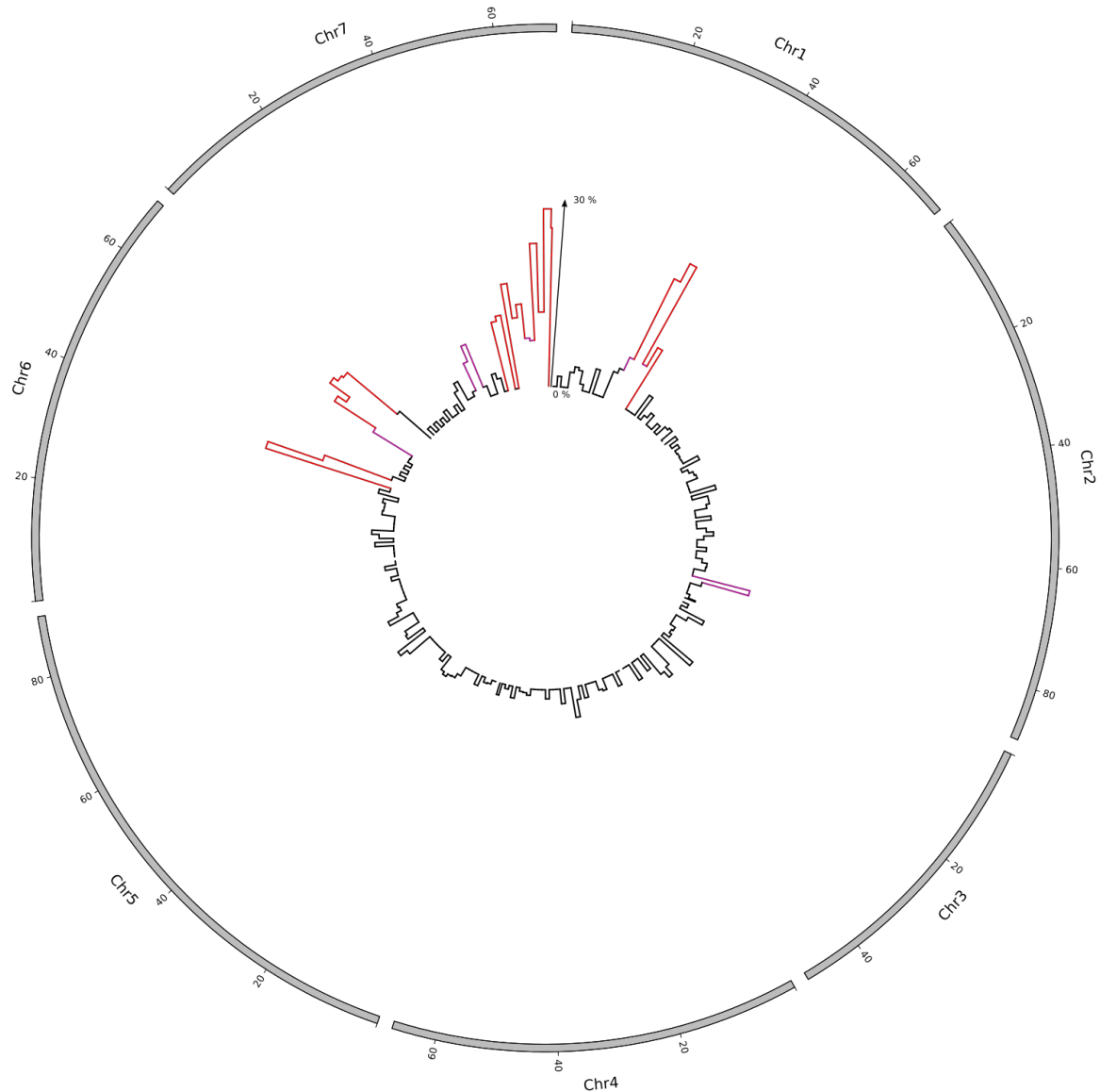
\* standard SEX-DETECTOR parameters (see Methods)

\*\* numbers and (%) are indicated

Another possibility that we explored is that the atypical sex-linked genes are true sex-linked genes, which are misplaced on the *C. sativa* assembly (i.e. they are incorrectly placed on other chromosomes than Chromosome 1). To check this, we mapped the *C. sativa* genes onto the *Rosa chinensis* reference genome (Raymond et al. 2018). *R. chinensis* has undergone few genome rearrangements since the diversification of the Rosaceae, and it belongs to the same order (Rosales) as *C. sativa*. If the atypical sex-linked genes map to the same genomic regions of *R. chinensis* as do the typical sex-linked genes (those mapping to *C. sativa* Chromosome 1), it would suggest that atypical sex-linked genes are true sex-linked genes. We mapped all the *C. sativa* genes onto the *R. chinensis* reference genome following the procedure described in the Method section below.



**Figure:** Distribution of the genes homologous to the *C. sativa* sex-linked genes in the *R. chinensis* genome. *R. chinensis* chromosomes are shown. In red are the windows for which the density is greater than 10% and in purple, windows for which density is greater than 7% but lower than 10%. A scale from 0 to 30% indicates the % of genes homologous to the *C. sativa* sex-linked genes present in the windows.



In order to identify the regions of *R. chinensis* homologous to the *C. sativa* sex chromosomes (hereafter called the SL-homologous regions), the *R. chinensis* genome was split in 2 Mb windows. The windows with more than 10% or 7% of genes homologous to *C. sativa* sex-linked genes defined the SL-homologous regions. We detected from 38 Mb to 50 Mb of such regions, using 10% and 7% threshold respectively. In those regions, we found 265 genes homologous to *C. sativa* sex-linked genes (including 31 atypical genes) and 304 genes homologous to *C. sativa* sex-linked genes (including 43 atypical genes) using 10% and 8% threshold respectively. As expected, there are several SL-homologous regions (see figure above), due to chromosomal rearrangements that have occurred since both species diverged. The degree of synteny of the sex-linked gene homologs is however high. We then tested whether the atypical sex-linked genes are enriched in the SL-homologous regions, which is

expected if they are true sex-linked genes (see above). We found that atypical sex-linked genes are over-represented on the SL-homologous regions compared to non-sex-linked genes (chi2 test p-value:  $2.14 \times 10^{-9}$  for the 10% threshold and  $1.73 \times 10^{-14}$  for the 7% threshold).

In conclusion, the pool of genes detected as sex-linked but mapping outside Chromosome 1 in the assembly of Grassa et al (2018) is most likely constituted by false positives in our analysis (permissive mapping) and by incorrectly placed genes in the assembly. This last type of error could arise if these genes are products of recent duplications, or because the sex chromosomes are often more difficult to assemble than the rest of the genome. As shown in Figure S4, there is quite some disagreement between the most recently published assemblies of the *C. sativa* genome, and erroneous placement has likely occurred in both. We are currently not able to quantify how many genes fall in each category (false positives vs assembly errors). Future improvements of the *C. sativa* genome should be able to resolve this issue.

## Methods

We did a blastp of *C. sativa* translated transcriptome against proteome of *R. chinensis* (the r1.1 predicted proteome available at <https://lipm-browsers.toulouse.inra.fr/pub/RchiOBHm-V2/>): blastp -max\_target\_seqs 1 -max\_hsp 1 -outfmt 6 -evalue 0.001. After the blastp we obtained 22,123 hits, of which:

- 502 sex-linked genes (among which 165 did not blast on Chromosome 1 of *C. sativa*)
- 20,529 inferred as autosomal or with undetermined segregation type by SEX-DETECTOR (=non-sex-linked genes)

Circos plots were done similarly as indicated in the Methods section of the main text. The genome of *R. chinensis* was split in windows of 2 Mb to analyse the density of mapped *C. sativa* sex-linked genes along this genome (bedtools makewindows). The sex-linked and non-sex-linked densities were determined with bedtools intersect. Then, the percent of sex-linked genes for each windows was calculated by dividing the number of sex-linked by the number of sex-linked and non-sex-linked genes.

## References

- Muyle, A., Käfer, J., Zemp, N., Mousset, S., Picard, F., & Marais, G. A. (2016). SEX-DETECTOR: a probabilistic approach to study sex chromosomes in non-model organisms. *Genome biology and evolution*, 8(8), 2530-2543.
- Muyle, A., Zemp, N., Fruchard, C., Cegan, R., Vrana, J., Deschamps, C., ... & Marais, G. A. (2018). Genomic imprinting mediates dosage compensation in a young plant XY system. *Nature plants*, 4(9), 677.
- Martin, H., Carpentier, F., Gallina, S., Godé, C., Schmitt, E., Muyle, A., ... & Touzet, P. (2019). Evolution of young sex chromosomes in two dioecious sister plant species with distinct sex determination systems. *Genome biology and evolution*, 11(2), 350-361.
- Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., ... & Carrere, S. (2018). The Rosa genome provides new insights into the domestication of modern roses. *Nature Genetics*, 50(6), 772.

**Table S1: Sequencing and mapping statistics**

Individuals	Total read numbers	% of mapped reads with BWA	% of mapped reads with Bowtie2	% of mapped reads with GSNAP (iteration #4)
Mother	48 020 391	84.18	83.20	88.36
Daughter 1	47 970 499	84.74	83.75	88.79
Daughter 2	48 038 555	83.53	82.51	88.93
Daughter 3	48 009 889	84.86	83.80	88.57
Daughter 4	48 168 799	84.51	83.43	88.99
Daughter 5	47 996 302	84.86	83.90	87.75
Father	47 889 741	79.96	79.00	85.12
Son 1	48 048 469	83.56	82.54	87.79
Son 2	47 890 377	78.31	77.37	83.42
Son 3	48 031 219	82.77	81.73	87.11
Son 4	47 965 502	78.33	77.45	83.76
Son 5	48 015 625	81.19	80.15	86.00
Mean female - male difference	60 583	3.70	3.66	2.99

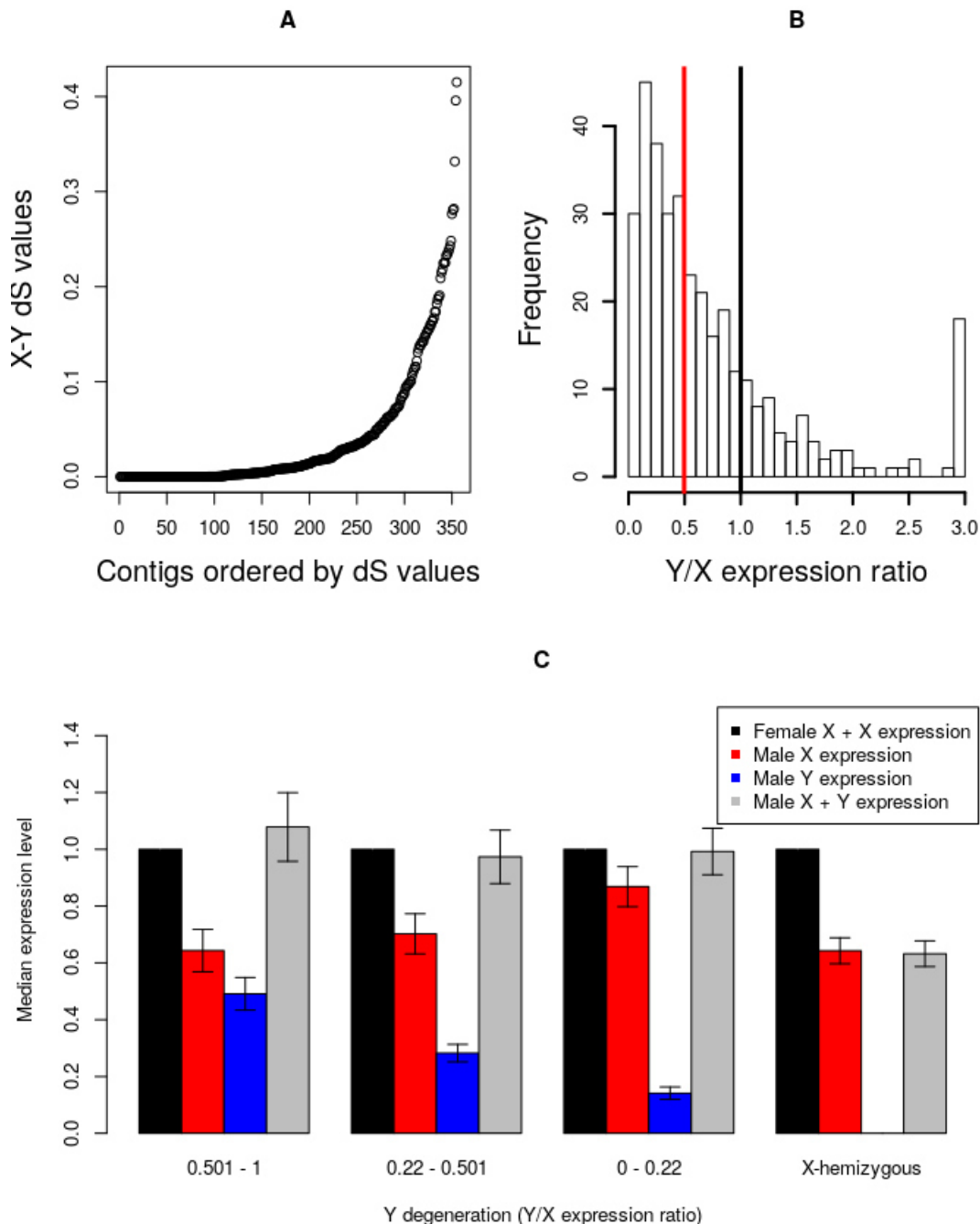
**Table S2: Mapping iterations and SEX-DETECTOR results.**

The Y genotyping error parameter (YGE) is shown for the different runs. YGE varies between 0 and 1; a high YGE value yields unreliable results and usually comes from mapping errors; acceptable values for YGE are around 5-10% (Muyle et al. 2016). Results are shown for gene sets with high to low minimum read numbers for SNP-calling; the former are supposed to comprise much less genotyping errors than the latter. For each run, the number of contigs for which SEX-DETECTOR produced an assignment (autosomal and sex-linked included) is indicated.

Minimum read# per SNP	BWA	GSNAP #1	GSNAP #2	GSNAP #3	GSNAP #4
150	0.19 (n=1345)	0.10 (n=1500)	0.05 (n=1533)	0.04 (n=1522)	0.03 (n=1527)
10	0.82 (n=2792)	0.79 (n=5074)	0.14 (n=9258)	0.07 (n=9532)	0.05 (n=9614)
3	0.84 (n=1790)	0.82 (n=3743)	0.82 (n=3976)	0.10 (n=11910)	0.07 (n=12081)

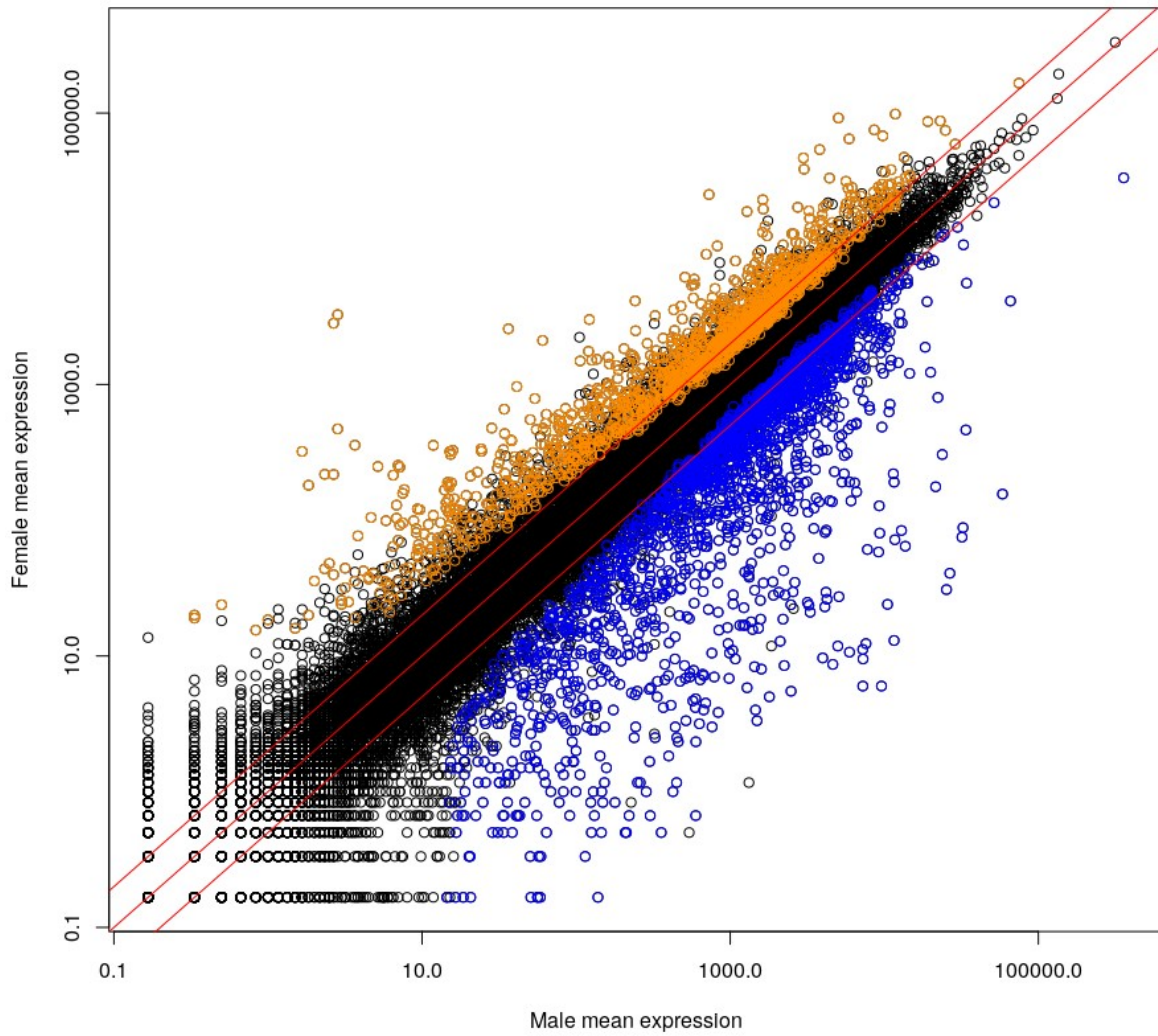
**Figure S1: patterns of molecular evolution of *C. sativa* sex chromosomes.**

A) X-Y  $d_s$  values for the XY gene pairs, B) Y/X expression ratio for the XY gene pairs, the black bar shows the expected value for no Y degeneration, the red bar shows the median observed here. Both are significantly different (Wilcoxon paired test  $p < 10^{-16}$ ). C) Dosage compensation in *C. sativa*. The expression levels of the X and Y alleles in males and females are shown for gene categories (n = 77 for each categories, except for X-hemizygous : n = 204) with different levels of Y degeneration (measured by the Y/X expression ratio). Sex-biased genes (with strong and significant differences in male and female expression) have been removed, as we do not expect them to exhibit dosage compensation (see Muyle et al. 2012, 2018). Sex-linked genes mapping to Chromosome 1 and elsewhere in the genome have been used to prepare this figure.



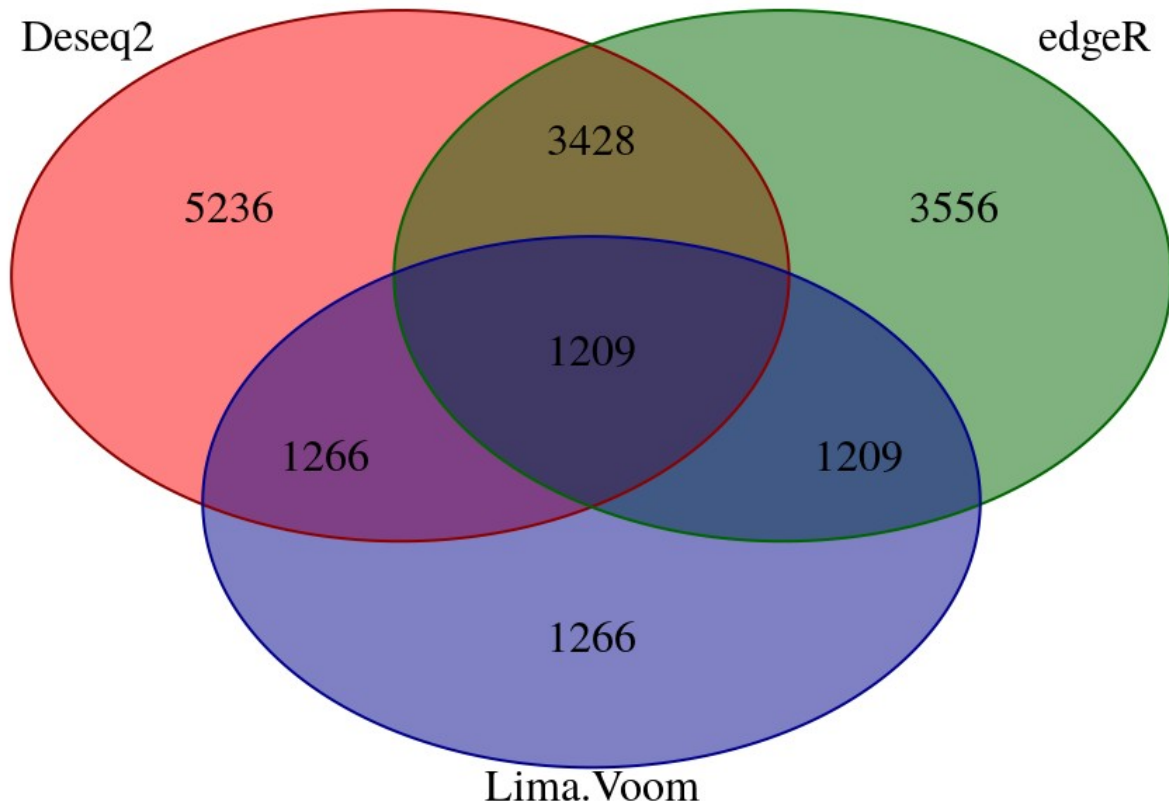
**Figure S2: Log-log plot of male expression versus female expression for all the *C. sativa* genes.**

The read lines indicate  $Y = X$  (central line) and two fold change. Genes with significant sex-biased expression are shown in orange (female-biased) or in blue (male-biased).



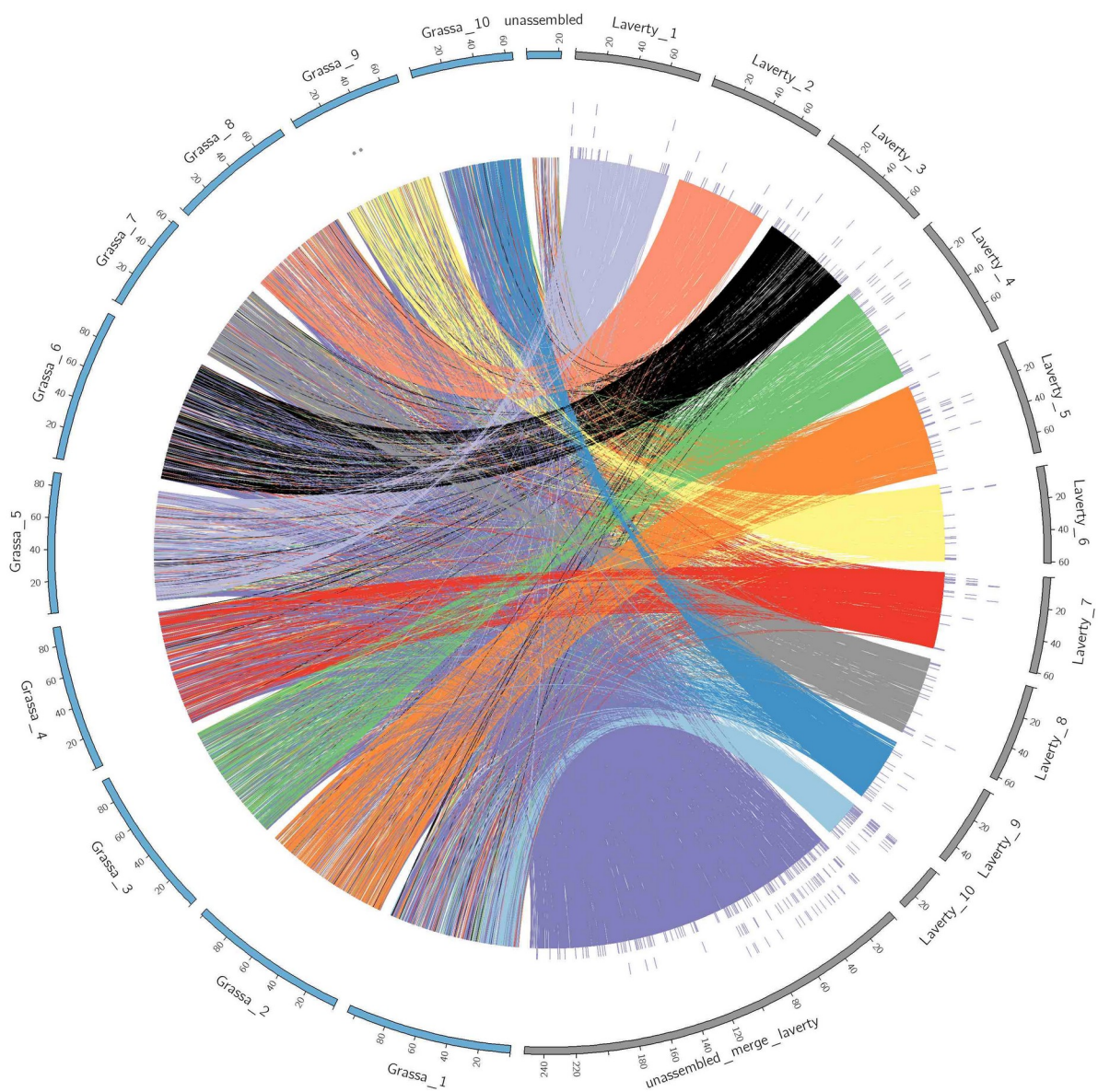
**Figure S3: Venn diagram showing the numbers of sex-biased genes found by DEseq2, EdgeR and limma-voom.**

Sex-biased the genes that had significant differences in expression between males and females in at least two of the three methods were retained.



**Figure S4: correspondence between the assemblies of Grassa et al. (2018) and Lavery et al. (2019).**

We blasted the 30,074 sequences derived from the draft genome of van Bakel et al. (2011) to both assemblies (blastn -max\_target\_seqs 1 -max\_hsp 1). For each gene, positions on both assemblies was used to make the links shown in the figure. Circos was used to represent graphically those links. Each Lavery et al. chromosome has a different color. The assembly of Grassa et al. is shown at the left and the one of Lavery et al. at the right. Sex-linked genes are indicated in the assembly of Lavery et al. by small bars; many sex-linked genes fall in the unassembled scaffolds in this assembly. Chromosome 1 in Grassa et al. corresponds to Chromosome 10 in Lavery et al. (one of the three candidates for being sex chromosomes mentioned in their paper). However, Chromosome 10 in Lavery et al. is much smaller than Chromosome 1 in Grassa et al., thus only part of the X Chromosome seems assembled in by Lavery et al. THCAS and CBDAS genes are indicated by two grey dots in the genome of Grassa et al.







# 4

## Chapter 4 : Development of genetic markers for sexing *Cannabis sativa* seedlings

Suite à l'identification de gènes XY chez *Cannabis sativa*, j'ai identifié des séquences génétiques suffisamment divergentes entre le X et le Y pour développer des amorces PCR Y-spécifiques. Dans un premier temps, des amorces Y-spécifiques ont été développées *in silico* avec l'aide de Hélène Henri (IE au laboratoire), puis ont été testées chez différents cultivars de chanvre par nos collaborateurs russes.

Parmi les amorces testées, 6 ont fourni des résultats positifs. Cependant, les tailles d'échantillons n'étaient pas assez conséquentes pour calculer un taux de sexage réussi de manière précise. De plus, les enjeux du sexage précoce de cannabis sont surtout pour les cultivars de producteurs de THC et de CBD plutôt que de chanvre. Nous avons donc écrit une note technique dans l'objectif de trouver des collaborateurs qui pourraient nous permettre de valider nos amorces sur de plus gros échantillons et idéalement sur de telles variétés.

Cette note technique n'a pas été envoyée pour publication dans un journal scientifique mais a été déposée sur biorxiv. Par conséquent, elle n'a pas été évaluée par les pairs. Le quatrième chapitre de cette thèse est une version reformatée du preprint présent sur biorxiv. Le contenu reste cependant inchangé. Nous prévoyons néanmoins de publier la description de ces amorces dans un journal une fois que les analyses auront été réalisées sur un grand nombre de cultivars.

# Development of genetic markers for sexing *Cannabis sativa* seedlings

Prentout D<sup>1</sup>, Razumova O<sup>2,2</sup>, Henri H<sup>1</sup>, Divashuk M<sup>2</sup>, Karlov G<sup>2</sup>, and Marais G.A.B<sup>1,4,\*</sup>

<sup>1</sup>Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Université de Lyon, Université Lyon 1, CNRS, Villeurbanne F-69622, France

<sup>2</sup>Laboratory of Applied Genomics and Crop Breeding, All-Russia Research Institute of Agricultural Biotechnology, Timiryazevskaya Str. 42, Moscow 127550, Russia

<sup>3</sup>N. V. Tsitsin Main Botanical Garden of Russian Academy of Sciences, Botanicheskaya Str., Moscow 127276, Russia

<sup>4</sup>Current adress: CIBIO, Biopolis, Campus de Vairao, Univ. Porto, Portugal

\*Equal contributions

## Abstract

*Cannabis sativa* is a dioecious plant with a XY system. Only females produce cannabinoids in large amount. Efficient male removal is an important issue for the cannabis industry. We have recently identified the sex chromosomes of *C. sativa*, which opens opportunities for developing universal genetic markers for early sexing of *C. sativa* plants. Here we selected six Y-linked markers and designed PCR primers, which were tested on five hemp cultivars both dioecious and monoecious. We obtained promising results, which need to be extended using a larger number of individuals and a more diverse set of cultivars, including THC producing ones.

## Introduction

*Cannabis sativa* is a dioecious plant with XY chromosomes (Divashuk et al., 2014). Only non-pollinated females produce THC and CBD (Small, 2015). Removing males as early as possible is thus a major issue for the THC/CBD industry. In *C. sativa*, however, sexual dimorphism is absent before the flowering stage as in many dioecious plants (Small, 2015; Barrett and Hough, 2013). Several methods have been developed to produce male-free crops (discussed in McKernan et al. (2020)). One of the most promising methods consists in identifying males before the flowering stage with Y-specific genetic markers, typically in *C. sativa* seedlings. A few genetic markers for sexing have been previously developed in *C. sativa* but accuracy or throughput need to be improved (discussed in Toth et al. (2020)). A recent study improved the throughput of MAD6, one of those genetic markers, using PACE - PCR Allele Competitive Extension (Törjék et al., 2002; Toth et al., 2020). The sex assay was conducted on 2,170 plants and 14 cultivars, and they correctly identified 98% of

the females and 100% of males (in a sub-sample of 270 plants). However, MAD6 and the other available markers are located in retrotransposons and they might not be present in all *C. sativa* cultivars. Despite all these advances, we still need to develop universal genetic markers for sexing *C. sativa* seedlings. In a recent study, we performed a genome-wide segregation analysis of *C. sativa* to identify sex-linked genes (Prentout et al., 2020). We ran SEX-DETECTOR on genotyping data of a *C. sativa* cross and identified SNPs that show sex-linkage when looking at allele transmission from parents to progeny (Muyle et al., 2016; Prentout et al., 2020). SEX-DETECTOR identified more >550 sex-linked genes (Prentout et al., 2020). Aligning those sex-linked genes onto a chromosome-level assembly of a *C. sativa* genome from Grassa et al. (2018), we found that the largest chromosome pair (number 1) was the sex chromosomes pair. Among the sex-linked genes that were identified by SEX-DETECTOR, 350 were XY gene pairs and the highest synonymous divergence between X and Y sequences reached 40%. These most strongly divergent XY gene pairs are interesting because they have a

very high chances to be present in all *C. sativa* populations/cultivars and thus offer a great opportunity to develop universal Y-linked genetic markers for early sexing of *C. sativa*. Here, we show the results on six such markers that have been tested on hemp cultivars.

## Methods

We used the XY gene pairs with the highest synonymous divergence identified in Prentout et al. (2020), for which CDS length was greater than 300 pb. We aligned the sequences with CLC workbench tool ‘create alignment’ (version 8.0.1, see <https://www.quiagenbioinformatics.com>) and uses the tool ‘Design Primer’ for primer design. For each gene pair, we aligned the Y-linked sequence from our cross (male  $\times$  female of hemp cultivar Zenitsa, see Prentout et al. (2020)) with three X-linked sequences, one of which was from our cross and the other two from a Purple Kush cultivar (THC producer) female and from a Finola cultivar (hemp) female (van Bakel et al., 2011). This increased the probability of detecting X/Y fixed differences (shared by all individuals of the species) instead of X/Y polymorphism (specific to some populations). To design the primers, we selected coding regions with high X-Y divergence. We kept those with at least two fixed differences between X-linked and Y-linked sequences, and avoided complementary forward and reverse primers to avoid association between our primers during PCR. We used the in silico PCR tool from the Van bakel lab (<http://genome.ccb.utoronto.ca/>) to test the size of the amplicons, and verify that a unique region of the genome was amplified. Our primers were tested in silico on both the Purple Kush and the Finola genomes. As both genomes are female genomes, if the X-linked primers resulted in amplification but the Y-

linked did not, we validated the primers as sufficiently divergent between the X-linked and Y-linked copies. We then tested our markers in vitro by PCR. Plant material was obtained from different cultivars and DNA was extracted. DNA isolation was performed from young leaves as described by Doyle and Doyle (1990) with some modifications. The extracting buffer contained 100 mM Tris-HCl (pH = 8.0), 20 mM EDTA (pH = 8.0), 2 M NaCl, 1.5% CTAB, 1.5% PVP and 0.2% B-mercaptoethanol. A 15 mM ammonium acetate solution in 75% ethanol was used for DNA washing. The PCR program for the all primers contained the following steps: 95 °C for 5 min followed by 35 cycles of 94 °C for 30 sec, 55 °C for 1 min, and 72 °C for 1 min and a final step of 72°C for 7 min. The markers were tested on three dioecious hemp cultivars (Zenitsa, Viktoria and Ekaterinodar) and two monoecious hemp cultivars (1147|16 and Maria). As the monoecious plants are XX (Razumova et al., 2014) the Y-linked markers primers are not supposed to amplify. They can thus serve as control.

## Results

Using the approach described in Methods, we selected thirteen XY gene pairs and twelve autosomal genes. We designed primer pairs for both X-linked and Y-linked sequences (XY gene pairs) for the autosomal sequence (autosomal genes). All the thirteen genetic Y-linked and twelve autosomal markers were validated in silico (see Methods). The primers were first tested in vitro on Zenitsa plant material (also used in Prentout et al. (2020)). From this preliminary test, we selected six Y-linked and one autosomal markers that amplified well with the primers that we designed and the PCR conditions that we set.

Table 1: Results are given for six Y-linked primer pairs and 1 autosomal (control) primer pair. Five different cultivars have been tested: three dioecious (Zenitsa, Viktoria, Ekaterinodar) and two monoecious (‘1447|16’ and Maria). Sample sizes are indicated with cultivars’ names. We had roughly 50:50 males/females in all cultivars. The percent of individuals that amplified is given for each marker in each cultivar. For monoecious cutlivars, plants are considered females because their genotype is XX. Therefore, a PCR in monoecious male is not applicable (na).

Cultivars	Y-linked marker 8		Y-linked marker 10		Y-linked marker 21		Y-linked marker 23		Y-linked marker 36		Y-linked marker 37		Autosomal marker 16	
	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Zenista (n = 26)	100	0	100	0	100	0	100	0	100	0	100	0	100	100
Viktoria (n = 8)	100	0	100	0	100	0	100	0	100	0	100	0	100	100
Ekaterinodar (n = 14)	80	11	100	0	100	0	0	0	80	11	80	0	100	100
1447 16 (XX) (n = 2)	na	0	na	0	na	0	na	0	na	0	na	0	na	100
Maria (XX) (n = 6)	na	0	na	0	na	0	na	0	na	0	na	0	na	100

1 shows the results of PCR experiments on those Y-linked and one autosomal markers in five different cultivars (three dioecious and two monoecious). Except for the Ekaterinodar cultivar, the Y-linked primers amplified in 100% of males and 0% of females and monoecious cultivars. The autosomal control markers amplified in all tested individuals. Amplicon size ranged from 174 to 305 bp.

## Discussion

We identified six very promising markers, which amplified only in males for the dioecious cultivars (except for Ekaterinodar) and did not amplify in monoecious cultivars. Results in Ekaterinodar were more ambiguous. We observed some amplification in one female for three markers: 8, 21 and 36 (see 1). Moreover, the markers 8, 21, 36 and 37 amplified only in 80% of the males for this cultivar. Marker 23 did not amplify in Ekaterinodar. Only marker 10 had amplification in 100% males and 0% females. It is unclear why results were not as good in this cultivar compared to others. It should be noted however that combining the results of the six markers identified correctly male and female plants in Ekaterinodar. More generally, sample sizes were small and the rates of success are thus only very rough. More tests are thus necessary. We need to test our markers on a much larger number of plants, from a much more diversified set of cultivars (e.g. including THC producers).

## References

Barrett, S. C. and Hough, J. (2013). Sexual dimorphism in flowering plants. *Journal of Experimental Botany*, 64(1):67–82.

Divashuk, M. G., Alexandrov, O. S., Razumova, O. V., Kirov, I. V., and Karlov, G. I. (2014). Molecular Cytogenetic Characterization of the Dioecious Cannabis sativa with an XY Chromosome Sex Determination System. *PLOS ONE*, 9(1):e85118. Publisher: Public Library of Science.

Grassa, C. J., Wenger, J. P., Dabney, C., Poplawski, S. G., Motley, S. T., Michael, T. P., Schwartz, C., and Weiblen, G. D. (2018). A complete Cannabis chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. preprint, Genomics.

McKernan, K. J., Helbert, Y., Kane, L. T., Ebling, H., Zhang, L., Liu, B., Eaton, Z., McLaughlin, S., Kingan, S., Baybayan, P., Concepcion, G., Jordan, M., Riva, A., Barbazuk, W., and Harkins, T. (2020). Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes. Technical report. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Muyle, A., Käfer, J., Zemp, N., Mousset, S., Picard, F., and Marais, G. A. (2016). SEX-DETECTOR: A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms. *Genome Biology and Evolution*, 8(8):2530–2543.

Prentout, D., Razumova, O., Rhoné, B., Badouin, H., Henri, H., Feng, C., Käfer, J., Karlov, G., and Marais, G. A. B. (2020). An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of Cannabis sativa. *Genome Research*, 30(2):164–172. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Small, E. (2015). Evolution and Classification of Cannabis sativa (Marijuana, Hemp) in Relation to Human Utilization. *The Botanical Review*, 81(3):189–294.

Toth, J. A., Stack, G. M., Cala, A. R., Carlson, C. H., Wilk, R. L., Crawford, J. L., Viands, D. R., Philippe, G., Smart, C. D., Rose, J. K. C., and Smart, L. B. (2020). Development and validation of genetic markers for sex and cannabinoid chemotype in Cannabis sativa L. *GCB Bioenergy*, 12(3):213–222. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcbb.12667>.

Törjék, O., Bucherna, N., Kiss, E., Homoki, H., Finta-Korpelová, Z., Bócsa, I., Nagy, I., and Heszky, L. E. (2002). Novel male-specific molecular markers (MADC5, MADC6) in hemp. *Euphytica*, 127(2):209–218.

van Bakel, H., Stout, J. M., Cote, A. G., Tallon, C. M., Sharpe, A. G., Hughes, T. R., and Page, J. E. (2011). The draft genome and transcriptome of Cannabis sativa. *Genome Biology*, 12(10):R102.



# 5






Chapter 5 : Plant genera  
*Cannabis* and *Humulus* share the  
same pair of well-differentiated  
sex chromosomes



À la suite des travaux sur les chromosomes sexuels de *Cannabis sativa*, nous avons voulu tester s'ils étaient homologues avec ceux de *Humulus lupulus*. Pour cela, nous avons débuté une collaboration avec une équipe slovène ayant la capacité de faire un croisement chez le houblon. Ce projet ayant vu le jour pendant ma thèse, j'ai été impliqué dès les premières réunions. J'ai donc participé à sa conception. Nos collaborateurs ont effectué le croisement et se sont chargé de le séquencer.

Ces travaux ont été principalement réalisé lors de la deuxième année de thèse et une partie de la troisième année à été consacrée à la rédaction de l'article, puis à sa publication en mai 2021 dans le journal *New Phytologist*. Cet article scientifique constitue le cinquième chapitre de ma thèse et été présenté lors d'une conférence internationale sur la diversité des chromosomes sexuels ainsi que pour des séminaires invité dans des laboratoires.

# Plant genera *Cannabis* and *Humulus* share the same pair of well-differentiated sex chromosomes

Djivan Prentout<sup>1</sup> , Natasa Stajner<sup>2</sup>, Andreja Cerenak<sup>3</sup>, Theo Tricou<sup>1</sup> , Celine Brochier-Armanet<sup>1</sup>, Jernej Jakse<sup>2</sup> , Jos Käfer<sup>1\*</sup>  and Gabriel A. B. Marais<sup>1,4\*</sup> 

<sup>1</sup>Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Université de Lyon, Université Lyon 1, CNRS, Villeurbanne F-69622, France; <sup>2</sup>Department of Agronomy, Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, Ljubljana SI-1000, Slovenia; <sup>3</sup>Slovenian Institute of Hop Research and Brewing, Cesta Zalskega Tabora 2, Zalec SI-3310, Slovenia; <sup>4</sup>LEAF- Linking Landscape, Environment, Agriculture and Food, Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa 1349-017, Portugal

## Summary

Author for correspondence:  
Djivan Prentout  
Email: djivan.prentout@univ-lyon1.fr

Received: 19 February 2021  
Accepted: 29 April 2021

New Phytologist (2021)  
doi: 10.1111/nph.17456

**Key words:** Cannabaceae, dioecy, dosage compensation, *Humulus lupulus*, sex chromosomes, Y degeneration.

- We recently described, in *Cannabis sativa*, the oldest sex chromosome system documented so far in plants (12–28 Myr old). Based on the estimated age, we predicted that it should be shared by its sister genus *Humulus*, which is known also to possess XY chromosomes.
- Here, we used transcriptome sequencing of an F<sub>1</sub> family of *H. lupulus* to identify and study the sex chromosomes in this species using the probabilistic method SEX-DETECTOR.
- We identified 265 sex-linked genes in *H. lupulus*, which preferentially mapped to the *C. sativa* X chromosome. Using phylogenies of sex-linked genes, we showed that a region of the sex chromosomes had already stopped recombining in an ancestor of both species. Furthermore, as in *C. sativa*, Y-linked gene expression reduction is correlated to the position on the X chromosome, and highly Y degenerated genes showed dosage compensation.
- We report, for the first time in Angiosperms, a sex chromosome system that is shared by two different genera. Thus, recombination suppression started at least 21–25 Myr ago, and then (either gradually or step-wise) spread to a large part of the sex chromosomes (c. 70%), leading to a degenerated Y chromosome.

## Introduction

Among > 15 000 dioecious angiosperm species (i.e. species with separate sexes; Renner, 2014), < 20 sex chromosome systems have been studied with genomic data (Ming *et al.*, 2011; Baránková *et al.*, 2020). Most plants with sex chromosomes exhibit male heterogamety, with XY chromosomes in males and XX chromosomes in females (Westergaard, 1958; Charlesworth, 2016). The portion of the Y chromosome that never recombines with the X experiences reduced selection, which results in an accumulation of deleterious mutations and transposable elements (Charlesworth & Charlesworth, 2000). This accumulation of transposable elements initially leads to an increase of the size of the Y chromosome, which becomes larger than the X (Ming *et al.*, 2011). When Y degeneration progresses, genetic material can be lost without fitness costs and the Y may shrink (Ming *et al.*, 2011). Therefore, after sufficient time of divergence, we may observe chromosome heteromorphy – a Y chromosome larger or smaller than the X chromosome, depending on the progress of degeneration (Ming *et al.*, 2011). Classically, heteromorphy was determined using light microscopy, which is rather imprecise and size differences of c. 10% could be considered

homomorphic (see Divashuk *et al.*, 2014). Although heteromorphy often corresponds to the later stages of sex chromosome evolution, it is nevertheless possible that sex chromosomes are homomorphic despite a large nonrecombining region and strong degeneration of the Y chromosome (e.g. Prentout *et al.*, 2020). Moreover, some systems do not evolve large nonrecombining region and stay homomorphic (Renner & Müller, 2021).

In plants, dioecy often is of recent origin (Renner, 2014; Käfer *et al.*, 2017), thus limiting the age of the sex chromosomes. Indeed, several rather recently evolved (< 10 Myr ago (Ma)) homomorphic sex chromosome systems with small nonrecombining regions have been described, as in *Carica papaya* and *Asparagus officinalis* (Wu & Moore, 2015; Harkess *et al.*, 2017). Heteromorphic sex chromosome systems also are found, with the Y being larger than the X, but recombination suppression also happened relatively recently (< 20 Ma), as in *Silene latifolia* and *Coccinia grandis* (Sousa *et al.*, 2013; Krasovec *et al.*, 2018; Fruchard *et al.*, 2020).

A few cases in which dioecy evolved longer ago also exist (Käfer *et al.*, 2017), but no strongly degenerated sex chromosomes have been described so far (Renner & Müller, 2021). Pucholt *et al.* (2017) described very young sex chromosomes in *Salix viminalis* despite ancestral dioecy for the sister genera *Salix*

\*These authors contributed equally to this work

and *Populus*. Thus, either the sex chromosomes evolved independently in different species, or there have been frequent turnovers. In the fully dioecious palm tree genus *Phoenix*, a sex-linked region evolved before the speciation of the 14 known species (Cherif *et al.*, 2016; Torres *et al.*, 2018). These sex chromosomes might be old, but do not appear to be strongly differentiated. A similar situation has been reported in the grapevine (*Vitis*) genus (Badouin *et al.*, 2020; Massonnet *et al.*, 2020), possibly because sex chromosome evolution is slowed down in such perennials with long generation time (Muyle *et al.*, 2017).

Thus, although homologous sex chromosomes are sometimes shared between species belonging to the same genus (e.g. *Silene* sect. *Melandrium*, *Phoenix*) (Cherif *et al.*, 2016; Bačovský *et al.*, 2020), homologous sex chromosomes between different genera have never been described in plants so far. This situation is in stark contrast to the situation in animals, for which several systems are > 100 Myr old and are shared by whole classes, for example, birds and mammals (Ohno, 1969; Fridolfsson *et al.*, 1998; Cortez *et al.*, 2014). Undoubtedly sex chromosomes have been less intensively studied in plants, yet there seem to be fundamental differences in the evolution of sex chromosomes in plants and animals (e.g. lack of strong sexual dimorphism in plants, discussed in Renner & Müller, 2021). However, the extent of the differences needs to be clarified and more plant sex chromosomes need to be studied.

Dioecy very likely evolved before the genera *Cannabis* and *Humulus* split, and might even be ancestral in the Cannabaceae family (Yang *et al.*, 2013; Zhang *et al.*, 2019). *Cannabis sativa* (marijuana and hemp) is a dioecious species with nearly homomorphic XY chromosomes (with homomorphy defined as above). These sex chromosomes have a large nonrecombining region and are estimated to have started diverging between 12 and 28 Ma (Peil *et al.*, 2003; Divashuk *et al.*, 2014; Prentout *et al.*, 2020).

As for *C. sativa*, cytological analyses of *Humulus lupulus* (hop) found a XY chromosome system with a large nonrecombining region, but the Y chromosome is smaller than the X (Shephard *et al.*, 2000; Karlov *et al.*, 2003; Divashuk *et al.*, 2011). The *H. lupulus* and *C. sativa* lineages split between 21 and 25 Ma (Divashuk *et al.*, 2014; Jin *et al.*, 2020), which is more recently than our higher bound estimate of the age of the *C. sativa* sex chromosomes (28 Ma; Prentout *et al.*, 2020). It thus is possible that the sex chromosomes of *C. sativa* and *H. lupulus* evolved from the same pair that already stopped recombining in their common ancestor, a question we address here.

As in many cultivated dioecious species, only female hop plants are harvested. Hop is used in beer brewing for its bitterness, and its production is increasing worldwide (Neve, 1991; King & Pavlovic, 2017), mostly because of the craft beer revolution (Barth-Haas, 2019; Mackinnon & Pavlovič, 2019). The molecule responsible for hop flower bitterness, lupulin, is concentrated in female ripe inflorescences, called cones (Okada & Ito, 2001). In pollinated cones, the presence of seed reduces their brewing quality; because *H. lupulus* is wind pollinated, a single male plant in the hop field or its vicinity can cause broad-scale damage to the crop (Thomas & Neve, 1976).

Usually, hop is not grown from seeds, so female-only cultures are easy to obtain, and there is no need for large-scale early sexing as in *C. sativa* (cf. Prentout *et al.*, 2020). However, for varietal improvement where controlled crosses are needed, knowing the sex early might be beneficial. In *H. lupulus*, sexing is reliable 1–2 years after sowing (Patzak *et al.*, 2002; Conway & Snyder, 2008). A few markers have been developed, but the use of Y-specific coding sequences may increase marker quality (Patzak *et al.*, 2002; Cerenak *et al.*, 2019).

Here we sequenced the transcriptome of 14 *H. lupulus* individuals. These individuals came from a cross, from which we sequenced the parents and six offspring of each sex. We used the probabilistic approach SEX-DETECTOR, which is based on allele segregation analysis within a cross, to identify sex-linked sequences (Muyle *et al.*, 2016). From these analyses on *H. lupulus* and our previous results on *C. sativa* (Prentout *et al.*, 2020) we describe for the first time well-differentiated sex chromosomes shared by two different genera in plants.

## Materials and Methods

### Biological material and RNA sequencing

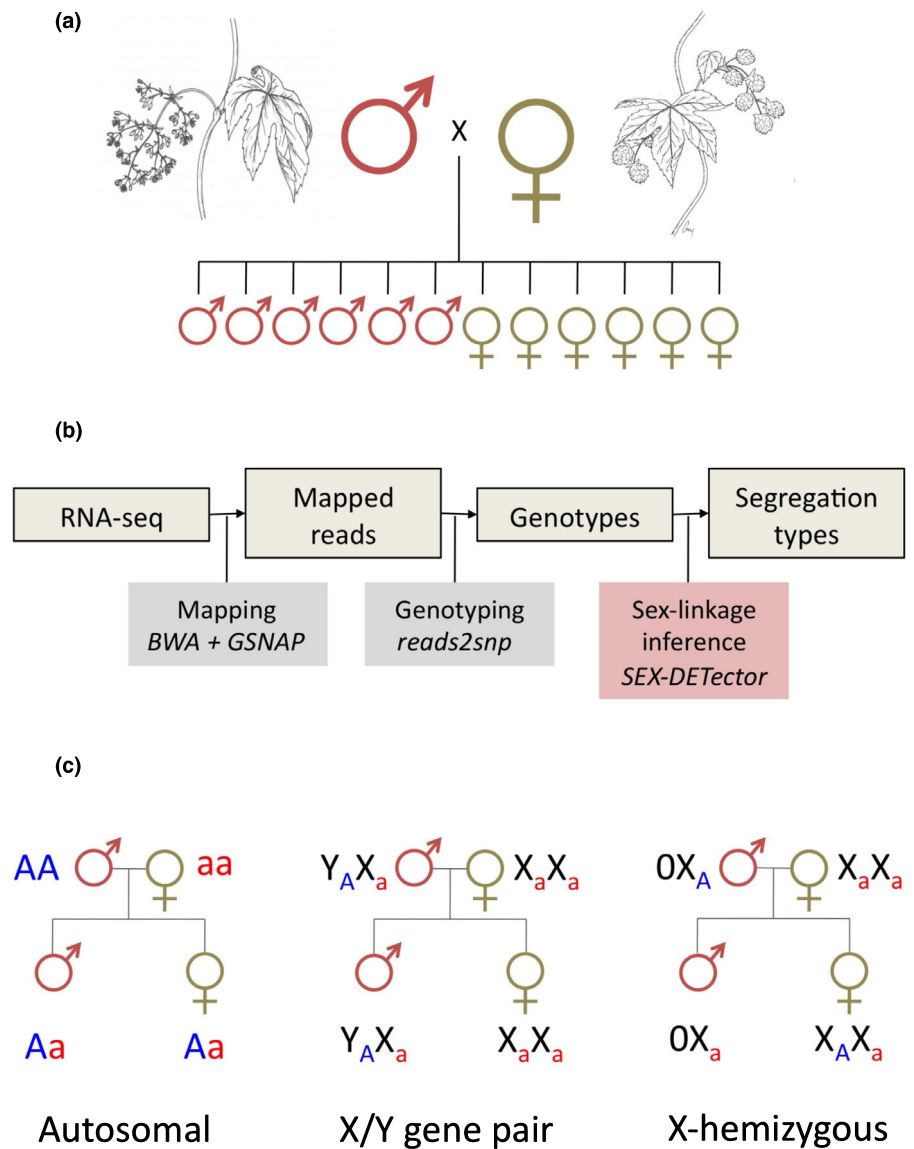
As indicated in Fig. 1a, we conducted a controlled cross for sequencing. The *Humulus lupulus* parents, cultivar 'Wye Target' (WT; female) and the Slovenian male breeding line 2/1 (2/1), as well as six female and six male F<sub>1</sub> siblings (Jakše *et al.*, 2013) were collected in July 2019 in the experimental garden of Slovenian Institute of Hop Research and Brewing, Žalec.

All offspring were phenotypically confirmed to carry either male or female reproductive organs and showed no anomalies in microsatellite genotyping data (Jakše *et al.*, 2013). Young leaves from the laterally developing shoots were picked, wrapped in aluminium foil and flash-frozen *in situ* in liquid N<sub>2</sub>. Later they were pulverized and stored at –80°C until RNA isolation.

Total RNA was isolated from 100 mg frozen tissue pulverized in liquid N<sub>2</sub> according to the protocol of Monarch Total RNA Miniprep Kit, including removal of DNA from the column with DNase I (New England Biolabs). Total RNA was quantified with QBIT 3.0, and quality was verified with the Agilent RNA Nano 6000 Kit to confirm appropriate sample RIN numbers. The total RNA samples were sent to Novogen for mRNA sequencing using Illumina's 100-bp paired end service. The data were submitted to the SRA database of the NCBI (BioSample accession SAMN17526021).

### Mapping, genotyping and SEX-DETECTOR

The bioinformatic pipeline is described schematically in Fig. 1b. First, the RNA-seq data were mapped to two different references: (1) the transcriptome *H. lupulus* (obtained from the annotated genome; Padgitt-Cobb *et al.*, 2019) and (2) the transcriptome assembly of *Cannabis sativa* that we also used for our previous *C. sativa* sex chromosome analysis (Supporting Information, Methods S1; Van Bakel *et al.*, 2011; Prentout *et al.*, 2020). For the mapping, we ran GSNAP (v.2019-09-12; Wu and Nacu, 2010; Wu *et al.*,



**Fig. 1** Schematic representation of the workflow used to detect sex-linkage. (a) Experimental design: six females and six males were obtained by a controlled cross, and all individuals (14) were sequenced. (b) Bioinformatic pipeline for the treatment of RNA-seq data. (c) Illustration of the underlying principles of the SEX-DETECTOR segregation analysis.

2016), an aligner that enables single nucleotide polymorphism (SNP)-tolerant mapping, with 10% mismatches allowed. This approach, already used for *C. sativa* analysis, was iterated several times by adding Y-specific SNPs to the references (and *H. lupulus*-specific SNPs while mapping on *C. sativa* reference; see Prentout *et al.*, 2020), which increased the number of mapped reads.

Then, SAMTOOLS (v.1.4; Li *et al.*, 2009) was used to remove unmapped reads and sort mapping output files for the genotyping. We genotyped individuals with READS2SNP (v.2.0.64; Gayral *et al.*, 2013), as recommended for SEX-DETECTOR (Muyle *et al.*, 2016) – by accounting for allelic expression biases, without filtering for paralogous SNPs, and only conserving SNPs supported by at least three reads for subsequent analysis.

We ran the XY model of SEX-DETECTOR on the genotyping data, using the SEM algorithm and a threshold for an assignment of 0.8. SEX-DETECTOR computes a posterior probability of being autosomal ( $P_A$ ), XY ( $P_{XY}$ ) and X-hemizygous ( $P_{X-hemi}$ ) for each SNP and for each gene (Fig. 1c). Thus, a gene with a  $P_A$  of  $\geq 0.8$  and at least one autosomal SNP without genotyping error is

classified as ‘autosomal’; a gene with  $P_{XY} + P_{X-hemi}$  of  $\geq 0.8$  and at least one sex-linked SNP without genotyping error is classified as ‘sex-linked’; otherwise, the gene is classified as ‘lack-of-information’. Among the sex-linked genes, we classified a gene as X-hemizygous if it fulfilled one of these two criteria: (1) the gene carried only X-hemizygous SNPs and at least one SNP without genotyping error, (2) the Y expression of the gene is detected only from positions with genotyping errors. A parameter that is important to optimize with SEX-DETECTOR is the Y-specific genotyping error rate ( $p$ ; see Muyle *et al.*, 2016). However, the quantity of Y-linked reads that map on a female reference diminishes with X–Y divergence; therefore, old and highly divergent sex chromosomes are more susceptible to mapping errors and thus genotyping errors.  $p$  is expected to be close to the whole transcriptome genotyping error rate ( $\epsilon$ ), but could be higher as a consequence of weak expression (resulting in less reads) of the Y-linked copies or to mapping on a divergent X reference. To reduce the gap between these two error rates, we ran four iterations with GSNAP, using at each time the SNPs file generated by SEX-DETECTOR. This SNPs file

contains *H. lupulus*-specific polymorphisms, initially absent from the *C. sativa* reference transcriptome, and increased the quantity of mapped reads by adding these SNPs to the reference and, thus, fitting it with the *H. lupulus* RNA-seq.

As detailed in Methods S1, we retained the mapping on *C. sativa* transcriptome assembly for downstream analysis. Indeed, the mapping of Y-linked reads and SEX-DETECTOR results obtained with a mapping on *C. sativa* reference transcriptome were more robust than those obtained with a mapping on *H. lupulus* reference transcriptome (Methods S1).

### Sex-linked gene positions on *C. sativa* genome

As the *H. lupulus* RNA-seq data were mapped on the *C. sativa* transcriptome, we determined the position of the transcript sequences from the *C. sativa* transcriptome assembly (van Bakel *et al.*, 2011) on a chromosome-level assembly of the *C. sativa* genome (Grassa *et al.*, 2018) with BLAST (v.2.2.30+; Altschul *et al.*, 1990). We selected the best hit with an e-value lower than  $10^{-4}$  to determine the position of the transcript on the genome. Then, we split each chromosome in windows of 2 Mb and computed the density of sex-linked genes and nonsex-linked genes per window using BEDTOOLS (v.2.26.0; Quinlan & Hall, 2010). Proportions of sex-linked genes were computed by dividing the number of sex-linked genes by the total number of genes (sex-linked, autosomal and undetermined) in the same window. For *C. sativa*, densities were already available from our previous analysis (Prentout *et al.*, 2020).

### Molecular clock and age of sex chromosomes

We used the translated reference transcripts (van Bakel *et al.*, 2011) to determine the X and Y Open Reading Frame (ORF) of nucleotide reference transcripts. For each XY gene pair, the *dS* values were estimated with codeml (PAML v.4.9; Yang, 2007) in pairwise mode. Then, we used two molecular clocks, derived from *Arabidopsis* species, to estimate the age of *H. lupulus* sex chromosomes (Koch *et al.*, 2000; Ossowski *et al.*, 2010). In the wild, *H. lupulus* flowers in the second or third year of development (Patzak *et al.*, 2002; Polley *et al.*, 1997), therefore, we took a generation time (GT) of 2 years, and used the molecular clocks as follows:  $(dS)/rate = dS/(1.5 \times 10^{-8})$  using the molecular clock from Koch *et al.*, (2000);  $(GT \times dS)/(2 \times \mu) = dS/(7 \times 10^{-9})$  using the clock from Ossowski *et al.*, (2010). Three different estimates of *dS* were used: the maximum *dS* value, the mean of the 5% highest *dS* values, and the mean of the 10% highest *dS* values.

### X and Y allele-specific expression analysis

In addition to identifying X and Y alleles, SEX-DETECTOR estimates their expression based on the number of reads (Muyle *et al.*, 2016). These estimates rely on counting reads spanning XY SNPs only and were normalized using the total read number in a library for each individual. We further normalized them by the median autosomal expression for each individual. The *C. sativa*

results presented here were generated in our previous analysis on *C. sativa* sex chromosomes (Prentout *et al.*, 2020).

### Correction of Y read mapping bias

The use of a female reference for the mapping of the reads might create mapping biases, resulting in the absence of Y reads in the most diverging parts of the genes. This issue may reduce the divergence detected and change the phylogenetic signal (Dixon *et al.*, 2019). If, within a same gene, regions that lack Y reads coexist with regions where the Y reads correctly mapped, we expect to see a signature similar to gene conversion (i.e. region-wise variation in divergence). Therefore, we ran GENECONV (v.1.81a; Sawyer, 1999) in pairwise and group mode with the multiple alignments used for the phylogeny (on 85 gene alignments before Gblock filtering, see below) in order to identify and remove regions with reduced divergence. We defined two groups, one for X and Y sequences in *H. lupulus*, and the other one for X and Y sequences in *C. sativa*. Then, we conserved only inner fragments and split the gene conversion regions from regions without gene conversion to obtain two subsets per gene. Thus, we obtained a subset of sequences corrected for the mapping bias, in addition to the set of genes not filtered with GENECONV.

### Phylogenetic analysis

We reconstructed gene families for genes identified as sex-linked in both *C. sativa* and *H. lupulus*. Then, we used BLASTP, filtering for the best hit (with an e-value threshold fixed at  $10^{-4}$ ), to find homologous sequences between *C. sativa* reference transcripts (the query sequence in BLASTP) (van Bakel *et al.*, 2011) and four outgroup transcriptomes (the subject sequence in BLASTP): *Trema orientalis* (Cannabaceae; van Velzen *et al.*, 2018), *Morus notabilis* (Moraceae; He *et al.*, 2013), *Fragaria vesca* ssp. *vesca* (Rosaceae; Shulaev *et al.*, 2011) and *Rosa chinensis* (Rosaceae; Raymond *et al.*, 2018). Gene families for which at least two outgroup sequences have been identified were kept, other gene families were discarded from subsequent analysis. Then, we added X and Y sequences reconstructed by SEX-DETECTOR to each gene family. To distinguish potential paralogous sequences or variants from alternative splicing, a blast of all sequences vs all sequences was realized. If two sequences from two distinct gene families matched with each other (with an e-value threshold fixed at  $10^{-4}$ ), then both families were removed from the dataset. Finally, we retrieved the corresponding nucleotide sequence of each protein, which constituted the dataset used for the phylogenetic analysis.

Using MACSE (v.2.03; Ranwez *et al.*, 2011), and before alignment, nonhomologous segments of  $\geq 60$  nucleotides within or 30 nucleotides at the extremity of a nucleotide sequence were trimmed if they displayed  $< 30\%$  of similarity with other sequences from the gene family. This step allowed the removal of misidentified outgroup sequences. Then, gene families with no remaining outgroup sequences were discarded. Finally, remaining families were aligned with MACSE, allowing sequences to be removed and realigned, one sequence at a time and over multiple iterations, to improve local alignment.

Nucleotide alignments were cleaned at the codon level using GBLOCKS (with default parameters) to conserve only codons shared by all sequences (v.0.91b; Castresana, 2000). For maximum-likelihood (ML) phylogenetic tree reconstruction, we used MODELFINDER in IQ-TREE (v.1.639; Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017) to select the best-fit substitution model for each alignment. Those models were then used in RAXML-NG (v.1.0.0; Kozlov *et al.*, 2019) to reconstruct gene family trees. The number of bootstrap replicates was estimated using AUTOMRE (Pattengale *et al.*, 2010) criterion (maximum 2000 bootstraps). The ML phylogenetic tree reconstruction was run on two datasets, one without removing potential mapping biases, and one with the potential mapping bias removed, as described above.

Bayesian phylogenies were built using PHYLOBAYES (v.3.4; Lartillot *et al.*, 2009) with the site-specific profiles CAT and the CAT-GTR models with a gamma distribution to handle across site rate variations. Two chains were run in parallel for a minimum of 500 cycles. The convergence between the two chains was checked every 100 cycles (with a burn-in equal to one fifth of the total length of the chains). Chains were stopped once all the discrepancies were  $\leq 0.1$  and all effective sizes were  $> 50$  and used to build a majority rule consensus tree.

### Statistics and linear chromosome representations

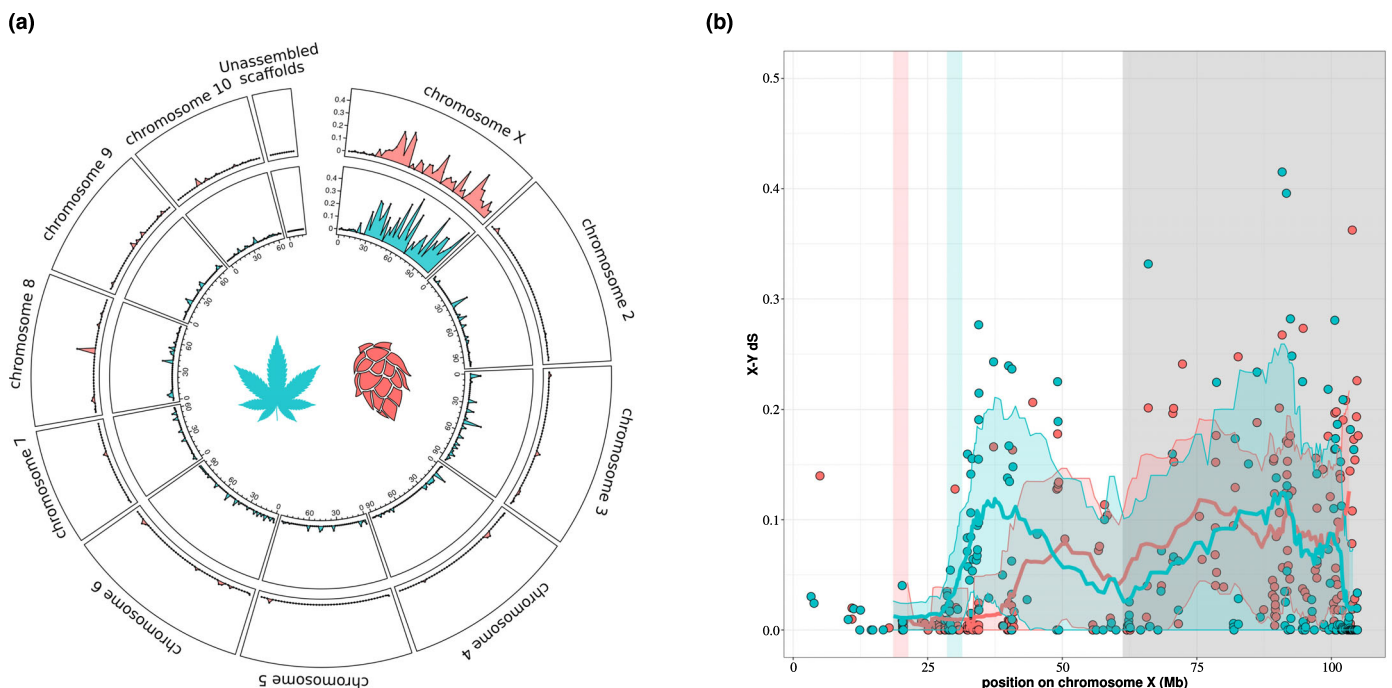
The statistical analyses were conducted with R (v.3.4.4; R Core Team, 2018). We report exact *P*-values when they are  $> 10^{-5}$ .

The representation of phylogenetic topologies, *dS* values on the first chromosome and the dosage compensation graphics have been done with GGPLOT2 (Wickham, 2011). For the circular representation of the sex-linked gene density along the *C. sativa* genome we used the CIRCLIZE package in R (GU *et al.*, 2014). The correspondence between names in Fig. 2a and names in the genome assembly is indicated in Table S1. We calculated confidence intervals for the median of a dataset of *n* observations by resampling 5000 times *n* values from the dataset (with replacement). The confidence intervals are then given by the quantiles of the distribution of median values obtained by resampling.

## Results

### Identification of sex-linked genes in *H. lupulus*

As mentioned in the Materials and Methods section, we used the mapping of the *H. lupulus* RNA-seq data on the *C. sativa* transcriptome assembly for downstream analysis. Of the 30 074 genes in the *C. sativa* reference transcriptome, 21 268 had detectable expression in our *H. lupulus* transcriptome data. The difference of properly-paired mapped reads between males (mean 32.3%) and females (mean 34.9%) was slightly significant (Wilcoxon's test two-sided *P*-value = 0.038; see Table S2), which may be explained by a reduced mapping efficiency of Y-linked reads on the female reference.



**Fig. 2** (a) *Humulus lupulus* sex-linked genes mapped on the *Cannabis sativa* genome (Grassa *et al.*, 2018). Inner graphs (in blue): *C. sativa* sex-linked gene density corrected by the total gene density in 2-Mb windows (from Prentout *et al.*, 2020). Outer graphs (in red): *H. lupulus* sex-linked gene density corrected by the total gene density in 2-Mb windows. (b) Synonymous divergence (*dS*) between X and Y copies of *H. lupulus* sex-linked genes (red) and those of *C. sativa* (blue) along the X chromosome of *C. sativa*. The curves represent the average *dS* with sliding windows (windows of 20 points), for *H. lupulus* (red) and *C. sativa* (blue). Confidence intervals (average  $\pm$  SD) are indicated around the *H. lupulus* curve (red area) and the *C. sativa* curve (blue area). The vertical red bar represents the putative Pseudo-Autosomal Boundary (PAB) in *H. lupulus*, the vertical blue bar represents the putative PAB in *C. sativa*, the grey area represents the region that stopped recombining in a common ancestor.

The sex-linked sequences from *H. lupulus* transcriptome data were identified with SEX-DETECTOR (Muyle *et al.*, 2016). It is important that genotyping error rate parameters  $\epsilon$  and  $p$  have similar values ( $\epsilon$ , whole transcriptome;  $p$ , Y chromosome) to obtain reliable SEX-DETECTOR outputs. At the fourth iteration of GSNAP mapping on *C. sativa* reference transcriptome  $\epsilon$  and  $p$  stabilized at 0.06 and 0.20, respectively (Table S3). Upon closer inspection, one *H. lupulus* male (#3) appeared to have many genotyping errors, as for some XY genes, this male was genotyped as both heterozygous (XY) and homozygous (XX), which increased the error rate  $p$ . The identification of Y SNPs with this individual RNA-seq data discarded the hypothesis of a mislabelled female or a XX individual that developed male flowers. A particularly strong Y reads mapping bias in this male may explain these observations. After removal of this male, the error rate  $p$  dropped to 0.10 (Table S3). A total of 265 sex-linked genes were identified in *H. lupulus*, which represents 7.8% of all assigned genes (autosomal genes + sex-linked genes; Table 1). The SEX-DETECTOR assignment output is provided in Table S4.

### *H. lupulus* and *C. sativa* sex chromosomes are homologous

Among 265 *H. lupulus* XY genes from the *C. sativa* transcriptome assembly (van Bakel *et al.*, 2011), 254 genes are present on the *C. sativa* chromosome-level genome assembly (Grassa *et al.*, 2018). As shown in Fig. 2a, 192 of these genes (75.6%) map on *C. sativa* chromosome number 1, which is the chromosome we previously identified as the X chromosome in *C. sativa* (Prentout *et al.*, 2020). Of the 265 sex-linked genes in *H. lupulus*, 112 also were detected as sex-linked in *C. sativa*, whereas 64 were detected as autosomal and 89 had unassigned segregation type (Prentout *et al.*, 2020).

The synonymous divergence ( $dS$ ) between X and Y copies of the sex-linked genes of *H. lupulus* was distributed in a similar way along the *C. sativa* sex chromosome as the values for this latter species, as shown in Fig. 2b. Although the sampling variation of these  $dS$  values is large, as expected (cf. Takahata & Nei, 1985), it can be observed that the larger values occurred in the region beyond 65 Mb.

**Table 1** Summary of the SEX-DETECTOR results.

	Number
All genes*	30 074
Expressed genes	21 268
Genes with SNPs used by SEX-DETECTOR	4472
Genes with undetermined segregation type class 1**	462
Genes with undetermined segregation type class 2***	354
Autosomal genes	3391
Sex-linked genes	265
XY genes	265
X-hemizygous genes	0

\*Transcripts from gene annotation of the *C. sativa* reference genome (van Bakel *et al.*, 2011).

\*\*Posterior probabilities < 0.8

\*\*\*Posterior probabilities > 0.8 but absence of single nucleotide polymorphisms (SNPs) without error.

### X-Y recombination likely stopped before the *Cannabis* and *Humulus* genera split

We reconstructed phylogenetic trees of genes detected as sex-linked in both species, including outgroup sequences from the order Rosales. For 27 of the 112 sex-linked genes present in both species, we could not identify any homologous sequences in the outgroup species and those genes were excluded from further analysis. For the remaining 85, we determined the topology of the gametologous sequences in the Cannabaceae, considering a node to be well-resolved when the bootstrap support exceeded 95%, or Bayesian support exceeded 0.95.

The three different methods for phylogenetic reconstruction provided consistent phylogenies (Table 2). More precisely, we observed three major topologies, as shown in Fig. 3: X copies of both species form a clade separated from a clade of Y sequences (topology I; Fig. 3a), the X and Y sequences of each species group together (topology II; Fig. 3b), or a paraphyletic placement of the X and Y sequences of *H. lupulus*, relative to *C. sativa* sequences (topology III; Fig. 3c). As shown in Table 2, we found that most genes had topology II, corresponding to recombination suppression after the split of the genera. A few genes, however, had topology I, which corresponds to genes for which recombination already was suppressed in a common ancestor of both species. As shown in Fig. 3(d), topologies I and III occurred mainly beyond 80 Mb, whereas topology II occurred all over the chromosome. Topology I is associated with higher synonymous divergence.

We identified 42 genes, of the 85 genes used for the phylogeny, with at least one fragment in at least one species that displayed reduced divergence (with a  $P$ -value < 0.05 in GENECONV output). Because this reduction of divergence may be caused by a mapping bias of Y reads, we ran the ML phylogenetic reconstruction method on regions with and without mapping bias (example in Fig. S1). As shown in Table 2 and Fig. 3(e), after mapping bias filter with GENECONV, the proportion of genes displaying topology I, indicating recombination suppression in a common ancestor, increased, whereas fewer genes with topology II were found, mainly in a restricted region corresponding to the region where recombination stopped independently in both species.

This leads us to define three regions on the X chromosomes of *C. sativa* and *H. lupulus* (with the *C. sativa* X chromosome as a reference). A region from *c.* 65 Mb to the end of the X chromosome that already stopped recombining in a common ancestor; from *c.* 20–30 Mb to *c.* 65 Mb, a part of the nonrecombining region that evolved independently in the two species; and from the beginning of the chromosome to *c.* 20–30 Mb, the pseudo-autosomal region (PAR), where few sex-linked genes are found.

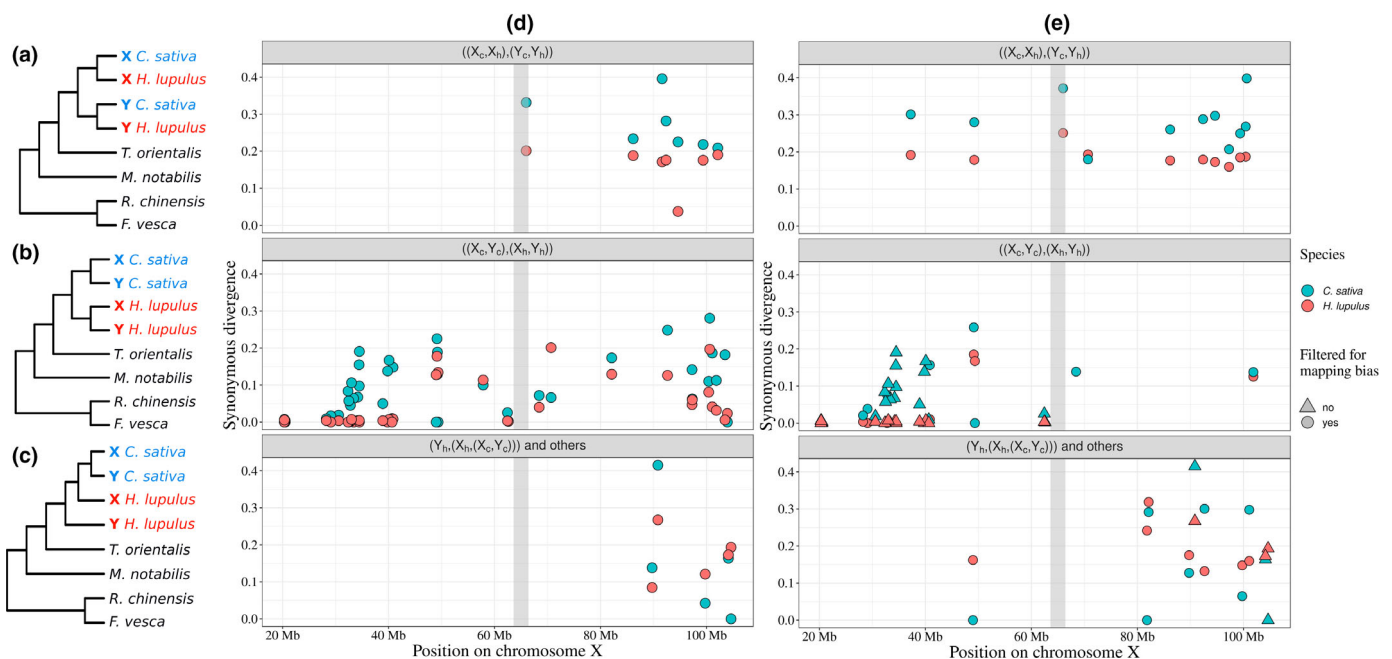
### Age of *H. lupulus* sex chromosomes

In order to estimate the age of the sex chromosomes, we used the maximum synonymous divergence between X and Y sequences and two molecular clocks, which were both derived from *Arabidopsis*. Because the sampling variance in  $dS$  values can be large, we used three ways to calculate the maximum  $dS$  value: the single highest  $dS$

**Table 2** Results of the phylogenetic reconstruction of sex-linked genes.

	Topology I ((X <sub>c</sub> , X <sub>h</sub> ), (Y <sub>c</sub> , Y <sub>h</sub> ))	Topology II ((X <sub>c</sub> , Y <sub>c</sub> ), (X <sub>h</sub> , Y <sub>h</sub> ))	Topology III (Y <sub>h</sub> , (X <sub>h</sub> , (X <sub>c</sub> , Y <sub>c</sub> )))	Other	Unresolved	Total
Maximum-likelihood (ML)	7	44	7	1	26	85
GTR (Bayesian)	4	45	4	8	24	85
CAT-GTR (Bayesian)	4	44	7	7	23	85
ML after GENECONV filtering	11	27	11	4	32	85

Phylogenetic trees with a bootstrap value equal or greater than 95% (and posterior probabilities higher than 0.95 for Bayesian reconstructions) at the node separating *Cannabis sativa* and *Humulus lupulus*, or Y and X sequences, are presented in the first four columns. Phylogenetic trees without such support are classified as 'unresolved'.



**Fig. 3** Distribution of the three topologies of the sex-linked genes on the X chromosome: (a) Topology I, XX-YY – arrest of recombination before the split of the two genera, (b) Topology II, XY-XY – arrest of recombination after the split of the two genera, (c) Topology III, Y-X-XY – *Humulus lupulus* X chromosome is closer to *Cannabis sativa* sequences than its Y counterpart. (d) Distribution of the topologies along the *C. sativa* X chromosome ('other' topology is included in the Y-X-XY topology panel), using the full gene sequences. For each gene, dots represent the *dS* values in *C. sativa* (blue) and *H. lupulus* (red). (e) Distribution of the topologies after filtering out possible mapping biases through geneconv. Triangles indicate that at least one segment was removed, dots indicate sequences for which no mapping bias was detected. The vertical grey bar (d,e) represents the putative boundary between the region that stopped recombining in a common ancestor and the region that stopped recombining independently in the two species.

value; the average of the 5% highest values; and the average of the 10% highest values. Furthermore, we calculated these on the raw alignments as well as the alignments with possible mapping biases removed. The different estimates are given in Table 3, and yielded values between 14.5 and 51.4 Ma. The minimum synonymous divergence between *C. sativa* and outgroup species *Morus notabilis* and *Rosa chinensis* was *c.* 0.45 and *c.* 0.65, respectively (Figs S2, S3), higher than the maximum synonymous divergence between sex-linked gene copies, indicating that the sex chromosomes probably evolved in the Cannabaceae family.

### Y gene expression

The Y over X expression ratio is a standard proxy for the degeneration of the Y chromosome. A Y/X expression ratio

close to 1 means no degeneration, a Y/X expression ratio close to 0.5 or below means strong degeneration. In *H. lupulus*, the median Y/X expression ratio was 0.637 (Fig. S4), which is significantly different from 1 (99<sup>th</sup> percentile of median distribution with 5000 samples in initial distribution = 0.673). The median was not different when considering all sex-linked genes (0.637) or only the sex-linked genes mapping on *C. sativa* X chromosome (0.639, *P* = 0.70, one-sided Wilcoxon rank sum test).

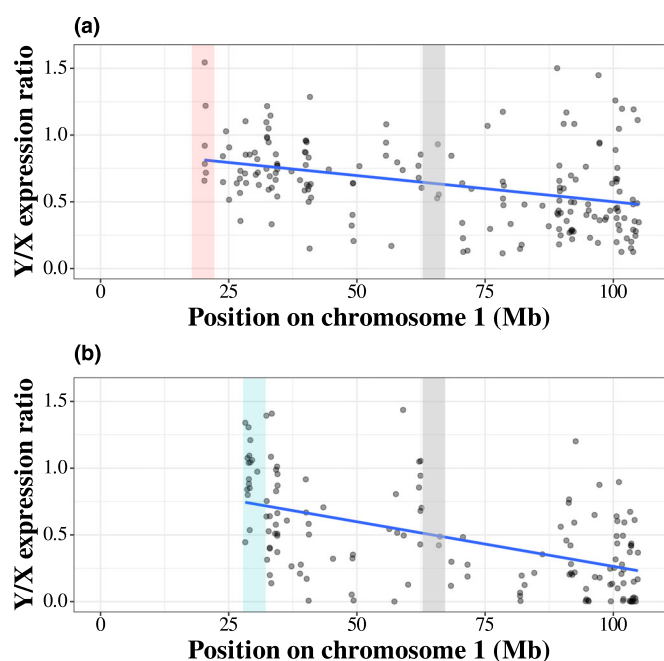
In both species, the reduced Y expression is correlated to the position on the X chromosome (linear regression: adjusted *R*<sup>2</sup> = -0.134, *P* < 10<sup>-5</sup> for *H. lupulus*; and adjusted *R*<sup>2</sup> = 0.278, *P* < 10<sup>-5</sup> for *C. sativa*). As shown in Fig. 4, the Y/X expression ratio decreased while moving away from the PAR in *H. lupulus*, and this also was confirmed in *C. sativa*.



**Table 3** Age estimates with two molecular clocks and different maximum synonymous divergence values (*dS* values).

	No filtering			Mapping bias filtering		
	<i>dS</i>	Age (Myr old) <sup>1</sup>	Age (Myr old) <sup>2</sup>	<i>dS</i>	Age (Myr old) <sup>1</sup>	Age (Myr old) <sup>2</sup>
Highest <i>dS</i>	0.362	51.4	24.0	0.362	51.4	24.0
Mean highest 5%	0.249	35.6	16.6	0.274	39.1	18.3
Mean highest 10%	0.217	31.0	14.5	0.214	34.4	16.1

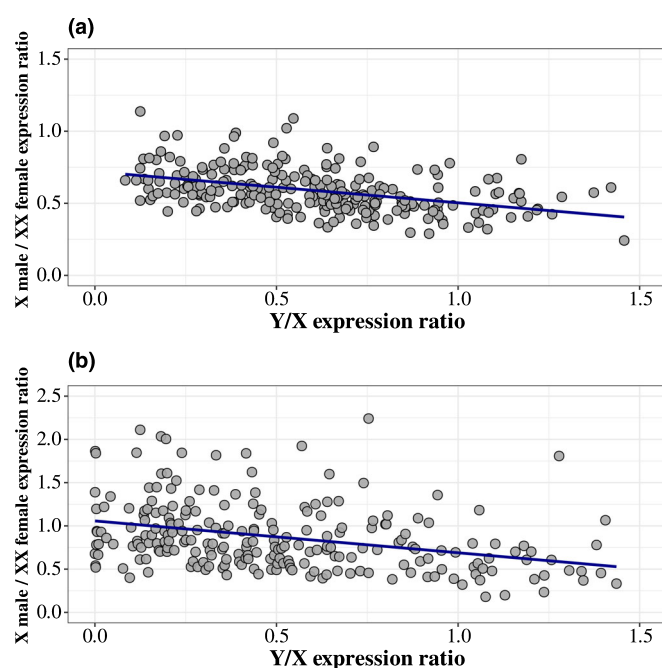
For each *dS* value, two ages were obtained using the molecular clocks of <sup>1</sup>Ossowski *et al.* (2010) and <sup>2</sup>Koch *et al.* (2000). Two alignment datasets were used, with or without filtering for possible mapping bias.



**Fig. 4** Y/X expression ratio along the nonrecombining region in *Humulus lupulus* (a) and *Cannabis sativa* (b). Each dot is the Y/X expression ratio for one gene in the nonrecombining region only (the linear regression result is indicated by the blue line). The vertical red bar represents the putative PAB in *H. lupulus*, the vertical blue bar represents the putative PAB in *C. sativa*, the vertical grey bar represents the putative boundary between the region that stopped recombining in a common ancestor and the region that stopped recombining independently in the two species.

### Dosage compensation

We tested whether the expression of the X chromosome changed following degeneration of the Y chromosome, a phenomenon called dosage compensation (Muyle *et al.*, 2017). We used the ratio of the male X expression over the female XX expression as a proxy for dosage compensation (Muyle *et al.*, 2012) and Y/X expression ratio as a proxy for Y degeneration. Genes with strong degeneration (Y/X expression ratio close to zero) displayed an increased expression of the X in males (linear regression: adjusted  $R^2 = 0.179$ ,  $P < 10^{-5}$  and adjusted  $R^2 = 0.097$ ,  $P < 10^{-5}$  for *H. lupulus* and *C. sativa*, respectively), as shown in Fig. 5. A dosage compensation pattern was found in both in *H. lupulus* and *C. sativa* in agreement with previous work (Prentout *et al.*, 2020).



**Fig. 5** The male X expression over female XX expression versus Y/X expression ratio for *Humulus lupulus* (a) and *Cannabis sativa* (b). Each black dot represents one gene. The blue line represents a linear regression.

### Discussion

We here identified the *Humulus lupulus* sex chromosomes, and found that they are homologous to those of *Cannabis sativa* (Prentout *et al.*, 2020), and that a part of these chromosomes had already stopped recombining in a common ancestor of the two species. Performing a segregation analysis with SEX-DETECTOR (Muyle *et al.*, 2016), we identified 265 XY genes in *H. lupulus*, among which 112 also were sex-linked in *C. sativa*. Mapping these genes on the chromosome-level assembly of *C. sativa* (Grassa *et al.*, 2018) suggested that the nonrecombining region is large in *H. lupulus*, as proposed before based on cytological studies (Divashuk *et al.*, 2011).

We identified three different regions on the sex chromosome, based on the distribution of sex-linked gene phylogenetic topologies and synonymous divergence between the X and Y copies on the *C. sativa* X chromosome: one region that had already stopped recombining in a common ancestor of *C. sativa* and *H. lupulus*, a region that independently stopped recombining in both species,

and the pseudo-autosomal region. Our results suggest the pseudo-autosomal boundary (PAB) in *H. lupulus* may be located around 20 Mb, whereas we estimated a PAB around 30 Mb in *C. sativa* (Prentout *et al.*, 2020); the nonrecombining region thus may be larger in *H. lupulus* than in *C. sativa*. With this estimation of the size of the nonrecombining region in *H. lupulus*, among the 3469 genes present on the X chromosome, 2045 genes would be located in this nonrecombining region (which represents 59.1% of all the genes on the X chromosome). However, a chromosome-level assembly of the *H. lupulus* genome would be needed to determine the exact position of the PAB in this species, as synteny might not be fully conserved. In addition, because we used one single cross, it is possible that we overestimated the size of the nonrecombining region as a consequence of linkage disequilibrium. Thus, genes around the PAB classified as sex-linked and for which we estimated a low *dS* value may still be recombining. An accurate estimation of the PAB, as has been done for example in *Silene latifolia*, would require much more offspring and data from several populations (Krasovec *et al.*, 2020).

Several sex-linked genes had topologies that were not compatible with either recombination suppression in a common ancestor or in each of the species independently. Strikingly, most of these topologies placed the *H. lupulus* Y sequence as an outgroup to the other sex-linked gene sequences. Whether this is the result of errors (e.g. long-branch attraction, mapping biases) remains to be investigated. Interestingly, genes with these ‘unexpected’ topologies all clustered (except for one gene) in a region of *c.* 25 Mb. This region is located at the extremity of the X chromosome, which, as we suggested, stopped recombining first. It is likely to observe a high rate of unexpected phylogenetic results in the region that stopped recombining first because the X–Y divergence should be the highest in this region, which could increase the mapping bias. Our approach to correct for the Y read mapping relies on GENECONV, which is known to have a high rate of false negatives (Lawson & Zhang, 2009). This also could explain the unexpected presence of some of the XY–XY genes in the older region.

X–Y gene conversion has been shown to affect only a few genes in animals (Katsura *et al.*, 2012; Trombetta *et al.*, 2014; Peneder *et al.*, 2017). Although we do not expect gene conversion for half of the genes that are sex-linked in both species, it is worth noting that a part of fragments identified by GENECONV may correspond to real gene conversion rather than mapping biases. Here again, assemblies of Y and X chromosomes in both species are required to determine the presence of true X–Y gene conversion.

The highest *dS* values and the genes with a topology indicating that recombination was already suppressed in the common ancestor are located in the same region (65 Mb to the end of the chromosome). These results suggest the presence of at least two strata in these sex chromosomes. We estimated that the youngest stratum is 10.1–29.4 Myr old in *H. lupulus*, and 15.9–19.8 Myr old in *C. sativa* (Table S5). However, although recombination suppression clearly did not occur for all of the sex-linked genes at the same time, we cannot determine the exact number of strata in the sex chromosomes of *C. sativa* and *H. lupulus*. It also is possible that recombination was suppressed gradually, with the

recombination suppression starting before the split of both genera and continuing afterwards. To clearly determine the number of strata, an identification of chromosomal inversions or significant differences in *dS* values along the sex chromosomes is required (Nicolas *et al.*, 2004; Lemaitre *et al.*, 2009; Wang *et al.*, 2012; reviewed in Wright *et al.*, 2016). Thus, X and Y chromosome assemblies for both *H. lupulus* and *C. sativa* are needed to exactly determine the number (and location) of strata in both species. Moreover, a Y chromosome assembly will allow the identification of Y-specific genes, which is not possible with SEX-DETECTOR and the data that we used.

We did not find X-hemizygous genes in *H. lupulus*. This is striking as 218 X-hemizygous genes (38% of all sex-linked genes) were found in *C. sativa* using the same methodology (Prentout *et al.*, 2020). A very low level of polymorphism could result in the inability of SEX-DETECTOR to identify X-hemizygous genes (Muyle *et al.*, 2016), but in that case SEX-DETECTOR also should have problems identifying autosomal genes, which was not the case here. Nonrandom X-inactivation in females could be an explanation, as the nonrandom expression of a single X allele in females would impede SEX-DETECTOR to identify X-linkage and X-hemizygous genes (Muyle *et al.*, 2016). We ran an Allele-Specific Expression (ASE) analysis, which did not support this hypothesis (Figs S5–S7). *Humulus lupulus* probably is an ancient polyploid that reverted to the ancestral karyotype (Padgitt-Cobb *et al.*, 2019). It is thus possible that the *H. lupulus* X chromosome comprises two copies of the ancestral X as some cytological data seem to suggest (Divashuk *et al.*, 2011). In this case, SEX-DETECTOR would manage to identify the XY gene pairs, but would fail to identify the X-hemizygous genes as these genes would exhibit unexpected allele transmission patterns (Fig. S8).

*Humulus lupulus* is a rare case of XY systems in plants in which the Y is smaller than the X (cf. Ming *et al.*, 2011). In *C. sativa*, both sex chromosomes have similar sizes (Divashuk *et al.*, 2014). If the size difference is caused by deletions of parts of the *H. lupulus* Y chromosome, which is the hypothesized mechanism in many species (cf. Ming *et al.*, 2011), we expect to observe that many XY gene pairs in *C. sativa* have missing Y copies in *H. lupulus*. As explained above, we did not detect any X-hemizygous genes. Furthermore, the XY gene pairs of *H. lupulus* were distributed uniformly on the *C. sativa* X chromosome, and no region appeared to be depleted in XY genes, which is not what we would observe if large deletions were present on the *H. lupulus* Y chromosome. The sex chromosome size differences observed in *H. lupulus* probably reflect complex dynamics, different from that of old animal systems with tiny Y chromosome resulting from large deletions (e.g. Skaletsky *et al.*, 2003; Ross *et al.*, 2005). The large size of the X chromosome in *H. lupulus* may be due to a full-chromosome duplication followed by a fusion (see above), whereas the Y chromosome has remained unchanged. Assemblies of the *H. lupulus* sex chromosomes will be needed to test these hypotheses.

Our estimates of the age of the *H. lupulus* sex chromosomes are larger than the estimates for *C. sativa*, although we found very similar X–Y maximum divergence in both species (higher bound age estimates are *c.* 50 Myr and *c.* 28 Myr old; highest *dS* values

are 0.362 and 0.415 in *H. lupulus* and *C. sativa*, respectively, see Prentout *et al.*, 2020). Of course, the molecular clocks that we used are known to provide very rough estimates as they derive from the relatively distant *Arabidopsis* genus, and are sensitive to potential differences in mutation rates between the annual *C. sativa* and the perennial *H. lupulus* (Neve, 1991; Petit & Hampe, 2006; Small, 2015; but see Krasovec *et al.*, 2018). Indeed, only one of these molecular clocks (which is based on the mutation rate) takes into account the generation time (2 yr in *H. lupulus* vs 1 yr in *C. sativa*). This produced age estimate is approximately two-fold greater than that from the other clock (based on the substitution rate), which was not the case with *C. sativa* (Prentout *et al.*, 2020). It is not known, however, if the generation time influences the substitution rate (Petit & Hampe, 2006). Furthermore, the short generation time in *C. sativa* probably is a derived trait, not reflecting the long-term generation time of the *Cannabis*–*Humulus* lineage, as the *Cannabis* genus is the only herbaceous genus in the Cannabaceae family (Yang *et al.*, 2013). Thus, the remarkable similarity between the highest *dS* values in both species indicates that the *C. sativa* and *H. lupulus* sex chromosomes have a similar age, as expected if they derive from the same common ancestor. Although it is not possible to estimate their age exactly with the current data, initial recombination suppression at least pre-dates the split between the genera, that occurred between 21 and 25 Myr ago (Ma) (Divashuk *et al.*, 2014; Jin *et al.*, 2020), and might even be 50 Myr old. We thus confirmed here that the XY system shared by *C. sativa* and *H. lupulus* is among the oldest plant sex chromosome systems documented so far (Prentout *et al.*, 2020).

Dioecy was inferred as the ancestral sexual system for the Cannabaceae, Urticaceae and Moraceae (Zhang *et al.*, 2019; note, however, that many monoecious Cannabaceae were not included). We found that the synonymous divergence between the Cannabaceae species and *Morus notabilis* was approximately 0.45, higher than the maximum divergence of the X and Y copies in the Cannabaceae. It remains possible that the sex chromosomes evolved before the split of the Cannabaceae and Moraceae families, because the oldest genes might have been lost or were not detected in our transcriptome data. There is, however, no report of whether or not sex chromosomes exist in Urticaceae and Moraceae (Ming *et al.*, 2011).

In order to estimate the Y expression, we counted the number of reads with Y SNPs. Therefore, the impact of a potential Y reads mapping bias should be weaker on Y expression analysis than on X–Y divergence analysis. We validated this assumption by removing genes with detected mapping bias from the analysis, which did not change the signal of Y expression reduction and dosage compensation (Table S6; Figs S9, S10). Dosage compensation is a well-known phenomenon in animals (e.g. Gu & Walters, 2017), but it has been documented only quite recently in plants (reviewed in Muyle *et al.*, 2017). Here we found evidence for dosage compensation in *H. lupulus*; this is not surprising as previous work reported dosage compensation in *C. sativa* and we showed here that both systems are homologous. *Cannabis sativa* and *H. lupulus* add up to the list of plant sex chromosome systems with dosage compensation (see Muyle *et al.*, 2017, for a

review; and Prentout *et al.*, 2020, and Fruchard *et al.*, 2020, for the latest reports of dosage compensation in plants). Further analyses are needed to determine whether this dosage compensation has been selected or is an outcome of regulatory feedback (Malone *et al.*, 2012; Krasovec *et al.*, 2019).

*Humulus lupulus* sex chromosomes, like those of *C. sativa*, are well-differentiated, with a large nonrecombining region. Both species show similar patterns of Y degeneration and dosage compensation, despite the fact that a large part of the nonrecombining region evolved independently in both species. These similarities, as well as the age of the chromosomes and the fact that they have been conserved since the most recent common ancestor of the two genera – a unique situation in plants so far – provide an exciting opportunity to test and elaborate hypotheses on sex chromosome evolution in plants.


## Acknowledgements


We thank Roberto Bacilieri for his help in setting up this collaboration and for discussions, Aline Muyle for advice on SEX-DETECTOR and Florian Bénétière for helpful suggestions regarding graphical representations. This work was performed using the computing facilities of the CC LBBE/PRABI; we thank Bruno Spataro and Stéphane Delmotte for cluster maintenance. Virtual machines from the Institut Français de Bioinformatique also were used to perform this work. This work received financial support from grant no. P4-0077 from ARRS (Slovenian Research Agency) to JJ and from an ANR grant (ANR-14-CE19-0021-01) to GABM.


## Author contributions


GABM, JK and DP conceptualized the study; GABM, JK, DP, NS and JJ formulated the methodology; DP, TT and CB-A operated the software; DP, TT and CB-A carried out the formal analysis; DP, NS, TT, CB-A, JJ, JK and GABM carried out the investigation; AC, NS and JJ supplied resources; DP, GABM, JK and TT wrote the original draft; all authors reviewed and edited the paper; DP and TT made the graphical representations; GABM and JK supervised the project; GABM administrated the project administration; and NS, JJ and GABM acquired funding. JK and GABM contributed equally to this work.


## ORCID

Jernej Jakše  <https://orcid.org/0000-0002-8907-1627>

Jos Käfer  <https://orcid.org/0000-0002-0561-8008>

Gabriel A. B. Marais  <https://orcid.org/0000-0003-2134-5967>

Djivan Prentout  <https://orcid.org/0000-0002-3088-3954>

Theo Tricou  <https://orcid.org/0000-0002-4432-2680>

## Data availability

The sequence data were deposited under the Bioproject with accession number PRJNA694508, BioSample SAMN17526021 (SRR13528971; SRR13528970; SRR13528969; SRR13528968;

SRR13528966; SRR13528965; SRR13528964; SRR13528967; SRR13528963; SRR13528962; SRR13528961; SRR13528960; SRR13528959; SRR13528958).

## References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.
- Báčovský V, Čegan R, Šimoníková D, Hříbová E, Hobza R. 2020. The formation of sex chromosomes in *Silene latifolia* and *S. dioica* was accompanied by multiple chromosomal rearrangements. *Frontiers Plant Science* 11: 205.
- Badouin H, Velt A, Gindraud F, Fluttre T, Dumas V, Vautrin S, Marande W, Corbi J, Sallet E, Ganofsky J *et al.* 2020. The wild grape genome sequence provides insights into the transition from dioecy to hermaphroditism during grape domestication. *Genome Biology* 21: 1–24.
- van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, Page JE. 2011. The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology* 12: 1–18.
- Baránková S, Pascual-Díaz JP, Sultana N, Alonso-Lifante MP, Balant M, Barros K, D'Ambrosio U, Malinská H, Peska V, Lorenzo IP *et al.* 2020. Sex-chrom, a database on plant sex chromosomes. *New Phytologist* 227: 1594–1604.
- Barth-Haas GmbH & Co. KG. 2019. Barth Report (1950-2019). Nuremberg. [WWW document] URL [https://www.barthhaas.com/fileadmin/user\\_upload/news/2019-07-23/barthreport20182019en.pdf](https://www.barthhaas.com/fileadmin/user_upload/news/2019-07-23/barthreport20182019en.pdf). Accessed: 4 May 2020.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17: 540–552.
- Čerenak A, Kolenc Z, Sehur P, Whittock SP, Koutoulis A, Beatson R, Buck E, Javornik B, Škof S, Jakše J. 2019. New male specific markers for hop and application in breeding program. *Scientific Reports* 9: 1–9.
- Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355: 1563–1572.
- Charlesworth D. 2016. Plant sex chromosomes. *Annual Review of Plant Biology* 67: 397–420.
- Cherif E, Zehdi-Azouzi S, Crabos A, Castillo K, Chabrilange N, Pintaud J-C, Salhi-Hannachi A, Glémin S, Aberlenc-Bertossi F. 2016. Evolution of sex chromosomes prior to speciation in the dioecious *Phoenix* species. *Journal of Evolutionary Biology* 29: 1513–1522.
- Conway S, Snyder R. 2008. *Humulus lupulus* – hops. *College seminar 235 food for thought: the science, culture, & politics of food*. [WWW document] URL [https://academics.hamilton.edu/foodforthought/Our\\_Research\\_files/hops.pdf](https://academics.hamilton.edu/foodforthought/Our_Research_files/hops.pdf). [accessed 4 May 2020].
- Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* 508: 488–493.
- Divashuk MG, Alexandrov OS, Kroupin PY, Karlov GI. 2011. Molecular cytogenetic mapping of *Humulus lupulus* sex chromosomes. *Cytogenetic and Genome Research* 134: 213–219.
- Divashuk MG, Alexandrov OS, Razumova OV, Kirov IV, Karlov GI. 2014. Molecular cytogenetic characterization of the dioecious *Cannabis sativa* with an XY chromosome sex determination system. *PLoS ONE* 9: e85118.
- Dixon G, Kitano J, Kirkpatrick M. 2019. The origin of a new sex chromosome by introgression between two stickleback fishes. *Molecular Biology and Evolution* 36: 28–38.
- Fridolfsson A-K, Cheng H, Copeland NG, Jenkins NA, Liu H-C, Raudsepp T, Woodage T, Chowdhary B, Halverson J, Ellegren H. 1998. Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proceedings of the National Academy of Sciences, USA* 95: 8147–8152.
- Fruchard C, Badouin H, Latrasse D, Devani RS, Muyle A, Rhoné B, Renner SS, Banerjee AK, Bendahmane A, Marais GAB. 2020. Evidence for dosage compensation in *Coccinia grandis*, a plant with a highly heteromorphic XY system. *Genes* 11: 787.
- Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N *et al.* 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genetics* 9: e1003457.
- Grassa CJ, Wenger JP, Dabney C, Poplawski SG, Motley ST, Michael TP, Schwartz CJ, Weiblen GD. 2018. A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. *BioRxiv*: 458083. doi: 10.1101/458083.
- Gu L, Walters JR. 2017. Evolution of sex chromosome dosage compensation in animals: a beautiful theory, undermined by facts and bedeviled by details (K Makova, Ed.). *Genome Biology and Evolution* 9: 2461–2476.
- Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize implements and enhances circular visualization in R. *Bioinformatics* 30: 2811–2812.
- Harkess A, Zhou J, Xu C, Bowers JE, Van der Hulst R, Ayyampalayam S, Mercati F, Riccardi P, McKain MR, Kakrana A *et al.* 2017. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature Communications* 8: 1–10.
- He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, Lee T-H, Wang X, Cai Q, Li D *et al.* 2013. Draft genome sequence of the mulberry tree *Morus notabilis*. *Nature Communications* 4: 1–9.
- Jakše J, Čerenak A, Radisek S, Satovic Z, Luthar Z, Javornik B. 2013. Identification of quantitative trait loci for resistance to Verticillium wilt and yield parameters in hop (*Humulus lupulus* L.). *Theoretical and Applied Genetics* 126: 1431–1443.
- Jin J-J, Yang M-Q, Fritsch PW, van Velzen R, Li D-Z, Yi T-S. 2020. Born migrants: historical biogeography of the cosmopolitan family Cannabaceae. *Journal of Systematics and Evolution* 58: 461–473.
- Käfer J, Marais GAB, Pannell JR. 2017. On the rarity of dioecy in flowering plants. *Molecular Ecology* 26: 1225–1241.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.
- Karlov GI, Danilova TV, Horlemann C, Weber G. 2003. Molecular cytogenetics in hop (*Humulus lupulus* L.) and identification of sex chromosomes by DAPI-banding. *Euphytica* 132: 185–190.
- Katsura Y, Iwase M, Satta Y. 2012. Evolution of genomic structures on mammalian sex chromosomes. *Current Genomics* 13: 115–123.
- King M, Pavlovic M. 2017. Analysis of hop use in craft breweries in Slovenia. *Journal of Agriculture Food and Development* 3: 21–26.
- Koch MA, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution* 17: 1483–1498.
- Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35: 4453–4455.
- Krasovec M, Chester M, Ridout K, Filatov DA. 2018. The mutation rate and the age of the sex chromosomes in *Silene latifolia*. *Current Biology* 28: 1832–1838.
- Krasovec M, Kazama Y, Ishii K, Abe T, Filatov DA. 2019. Immediate dosage compensation is triggered by the deletion of y-linked genes in *Silene latifolia*. *Current Biology* 29: 2214–2221.
- Krasovec M, Zhang Y, Filatov DA. 2020. The location of the pseudoautosomal boundary in *Silene latifolia*. *Genes* 11: 610.
- Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25: 2286–2288.
- Lawson MJ, Zhang L. 2009. Sexy gene conversions: locating gene conversions on the X-chromosome. *Nucleic Acids Research* 37: 4570–4579.
- Lemaitre C, Braga MDV, Gautier C, Sagot M-F, Tannier E, Marais GAB. 2009. Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biology and Evolution* 1: 56–66.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Mackinnon D, Pavlovič M. 2019. Global hop market analysis within the International Hop Growers' Convention. Hmeljarski Bilten 26: 99–108.
- Malone JH, Cho D-Y, Mattiuzzo NR, Artieri CG, Jiang L, Dale RK, Smith HE, McDaniel J, Munro S, Salit M *et al.* 2012. Mediation of *Drosophila*

- autosomal dosage effects and compensation by network interactions. *Genome Biology* 13: 1–17.
- Massonnet M, Cochetel N, Minio A, Vondras AM, Lin J, Muyle A, Garcia JF, Zhou Y, Delle Donne M, Riaz S *et al.* 2020. The genetic basis of sex determination in grapes. *Nature Communications* 11: 1–12.
- Ming R, Bendahmane A, Renner SS. 2011. Sex chromosomes in land plants. *Annual Review of Plant Biology* 62: 485–514.
- Muyle A, Käfer J, Zemp N, Mousset S, Picard F, Marais GA. 2016. SEX-DETECTOR: a probabilistic approach to study sex chromosomes in non-model organisms. *Genome Biology and Evolution* 8: 2530–2543.
- Muyle A, Shearn R, Marais GA. 2017. The evolution of sex chromosomes and dosage compensation in plants. *Genome Biology and Evolution* 9: 627–645.
- Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais GAB. 2012. Rapid *de novo* evolution of x chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *PLoS Biology* 10: e1001308.
- Neve RA. 1991. *Hops*. London, UK: Chapman and Hall.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–274.
- Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Monéger F. 2004. A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *PLoS Biology* 3: e4.
- Ohno S. 1969. Evolution of sex chromosomes in mammals. *Annual Review of Genetics* 3: 495–524.
- Okada Y, Ito K. 2001. Cloning and Analysis of Valerophenone Synthase Gene Expressed Specifically in Lupulin Gland of Hop (*Humulus lupulus* L.). *Bioscience, Biotechnology & Biochemistry* 65: 150–155.
- Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327: 92–94.
- Padgett-Cobb LK, Kingan SB, Wells J, Elser J, Kronmiller B, Moore D, Concepcion G, Peluso P, Rank D, Jaiswal P. 2019. A phased, diploid assembly of the Cascade hop (*Humulus lupulus*) genome reveals patterns of selection and haplotype variation. *BioRxiv*: 786145. doi: 10.1101/786145.
- Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. 2010. How many bootstrap replicates are necessary? *Journal of Computational Biology* 17: 337–354.
- Patzak J, Nesvadba V, Chmelarsky I, Vejř P, Skupinová S. 2002. Identification of sex in F1 progenies of hop (*Humulus lupulus*) by molecular marker. *Plant, Soil & Environment* 48: 318–321.
- Peil A, Flachowsky H, Schumann E, Weber WE. 2003. Sex-linked AFLP markers indicate a pseudoautosomal region in hemp (*Cannabis sativa* L.). *Theoretical and Applied Genetics* 107: 102–109.
- Peneder P, Wallner B, Vogl C. 2017. Exchange of genetic information between thierian X and Y chromosome gametologs in old evolutionary strata. *Ecology and Evolution* 7: 8478–8487.
- Petit RJ, Hampe A. 2006. Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution & Systematics* 37: 187–214.
- Prentout D, Razumova O, Rhoné B, Badouin H, Henri H, Feng C, Käfer J, Karlov G, Marais GAB. 2020. An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of *Cannabis sativa*. *Genome Research* 30: 164–172.
- Polley A, Ganai MW, Seigner E. 1997. Identification of sex in hop (*Humulus lupulus*) using molecular markers. *Genome* 40: 357–361.
- Pucholt P, Wright AE, Conze LL, Mank JE, Berlin S. 2017. Recent sex chromosome divergence despite ancient dioecy in the Willow *Salix viminalis*. *Molecular Biology and Evolution* 34: 1991–2001.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.
- R Core Team. 2018. *R: A language and environment for statistical computing, v.3.4.4*. Vienna, Austria: R Foundation for Statistical Computing. [WWW document] URL <https://www.R-project.org/>.
- Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: multiple alignment of coding Sequences accounting for frameshifts and stop codons. *PLoS ONE* 6: e22594.
- Raymond O, Guozy J, Just J, Badouin H, Verdenaud M, Lemainque A, Vergne P, Moja S, Choisine N, Pont C *et al.* 2018. The Rosa genome provides new insights into the domestication of modern roses. *Nature Genetics* 50: 772–777.
- Renner SS. 2014. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *American Journal of Botany* 101: 1588–1596.
- Renner SS, Müller NA. 2021. Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. *Nature Plants* 1–11.
- Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP *et al.* 2005. The DNA sequence of the human X chromosome. *Nature* 434: 325–337.
- Sawyer S. 1999. GENECONV: A computer package for the statistical detection of gene conversion. [WWW document] URL <http://www.math.wustl.edu/~sawyer>.
- Shepherd HL, Parker JS, Darby P, Ainsworth CC. 2000. Sexual development and sex chromosomes in hop. *New Phytologist* 148: 397–411.
- Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP *et al.* 2011. The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics* 43: 109–116.
- Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier LD, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T *et al.* 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423: 825–837.
- Small E. 2015. Evolution and classification of *Cannabis sativa* (Marijuana, Hemp) in relation to human utilization. *The Botanical Review* 81: 189–294.
- Sousa A, Fuchs J, Renner SS. 2013. Molecular cytogenetics (FISH, GISH) of *Coccinia grandis*: A ca. 3 myr-old species of cucurbitaceae with the largest Y/autosome divergence in flowering plants. *Cytogenetic and Genome Research* 139: 107–118.
- Takahata N, Nei M. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110: 325–344.
- Thomas GG, Neve RA. 1976. Studies on the effect of pollination on the yield and resin content of hops (*Humulus lupulus* L.). *Journal of the Institute of Brewing* 82: 41–45.
- Torres MF, Mathew LS, Ahmed I, Al-Azwani IK, Krueger R, Rivera-Núñez D, Mohamoud YA, Clark AG, Suhre K, Malek JA. 2018. Genus-wide sequencing supports a two-locus model for sex-determination in *Phoenix*. *Nature Communications* 9: 1–9.
- Trombetta B, Sellitto D, Scozzari R, Cruciani F. 2014. Inter- and intraspecies phylogenetic analyses reveal extensive X-Y gene conversion in the evolution of gametologous sequences of human sex chromosomes. *Molecular Biology and Evolution* 31: 2108–2123.
- van Velzen R, Holmer R, Bu F, Rutten L, van Zeijl A, Liu W, Santuari L, Cao Q, Sharma T, Shen D *et al.* 2018. Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proceedings of the National Academy of Sciences, USA* 115: E4700–E4709.
- Wang J, Na J-k, Yu Q, Gschwend Ar, Han J, Zeng F, Aryal R, VanBuren R, Murray Je, Zhang W *et al.* 2012. Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proceedings of the National Academy of Sciences, USA* 109: 13710–13715.
- Westergaard M. 1958. The mechanism of sex determination in dioecious flowering plants. *Advances in Genetics* 9: 217–281.
- Wickham H. 2011. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* 3: 180–185.
- Wright AE, Dean R, Zimmer F, Mank JE. 2016. How to make a sex chromosome. *Nature Communications* 7: 1–8.
- Wu M, Moore RC. 2015. The evolutionary tempo of sex chromosome degradation in *Carica papaya*. *Journal of Molecular Evolution* 80: 265–277.
- Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873–881.
- Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology* 1418: 283–334.
- Yang M-Q, van Velzen R, Bakker FT, Sattarian A, Li D-Z, Yi T-S. 2013. Molecular phylogenetics and character evolution of Cannabaceae. *Taxon* 62: 473–485.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Zhang Q, Onstein RE, Little SA, Sauquet H. 2019. Estimating divergence times and ancestral breeding systems in *Ficus* and *Moraceae*. *Annals of Botany* 123: 191–204.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

**Fig. S1** Example of genes whose topology changed with the mapping bias filtering.

**Fig. S2** Histogram of ( $dS$ ) between *C. sativa* and *M. notabilis*.

**Fig. S3** Histogram of  $dS$  between *C. sativa* and *R. chinensis*.

**Fig. S4** Histogram of the Y/X expression ratio.

**Fig. S5** Histogram of the Allele-specific expression analysis for the parents.

**Fig. S6** Histogram of the Allele-specific expression analysis for the daughters in the nonrecombining region.

**Fig. S7** Histogram of the Allele-specific expression analysis for the daughters out of the nonrecombining region.

**Fig. S8** SEX-DETECTOR inference errors due to WGD in *H. lupulus*.

**Fig. S9** Y/X expression ratio along the sex chromosome without genes with a detected mapping bias.

**Fig. S10** Dosage compensation analysis without genes with a detected mapping bias.

**Methods S1** Methods for the RNA-seq mapping on *H. lupulus* reference, for the estimation of the synonymous divergence with two outgroup species and for the Allele-specific expression analysis.

**Table S1** Summary of chromosome names in Fig. 2(a) and in assembly fasta file.

**Table S2** Statistics of mapping on *H. lupulus* and *C. sativa* references.

**Table S3** Summary of SEX-DETECTOR genotyping errors and inferences.

**Table S4** SEX-DETECTOR assignment file output.

**Table S5** Age estimates of the youngest strata in *H. lupulus* and *C. sativa*.

**Table S6** Expression analysis statistics summary.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

## **New Phytologist Supporting Information**

**Article title: Plant genera *Cannabis* and *Humulus* share the same pair of well-differentiated sex chromosomes**

**Authors: Djivan Prentout, Natasa Stajner, Andreja Cerenak, Theo Tricou, Celine Brochier-Armanet, Jernej Jakse, Jos Käfer and Gabriel AB Marais**

**Article acceptance date: 29 April 2021**

## **Methods S1 - Supplementary Materials and Methods**

### **RNA-seq mapping on assemblies, genotyping and SEX-DETECTOR analyses**

The RNA-seq data were mapped on two references: the *Humulus lupulus* transcriptome (obtained from the genome annotation; Padgitt-Cobb *et al.*, 2019), and the *Cannabis sativa* draft transcriptome (van Bakel *et al.*, 2011). For each reference we used GSNAP with the parameters described in the main text, except that we allowed 5% of mismatches for mapping on *H. lupulus* reference instead of 10% for mapping on *C. sativa* reference. Genotyping and SEX-DETECTOR analyses were performed in identical ways (as described in the main text).

### **Synonymous divergence with outgroup species**

Synonymous divergence was estimated between *C. sativa* and *Rosa chinensis*, and *C. sativa* and *Morus notabilis*. We computed the  $dS$  with homologous sequences used for phylogenetic analyses ( $n = 85$ ). The Open Reading Frame (ORF) was determined during the alignment step of the phylogenetic pipeline since we used protein coding sequences as guide for the nucleotide alignment. We computed  $dS$  with *codeml* as described in the main text.

### **Allele-specific Expression (ASE) analysis**

We computed the ASE by counting the level of expression for each SNP with GATK (McKenna *et al.*, 2010). We retained SNPs present in at least 2 offspring and with at least 10 reads. Then, we computed the mean expression of both alleles for each genes with at least 10 SNPs. Finally, we computed the ratio: allele 1 expression / allele 2 expression.

# Results

## Chromosome names

Table S1. Correspondence between chromosome names used for the Fig 2a and used for the chromosome-level assembly in Grassa *et al.* (2018).

Chromosome name in Fig. 2a	Chromosome name in Grassa <i>et al.</i> (2018)
Chromosome X	LR213627.1
Chromosome 2	LR213628.1
Chromosome 3	LR213629.1
Chromosome 4	LR213630.1
Chromosome 5	LR213631.1
Chromosome 6	LR213632.1
Chromosome 7	LR213633.1
Chromosome 8	LR213634.1
Chromosome 9	LR213635.1
Chromosome 10	LR213636.1

## Mapping results

Table S2. Mapping results on several references at fourth iteration of GSNAP. Indicated are the library sizes, and the total numbers and percentages of properly paired reads.

Assembly	Father	Male 1	Male 2	Male 3	Male 4	Male 5	Male 6	Mother	Female 1	Female 2	Female 3	Female 4	Female 5	Female 6
Library size (in million reads)	62.6	84.2	85.1	95.7	94.6	69.6	62.5	64.8	61.6	92.7	86.0	70.0	83.2	74.7
<i>H. lupulus</i> transcriptome (in million reads)	48.5	63.5	61.6	67.7	49	46.7	46.4	44.9	46.3	68.0	66.1	51.7	59.4	56.6
<i>H. lupulus</i> transcriptome (%)	77.3	75.4	72.3	70.7	75.9	67	74.3	69.3	75.4	73.4	76.9	73.8	71.3	73.8
<i>C. sativa</i> transcriptome (in million reads)	22.0	28.5	27.2	29.8	20.2	20.3	20.6	20.7	21.6	32.0	31.1	26.8	27.8	26.4
<i>C. sativa</i> transcriptome (%)	35.1	33.8	32.0	31.2	31.2	29.8	33.0	31.9	35.0	34.5	36.2	38.2	33.6	35.4



## SEX-DETECTOR outputs

Table S3. SEX-DETECTOR genotyping errors and SEX-DETECTOR inferences, for mapping on *C. sativa* transcriptome (30,074 genes) and *H. lupulus* transcriptome (26,149 genes) references, at first and fourth GSNAP iterations. Both genotyping error rates are presented :  $\varepsilon$  – Genotyping error rate for whole transcriptome;  $p$  – Y genotyping error rate. Numbers of autosomal genes, XY genes and X-hemizygous genes are presented for each mapping.

	<b>E</b>	<b>P</b>	<b>#Autosomal genes</b>	<b>#XY genes</b>	<b>#X-hemizygous genes</b>
<i>H. lupulus</i> genome 1 <sup>st</sup> iteration	0.05	0.88	970	0	0
<i>H. lupulus</i> genome 4 <sup>th</sup> iteration	0.05	0.85	998	0	0
<i>C. sativa</i> 1 <sup>st</sup> iteration	0.05	0.86	2948	0	0
<i>C. sativa</i> 4 <sup>th</sup> iteration	0.06	0.20	3103	122	0
<i>C. sativa</i> 4 <sup>th</sup> iteration minus male 3	0.06	0.1	3391	265	0

## Choice of the assembly

As explained in the main text, the optimization of the parameter  $p$  is crucial to correctly interpret SEX-DETECTOR outputs. The Table S3 shows that GSNAP mapping on *C. sativa* transcriptome was the best way to reduce  $p$  to a reasonable value. Moreover, SEX-DETECTOR didn't identify sex-linked genes with the mapping on *H. lupulus* reference, which is likely explained by the high value of  $p$ . Furthermore, the authors of the *H. lupulus* genome assembly put forward the hypothesis of two Whole Genome Duplications (WGD) to explain the size of *H. lupulus* genome (>3Gb; Padgitt-Cobb *et al.*, 2019). Although they filtered these duplications in the “deduplicated” annotation, a BUSCO analysis of the assembly revealed that 28% of genes are still duplicated (Padgitt-Cobb *et al.*, 2019). The absence of sex-linked genes identification, the high value of  $p$ , and the high rate of duplicated genes in *H. lupulus* assembly convinced us to favour *C. sativa* reference transcriptome despite a lower mapping quality.

PK09610, position = 49217204 bp

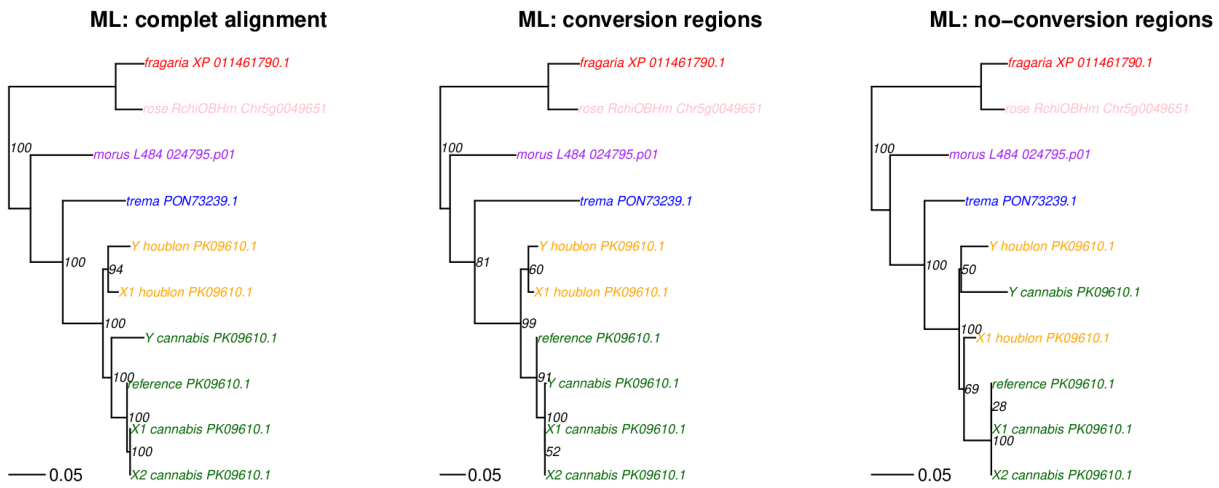


Figure S1. Phylogenetic results with maximum likelihood (ML) method for the gene PK11270 (on chromosome 1) for which geneconv identified gene conversion in a region representing around 50% of the gene length. On the left: the topology obtained for the whole sequence, in the middle: topology obtained for the region identified by geneconv as gene conversion, on the right: topology obtained for the region of the gene without gene conversion.

Synonymous divergence with *M. notabilis* and *R. chinensis*

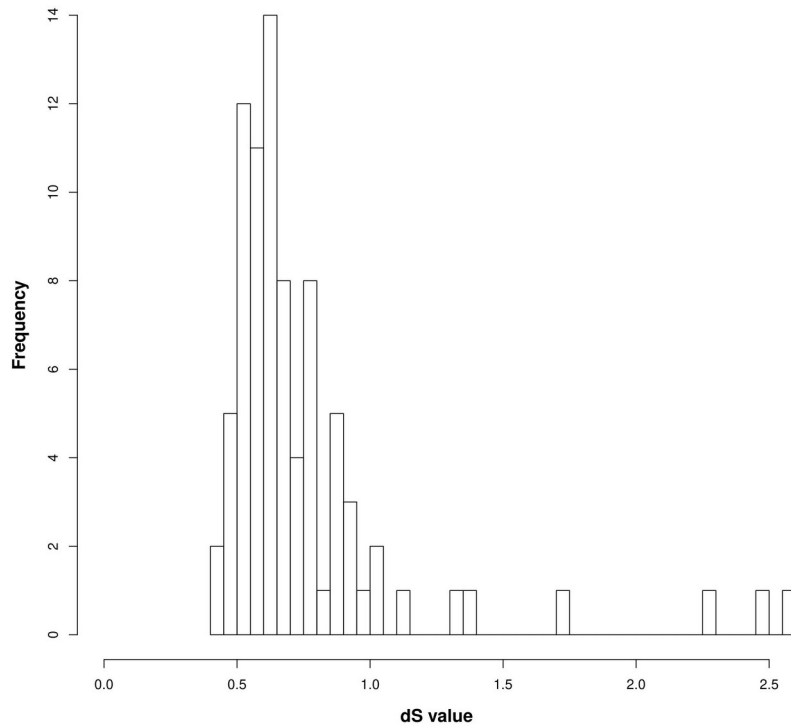


Figure S2. Histogram of synonymous divergence ( $dS$ ) between *C. sativa* and *M. notabilis*. The lowest value is 0.437.

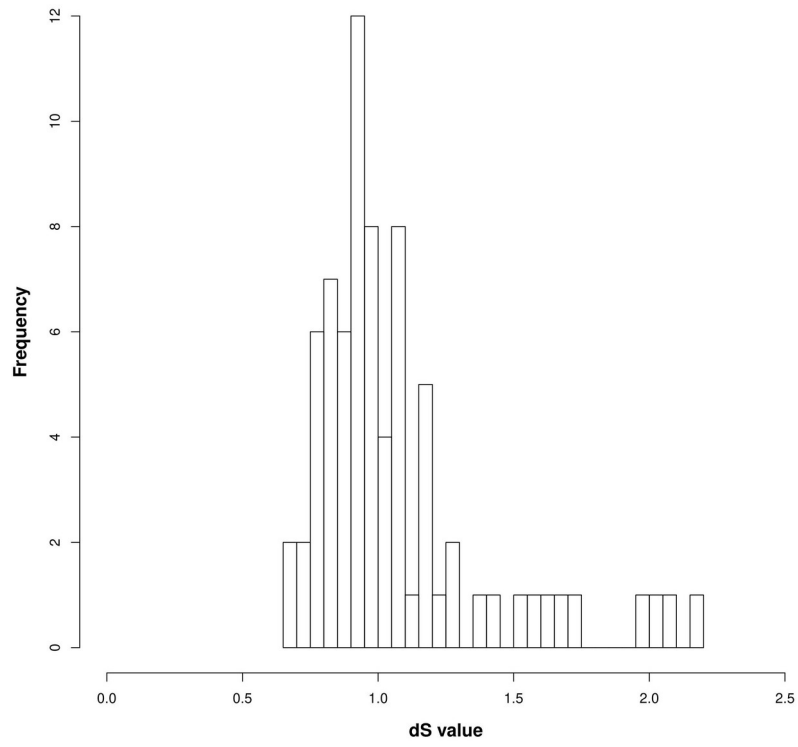


Figure S3. Histogram of synonymous divergence ( $dS$ ) between *C. sativa* and *R. chinensis*. The lowest value is 0.673.

## Expression results and ASE analysis

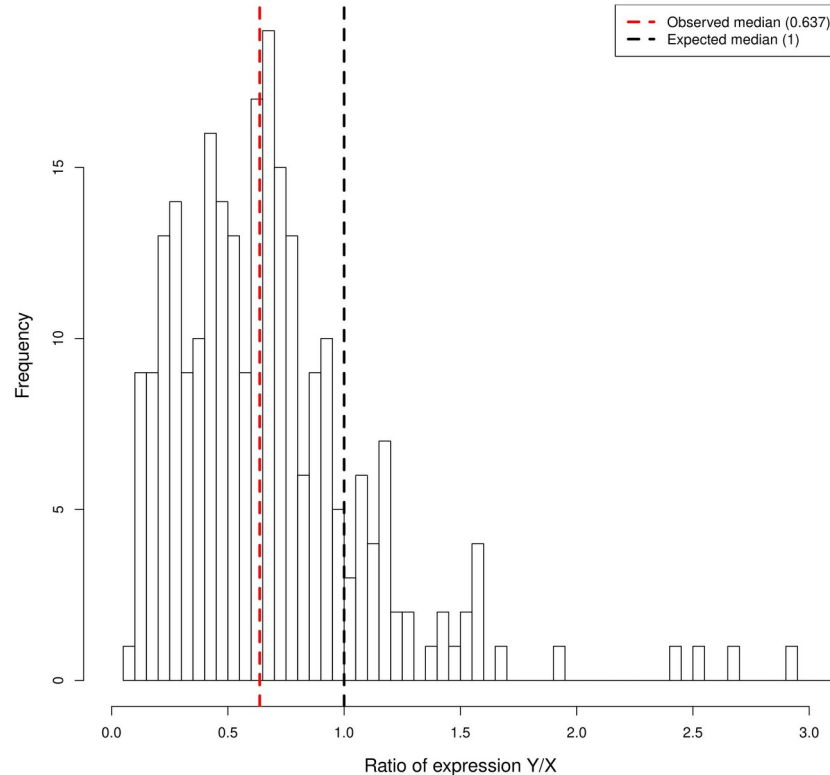


Figure S4. Histogram of the Y/X expression ratio. The dotted red bar represents the observed median (=0.637), and the dotted black bar the expected median of the ratio without Y degeneration (=1).

We quantified the ASE in the non-recombining region to determine the presence of X Chromosome Inactivation (XCI) in females. We computed the ratios of expression of one allele (A1) over the other one (A2) for each gene. We represented histograms of these ratios for genes in the non-recombining region and for the whole transcriptome. Figure S5 shows these histograms for the parents, Figure S6 for the daughters in the non-recombining region and Figure S7 for the daughters out of the non-recombining region.

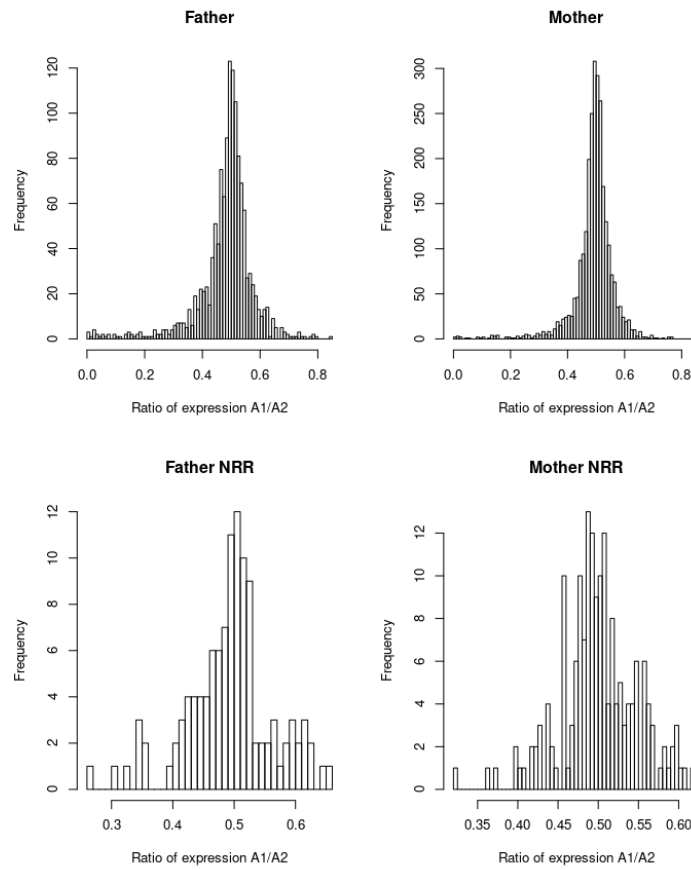


Figure S5. Histogram of ratio of A1 expression / A2 expression. Histograms are presented for the two parents. On the top, the ASE for the whole transcriptome, on the bottom, the ASE for the non recombining region (NRR) only.

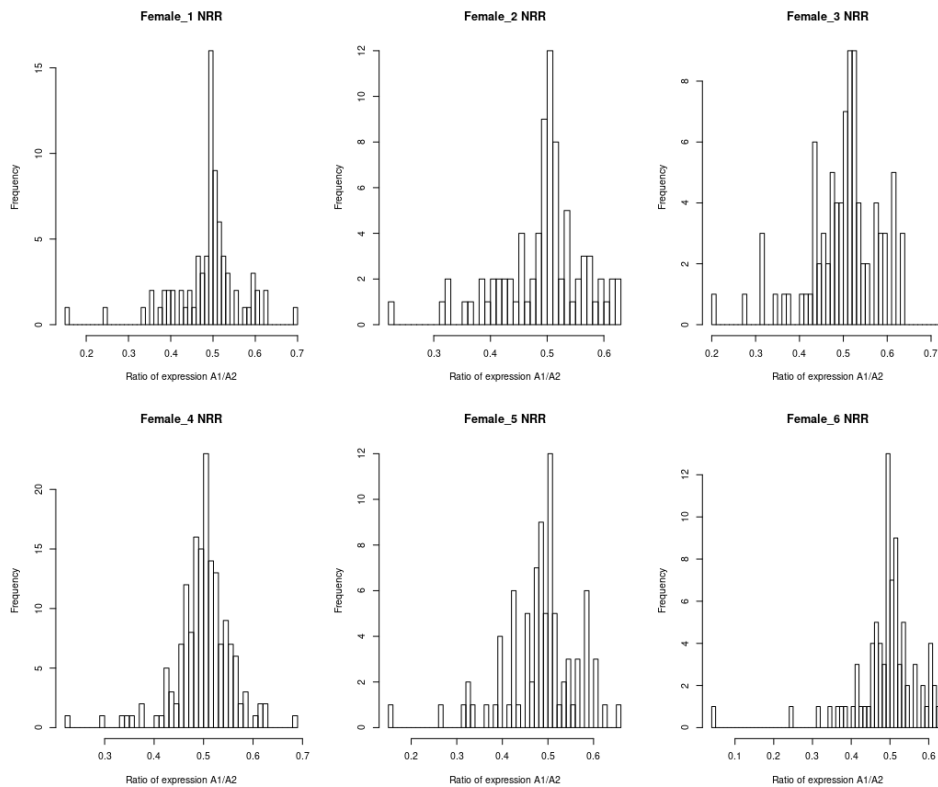


Figure S6. Histogram of ratio of A1 expression / A2 expression in daughters for genes in the non-recombining region (NRR).

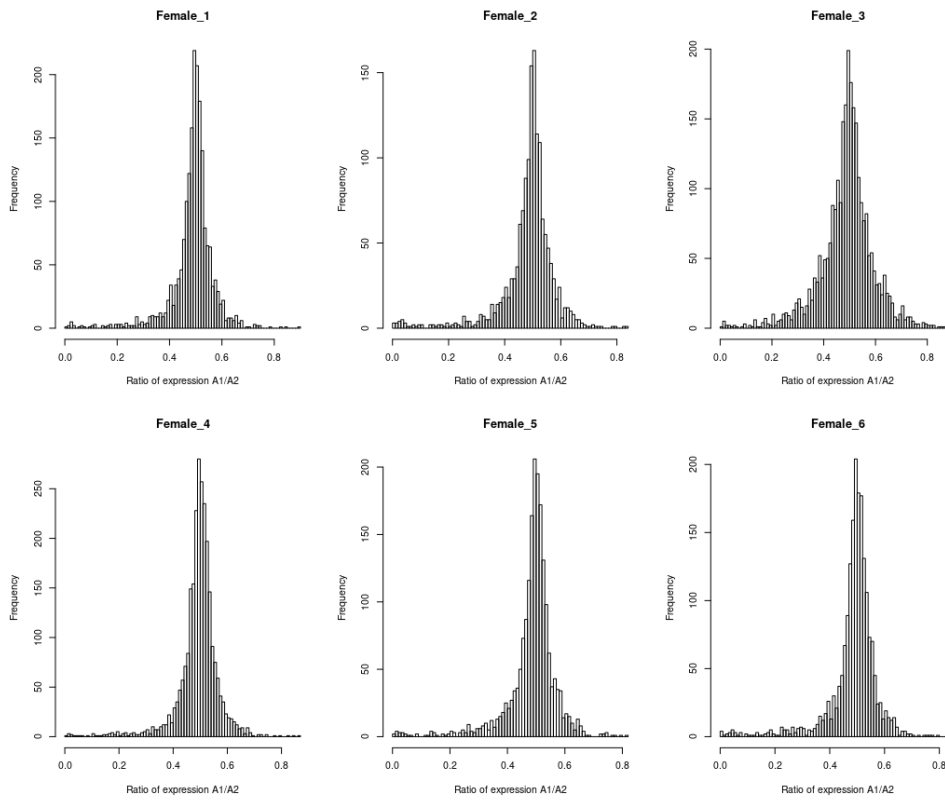


Figure S7. Histogram of ratio of A1 expression / A2 expression in daughters for genes out of the non-recombining region.

The three latter Figures don't support the hypothesis of an XCI since the mode of the allele expression ratio distribution is close to 0.5 in the non-recombining region. This ratio indicates that both alleles are expressed in same proportion, which is not expected with XCI.

## Duplications impact on SEX-DETECTOR inferences

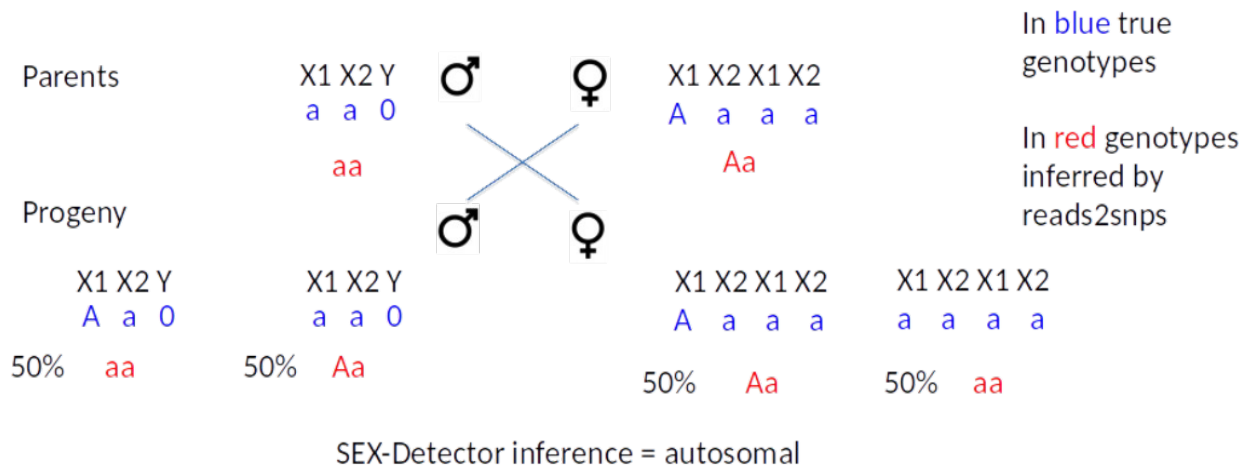


Figure S8: Hypothesis to explain the absence of X-hemizygous genes. If the *H. lupulus* X chromosome comprised two X1 + X2 chromosomes (from a possible WGD, see Padgitt-Cobb *et al.*, 2019) whose reads map to the *C. sativa* X chromosome, then X-hemizygous genes will be impossible to detect. This is because ploidy differences between males and females is no longer detectable as shown in the example above. In this example, X-hemizygous genes will have an autosomal-like segregation. We should observe an excess of autosomal genes on the X-specific region XSR in *H. lupulus* (compared to *C. sativa*). The X-hemizygous genes detected in *C. sativa* should have a XY or autosomal segregation type in *H. lupulus*.

The example in Figure S8 shows that, under the hypothesis of WGDs in *H. lupulus*, some SNPs in X-hemizygous genes may be classified as autosomal by SEX-DETECTOR. Thus, genomic data or cytological analysis are needed to validate this hypothesis.

## Reduction of Y expression and dosage compensation

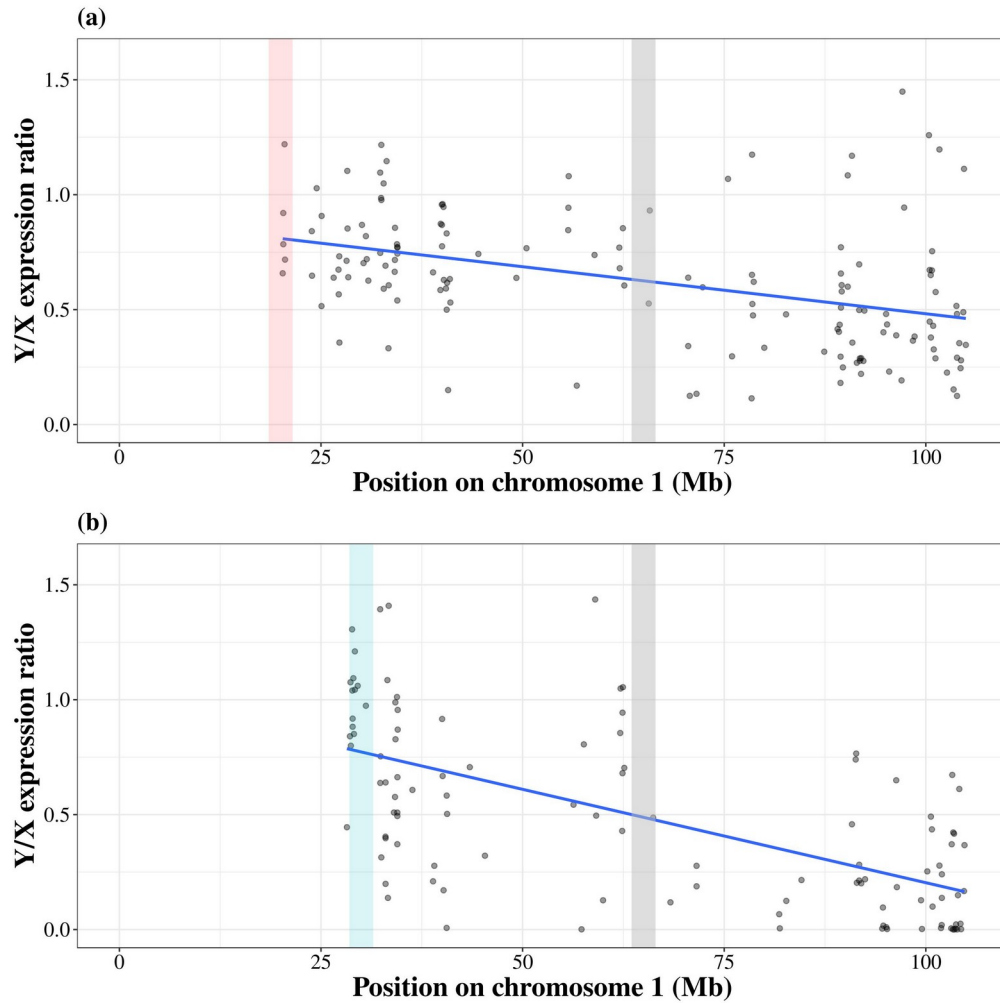


Figure S9. Y/X expression ratio along the X chromosome in *H. lupulus* **(a)**, and *C. sativa* **(b)** for genes without the detection of mapping bias with geneconv. The grey dots represent the Y/X expression ratio for each gene in the non-recombining region only. The blue line represents a linear regression (adjusted  $R^2=0.174$ ,  $p$ -value $<10^{-5}$ , and adjusted  $R^2=0.391$ ,  $p$ -value $<10^{-5}$  for *H. lupulus* and *C. sativa*, respectively). The vertical red bar represents the putative Pseudo-Autosomal Boundary (PAB) in *H. lupulus*, the vertical blue bar represents the putative PAB in *C. sativa*, the vertical grey bar represents the putative boundary between the region that stopped recombining in the common ancestor and the region that stopped recombining independently in the two species.



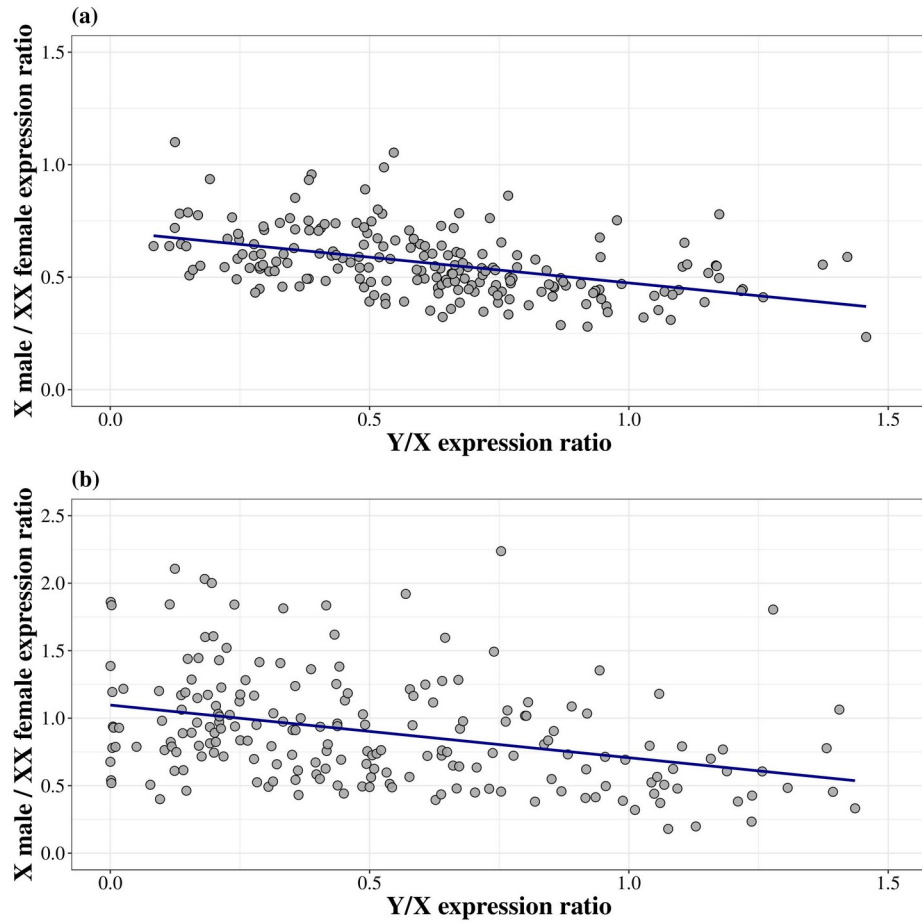


Figure S10. The male X expression over female XX expression versus Y/X expression ratio for *H. lupulus* (a) and *C. sativa* (b). Each black dot represents one gene. The blue line represents a linear regression (adjusted  $R^2=0.199$ ,  $p\text{-value}<10^{-5}$ , and adjusted  $R^2=0.102$ ,  $p\text{-value}<10^{-5}$  for *H. lupulus* and *C. sativa*, respectively).

We repeated the expression level analyses using a dataset without the genes with possible mapping bias. The adjusted  $R^2$  is always greater when removing genes for which we identified a mapping bias, as summarized in Table S6. This results confirmed that the reduction of the Y expression and the dosage compensation for genes with a Y expression strongly reduced, are not induced by the mapping bias of Y divergent sequences.

Table S5. Age estimates (in millions of years, My) of the youngest stratum in *H. lupulus* (20Mb-65Mb) and *C. sativa* (30Mb-65Mb) with two molecular clocks and different maximum  $dS$  values. We took a generation time of 2 years for *H. lupulus* and 1 year for *C. sativa*. For each  $dS$  value, two ages were obtained using the molecular clocks of <sup>1</sup>Ossowski *et al.* (2010) and <sup>2</sup>Koch *et al.* (2000).

	<i>H. lupulus</i>			<i>C. sativa</i>		
	$dS$	age (My) <sup>1</sup>	age (My) <sup>2</sup>	$dS$	age (My) <sup>1</sup>	age (My) <sup>2</sup>
Highest $dS$	0.206	29.4	13.7	0.277	19.8	18.5
Mean highest 5%	0.178	25.4	11.9	0.253	18.1	16.9
Mean highest 10%	0.152	21.7	10.1	0.239	17.1	15.9

Table S6. Summary of adjusted R<sup>2</sup> for Y/X expression ratio and dosage compensation analyses for the two datasets. Dataset 1 = all genes; and Dataset 2 = genes without mapping bias identification only.

	<i>H. lupulus</i>		<i>C. sativa</i>	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
Dosage compensation analysis	0.179	0.199	0.097	0.102
Y/X expression ratio analysis	0.134	0.174	0.278	0.391

## References

- van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, Page JE. 2011.** The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology* **12**, 1-18.
- Grassa, C. J., Wenger, J. P., Dabney, C., Poplawski, S. G., Motley, S. T., Michael, T. P., et al 2018.** A complete Cannabis chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. *BioRxiv*, 458083.
- Koch MA, Haubold B, Mitchell-Olds T. 2000.** Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution* **17**: 1483–1498.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A., et al 2010.** The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303.
- Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010.** The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. *Science* **327**: 92–94.
- Padgitt-Cobb, L. K., Kingan, S. B., Wells, J., Elser, J., Kronmiller, B., Moore, D., et al. 2019.** A phased, diploid assembly of the Cascade hop (*Humulus lupulus*) genome reveals patterns of selection and haplotype variation. *BioRxiv*, 786145.



# 6

## Discussion

# 1 Résumé des principaux résultats de thèse

Au cours de cette thèse j'ai analysé des données transcriptomiques de plantes dioïques afin d'apporter des éléments de compréhension concernant l'évolution des chromosomes sexuels chez les plantes. Premièrement, j'ai voulu tester l'existence d'une paire de chromosomes sexuels chez *Silene acaulis* ssp *exscapa*. Cette analyse, présentée dans le premier axe, a confirmé l'intérêt de ce modèle biologique pour étudier la formation de chromosomes sexuels. Bien que nos résultats ne permettent pas de répondre clairement à la question, ils montrent que si une région non-recombinante existe, elle sera probablement petite et récente. Deuxièmement, j'ai décrit une paire de chromosomes sexuels homologues entre deux espèces qui ont divergé il y a plus de 20 millions d'années, à savoir, *Cannabis sativa* et *Humulus lupulus*. C'est la première fois qu'une paire de chromosomes sexuels homologues à deux d'espèces appartenant à des genres différents est décrite chez les plantes. Cette paire de chromosomes sexuels est aussi unique de par son âge et sa dégénérescence au regard des systèmes actuellement décrits chez les plantes.

## 2 Perspectives

### 2.1 *Silene acaulis*

Concernant les analyses faites chez *S. acaulis* ssp *exscapa*, il reste à déterminer si les gènes identifiés comme potentiellement non-recombinants sont des vrais ou des faux positifs. Comme discuté dans le chapitre 2, un génome assemblé au niveau chromosomique pourrait aider à déterminer si une région non-recombinante existe chez *S. acaulis* ssp *exscapa*. Au début de ma thèse, nous avons organisé une réunion avec différentes personnes travaillant sur *S. acaulis* en Europe. Au sein de ce groupe de spécialistes de *S. acaulis*, des collaborateurs d'une équipe de l'ETH de Zurich, Alex Widmer et Martin Fischer, sont justement en train d'assembler un génome de référence. Pour ce faire, ils ont généré des données de séquençage de lectures courtes et longues. Ce sont des collaborateurs avec qui j'ai eu des échanges réguliers au cours de ma thèse. En effet, leurs travaux concernent principalement les liens génétiques entre sous-espèces du complexe *S. acaulis*, mais aussi l'étude de la génétique du sexe chez *S. acaulis* ssp *exscapa*. Nos échanges portent surtout sur cette deuxième partie de leurs travaux. J'ai par exemple analysé des données RAD-seq d'une population de *S. acaulis* dioïque séquençée par cette équipe. Cette analyse n'a pas permis de conclure sur la présence d'une région non-recombinante et n'a pas été présentée dans cette thèse. De leur côté, ils n'ont pas non plus réussi à déterminer si une paire de chromosomes sexuels existe chez cette sous-espèce, ce qui souligne la difficulté de travailler sur ce système avec les données actuelle-

ment disponibles. Les technologies de séquençage qu'ils utilisent pour l'assemblage devraient permettre d'obtenir un génome de référence de bonne qualité, voire au niveau chromosomique. Une fois l'assemblage disponible, il sera possible de réaliser une partie des analyses discutées dans le second chapitre.

Par ailleurs, ma troisième année de thèse a été en bonne partie dédiée à un projet d'analyse comparative de gènes à expression différentielle entre mâles et femelles chez les plantes. Neufs espèces de silènes dioïques (dont *S. acaulis* ssp *exscapa*) issues de trois événements indépendants de transition vers la dioécie sont comparées. Cette analyse permet, par exemple, de tester s'il y a convergence des gènes dont l'expression est différentielle en fonction du sexe. Aussi, les événements d'émergence de la dioécie ayant des âges différents, je vais pouvoir étudier quels sont les premiers gènes dont l'expression devient différentielle suite à une transition vers la dioécie. Ce projet, toujours en cours, devrait en partie permettre de mieux comprendre les mécanismes génétiques qui régulent le sexe chez *S. acaulis* ssp *exscapa*.

## 2.2 *Cannabaceae*

Mes travaux ont mis en évidence des paires de chromosomes sexuels âgées de plusieurs dizaines de millions d'années avec une dégénérescence du chromosome Y importante pour des plantes. Aucun système de plus de 20 millions d'années avec une région non-recombinante de 70Mb n'avait été identifié jusqu'à présent (voire la Table 2 dans l'introduction)(résumé dans Carey, Yu et Harkess (2021); Renner et Müller (2021); Charlesworth (2021b)). Il est probable que la moitié de la région aujourd'hui non-recombinante l'était déjà chez le dernier ancêtre commun de ces deux espèces. Il reste tout de même des analyses à réaliser pour mieux caractériser l'évolution des chromosomes sexuels de *Cannabis sativa* et de *Humulus lupulus*. De plus, les gènes du déterminisme du sexe restent inconnus chez ces deux espèces. À l'instar des chromosomes sexuels, les gènes du déterminisme du sexe chez les plantes restent généralement mal décrits bien qu'un nombre croissant d'études aient été publiées ces dernières années (*e.g.* différentes espèces du genre *Populus*; *Asparagus officinalis*; *Actinidia chinensis*; *Vitis vinifera* ssp *vinifera*)(Müller et al., 2020; Harkess et al., 2020; Akagi et al., 2019; Massonnet et al., 2020; Badouin et al., 2020; Zou et al., 2021). Au-delà des intérêts fondamentaux d'étudier les gènes du déterminisme du sexe chez les plantes, les applications sont multiples dans l'agronomie (résumé dans Heikrujam et al. (2015)). Les perspectives pour *C. sativa* et *H. lupulus* sont énumérées indépendamment dans les deux sous-sections ci-dessous.

### 2.2.1 *Cannabis sativa*

Les analyses faites sur *Cannabis sativa* ouvrent plusieurs perspectives.

Tout d'abord, la caractérisation de l'évolution des chromosomes sexuels de *C. sativa* n'est pas encore terminée. Il serait intéressant de déterminer si des strates évolutives sont présentes

sur cette paire de chromosomes sexuels. Par ailleurs, connaître la dynamique d'insertion d'éléments transposables dans la région non-recombinante du chromosome Y permettrait d'en savoir plus sur l'homomorphie chromosomique de ce système. Expliquer cette homomorphie chromosomique dans un système aussi ancien serait une première chez les plantes. Enfin, l'analyse SEX-DETECTOR a été réalisée à partir de données de RNA-seq. Ceci limite donc les gènes étudiés à ceux qui sont exprimés dans le tissu lors de l'échantillonnage. De plus, SEX-DETECTOR classe des gènes dans la partie de la région pseudo-autosomale proche de la région non-recombinante (région appelée Pseudo-Autosomal Boundary - PAB) comme liés au sexe (Muyle et al., 2016). Ceci conduit généralement à une surestimation de la taille de la région non-recombinante. Un assemblage du chromosome Y fournirait les informations nécessaires pour clarifier ces derniers points.

Ensuite, le mécanisme génétique responsable du déterminisme du sexe chez *C. sativa* reste inconnu. Son identification possède des intérêts à la fois fondamentaux et appliqués. Très peu de gènes responsables du déterminisme du sexe ont été identifiés chez les plantes jusqu'à présent. Continuer à en identifier de nouveaux est donc primordial d'un point de vue fondamentale. Par ailleurs, une meilleure compréhension du mécanisme qui détermine le sexe pourrait être utile dans l'industrie du cannabis thérapeutique et récréatif (voir ci-dessous). Afin d'identifier le ou les gènes responsables du déterminisme du sexe, nous avons débuté une collaboration avec un groupe du University College of Dublin. Cette équipe de recherche, dirigée par Rainer Melzer, fait dorénavant partie des groupes spécialistes de la génétique du développement chez *C. sativa* (Schilling et al., 2020; Dowling, Melzer et Schilling, 2021). Leurs travaux portent principalement sur la génétique de la floraison et, depuis notre collaboration, sur le déterminisme génétique du sexe. Ce groupe possède une expertise expérimentale qui est complémentaire à notre expertise bio-informatique. Grâce à mes travaux de thèse j'ai pu sélectionner un ensemble de gènes qui seraient potentiellement responsables du déterminisme du sexe. Les critères de sélection ont été fixés pour obtenir des gènes qui supportent l'hypothèse d'un déterminisme X / autosomes (*e.g.* X-hémizygote, expression non compensée chez les mâles, présent dans la région la plus ancienne, fonction en lien avec les phytohormones liées au développement du sexe). Ces gènes sont actuellement testés par nos collaborateurs. Enfin, les amorces spécifiques du chromosome Y que nous avons développées nécessitent d'être validées sur un plus grand nombre de variétés et d'individus (voir le chapitre 4) (Pren-tout et al., 2020a). Une collaboration a été mise en place avec la société coopérative agricole Hemp'it. Cette société française est spécialisée dans la production et la commercialisation de semences de chanvre. Grâce à cette collaboration, nous avons accès à des centaines d'individus provenant de dizaines de fonds génétiques différents, ce qui fournira une puissance statistique suffisante pour la validation (sur les variétés testées). Un protocole d'extraction d'ADN permettant de tester les amorces a été développé au sein du laboratoire (par Hélène Henri). Les essais sur quelques échantillons ont permis de sélectionner 2 paires d'amorces PCR Y-spécifiques et une paire d'amorces autosomales. Ces trois paires d'amorces ont des tailles différentes ce qui permet de faire des PCR en multiplex. Ce kit d'amorces multi-

plex a parfaitement fonctionné sur quelques individus mâles et femelles (n=6). Nous allons ainsi pouvoir commencer la validation à grande échelle. Une fois ces amorces validées sur le chanvre, nous pourrions envisager une application sur des variétés de marijuana.

### 2.2.2 *Humulus lupulus*

Certaines analyses réalisées sur les chromosomes sexuels de *Cannabis sativa* n'ont pas pu l'être sur ceux de *Humulus lupulus*. Il n'était par exemple pas possible de déterminer la taille de la région non-recombinante chez cette espèce. La perte de gènes n'a pas pu non plus être estimée chez *H. lupulus*. Ceci s'explique par deux raisons principales. Premièrement, l'absence de transcriptome de référence de bonne qualité et d'un génome assemblé au niveau chromosomique m'a contraint à analyser les données de *H. lupulus* avec les références disponibles chez *C. sativa*. Deuxièmement, le génome de *H. lupulus* est environ 4 fois plus gros que celui de *C. sativa*. Cette différence pourrait être expliquée par deux duplications complètes de génome depuis la séparation entre les genres *Humulus* et *Cannabis* (Padgitt-Cobb et al., 2021). Ces deux duplications complètes complexifient les analyses génomiques chez cette espèce.

Un génome assemblé au niveau chromosomique, incluant le chromosome Y, offrirait donc plusieurs perspectives à mon travail. Ceci permettrait dans un premier temps d'approfondir l'analyse évolutive des chromosomes sexuels pour les mêmes raisons que celles précédemment évoquées dans la section sur *C. sativa* (nombre de strates évolutives, nombre d'insertions d'éléments transposables, perte de matériel génétique, ...). Nos données semblent ne pas supporter l'hypothèse de grosses délétions. En effet, des gènes XY ont été retrouvés sur l'ensemble du chromosome X de *C. sativa*. Un assemblage du chromosome Y de *H. lupulus* est nécessaire pour déterminer si cette hétéromorphie chromosomique s'explique par une perte de matériel génétique ou par une absence de duplication du chromosome Y chez *H. lupulus*. Ce système deviendrait le premier système hétéromorphique avec un chromosome Y plus petit à être décrit avec des données génomiques chez les plantes. Une gymnosperme, *Cycas revoluta*, possède aussi une paire de chromosomes sexuels avec un Y plus petit que le X mais n'a pas été étudiée avec des données génomiques (Segawa, Kishi et Tatuno, 1971). L'assemblage du chromosome X est nécessaire pour déterminer la taille de la région non-recombinante chez *H. lupulus*. Comme discuté dans le cinquième chapitre, l'absence de gène X-hemizygotes dans les résultats de SEX-DETECTOR est probablement plus expliquée par des biais méthodologiques que par une absence biologique. Un assemblage du génome de *H. lupulus* est nécessaire pour le déterminer.

Enfin, la production du houblon pourrait être améliorée avec le développement de marqueurs spécifiques du Y. La propagation végétative fonctionne bien chez cette plante, il est donc relativement simple de conserver uniquement des femelles dans les cultures. En revanche, un sexage précoce des individus pourrait faciliter les croisements contrôlés et l'amélioration variétale. Tout comme pour le cannabis, mes travaux sur les chromosomes sexuels du houblon offrent cette perspective. Aucune collaboration n'a été mise en place pour le moment, mais des échanges à ce sujet ont eu lieu avec nos collaborateurs slovènes (les collaborateurs



impliqués dans l'article du cinquième chapitre de thèse).

### 3 Apports de mes travaux pour la compréhension de l'évolution des chromosomes sexuels

Depuis quelques années, les progrès des technologies de séquençage et le développement de méthodes pour identifier des paires de chromosomes sexuels ont permis de limiter les coûts et les efforts nécessaires à l'identification de régions non-recombinantes (résumé dans Carey, Yu et Harkess (2021)). Ainsi, parmi la trentaine de paires de chromosomes sexuels évoquées dans l'introduction, une dizaine a été identifiées chez les plantes depuis Janvier 2020 (*e.g. Salix nigra, Amborella trichopoda, Salix Viminalis, Solanum appendiculatum, Coccinia grandis, Cannabis sativa, Humulus lupulus, Ginkgo biloba*, etc) (Sanderson et al., 2021; Käfer et al., 2021; Almeida et al., 2020; Wu et al., 2021; Fruchard et al., 2020; Prentout et al., 2020b, 2021; Liao et al., 2020). Mes travaux participent donc significativement à l'enrichissement de systèmes décrits chez les plantes. Plusieurs revues ont aussi été publiées ces deux dernières années, soulignant la quantité d'informations qui nécessitent d'être synthétisées (Baránková et al., 2020; Renner et Müller, 2021; Charlesworth, 2021b). L'article sur les chromosomes sexuels de *C. sativa* est le seul à avoir été publié avant la parution des récentes revues et est cité dans chacune d'elle (Baránková et al., 2020; Charlesworth, 2021b; Renner et Müller, 2021). Le nombre croissant d'études sur les chromosomes sexuels a mis en évidence une diversité de schémas évolutifs plutôt qu'une unité (résumé dans : Bachtrog et al. (2014); Furman et al. (2020); Charlesworth (2021a)). En effet, depuis plusieurs années, émerge l'idée qu'un modèle universel d'évolution des chromosomes sexuels n'est plus envisageable (Bachtrog et al., 2014; Furman et al., 2020). Les chromosomes sexuels de plantes ont beaucoup été étudiés pour tenter de comprendre les stades précoces de l'évolution des chromosomes sexuels, car ils sont généralement plus jeunes que chez les animaux (résumé dans Charlesworth (2019, 2021b)). L'absence de données empiriques d'une paire de chromosomes sexuels ancienne et fortement dégénérée chez les plantes laissait penser que ceux-ci n'existaient que chez certains animaux. Cependant, mes travaux sur l'évolution des chromosomes sexuels dans la famille des Cannabaceae ont montré qu'il existe des chromosomes sexuels stables chez les plantes, qui peuvent être partagés entre espèces divergeant depuis plusieurs dizaines de millions d'années. La dégénérescence que nous avons observée dénote aussi de celles décrites dans les autres modèles biologiques. L'évolution de la paire de chromosomes sexuels homologue entre *C. sativa* et *H. lupulus* ressemble donc plus aux évolutions précédemment décrites chez certains animaux que chez les plantes. Mes travaux offrent donc des arguments contre un modèle spécifique aux plantes et un autre pour les animaux. La trajectoire évolutive d'une paire de chromosomes

sexuels ne s'explique pas simplement par le règne auquel appartient l'espèce, mais par des causes multiples. Nous pouvons faire l'hypothèse que des caractéristiques génomiques (dynamique des éléments transposables, taux de mutation, taux de recombinaison), des traits d'histoire de vie (dimorphisme sexuel, nombre de gènes exprimés en phase haploïde) ou encore la taille efficace peuvent induire des trajectoires évolutives différentes. Cette nouvelle interprétation de l'évolution des chromosomes sexuels chez les plantes souligne l'importance de continuer à décrire avec des données empiriques d'autres paires de chromosomes sexuels. La formation d'une région non-recombinante et les premières phases de son expansion restent probablement les étapes les moins bien comprises du modèle d'évolution d'une paire de chromosomes sexuels (résumé dans Charlesworth (2021a)). Comme présenté en introduction, l'intérêt porté à ces étapes par la communauté connaît un renouveau depuis quelques années, aussi bien d'un point de vue théorique (Jeffries et al., 2021; Jay, Tezenas et Giraud, 2021; Lenormand et Roze, 2021), que d'un point de vue empirique (Charlesworth et al., 2021; Munby et al., 2021). Les résultats obtenus chez *S. acaulis* ssp *exscapa* ne permettent pas de tester les différentes hypothèses de formation et d'expansion de la région non-recombinante. En revanche, le signal XY présent dans les données confirme que cette plante semble être intéressante pour mieux comprendre les étapes précoces du modèle d'évolution des chromosomes sexuels.

## 4 Identifier des chromosomes sexuels chez des espèces non-modèles

Durant ma thèse j'ai travaillé sur des espèces que l'on peut qualifier de non-modèles. Ceci est d'autant plus vraie pour *Silene acaulis* que pour *Cannabis sativa* et *Humulus lupulus*. En effet, des analyses cytologiques avaient déjà été publiées chez les deux espèces de Cannabaceae. Nous avons donc connaissance de la taille des génomes, du nombre de chromosomes ainsi que du sexe qui est hétérogamétique pour ces espèces. Toutes ces informations étaient inconnues pour *Silene acaulis*. Par ailleurs, seule *C. sativa* avait un transcriptome de référence de bonne qualité qui avait été publié (les génomes assemblés au niveau chromosomique de *C. sativa* ont été publiés seulement à la fin de nos analyses). Une solution lorsque l'on travail sur une espèce non-modèle peut être d'utiliser des ressources développées pour une espèce proche. En effet, outre la dimension financière, assembler un génome de référence de bonne qualité est un projet de recherche à part entière. Cette solution était difficilement envisageable pour les analyses de cette thèse car les plantes pour lesquelles un génome de bonne qualité avait été publié étaient toutes trop divergentes de mes espèces d'études.

En fonction des connaissances a priori du système d'étude et des ressources disponibles, il est important de bien définir son protocole expérimental.

Connaître le sexe hétérogamétique est important pour choisir la méthode d'identification des chromosomes sexuels. Pour des méthodes se basant sur l'identification de SNPs (SEX-

DETECTOR, SDpop, densité en SNP) il est impératif que les lectures des gamétologues soient alignées ensemble. Il est donc préférable d'utiliser une référence construite à partir d'un individu homogamétique pour ne pas avoir d'assemblages indépendants de la région X (ou Z) et de la région Y (ou W). Au contraire, pour les méthodes basées sur les ratios de couvertures *mâles/femelles*, il est impératif d'aligner indépendamment les séquences gamétologues pour identifier des différences entre sexes. Il est donc préférable d'utiliser un assemblage d'un individu hétérogamétique. Les régions Y-spécifiques (ou W-spécifiques) seront assemblées indépendamment des séquences X (ou Z). Dans le cas où l'hétérogamétie est inconnue, il est possible d'utiliser la référence d'une espèce proche qui n'a pas de chromosomes sexuels. Ainsi les séquences X et Y (ou Z et W) devraient toutes être alignées sur la séquence homologue de l'espèce proche. Il faut noter qu'avec une telle approche, les gènes trop divergents entre l'espèce étudiée et celle de référence ne pourront pas être analysés.

Le choix de la technologie de séquençage est aussi une étape importante lors de la conception de l'analyse. Un séquençage ARN sera plus économique, probablement moins lourd à analyser et permettra d'accéder à l'expression des gènes. En revanche, un séquençage ADN donnera accès à l'ensemble des gènes sans avoir besoin de séquencer plusieurs tissus et plusieurs stades de développement. Il permet aussi d'avoir les séquences non-codantes qui peuvent être utiles pour identifier une région non-recombinante (discuté dans le chapitre 2). De plus, utiliser de l'ADN permet de réduire certains bruits générés par les données d'expression. Par exemple des expressions sexe-spécifiques.

Les travaux sur *S. acaulis* ont confirmé la difficulté de travailler sur une espèce avec une paire de chromosomes sexuels probablement récente, donc avec une faible divergence. Plusieurs facteurs peuvent expliquer cette difficulté. La région non-recombinante est supposée être de petite taille, donc inclure peu de gènes, et la divergence entre les séquences gamétologues devrait être faible. Dans ces conditions, une méthode utilisant des ratios de couvertures risque de ne pas pouvoir identifier le peu de gamétologues dont le niveau de divergence est faible. Une méthode basée sur l'identification de SNPs semble donc être plus cohérente. Par ailleurs, utiliser un séquençage ARN aura pour avantage de coupler l'identification des SNPs avec une analyse d'expression. En revanche, un séquençage ADN aura pour avantage d'inclure tous les gènes de la région non-recombinante, ce qui peut ne pas être le cas avec des données transcriptomiques. De plus, des faux-positifs générés par le bruit transcriptionnel seront plus problématiques pour des systèmes avec peu de gènes. Pour de vieux systèmes avec des milliers de gènes, les quelques faux-positifs générés avec des données ARN auront des conséquences bien moins significatives. De manière générale, l'ensemble des méthodes nécessitent d'identifier des différences entre sexes. Trouver une région non-recombinante très jeune est donc généralement compliqué. Les travaux sur *S. acaulis* suggèrent que pour contourner ce problème il est nécessaire de combiner différentes approches.

Travailler sur des chromosomes sexuels qui divergent depuis plusieurs millions d'années comporte aussi des difficultés. Lorsque la divergence entre les copies X et Y devient trop importante, l'alignement des séquences Y sur une référence X est méthodologiquement limité.

Pour cette raison, les méthodes d'identification de SNPs nécessitant un alignement sur une référence homogamétique ont souvent été considérées comme moins adaptées que celles basées sur les ratios de couvertures. En revanche, mes travaux présentés dans les chapitres 3 et 5 ont montré qu'il est possible de conduire une analyse SEX-DETECTOR sur une paire de chromosomes sexuels assez divergente (divergence synonyme maximale autour de 0.4). Pour cela, nous avons utilisé l'outil SNP tolérant de GSNAP (voir matériel et méthodes des chapitres 3 et 5 pour plus de détails). Grâce à cet outil, il était possible d'aligner des séquences Y sur une référence X qui n'étaient pas alignées avec d'autres outils. Une méthode a aussi été trouvée pour filtrer, au sein d'un gène, les régions où les séquences Y n'avaient pas été alignées à cause de la divergence (méthode présentée et discutée dans le chapitre 5). Ainsi, nos méthodes montrent qu'il est possible d'étudier de vieux chromosomes sexuels sans avoir besoin de générer et d'assembler des données de séquençage ADN. Pour rappel, l'identification des gènes de la région non-recombinante de *C. sativa* a été faite à partir de données de séquençage RNA-seq (50pb single-end) alignées sur un transcriptome de référence. Ce résultat permet donc d'envisager, à faibles coûts, l'identification de nouveaux systèmes de chromosomes sexuels fortement dégénérés. Par ailleurs, au moment de la publication de l'article sur les chromosomes sexuels de *C. sativa*, trois génomes étaient publiés, dont deux au niveau chromosomique (van Bakel et al., 2011; Laverty et al., 2019; Grassa et al., 2021). Cependant, aucun n'avait identifié la paire de chromosomes sexuels. Bien que l'identification des chromosomes sexuels n'était pas l'objectif principal de ces travaux, ceci montre qu'il n'est pas trivial de trouver la paire de chromosomes sexuels avec un assemblage du génome. La principale raison est la difficulté d'assembler le chromosome Y, notamment à cause des éléments transposables. L'accumulation d'éléments transposables rend l'assemblage des chromosomes Y de vieux systèmes compliqué. Notre approche pourrait donc être plus adaptée et bien moins onéreuse.

Enfin, les résultats obtenus chez *H. lupulus* et *S. acaulis* montrent qu'il est préférable d'utiliser une référence bien assemblée même si elle provient d'une espèce différente. Des assemblages *de novo* ont été générés pour ces analyses mais n'ont pas été utilisés. Bien que le signal soit plus faible avec l'utilisation d'une référence d'une autre espèce, la qualité et la robustesse des résultats justifient son utilisation.

## 5 Conclusions

Pour conclure, les travaux présentés dans ce manuscrit de thèse fournissent des informations empiriques uniques pour des stades d'évolution des chromosomes sexuels particulièrement peu étudiés chez les plantes.

Les chromosomes sexuels des Cannabaceae confirment l'existence de systèmes anciens et fortement dégénérés chez les plantes, alors que les analyses faites chez *Silene acaulis* confirment l'intérêt de cette espèce pour les stades précoces des chromosomes sexuels. Par ailleurs, cette

thèse appuie la nécessité de continuer à étudier de nouvelles paires de chromosomes sexuels pour comprendre les limites du modèle actuel. De plus, des approches empiriques sont nécessaire pour confirmer ou infirmer les nombreux modèles théoriques sur la formation des régions non-recombinantes.

## Références

- Akagi, T., S. M. Pilkington, E. Varkonyi-Gasic, et al. 2019, Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nature Plants* **5** :801–809.
- Almeida, P., E. Proux-Wera, A. Churcher, et al. 2020, Genome assembly of the basket willow, *Salix viminalis*, reveals earliest stages of sex chromosome expansion. *BMC Biology* **18** :78.
- Bachtrog, D., J. E. Mank, C. L. Peichel, et al. 2014, Sex Determination : Why So Many Ways of Doing It ? *PLOS Biology* **12** :e1001899.
- Badouin, H., A. Velt, F. Gindraud, et al. 2020, The wild grape genome sequence provides insights into the transition from dioecy to hermaphroditism during grape domestication. *Genome Biology* **21** :223.
- Baránková, S., J. P. Pascual-Díaz, N. Sultana, et al. 2020, Sex-chrom, a database on plant sex chromosomes. *New Phytologist* **227** :1594–1604.
- Carey, S., Q. Yu, et A. Harkess. 2021, The Diversity of Plant Sex Chromosomes Highlighted through Advances in Genome Sequencing. *Genes* **12** :381.
- Charlesworth, D. 2019, Young sex chromosomes in plants and animals. *New Phytologist* **224** :1095–1107.
- . 2021a, The timing of genetic degeneration of sex chromosomes. *Philosophical Transactions of the Royal Society B : Biological Sciences* **376** :20200093.
- . 2021b, When and how do sex-linked regions become sex chromosomes ? *Evolution* **75** :569–581.
- Charlesworth, D., R. Bergero, C. Graham, J. Gardner, et K. Keegan. 2021, How did the guppy Y chromosome evolve ? *PLOS Genetics* **17** :e1009704.
- Dowling, C. A., R. Melzer, et S. Schilling. 2021, Timing is everything : the genetics of flowering time in *Cannabis sativa*. *The Biochemist* **43** :34–38.
- Fruchard, C., H. Badouin, D. Latrasse, R. S. Devani, A. Muyle, B. Rhoné, S. S. Renner, A. K. Banerjee, A. Bendahmane, et G. A. B. Marais. 2020, Evidence for Dosage Compensation in *Coccinia grandis*, a Plant with a Highly Heteromorphic XY System. *Genes* **11** :787.

- Furman, B. L. S., D. C. H. Metzger, I. Darolti, A. E. Wright, B. A. Sandkam, P. Almeida, J. J. Shu, et J. E. Mank. 2020, Sex Chromosome Evolution : So Many Exceptions to the Rules. *Genome Biology and Evolution* **12** :750–763.
- Grassa, C. J., G. D. Weiblen, J. P. Wenger, C. Dabney, S. G. Poplawski, S. Timothy Motley, T. P. Michael, et C. J. Schwartz. 2021, A new Cannabis genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytologist* **230** :1665–1679.
- Harkess, A., K. Huang, R. van der Hulst, B. Tissen, J. L. Caplan, A. Koppula, M. Batish, B. C. Meyers, et J. Leebens-Mack. 2020, Sex Determination by Two Y-Linked Genes in Garden Asparagus[OPEN]. *The Plant Cell* **32** :1790–1796.
- Heikrujam, M., K. Sharma, M. Prasad, et V. Agrawal. 2015, Review on different mechanisms of sex determination and sex-linked molecular markers in dioecious crops : a current update. *Euphytica* **201** :161–194.
- Jay, P., E. Tezenas, et T. Giraud. 2021, A deleterious mutation-sheltering theory for the evolution of sex chromosomes and supergenes. Tech. rep.
- Jeffries, D. L., J. F. Gerchen, M. Scharmann, et J. R. Pannell. 2021, A neutral model for the loss of recombination on sex chromosomes. *Philosophical Transactions of the Royal Society B : Biological Sciences* **376** :20200096.
- Käfer, J., A. Bewick, A. Andres-Robin, et al. 2021, A derived ZW chromosome system in *Amborella trichopoda*, representing the sister lineage to all other extant flowering plants. *New Phytologist* **n/a**.
- Laverty, K. U., J. M. Stout, M. J. Sullivan, et al. 2019, A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci. *Genome Research* **29** :146–156.
- Lenormand, T. et D. Roze. 2021, Y recombination arrest and degeneration in the absence of sexual dimorphism. Tech. rep.
- Liao, Q., R. Du, J. Gou, et al. 2020, The genomic architecture of the sex-determining region and sex-related metabolic variation in *Ginkgo biloba*. *The Plant Journal* **104** :1399–1409.
- Massonnet, M., N. Cochetel, A. Minio, et al. 2020, The genetic basis of sex determination in grapes. *Nature Communications* **11** :2902.
- Munby, H., T. Linderoth, B. Fischer, et al. 2021, Differential use of multiple genetic sex determination systems in divergent ecomorphs of an African crater lake cichlid. *bioRxiv* 2021.08.05.455235.
- Muyle, A., J. Käfer, N. Zemp, S. Mousset, F. Picard, et G. A. Marais. 2016, SEX-DETECTOR : A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms. *Genome Biology and Evolution* **8** :2530–2543.

- Müller, N. A., B. Kersten, A. P. Leite Montalvão, et al. 2020, A single gene underlies the dynamic evolution of poplar sex determination. *Nature Plants* **6** :630–637.
- Padgitt-Cobb, L. K., S. B. Kingan, J. Wells, et al. 2021, A draft phased assembly of the diploid Cascade hop (*Humulus lupulus*) genome. *The Plant Genome* **14** :e20072.
- Prentout, D., O. Razumova, H. Henri, M. Divashuk, G. Karlov, et G. A. Marais. 2020a, Development of genetic markers for sexing *Cannabis sativa* seedlings. Tech. rep.
- Prentout, D., O. Razumova, B. Rhoné, H. Badouin, H. Henri, C. Feng, J. Käfer, G. Karlov, et G. A. B. Marais. 2020b, An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of *Cannabis sativa*. *Genome Research* **30** :164–172.
- Prentout, D., N. Stajner, A. Cerenak, T. Tricou, C. Brochier-Armanet, J. Jakse, J. Käfer, et G. A. B. Marais. 2021, Plant genera *Cannabis* and *Humulus* share the same pair of well-differentiated sex chromosomes. *New Phytologist* **231** :1599–1611.
- Renner, S. S. et N. A. Müller. 2021, Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. *Nature Plants* **7** :392–402.
- Sanderson, B. J., G. Feng, N. Hu, et al. 2021, Sex determination through X–Y heterogamety in *Salix nigra*. *Heredity* **126** :630–639.
- Schilling, S., C. A. Dowling, J. Shi, L. Ryan, D. Hunt, E. O'Reilly, A. S. Perry, O. Kinane, P. F. McCabe, et R. Melzer. 2020, *The Cream of the Crop : Biology, Breeding and Applications of Cannabis sativa*. preprint, Preprints.
- Segawa, M., S. Kishi, et S. Tatuno. 1971, SEX CHROMOSOMES OF *CYCAS REVOLUTA*. *The Japanese Journal of Genetics* **46** :33–39.
- van Bakel, H., J. M. Stout, A. G. Cote, C. M. Tallon, A. G. Sharpe, T. R. Hughes, et J. E. Page. 2011, The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology* **12** :R102.
- Wu, M., D. C. Haak, G. J. Anderson, M. W. Hahn, L. C. Moyle, et R. F. Guerrero. 2021, Inferring the Genetic Basis of Sex Determination from the Genome of a Dioecious Nightshade. *Molecular Biology and Evolution* **38** :2946–2957.
- Zou, C., M. Massonnet, A. Minio, et al. 2021, Multiple independent recombinations led to hermaphroditism in grapevine. *Proceedings of the National Academy of Sciences* **118**.





