Sex chromosome evolution in plants : methodological developments and NGS data analysis in the Silene genus

Aline Muyle

Université Claude Bernard Lyon 1 - UCBL

Ecole doctorale Évolution Écosystèmes Microbiologie Modélisation - E2M2

Évolution des chromosomes sexuels chez les plantes: développements méthodologiques et analyses de données NGS de Silènes.

Aline MUYLE

Doctorat en Biologie, sous la direction de Gabriel MARAIS

Date de soutenance: 3 septembre 2015
Numéro d'ordre: 109 - 2015

Membres du Jury:
Rapporteurs: John Pannell, Nicolas Galtier.
Examinateurs: Xavier Vekemans, Maud Tenaillon.
Professeur Université Lyon 1: Dominique Mouchiroud

Sex chromosome evolution in plants: methodological developments and NGS data analysis in the *Silene* genus.

Laboratoire de rattachement:
UMR CNRS 5558 - LBBE
“Biométrie et Biologie Évolutive”
UCB Lyon 1 - Bât. Grégor Mendel
43 bd du 11 novembre 1918
69622 VILLEURBANNE cedex

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est - Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud - Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme Caroline FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur Georges TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. Jean-Claude PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Acknowledgements

Cette thèse est le fruit d'un travail collaboratif, aussi je remercie toutes les personnes impliquées dans le contenu de ces pages. En particulier mon directeur de thèse Gabriel Marais qui a orchestré mon travail et m'a donné beaucoup de liberté, Franck Picard qui m'a guidée jusqu'à convergence et au delà et Niklaus Zemp et Alex Widmer sans qui je n'aurais eu que très peu de données à analyser, et avec qui il est toujours palpitant d'étudier le monde vivant.

Merci à Rylan qui a transformé ma vie.

Je remercie ma mère, mon père, mon frère, mes tantes Monique et Clotilde et ma cousine Lara pour le soutien qu'ils et elles m'ont apporté ainsi que les bons moments, Kerry et Robert pour leur gentillesse et Stéphanie pour la détente !

CV

Training

April 2013: Wellcome Trust-EMBL-EBI Advanced Course on Computational Molecular Evolution, Sanger Centre, Hinxton, Britain.

2011-2012: Agrégation in Biology and Geology at ENS Lyon, France (national rank : 2nd).

2009-2011: Master's degree in Biosciences at ENS Lyon, France (rank : 1st).

2008-2009: Bachelor of science in fundamental Biology at ENS Lyon, France.

2006-2008: Preparatory classes in Science at Lycée du Parc, Lyon, France. Equivalent to the first and second year of a Bachelor of Science with a national competition at the end.

2005-2006: A-levels in Science at Lycée de Die, France.

Research experiences

09/2012–present: PhD under the supervision of Dr Gabriel Marais on the joint evolution of dioecy and sex chromosomes in Silene.

02/2011–05/2011: Second year master's internship under the supervision of Gabriel Marais (LBBE, Lyon, France). Search of new sex-linked genes in the dioecious plant Silene latifolia using New Sequencing Technologies and study of the evolution of the sex- chromosomes.

09/2010–12/2010: Second year master's internship under the supervision of Sylvain Glémin (ISEM, Montpellier, France). Study of the effect of mating systems on the efficacy of selection in Angiosperms.

02/2010–05/2010: First year master's internship under the supervision of Sylvain Glémin (ISEM, Montpellier, France). Study of the role of biased gene conversion and selection on the evolution of GC-content in the Oryza genus (rice).

05/2009–07/2009: Bachelor internship under the supervision of Mathilde Dufay (GEPV, Lille, France). Empirical study of gynodioecy maintenance in Beta vulgaris (beet).

University teaching (64h/year during the PhD)

Licence 1 Practicals on mathematics and statistics for Biology (TDs MATHSV).

Licence 2 Students' professional projects interviews (PEL4). Practicals on population genetics (TPs Génétique 2).

Licence 3 Practicals on molecular evolution (at UCBL and ENS), creation of new topics.

Master 1 Practicals on molecular evolution (at ENS).

Master 2 Lecture on host-parasite coevolution (Préparation à l'Agrégation, ENS). Oral competition trainings (Leçons d'Agrégation). Botany field trips.

Conferences

Posters

- SEX-DETector : A model-based method for detecting sex chromosomes and studying sex determination in non-model organisms using NGS data. 11th RECOMB - Comparative Genomics, Lyon, October 2013.
- SEX-DETector : A model-based method for detecting sex chromosomes and studying sex determination in non-model organisms using NGS data. Recent advances on the evolution of sex and genetic systems, Roscoff, May 2013.
- Dosage compensation in the XY chromosomes of Silene latifolia. Genetics, Epigenetics and Evolution of Sex Chromosomes Conference, Paris, Juin 2011.

Talks

- Evolution of dosage compensation in the dioecious plant Silene latifolia. SMBE, Fitch Symposium. Vienna, Austria, July 2015.
- SEX-DETector : A framework for identifying sex-linked genes using RNAseq data. BGE (Bioinformatique pour la Génomique Environnementale), Lyon, May 2014.
- SEX-DETector : A cheap and efficient approach to identify sex-linked genes in non-model organisms using NGS data. PAG XXII (Plant and Animal Genomes), San Diego, January 2014.
- Estimating gene loss in the Y chromosome of the dioecious plant Silene latifolia. Plant Genome Evolution, Amsterdam, September 2013.
- Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. GDRE Comparative Genomics meeting, November 2012.
- GC-biased gene conversion and selection affect GC content in the Oryza genus (rice). GDR génomique des populations, Toulouse, November 2010.

Other scientific experiences

09/2013–07/2014 Weekly organisation of the Journal Club of the department in LBBE (Lyon)

09/2013–12/2013 Co-supervision of Cécile Fruchard's Master internship.

06/2014–12/2014 Co-supervision of Paul Jay's Master internship.

Résumé (French)

Mots clés: *Silene latifolia*, chromosomes sexuels, RNA-seq, BACs, gènes liés au sexe, compensation de dosage, perte de gènes Y, expression biaisée pour le sexe.

Malgré leur importance dans le déterminisme du sexe chez de nombreux organismes, les chromosomes sexuels ont été étudiés chez quelques espèces seulement du fait du manque de séquences disponibles. En effet, le séquençage et l'assemblage des chromosomes sexuels est rendu très difficile par leurs abondantes séquences répétées. Durant cette thèse, une méthode probabiliste a été développée pour inférer les gènes liés au sexe à partir de données RNA-seq chez une famille. Des tests de cette méthode appelée SEX-DETector sur des données réelles et simulées suggèrent qu'elle fonctionnera sur une grande variété de systèmes. La méthode a inféré ~1300 gènes liés au sexe chez *Silene latifolia*, une plante dioïque qui possède des chromosomes sexuels XY pour lesquels quelques données de séquences sont disponibles (dont certaines obtenues lors de cette thèse par séquençage de BACs). Les gènes du Y sont moins exprimés que ceux du X chez *S. latifolia*, mais le statut de la compensation de dosage (un mécanisme qui corrige la sous-expression des gènes liés au sexe chez les mâles) est encore controversé. L'analyse des nouveaux gènes liés au sexe inférés par SEX-DETector a permis de confirmer la compensation de dosage chez S. *latifolia*, qui est effectuée par la surexpression du X maternel, possiblement via un mécanisme épigénétique d'empreinte. Les données ont également été utilisées pour étudier l'évolution de l'expression biaisée pour le sexe chez *S. latifolia* et ont révélé que la majorité des changements de niveaux d'expression ont eu lieu chez les femelles. Les implications de nos résultats concernant l'évolution de la dioécie et des chromosomes sexuels sont discutés.

Abstract

Key words: *Silene latifolia*, sex chromosomes, RNA-seq, BACs, sex-linked genes, dosage compensation, Y gene loss, sex-biased expression.

In many organisms, sexes are determined by sex chromosomes. However, studies have been greatly limited by the paucity of sex chromosome sequences. Indeed, sequencing and assembling sex chromosomes are very challenging due to the large quantity of repetitive DNA that these chromosomes comprise. In this PhD, a probabilistic method was developed to infer sex-linked genes from RNA-seq data of a family (parents and progeny of each sex). The method, called SEX-DETector, was tested on simulated and real data and should perform well on a wide variety of sex chomosome systems. This new method was applied to *Silene latifolia*, a dioecious plant with XY system, for which partial sequence data on sex chromosomes are available (some of which obtained during this PhD by BAC sequencing), SEX-DETector returned ~1300 sex-linked genes. In *S. latifolia*, Y genes are less expressed than their X counterparts. Dosage compensation (a mechanism that corrects for reduced dosage due to Y degeneration in males) was previously tested in *S. latifolia*, but different studies returned conflicting results. The analysis of the new set of sex-linked genes confirmed the existence of dosage compensation in *S. latifolia*, which seems to be achieved by the hyperexpression of the maternal X chromosome in males. An imprinting mechanism might underlie dosage compensation in that species. The RNA-seq data were also used to study the evolution of differential expression among sexes in *S. latifolia*, and revealed that in this species most changes have affected the female sex. The implications of our results for the evolution of dioecy and sex chromosomes in plants are discussed.

Résumé détaillé (French)

Évolution des systèmes de reproduction chez les plantes

Les plantes ont des systèmes de reproduction extrêmement variés, pouvant aller de l'autofécondation presque totale (où un même individu, à la fois mâle et femelle, féconde ses propre ovules avec son pollen, comme par exemple chez *Arabidopsis thaliana*) à une allofécondation totale (où des individus différents se fécondent les uns les autres, c'est le cas des plantes dioïques qui présentent des individus mâles et des individus femelles, comme par exemple *Silene latifolia*, le compagnon blanc). Les plantes peuvent aussi présenter des systèmes dits mixtes, où à la fois autofécondation et allofécondation existent chez des individus hermaphrodites d'une même espèce. De nombreux mécanismes favorisant soit autofécondation soit allofécondation ont évolué chez les plantes. La cléistogamie est un processus au cours duquel les fleurs restent closes à maturité, et seule une autofécondation est possible, comme par exemple chez la violette à l'automne, ce qui permettrait une assurance reproductive. Certaines plantes présentent de l'hétérostylie, un système dans lequel une même espèce possède des morphes de fleurs différents (styles longs et étamines courtes et inversement), chaque individu ne possède qu'un type de morphe floral, ce qui empêche l'autofécondation car seuls des morphes différents peuvent se féconder, c'est le cas, par exemple, des primevères. Chez les plantes auto-incompatibles il existe un mécanisme génétique de reconnaissance et de rejet du pollen portant des allèles identiques aux allèles du pistil, empêchant l'autofécondation. Les plantes gynodioïques possèdent des individus hermaphrodites et des individus femelles au sein d'une même espèce (comme par exemple chez *Silene vulgaris*, la silène enflée). L'allofécondation est obligatoire pour les femelles des espèces gynodioïques. De la même façon, l'allofécondation est obligatoire pour les mâles des espèces androdioïques, où mâles et hermaphrodites cohabitent. Les plantes sont également capables de reproduction asexuée.

Il est admis que le système de reproduction ancestral des angiospermes est l'hermaphrodisme, et différentes voies évolutives existent entre les systèmes de reproduction. Une plante hermaphrodite allogame peut évoluer vers l'autogamie si le coût de la dépression de consanguinité (dû à l'exposition des mutations récessives délétères chez les autogames suite à leur homozygotie) est inférieur aux avantages directs de l'autogamie (assurance reproductive, pouvoir colonisateur et avantage de transmission des gènes par rapport à l'allogamie: une autogame transmet ses gènes via ses ovules, son pollen autofé-

condant et son pollen allofécondant, alors qu'une allogame ne transmet ses gènes que via ses ovules et son pollen allofécondant). La dioécie peut évoluer depuis l'hermaphrodisme via la gynodioécie ou la monoécie, et la dioécie elle-même peut évoluer en androdioécie (Figure 1).

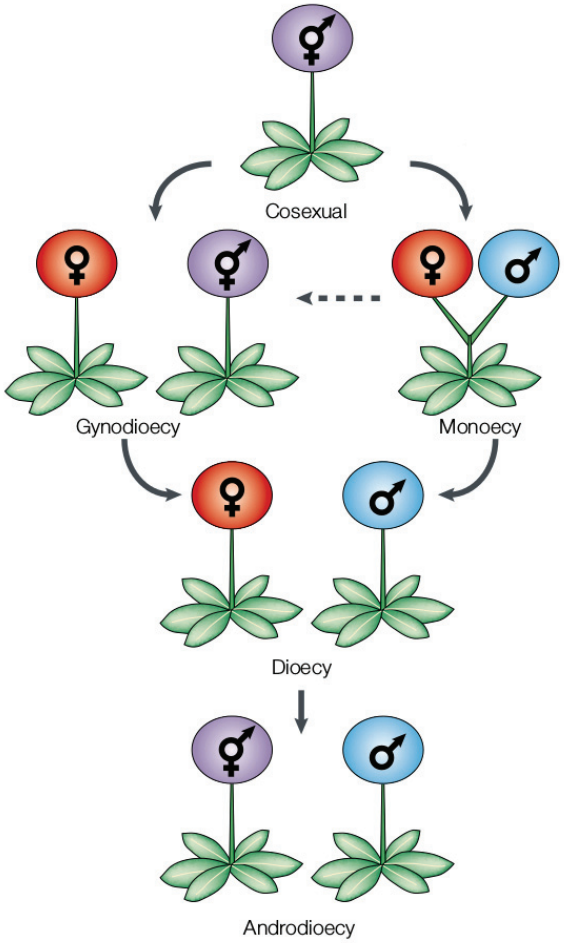


Figure 1: Transitions entre systèmes de reproduction chez les plantes (tiré de Barrett, 2002).

Conséquences génomiques des systèmes de reproduction chez les plantes

Les systèmes de reproduction déterminent comment le matériel génétique est transmis d'une génération à l'autre et influencent par là même l'évolution des génomes des plantes (comme expliqué dans le Chapitre 1). J'ai étudié, lors de l'un de mes stages de Master sous la direction de Sylvain Glémin, l'effet de l'autogamie sur l'évolution génomique des angiospermes, un travail que j'ai poursuivi en thèse et qui a abouti à une publication dans *Journal of Evolutionary Biology* (Appendix A). Les théories évolutives prédisent que la sélection est moins efficace chez les autogames par rapport aux allogames. En effet, chez les plantes autogames le génome est à l'état homozygote ce qui diminue la taille efficace par deux car, à chaque génération, un allèle est automatiquement transmis avec son allèle identique et le processus de coalescence revient à piocher des individus et non pas des allèles, même si l'on a affaire à des espèces diploïdes. La taille efficace est également diminuée suite à des phénomènes d'entraînement qui sont la cause de l'absence

de recombinaison (même si la recombinaison a physiquement lieu chez les autogames, elle est sans effet car elle a lieu entre des régions homozygotes et donc identiques génétiquement et ne peut pas casser la liaison entre les allèles des différents gènes d'une région génomique). Cette diminution de la taille efficace entraîne une augmentation de l'intensité de la dérive génétique et diminue l'efficacité de la sélection. On s'attend donc à ce que les plantes autogames accumulent plus de mutations délétères que les allogames. Nous n'avons pas pu observer cela chez les angiospermes que nous avons étudiées, peut-être parce que l'autogamie est trop récente pour permettre la détection d'une accumulation de mutations délétères, ou alors parce que la plupart des mutations sont plutôt très délétères et donc efficacement éliminées par la sélection, même chez les plantes autogames.

La conversion génique biaisée vers GC (BGC) est un mécanisme ayant lieu lors de la recombinaison, qui provoque préférentiellement le remplacement des allèles A ou T par des allèles C ou G. Lors d'un stage de Master sous la direction de Sylvain Glémin, j'ai pu mettre en évidence de la BGC chez le riz, un travail publié dans *Molecular Biology and Evolution* (Appendix C). Étant donné que la recombinaison n'est pas efficace chez les autogames, on s'attend à ce que la BGC soit moins forte chez les autogames que chez les allogames, mais les données de séquences actuelles n'ont pas encore permis de vérifier clairement cet attendu théorique.

La dioécie est un mode de reproduction plutôt rare chez les angiospermes (environ 5%), et les études phylogénétiques suggèrent qu'elle est souvent d'origine évolutive récente. Une étude de comparaison de richesse en espèces entre clades dioïques et non dioïques a mené les auteurs à faire l'hypothèse que la dioécie serait une impasse évolutive (Heilbuth, 2000). Cependant cette hypothèse a été contredite par des résultats de l'équipe de Gabriel Marais qui suggèrent que les plantes dioïques diversifient plus que les plantes non dioïques (Käfer et al., 2014).

Chez les plantes dioïques, le sexe des individus peut être déterminé par des chromosomes sexuels qui sont une paire particulière de chromosomes possédant un matériel génétique différent l'un de l'autre et ne recombinant pas.

Évolution des chromosomes sexuels chez les plantes

Comme expliqué dans le chapitre 2, il existe trois types de chromosomes sexuels: femelle hétérogamétiques (les femelles sont ZW et les mâles ZZ), mâle hétérogamétiques (les mâles sont XY et les femelles XX) et haplo-diploïdes (sporophytes UV, femelles gamétophytes U et mâles gamétophytes V).

Il existe des régions spécifiques au sexe mâle sur les chromosomes Y, spécifiques au sexe femelle sur les chromosomes W et spécifiques à chaque sexe sur les chromosomes U et V. Ces régions spécifiques au sexe ne recombinent pas avec leurs homologues X, Z ou U/V, alors que les chromosomes X recombinent entre eux chez les femelles et les chromosomes Z chez les mâles. Les chromosomes X/Y, Z/W et U/V recombinent entre

eux en dehors des régions spécifiques au sexe, c'est-à-dire au niveau des régions pseudoautosomales, vestige du fait que les chromosomes sexuels étaient au départ une paire d'autosomes ordinaires. L'absence de recombinaison des régions spécifiques au sexe entraîne leur dégénérescence génétique car les allèles sont transmis en blocs et la sélection agit sur des groupes de liaison au lieu d'agir sur des allèles (effets Hill-Robertson).
Cette dégénérescence s'observe à différents niveaux de l'évolution des chromosomes sexuels comme par exemple la perte de gènes sur les régions spécifiques au sexe des chromosomes Y, W, U et V, ainsi que la diminution du niveau d'expression des gènes en comparaison aux chromosomes X et Z ou en comparaison aux autosomes. Cette diminution du niveau d'expression et cette perte de gène entraînent des inégalités entre sexe au niveau de la quantité de transcrits et protéines produits, avec le sexe hétérogamétique (XY et ZW) qui a moins d'expression que le sexe homogamétique (XX et ZZ). Un mécanisme appelé la compensation de dosage a évolué chez certaines espèces et permet de rétablir des niveaux d'expression similaires entre mâles et femelles, mais surtout similaires à l'expression ancestrale des chromosomes sexuels, lorsqu'ils étaient encore une paire d'autosomes.
Jusqu'à peu la compensation de dosage n'était connue que chez certains animaux (mammifères, drosophiles, *C. elegans* et quelques autres). L'existence de la compensation de dosage chez la plante *Silene latifolia*, qui possède des chromosomes sexuels XY, est controversée. Lors de mon stage de Master sous la direction de Gabriel Marais, et en collaboration avec l'équipe d'Alex Widmer à Zurich, j'ai pu mettre en évidence de la compensation de dosage chez *S. latifolia*, avec l'expression du X unique des mâles qui augmentait lorsque l'expression du Y diminuait du fait de la dégénerescence, ce qui permettait de maintenir des niveaux d'expression similaires entre mâles et femelles. Ce travail a été publié dans *Plos Biology* (Appendix B). Cependant d'autres travaux qui se sont focalisés sur les gènes X-hémizygotes (les gènes pour lesquels la copie Y a été perdue) n'ont pas observé de compensation de dosage chez *S. latifolia* (Bergero et al., 2015; Chibalina and Filatov, 2011), contrairement à mes travaux de Master qui se focalisaient sur les gènes ayant conservé une copie Y transcrite. L'un des objectifs de ma thèse était de clarifier cette controverse.
Un autre objectif de ma thèse était d'étudier la perte de gènes Y chez *S. latifolia*, ce qui aurait permis de tester si, conformément à l'attendu théorique, les chromosomes Y de plantes perdent moins rapidement leurs gènes que les animaux. En effet, une grande partie du génome est exprimée dans les grains de pollen (contrairement aux spermatozoïdes), ce qui rendrait la sélection plus efficace pour limiter la perte de gènes Y, du moins pour les gènes Y exprimés dans le pollen. Conformément à cet attendu théorique des travaux ont montré que les gènes Y exprimés dans le pollen accumulent moins de mutations délétères que les autres gènes Y (Chibalina and Filatov, 2011), et le pourcentage de gènes perdus chez *S. latifolia* est faible compte tenu de l'âge du système (~14.5% Bergero and Charlesworth, 2011; Bergero et al., 2015). Cependant aucune comparaison

formelle aux systèmes animaux n'a été menée.
Afin de réaliser ces études il était nécessaire de disposer de séquences liées au sexe, c'est à dire localisées sur les régions qui ne recombinent pas entre le X et le Y.

Obtention de séquences de gènes liés au sexe

Le séquençage et l'assemblage des régions non-recombinantes des chromosomes sexuels sont rendus très difficiles par leurs abondantes séquences répétées. Une solution est de recourir à une carte physique et de séquencer des clones de chromosomes artificiels bactériens (BACs). Nous l'avons appliquée à *S. latifolia* sur une petite portion des chromosomes X et Y, en collaboration avec les équipes de Roman Hobza et Alex Widmer, ce qui a permis d'identifier de façon fiable de nouveaux gènes liés au sexe (Chapitre 3). Nous avons pu estimer un pourcentage de perte de gène sur le Y en combinant les données BACs aux données RNA-seq déjà disponibles chez *S. latifolia*. Bien que nos données étaient insuffisantes, elles suggèrent que l'estimation de la perte de gènes sur le Y uniquement à l'aide de données RNA-seq est biaisée et sous-estimée. Malheureusement la stratégie reposant sur le séquençage de BACs est laborieuse et extrêmement coûteuse et nous avons par la suite favorisé d'autres approches.
D'autres méthodes ont été développées pour les espèces non modèles qui sont moins coûteuses et reposent sur le séquençage d'un génome mâle et d'un génome femelle, cependant ces méthodes requièrent que le génome soit facilement assemblable et que les chromosomes sexuels soient très divergents. D'autres méthodes reposent sur le séquençage d'un génome et la réalisation d'une carte génétique pour identifier les régions liées au sexe, mais encore une fois cela requiert un génome bien assemblé. Une solution pour s'affranchir du problème des régions répétées est d'analyser des données de RNA-seq au lieu de DNA-seq, de cette façon on a également directement accès aux gènes et à leur niveau d'expression dans le(s) tissus échantillonné(s). Cette stratégie a fait ses preuves avec le séquençages de croisements par RNA-seq (parents et descendants de chaque sexe), cependant la méthodologie appliquée dans les précédentes études reposait sur des filtres empiriques et limitait l'application de cette stratégie à d'autres jeux de données et d'autres espèces (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011). L'un des buts de ma thèse était donc de développer une méthode reposant sur un modèle probabiliste qui inférerait les gènes liés au sexe, à partir de données RNA-seq sur un croisement et qui fonctionnerait sur tous les types de chromosomes sexuels (XY, ZW, UV), quelque soit l'espèce ou le jeu de donnée.
Cette méthode s'appelle SEX-DETector (Chapitre 4), je l'ai développée avec l'aide de Gabriel Marais, Franck Picard et Sylvain Mousset au LBBE à Lyon et l'équipe d'Alex Widmer à Zurich nous a fourni les données de séquences pour tester la méthode. Elle repose sur un modèle probabiliste de transmission des allèles des parents aux descendants et intègre des paramètres de taille des chromosomes sexuels, niveau de dégénérescence, composition en base de l'espèce, niveau d'hétérozygotie et erreurs de génotypage qui sont notamment plus fréquentes sur les allèles Y et W du fait de leur faible expression. Ce

qui permet à la méthode de s'adapter à toute espèce et type de chromosomes sexuels. Un modèle a également été développé pour le système UV. La méthode a été testée sur des données de séquences réelles et simulées et donne de meilleurs résultats que les précédentes méthodes empiriques, notamment grâce au modèle probabiliste qui permet de calculer pour chaque gène une probabilité d'être lié au sexe et de filtrer sur cette probabilité, plutôt que d'utiliser des filtres arbitraires.
J'ai analysé un jeu de donnée RNA-seq sur un croisement de *S. latifolia* fourni par l'équipe d'Alex Widmer à Zurich, ce qui m'a permis d'inférer environ 1300 gènes liés au sexe chez cette espèce. Ces inférences ont été utilisées par la suite pour des analyses biologiques.

Box 0.1 : Matériel et Méthode de la thèse

Durant ma thèse j'ai fait du développement méthodologique (écriture d'un modèle probabiliste et d'un programme), des analyses de données en bioinformatiques (principalement de séquences RNA-seq ce qui a impliqué de l'assemblage de novo, du mapping, du génotypage et des analyses de niveaux d'expression), et des analyses évolutives et statistiques.
J'ai par ailleurs fait du terrain pour échantillonner des individus de *Silene acaulis*, une plante dioïque alpine, dans le but d'étudier la présence de chromosomes sexuels chez cette espèce. J'ai également fait des cultures de silènes en serre à partir de graines échantillonnées par l'équipe de Gabriel Marais sur le terrain ou envoyées par des collègues européens. J'ai préparé les échantillons pour le séquençage RNA-seq de ces plantes pour constituer un jeu de donnée permettant de comparer des silènes ayant des systèmes de reproduction contrastés.

Évolution de la compensation de dosage chez *Silene latifolia*

Grâce au nouveau jeu de donnée dont je disposais chez *S. latifolia* j'ai pu étudier d'avantage la question de la compensation de dosage chez cette espèce pour tenter d'expliquer les différentes conclusions auxquelles les précédentes études étaient arrivées (Chapitre 5). Je disposais d'inférences de gènes X/Y et de gènes X-hémizygotes avec leurs niveaux d'expression pour les copies X dans chaque sexe et Y chez les mâles, ainsi que des niveaux d'expression chez une espèce proche sans chromosomes sexuels, *Silene vulgaris*, ce qui donnait une estimation du niveau d'expression ancestral des chromosomes sexuels de *S. latifolia*, lorsqu'ils étaient encore une paire d'autosomes. J'ai pu observer que lorsque le Y dégénère et que son expression diminue, l'expression du X chez les mâles augmente ce qui permet de maintenir des niveaux d'expression similaires entre les gènes X/Y des mâles de *S. latifolia* et leurs homologues autosomaux chez *S. vulgaris*. Cela confirme qu'il existe une compensation de dosage chez *S. latifolia*. En accord avec les précédentes études, les gènes X-hémizygotes étaient principalement non-compensés, ce qui suggère que ces gènes ne sont pas sensibles au dosage et pourrait expliquer pourquoi ils ont perdu leur copie Y. Conformément à cela, une analyse GO a révélé que les gènes X-hémizygotes étaient significativement appauvris en gènes impliqués dans le complexe ribosomal, un exemple type de complexe protéique qui requiert une stœchiométrie bien réglée entre les différents composants pour permettre son fonctionnement optimal. Les

gènes X-hémizygotes hautement exprimés étaient cependant compensés, ce qui suggère qu'ils sont plus sensibles au dosage. L'augmentation de l'expression du X chez les mâle en comparaison à l'expression autosomale chez *S. vulgaris* a également eu lieu chez les femelles, à un niveau plus restreint, et est uniquement causée par l'augmentation de l'expression du X maternel. Ainsi il semble que la compensation de dosage soit le fruit d'un mécanisme épigénétique d'empreinte chez *S. latifolia* qui entraîne la surexpression du chromosome X maternel, à la fois chez les mâles et les femelles. Cette augmentation chez les femelles est potentiellement délétère et suggère que le système est jeune et probablement pas encore complètement optimisé.

Évolution des gènes avec une expressio biaisée pour le sexe chez *Silene latifolia*

Les inférences de SEX-DETector et les niveaux d'expression des copies X et Y que j'ai générés ont aussi été utilisés pour une étude sur l'évolution des gènes biaisés pour le sexe chez *S. latifolia* (Chapter 6), réalisée principalement par l'équipe d'Alex Widmer à Zurich et pour laquelle nous avons contribué. Les résultats montrent que les conflits entre sexes sont principalement résolus par des changements de niveau d'expression chez les femelles de *S. latifolia*.

Conclusion

En conclusion, ma thèse a permis le développement d'une méthode pour inférer les gènes liés au sexe qui fonctionne sur tous types de chromosomes sexuels. Cela a permis de confirmer la présence de compensation de dosage chez *S. latifolia* et d'en éclaircir le mécanisme.

Je vais poursuivre mes travaux de recherche avec un post-doctorat dans l'équipe de Gabriel Marais. Des projets sont actuellement en cours pour appliquer SEX-DEtector à de nombreuses espèces de plantes (ANR NGSex).

J'ai pour projet d'utiliser de nouvelles données DNA-seq qui permettront d'estimer plus précisément le pourcentage de perte de gènes Y chez *S. latifolia*, sans les biais du RNA-seq, ce qui rendra possible le test de l'hypothèse selon laquelle les plantes perdent moins de gènes Y que les animaux.

Par ailleurs, lors de ma thèse j'ai constitué un jeu de donnée RNA-seq de trois espèces de silènes (Box 0.1) avec des systèmes de reproduction contrastés (*Silene latifolia* dioïque, *Silene vulgaris* gynodioïque et *Silene viscosa* hermaphrodite) avec plusieurs individus échantillonnés pour chaque espèce sur leur aire de répartition. Cela me permettra, avec les inférences de SEX-DETector, de tester si le chromosome X de *S. latifolia* évolue plus rapidement que le reste du génome, comme observé chez certaines espèces d'animaux. Je pourrais également tester si les systèmes de reproduction impactent l'évolution des plantes, et notamment si la dioécie est liée à une accumulation de mutations délétères, comme suggéré dans la littérature. Cette étude sera la première reposant sur des données

génomiques pour tester l'hypothèse de l'impasse évolutive de la dioécie.

Références

Barrett, S. C. H. (2002). The evolution of plant sexual diversity. eng. *Nature Reviews. Genetics* 3(4), 274–284. ISSN: 1471-0056. DOI: 10.1038/nrg776.

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Qiu, S., and Charlesworth, D. (2015). Gene Loss from a Plant Sex Chromosome System. ENG. *Current biology: CB*. ISSN: 1879-0445. DOI: 10.1016/j.cub.2015.03.015.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Heilbuth, J. C. (2000). Lower species richness in dioecious clades. *The American Naturalist* 156(3), 221–241.

Käfer, J., Boer, H. J. de, Mousset, S., Kool, A., Dufay, M., and Marais, G. A. B. (2014). Dioecy is associated with higher diversification rates in flowering plants. en. *Journal of Evolutionary Biology* 27(7), 1478–1490. ISSN: 1010061X. DOI: 10.1111/jeb.12385.

Contents

List of Figures

List of Tables

Part I

General Introduction

Chapter 1

Breeding systems and genome evolution in plants

The introductory part starts by a general introductive chapter on the consequences of plant breeding systems on genome evolution. It corresponds to a book chapter in the Encyclopedia of Evolutionary Biology, which was written following an invitation by EEB editor Hiroshi Akashi. The manuscript was submitted in December 2014 and is now accepted. It was for me an opportunity to review a field for which I have a strong interest. I wrote the manuscript with the help of Gabriel Marais and he made the figures.

Breeding systems and genome evolution in plants

Aline Muyle[1], Gabriel AB Marais[1]

1.1 Keywords

Breeding system, selfing, outcrossing, asexuality, dioecy, nucleotide diversity, selection, genome size, transposable element, base composition.

1.2 Abstract

Plants have very diverse breeding systems. Breeding systems determine how DNA is transmitted from one generation to the next and can influence many evolutionary forces shaping the genomes. In this chapter, we review the theoretical predictions on how breeding systems should affect plant genome evolution and also how data fits with these predictions. We focus on the transition to selfing, asexuality and dioecy, for which theory and data are available.

1.3 Introduction

Breeding systems exhibit the most extraordinary variety in plants, and particularly in angiosperms where most of our knowledge comes from (but see Bainard et al., 2013). Animals are comparatively very uniform with gonochorism being the general rule (~95%). Plants can either reproduce sexually or asexually or combine both strategies (Neiman et al., 2014). Sexually reproducing plants can either outcross, self-fertilize or have mixed mating systems. A large diversity of mechanisms have evolved in plants that promote either outcrossing or selfing. As selfing is possible only in hermaphrodites, dioecy is an obvious way of avoiding it. Outcrossing is also obligatory for females in gynodioecious species and for males in androdioecious species. Other mechanisms prevent selfing from happening in hermaphroditic species, such as heterostyly in which individuals have different flower morphologies, or self-incompatibility in which self-pollen is rejected.
Breeding systems are highly labile in plants and transitions from one to another are common (see Box 1.3 for possible pathways). See Table 1.1 for the percentages of the breeding systems discussed in this chapter in angiosperms (Hojsgaard et al., 2014; Igic and Kohn, 2006; Renner, 2014; Renner and Ricklefs, 1995; Richards, 1997).

[1]Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France

Box 1.1 : Glossary

- **Apomixis or asexuality or clonality:** asexual species only reproduce via mitosis, daughters are genetically identical to mothers unless a mutation occurs. In apomixis a seed develops from an unfertilized ovule.
- **Sexuality:** sexual species reproduce through the alternation of meiosis and syngamy (fusion of gametes). During meiosis, chromosomes recombine and then segregate in different gametes.
- **Hermaphroditism:** in hermaphroditic species, individuals carry both male and female reproductive organs and produce both ovules and pollen grains.
- **Dioecy:** in dioecious species, sexes are separated, females produce only ovules and males produce only pollen.
- **Gonochorism:** separated sexes in animals, individuals are either females (producing only ovules) or males (producing only sperm).
- **Outcrossing:** Outcrossing species reproduce through the fusion of gametes from distinct individuals at each generation.
- **Selfing:** Selfing species are sexual and undergo meiosis, but (in the case of pure selfing) fertilization only occurs between gametes produced by the same hermaphroditic individual.
- **Gynodioecy:** a breeding system characterized by the co-occurrence of hermaphroditic and female individuals in the same species, often as the result of nucleo-cytoplasmic interactions.
- **Androdioecy:** a breeding system characterized by the co-occurrence of hermaphroditic and male individuals in the same species.
- **Heterostyly:** in heterostylous species, obligate outcrossing is enforced by morphological differences in the length or shape of the pistil and stamen. Two or three flower morphs coexist in a species and each individual expresses only one morph for all its flowers. The pollen from a given morph cannot fertilize a flower of the same morph. Is opposed to homostyly.
- **Self-incompatibility:** a genetic system that prevents self-fertilization in hermaphrodites through recognition and rejection of pollen expressing the same allelic specificity as that expressed in the pistils, is opposed to self-compatibility.
- **Synonymous:** refers to a sequence modification that does not result in a change at the protein level, is opposed to non-synonymous.
- **B chromosomes:** supernumerary and facultative chromosomes that do not obey the ordinary Mendelian laws of inheritance.
- **Genetic drift:** the random sampling of genetic variants in a finite population that leads to the random loss of some variants.
- **Selection:** the process by which genetic variants become more or less frequent in a population according to their effect on fitness. Positive selection favors the fixation of advantageous alleles. Purifying selection favors the elimination of deleterious alleles. Balancing selection maintains diversity at a locus.
- **Recombination:** the process by which two DNA molecules exchange genetic information, resulting in the production of a new combination of alleles.
- **Selection on codon usage:** in some species, natural selection may favor some synonymous codons over others for encoding an amino acid. The preferred codons confer greater efficacy and/or accuracy of translation. This may generate a biased use of synonymous codons (i.e. codon usage bias).
- **Linkage disequilibrium:** the non-random association of alleles along a portion of chromosome in a population or species.

In this chapter, we will focus on the consequences of breeding systems on genome evolution in plants, including their effects on nucleotide diversity (both neutral and selective), genome base composition, genome structure and genome size. We will focus on the nuclear genome.

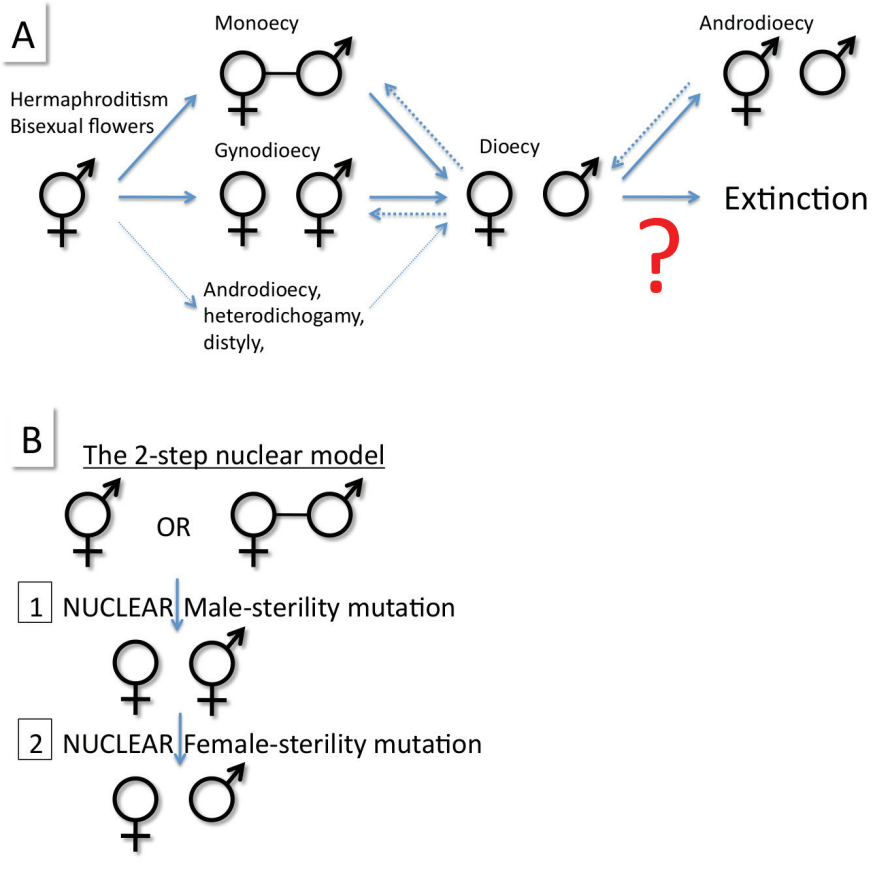
Box 1.2 : Nomenclature

- N_e**:** The effective size of a population, determines the rate of change in allele frequencies of a population caused by genetic drift.
- μ**:** mutation rate per locus per generation.
- s**:** the selection coefficient, measures the relative fitness of a genotype compared to another, 0 for mutations without effect on fitness, >0 for beneficial mutations that increase fitness and <0 for deleterious mutations. The higher the absolute value of the coefficient, the stronger the effect of the mutation on fitness.
- K_a/K_s**:** the ratio of non-synonymous to synonymous substitution rates. Because selection acts primarily on proteins and not DNA sequences, synonymous changes are often treated as selectively neutral. K_a/K_s ratio is a proxy of the intensity and form of selection acting on protein sequences (~ 1 for neutral evolution, >1 for positive selection and <1 for negative selection).
- π_S**:** population synonymous nucleotide diversity, the chance that two randomly chosen sequences will differ with a synonymous mutation at a given locus. Is assumed to reflect the neutral genetic diversity.
- π_N/π_S**:** the ratio of non-synonymous to synonymous polymorphism rate. Similar to the K_a/K_s ratio, π_N/π_S captures the intensity and form of selection acting on protein sequences, but in polymorphism data and hence at a shorter time scale.
- **TEs:** transposable elements, DNA sequence parasites that increase their copy number in genomes through copy/paste (retrotransposons) or cut/paste (transposons) processes.
- **BGC:** biased gene conversion. Gene conversion is the process by which a portion of DNA molecule is copy-pasted onto the other DNA molecule during recombination. In some organisms, this process is biased in favor of G or C alleles, which are more frequently copy-pasted than A or T alleles.
- F_{IS}**:** the inbreeding coefficient, stands for the excess (F_{IS} values between -1 and 0) or deficit (F_{IS} values between 0 and 1) of heterozygotes occurring in a population as compared to expectations under Hardy-Weinberg equilibrium.
- **GC3:** GC content at third codon position.
- **GC*:** equilibrium GC content, reached by a genome if estimated mutation and substitution rates remain constant.

Table 1.1: Percentages of the breeding systems discussed in this chapter in angiosperms.

breeding system	% of angiosperms
Dioecy	~6%
Autogamy	~27%
Apomixis	~0.2%

Box 1.3 : Plant breeding system transitions



A) In plants, dioecy can evolve from the ancestral hermaphroditic state (both male and female organs inside flowers of an individual) either through monoecy (male and female flowers on the same plant) or through gynodioecy (female and hermaphrodite individuals in the same species) (Barrett, 2002; Renner, 2014). Other pathways may exist but are thought to represent a minority of cases: through androdioecy (male and hermaphrodite individuals in the same species), heterodichogamy (temporal separation of male and female maturity in hermaphrodite individuals) and distyly (presence of two style morphs inside a species, morphs are carried by different individuals).
Dioecy was hypothesised to lead to extinction (Section 1.4), but can also evolve into androdioecy, which is possibly facilitated by the presence of inconstant males (males that happen to produce some seeds).
At the genetic level, the transition through gynodioecy is thought to happen through a first nuclear male-sterility mutation followed by a female-sterility mutation, as shown in **(B)**.

1.4 Changes in breeding systems and their effects on population genetics parameters

1.4.1 Effects on heterozygosity levels

When selfing evolves from outcrossing, homozygosity is almost complete after a few generations ($F_{IS} \sim 1$). This reveals recessive mutations, which effects were hidden in het-

erozygous outcrossers ($F_{IS} \sim 0$). In asexuals, due to clonal transmission of the genome, any new mutation appearing in a genotype endlessly remains at the heterozygous state, unless the exact same mutation occurs on the homologous chromosome (Meselson effect). Heterozygosity is thus high in asexuals ($F_{IS} \sim -1$, Glémin and Galtier, 2012).

1.4.2 Effects on effective recombination

The efficacy of recombination depends on both crossing over and outcrossing rates (Glémin and Galtier, 2012). In asexuals, there are no crossovers so that the whole genome is non-recombining. In selfers, due to high homozygosity levels, homologous chromosomes are virtually identical inside an individual. Consequently, crossovers cannot break linkage between alleles in a selfing population and recombination is mostly inefficient (Glémin and Galtier, 2012; Nordborg, 2000).

1.4.3 Effects on effective population size

With pure selfing, N_e is decreased by half. Full homozygosity reduces the coalescent process to sampling diploid individuals (n) instead of sampling alleles ($2n$), since in selfers, two copies of the same allele are found in an individual. With partial selfing, $N_e = 2N/(1+F)$ with N the population size and F the expected equilibrium inbreeding coefficient (Nordborg and Donnelly, 1997; Pollak, 1987). N_e is also decreased by half in asexuals due to a similar process of genotype sampling during the coalescent process (Haag and Roze, 2007).

N_e is further reduced in selfers and asexuals due to inefficient or absent recombination, which results in selective interference among genetically linked loci (reviewed in Charlesworth, 2009; Gordo and Charlesworth, 2001). For instance, a strongly deleterious mutation will be removed by selection along with the mutations genetically linked to it, even if these are slightly advantageous ("ruby in the rubbish", Charlesworth et al., 1993; Peck, 1994). Similarly, mutations on a haplotype carrying a strongly advantageous mutation will reach fixation by genetic hitch-hiking, even though they are slightly deleterious (Maynard Smith and Haigh, 1974). These processes increase genetic drift and decrease genetic variation since more alleles are lost than would be without linkage, resulting in a reduced N_e. In small selfing and asexual populations, genomes without mutations cannot be regenerated by recombination of two genomes carrying different mutations and genomes tend to accumulate mutations until they are removed by selection, reducing N_e further (Muller's ratchet, Heller and Maynard Smith, 1978; Muller, 1964).

The ability of selfers to found new populations from a single seed (Baker, 1955) combined with their reduced pollen migration (due to reduced allocation to male function) leads to small and isolated populations where drift is important and recurrent bottleneck events take place, which further reduces N_e (metapopulation dynamics, Ingvarsson, 2002; Pannell and Charlesworth, 2000).

In angiosperms, it has been suggested that dioecy is an evolutionary dead-end (Heilbuth, 2000). Dioecy may bring some handicaps. Only females produce seeds, which may cause less efficient dispersal of individuals and lead to isolated populations with reduced N_e (the seed-shadow handicap, Heilbuth et al., 2001). Also, fertilization could be reduced: sexual selection drives male flowers to attract pollinators as much as possible, to the detriment of females if pollinators happen to be rare, decreasing N_e further (Vamosi and Otto, 2002). However, these theories were recently challenged by a study, which found higher (and not lower) diversification rates in dioecious versus non-dioecious clades (Käfer et al., 2014).

1.4.4 Effects on the efficacy of selection

The outcome of natural selection depends on the product $N_e s$ (Kimura, 1962). When N_e is reduced, mutations with low absolute values of s will behave almost neutrally. Weakly deleterious alleles may be fixed while weakly advantageous ones may be lost because of genetic drift overcoming selection. Because of reduced N_e, selfers, asexuals and perhaps dioecious species are expected to experience a reduced efficacy of selection assuming that many newly arisen mutations have low selection coefficients.
The next sections review the real data that tested these theoretical expectations (summarised in Figure 1.1).

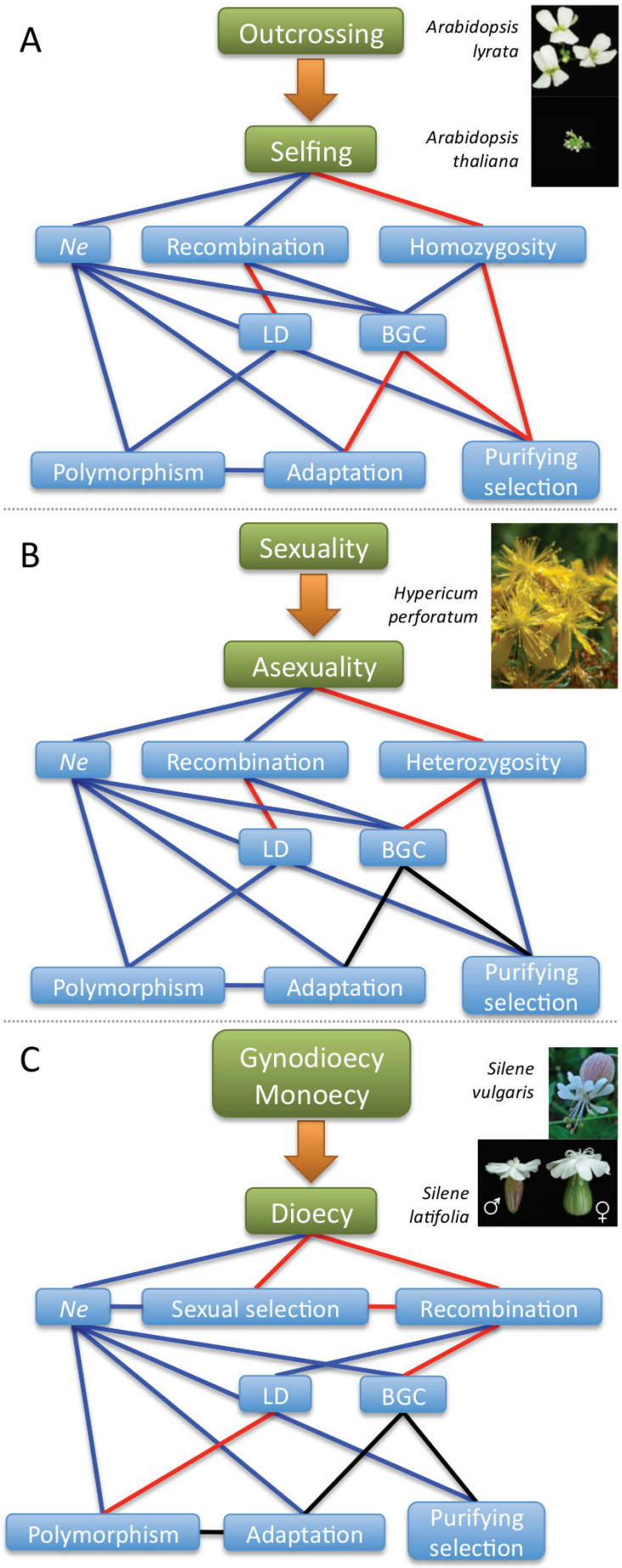


Figure 1.1: Summary of the theoretical expectations for the effects of three breeding systems on genome evolution in plants. Transitions to **A)** selfing, **B)** asexuality and **C)** dioecy. Blue lines indicate a negative effect, red lines a positive one. Examples of plant taxa with the studied transitions are shown. *Hypericum perforatum* includes both apomictic and sexual forms.

1.5 Breeding systems and within-species neutral diversity

Neutral nucleotide diversity is directly linked to the effective population size ($\pi_S = 4N_e\mu$) so that higher levels of within-population neutral diversity are expected in outcrossers compared to selfers and in sexuals compared to asexuals (reviewed in Charlesworth and Wright, 2001; Glémin and Galtier, 2012. No studies have compared dioecious and non-dioecious species diversity so far.

1.5.1 Selfing, asexuality and within-species neutral diversity

Consistent with the theoretical expectations, selfing populations were found to be less diverse than closely related outcrossing populations and asexual species were found to have a lower diversity compared to their sexual counterparts (Table 1.2). Heterozygosity is, on the other hand, high in asexual individuals as discussed previously (Meselson effect). Looking at levels of species-wide diversity may blur differences among selfing and outcrossing species because selfing increases population differentiation through reduced pollen gene flow (Charlesworth et al., 1997). Extinction-recolonization dynamics, however, can reduce more strongly $N_e s$ in structured selfing populations (Ingvarsson, 2002). The levels of neutral diversity in selfing and asexual species also depend on whether the change in breeding system has a single or multiple origins (Ness et al., 2010) and whether partial sexuality is maintained in asexuals (Cosendai et al., 2013; Eckert and Barrett, 1993).

Table 1.2: Summary of studies comparing neutral nucleotide diversity between selfing and outcrossing or asexual and sexual species. * yes if neutral diversity lower in selfers compared to outcrossers or asexuals compared to sexuals. Note that allozyme diversity is not neutral.

Taxonomic group	Groups compared	Dataset	diversity*	comment	References
		Selfing versus outcrossing			
Aegilops	5 selfers / 1 outcrosser	RFLP 52 genes	yes		Dvorák et al., 1998
angiosperms		Meta-analysis allozyme diversity	yes		Hamrick and Godt, 1990
angiosperms		Meta-analysis allozyme diversity	yes	high variance in species wide diversity levels in selfers	Schoen and Brown, 1991
angiosperms		Meta-analysis allozyme diversity	yes		Hamrick and Godt, 1996
angiosperms	11 selfers / 12 outcrossers	Meta-analysis allozyme, microsatellite	yes		Charlesworth and Pannell, 2001
angiosperms	15 selfers / 15 outcrossers	Meta-analysis allozyme diversity	yes		Charlesworth, 2003
angiosperms and gymnosperms	10 selfers / 38 outcrossers	DNA markers	yes	low diversity within selfing populations but high among population diversity	Nybom, 2004
angiosperms	29 selfers / 42 outcrossers	Meta-analysis	yes		Glémin et al., 2006
Arabidopsis	1 selfer / 1 outcrosser	5 nuc. Loci	yes		Wright et al., 2003
Arabidopsis/Capsella	1 selfer / 1 outcrosser	257 and 483 nuc genes	yes		Slotte et al., 2010
Arabidopsis/Capsella	2 selfers / 2 outcrossers	780, 354 / 120, 346 exons and 821 / 41introns	yes		Qiu et al., 2011

Continued on next page

Table 1.2 – continued from previous page

Taxonomic group	Groups compared	Dataset	diversity*	comment	References
Capsella	1 selfer / 1 outcrosser	39 nuc genes, 14 / 20 individuals	yes		Foxe et al., 2009
Capsella	1 selfer / 1 outcrosser	1 mit. 17 nuc. Genes, 25 / 7 individuals	yes		Guo et al., 2009
Capsella	1 selfer / 1outcroser	genome/transcriptome	yes		Slotte et al., 2013
Capsella	1 selfer / 1outcroser	6 / 5 transcriptomes	yes		Brandvain et al., 2013
Clarkia xantiana	1 selfer / 1 outcrosser	8 nuc. 3 cp. genes, 80 / 75 individuals	yes		Pettengill and Moeller, 2012
Collinsia	1 selfer / 1 outcrosser	17 nuc genes	yes		Hazzouri et al., 2013
Eichhornia paniculata	outcrossing and selfing populations	allozyme diversity	yes		Barrett and Husband, 1990
Eichhornia paniculata	trimorphic, dimorphic, and monomorphic populations	10 EST, 225 individuals	yes		Ness et al., 2010
Juan Fernandez Archipelago (Chile)	6 selfers / 12 outcrossers	allozyme diversity	yes		Crawford et al., 2001
Leavenworthia	1 outcrosser / 4 selfers	allozyme diversity	yes		Charlesworth and Yang, 1998
Leavenworthia	2 selfers, 1 outcrosser, outcrossing and selfing populations	6 loci	yes		Liu et al., 1999, 1998
Leavenworthia alabamica	outcrossing and selfing populations	8 nuc. Genes	yes		Busch et al., 2011

Continued on next page

Table 1.2 – continued from previous page

Taxonomic group	Groups compared	Dataset	diversity*	comment	References
Lycopersicon pimpinellifolium	outcrossing and selfing populations	allozyme diversity	yes		Rick et al., 1977
Lycopersicon	3 selfers, 2 mixed, 4 outcrossers	RFLP 36 loci	yes		Stephan and Langley, 1998
Lycopersicon	2 selfers / 3 outcrossers	5 loci	yes		Baudry et al., 2001
Mimulus	1 selfer / 1 outcrosser	2 nuc loci	yes		Sweigart and Willis, 2003
Miscanthus	1 selfer / 1 outcrosser	Adh1 locus	yes		Chiang et al., 2003
		Asexuality versus sexuality			
angiosperms	22 asexuals	allozyme diversity	yes	most genotypes occur in a single population	Ellstrand and Roose, 1987
angiosperms		allozyme diversity	yes		Hamrick and Godt, 1990
Elodea	3 asexuals	AFLPs	yes		Lambertini et al., 2010
Grevillea renwickiana	1 asexual	10 microsatellite	yes		James and McDougall, 2014
Zeuxine/Eulophia	1 apomictic / 2 outcrossers	RAPDs	yes		Sun and Wong, 2001

1.5.2 Self-incompatibility and the S-locus neutral diversity

In self-incompatible species, self-pollen is recognized and rejected by the pistil through protein interactions. The "pollen" and "pistil" genes are found in the tightly linked S-locus, which is under balancing selection: pollen carrying rare alleles are rejected at lower rates compared to pollen carrying common alleles, which prevents the loss of rare alleles through drift. The S-locus is indeed highly diverse (Castric and Vekemans, 2004).

1.6 Breeding systems and the global efficacy of natural selection

When selection is reduced, species are expected to accumulate more deleterious mutations. Assuming that deleterious mutations are much more frequent than beneficial ones, both π_N/π_S and K_a/K_s should be higher in selfers compared to outcrossers, asexuals compared to sexuals and perhaps in dioecious compared to hermaphroditic species. Codon usage bias (if under selection) should also be weaker (Marais et al., 2004).

1.6.1 The transition from outcrossing to selfing

Studies based on divergence (K_a/K_s) failed to detect reduced selection in selfers compared to outcrossers (Table 1.3). This could be due to the purging effect of homozygosity that reveals recessive mutations in selfers and increases the efficacy of purifying selection (Charlesworth, 1992; Glémin, 2007). It has also been suggested that rare events of outcrossing in partially selfing plants could be enough to avoid the accumulation of deleterious mutations, and this could be promoted by the high recombination rates observed in selfers (Wright et al., 2008). Moreover, selfing will increase K_a/K_s only if mutations have mild effects on fitness, which might not be the case: very little is known about the shape of the distribution of fitness effects of mutations (Glémin and Muyle, 2014). In agreement with this idea, selection on codon usage (involving small selective coefficients) is weaker in selfers than in outcrossers (Table 1.3).

Table 1.3: Summary of studies comparing molecular evolution between mating systems: selfing and outcrossing or asexual and sexual species or dioecious and hermaphrodite (modified from Glémin and Muyle, 2014).

Taxonomic group	Groups compared	Dataset	K_a/K_s*	π_N/π_S*	Positive selection*	Codon usage*	References
Selfing versus outcrossing							
Angiosperms	29 selfers / 42 outcrossers	Meta-analysis (polymorphism)		yes	unclear		Glémin et al., 2006
Angiosperms	290 selfers / 426 outcrossers	2 cytoplasmic genes matK, rbcL	no				Glémin and Muyle, 2014
Arabidopsis/-Drosophila	1 selfer / 1 outcrosser + sister species each	12 / 34 genes		yes	yes		Bustamante et al., 2002
Arabidopsis	1 selfer / 1 outcrosser	23 nuc genes + 1 chloro gene	no			no	Wright et al., 2002
Arabidopsis	1 selfer / 1 outcrosser	675 loci / 62 large exons nuc genes, 13 orthologous genes		Not on 13 orthologs	no		Foxe et al., 2008
Arabidopsis/Brassica	1 selfer / 2 outcrossers	185 and 83 nuc genes				no	Wright et al., 2007
Arabidopsis/-Capsella	1 selfer / 1 outcrosser + sister species each	257 and 483 nuc genes	no	yes	yes		Slotte et al., 2010
Arabidopsis/-Capsella	2 selfers / 2 outcrossers	780, 354 / 120, 346 exons and 821 / 41introns				yes	Qiu et al., 2011
Capsella	1 selfer / 1 outcrosser	Complete genome / 11 transcriptomes		yes			Slotte et al., 2013

Continued on next page

Table 1.3 – continued from previous page

Taxonomic group	Groups compared	Dataset	K_a/K_s*	π_N/π_S*	Positive selection*	Codon usage*	References
Capsella	1 selfer / 1 outcrosser	Complete genome / 11 transcriptomes		yes (after removing shared ancestral polymorphism)			Brandvain et al., 2013
Collinsia	1 selfer / 1 outcrosser	17 nuc genes + transcriptomes		Not on 17 genes, yes on transcriptomes		no	Hazzouri et al., 2013
Eichhornia	3 selfers / 1 outcrosser	7890 nuc genes (transcriptomes)		yes		yes	Ness et al., 2012
Triticeae	2 selfers / 2 outcrossers	52 nuc genes + 1 chloro gene	no		yes	yes	Haudry et al., 2008
Triticeae	9 selfers / 10 outcrossers	27 nuc genes	no				Escobar et al., 2010
		Asexuality versus sexuality					
Boechera spatifolia	2 asexuals / 9 mixed / 19 sexual	9126 SNPs, 14 SSR, traits			yes		Lovell et al., 2014
Oenothera	16 asexuals / 16 sexuals	1 nuc defense gene (chiB), 5 conserved nuc genes	3 out of 5 genes		Yes for ChiB		Hersch-Green et al., 2012
Ranunculus auricomus	2 asexuals / 3 sexuals	1231 ORFs (transcriptomes)	not genome wide, yes on sexual reproduction genes				Pellino et al., 2013

Continued on next page

Table 1.3 – continued from previous page

Taxonomic group	Groups compared	Dataset	K_a/K_s*	π_N/π_S*	Positive selection*	Codon usage*	References
			Dioecy versus hermaphroditism				
Silene	3 5Myo dioecious and 3 recent dioecious / 1 gynodioecious / 1 outgroup	16 nuc genes, 4 chloro genes, 25 ESTs	yes in 5Myo dioecious species				Käfer et al., 2013

Contrasting with results on K_a/K_s, most studies based on polymorphism (π_N/π_S) detected reduced selection in selfers: in *Arabidopsis*, *Capsella*, *Collinsia*, *Eichhornia* and in an angiosperm meta-analysis of about thirty species (Table 1.3). This suggests that selfing does result in selection being reduced but is probably of such recent origin that reduced selection can only be detected over short time scales (i.e. in polymorphism and not divergence data, Figure 1.2).

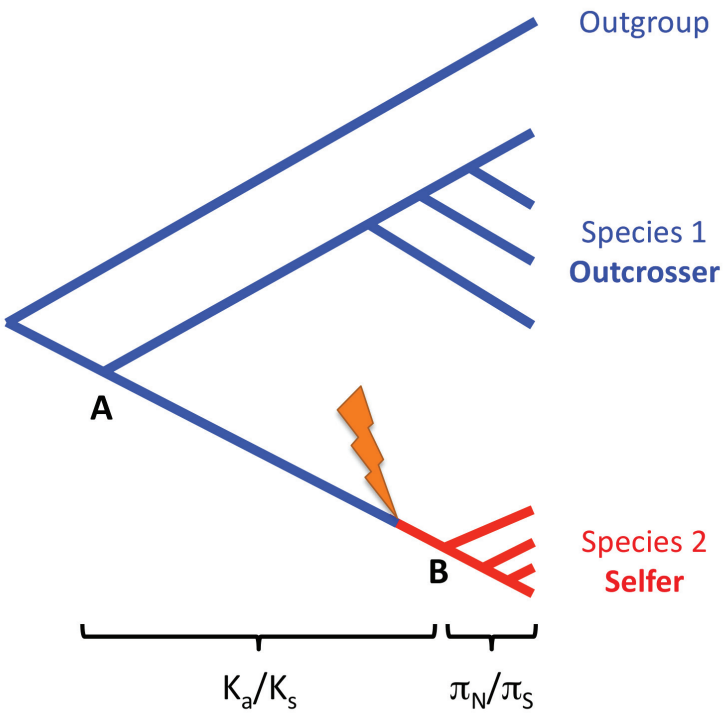


Figure 1.2: Timing of the transition from outcrossing to selfing and its effects on divergence and polymorphism data. The transition is depicted by an orange lightning. If the transition is recent and a little older than the coalescent point of all mutations in the selfing population (node B), it will affect the polymorphism data (studied with π_N/π_S) strongly and the divergence data (studied with K_a/K_s) only weakly. This is because the divergence data will mainly reflects the evolution that took place from node A to B, whereas polymorphism will reflect the evolution that took place from node B to now (adapted from a figure by Glémin and Marais, Thomas et al., 2015).

More efficient selection on recessive mutations in selfers than outcrossers could make positive selection and adaptation faster in selfers than in outcrossers (Charlesworth, 1992; Glémin, 2007). However, this is true for new beneficial mutations, and not for standing variation. When adaptation is underlain by standing variation, it is more efficient in outcrossers than in selfers (Glémin and Ronfort, 2013). In agreement with this idea, positive selection seems less efficient in selfers than in outcrossers (Table 1.3).

1.6.2 The transition from sexuality to asexuality

Very few studies have so far compared sexual and asexual plant species (Table 1.3). Current data suggests adaptation is weak in asexuals. However, K_a/K_s analysis revealed no clear trend of reduced selection. No π_N/π_S analysis is available so that it is not possible

to disentangle the possible effect of recent switches to asexuality from no accumulation of deleterious mutations.

1.6.3 The transition from hermaphroditism to dioecy

The only study that tested the hypothesis of reduced selection in dioecious species using K_a/K_s is consistent with expectations when dioecy was several millions years old (Table 1.3). How the expectations are affected by the ecology of the species remains to be explored: for example wind-pollinated dioecious species will not be affected by pollinator competition between sexes and the reduction in $N_e s$ should not be as strong as in insect-pollinated species.

1.7 Breeding systems and the evolution of genome base composition

BGC is a molecular process associated with recombination that drives genomic base composition towards a high GC content (Figure 1.3, Lesecque et al., 2013; Marais, 2003), which was shown to affect plant genomes (Glémin et al., 2014; Muyle et al., 2011; Serres-Giardi et al., 2012). Because BGC will operate when recombination occurs in heterozygous loci, it will increase GC content in outcrossing and not in purely selfing species (Marais et al., 2004). Because of the absence of recombination in asexual species, BGC should not affect GC content.

Observations fitted expectations for non coding DNA (higher GC content in outcrossing Poaceae, Glémin et al., 2006), unless a too small intron dataset was used (Brassicaceae, Qiu et al., 2011). But studies using coding sequences gave conflicting results in Triticeae (Escobar et al., 2010; Haudry et al., 2008), or supported BGC but could not distinguish it from mutational biases in other groups (Hazzouri et al., 2013; Wright et al., 2007). More studies comparing selfing and outcrossing species are required, with appropriate methods that allow to distinguish BGC from mutational biases (GC* or derived allele frequency spectra) and large non-coding datasets in order to distinguish BGC from selection on codon usage. Similar studies also have to be carried out in asexual and sexual species.

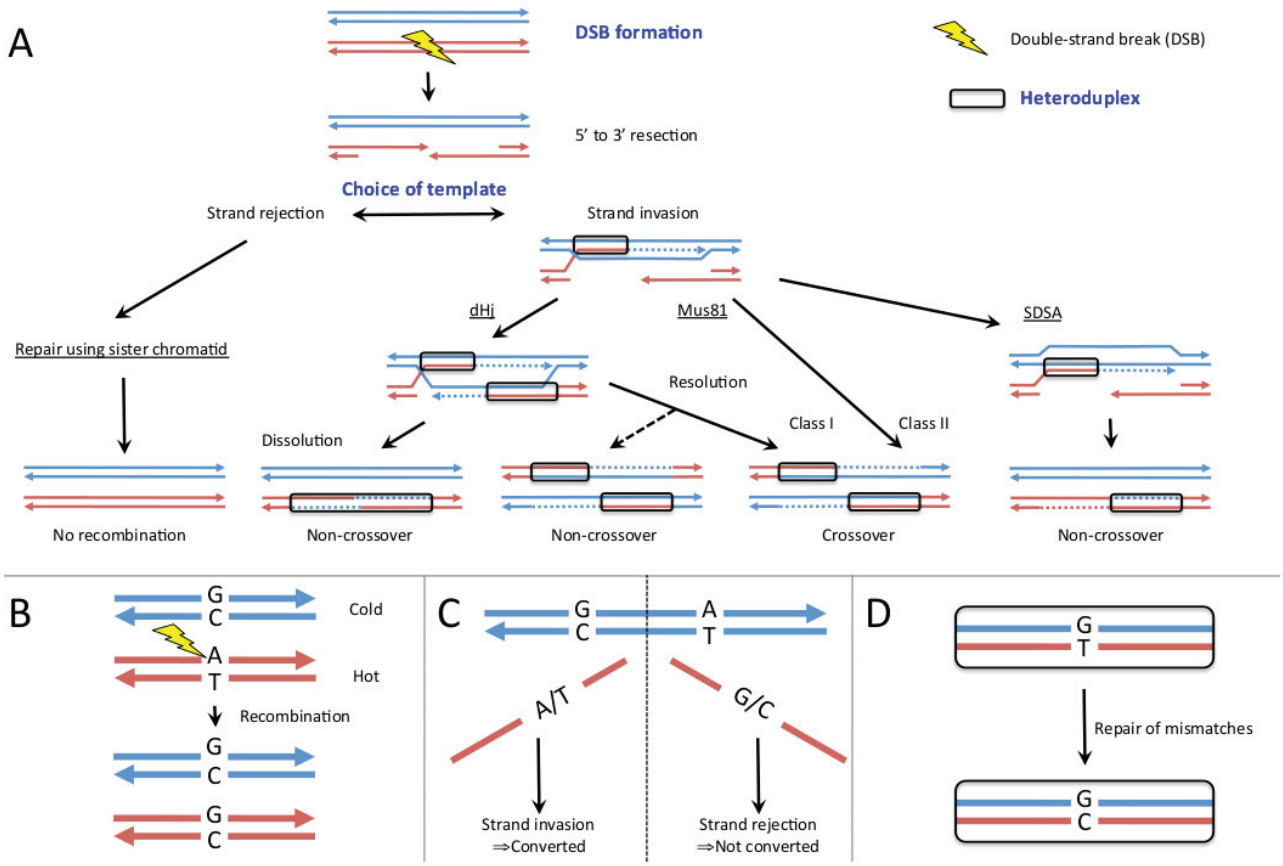


Figure 1.3: Possible mechanisms for biased gene conversion (BGC), figure adapted from Lesecque et al., 2013. Different pathways for double-strand break (DSB) repair during meiotic recombination are shown in **A)** Blue and red chromosomes indicate chromosomes from different parental origin. Recombination starts with the DSBs on one chromosome. Broken DNA is resected forming single-stranded DNA, which will be used to select a template for repairing the breaks. The mismatch repair (MMR) triggers the choice of template: if the degree of similarity between the single-stranded DNA and the potential template is sufficient, the template is invaded and a D-loop is formed. Otherwise, repair takes place by using the sister chromatin, no recombination happens. Recombination can take place by several pathways: the repair of the breaks by the MMR can lead to the formation of a double-Holliday junction (dHj) intermediate. Resolution of this intermediate may result in crossovers (COs) of class 1, i.e. showing CO interference (a mechanism by which 1 CO / arm / meiosis is ensured). Another pathway (Mus81) produces non-interfering COs. And the synthesis-dependant strand annealing (SDSA) pathway gives non-crosser (gene conversion) events only. Heteroduplexes (DNA with strands from different origins) can be found at various stages of the recombination pathway; they may include mismatches that are detected and repaired by different pathways (MMR and base-excision repair BER). BGC may result from different possible biases. **B)** Initiation bias. If A/T alleles are recombinationally hotter than G/C alleles; G/C alleles will get copy-pasted onto the A/T alleles (recombination hot spots get converted by recombination cold spot, see the "hot spot paradox", Lesecque et al., 2014), and this will increase GC content. **C)** Choice of template bias. If the chance of the single-stranded DNA to invade the homologous chromosome is higher when it's GC-rich; this will also create BGC. **D)** Repair bias. If repair of mismatches in heteroduplexes is biased in favor of G/C alleles; this may also generate BGC. A recent study in yeast suggests that BGC is found mostly associated with long gene conversion tracts and CO and may be due to MMR (Lesecque et al., 2013). While evidence for BGC have been found in many organisms (Pessia et al., 2012), in most cases the exact mechanisms remains unknown and could be one of the three mechanisms outlined here.

1.8 Breeding systems and the evolution of chromosomes and genome structure

Due to inefficient recombination, strong linkage disequilibrium is expected in selfers and was observed in the highly selfing *Arabidopsis thaliana* (Kim et al., 2007; Miyashita et al., 1999; Nordborg et al., 2005; Wright et al., 2008). In the outcrossing *Collinsia linearis*, linkage disequilibrium decayed rapidly (over less than 1 kbp), whereas in the selfing *Collinsia rattanii* it could extend to several kbp (Hazzouri et al., 2013).
Chromosomal rearrangements, when heterozygous, are deleterious as they cause chromosome disjunction and result in unbalanced gametes (Lande, 1985). However selfing is associated with almost complete homozygosity and selection against chromosomal rearrangements will be weak. Fast karyotypic evolution is expected in selfers (Charlesworth, 1992). For example, *A. thaliana*'s karyotype includes 5 chromosome pairs and not 8, the standard karyotype in Brassicaceae (Hu et al., 2011).
The mechanisms for sex determination is unknown in most of the ~15,000 dioecious species; sex chromosomes have been described in only 40 of them (Ming et al., 2011). In male heterogametic systems males are XY and females XX, in female heterogametic systems females are ZW and males are ZZ, and in some species with a haplodiploid life cycle the gametophytes have separated sexes and females are U and males are V (Bachtrog et al., 2011). Y, W, U and V chromosomes do not recombine which strongly impacts their molecular evolution. Self-incompatibility S-loci are also non-recombining and have similar properties (Castric and Vekemans, 2004).

1.9 Breeding systems and the evolution of genome size

Three major mechanisms influence genome size: ploidy levels, genomic parasite dynamics (TEs and B chromosomes) and insertion/deletion events. We will review the effect of breeding systems on each mechanism impacting genome size and then consider the global effect of breeding systems on genome size.

1.9.1 Breeding systems and ploidy levels

Virtually all apomictic plants are polyploid with some exceptions of recovered diploidy in apomictic hybrids (usually polyhaploid, see Asker and Jerling, 1992; Koltunow and Grossniklaus, 2003). This results in larger genomes in asexual compared to sexual plants as observed in *Hypericum* (Matzk et al., 2003). Whether apomixis causes polyploidy, or whether polyploidy causes apomixis by disrupting developmental processes is, however, still unclear (Bhat et al., 2005).
Gametophytic self-incompatibility systems can be directly broken down by polyploidy through the formation of diploid pollen grain (Robertson et al., 2011). The inbreeding

depression following self-incompatibility breakdown was proposed to trigger the evolution of dioecy in the Solanaceae family and 12 other genera (Miller, 2000). This proposed association was further tested in 22 genera (i.e. a small fraction of the dioecious genera) and hermaphroditism is indeed more common among diploid than polyploid species, whereas gender dimorphism (dioecy and gynodioecy) is more frequent among polyploid species (Ashman et al., 2013). In most dioecious genera though, the transition to dioecy is probably not accompanied by polyploidization.
Liverworts are ancestrally dioecious and it has been suggested that polyploidy could happen more often in hermaphrodites. However, a test on 67 liverwort species, controlling for phylogenetic inertia, did not support this hypothesis (Bainard et al., 2013).

1.9.2 Breeding systems and insertion/deletion

A. thaliana has smaller introns than *A. lyrata* because of different indel dynamics in both species (Wright et al., 2002). A whole genome alignment of *A. thaliana* and *A. lyrata* showed that more than 50% of the *A. lyrata* genome is missing from the *A. thaliana* genome (Hu et al., 2011), explaining the large difference in genome size between both species (*A. thaliana*: 125 Mb, *A. lyrata*: 207 Mb). Large deletions of TEs and intergenic regions, probably selected for, resulted in genome streamlining in *A. thaliana*.

1.9.3 Breeding systems and TE dynamics

TE amplification is predominantly deleterious when insertion occurs in or near functional regions or when those are lost by ectopic recombination between TE copies (Charlesworth and Langley, 1989). The number of TEs in a population is determined by the balance between the capacity to invade the host genome through transposition and the efficiency of the host defense mechanisms against TEs. Old TE elements end up being inactivated by the host defense mechanisms. A TE element will survive if it can jump from a host to another through horizontal gene transfer and contaminate new hosts, which do not have yet efficient defense mechanisms against that element (Schaack et al., 2010).
In asexual species, TE do not have that chance and they will remain confined to a single host (Hickey, 1982). TE transposition is expected to decline on the long run (Bestor, 1999; Wright and Finnegan, 2001). However, reduced selection against TEs in asexuals might favor TE accumulation. The outcome of these two opposed effects is not easy to predict. It is not clear whether fewer or more TEs should be found in asexuals than in sexuals.
In *Hypericum*, the mean DNA content per chromosome for old asexual speciesis higher than that of closely related sexual species, suggesting an accumulation of TEs in asexuals (Matzk et al., 2003). In a study comparing four pairs of closely related sexuals and asexuals for three TEs, no reduced selection against TEs was found in asexuals using K_a/K_s, suggesting their transposition capacity has not decreased. The recent origin of the asexuals could, however, explain the results, as shown by simulations(Docking et al., 2006).

Similar expectations hold for selfing species: TE transposition should decline because of confinement to a lineage but this could be blurred by a globally reduced selection in selfers. But the dynamics of TEs in selfers are even more difficult to predict than in asexuals. Because of a high level of homozygosity, purifying selection is expected to purge more efficiently the deleterious recessive mutations from the population in selfers than in outcrossers. However, ectopic exchanges among TE copies will probably be rare in homozygotes (Montgomery et al., 1991), and selection will not act against TEs this way in selfers. It is not clear which one of these many opposing effects should prevail (Boutin et al., 2012; Charlesworth and Charlesworth, 1995; Morgan, 2001; Wright and Schoen, 1999).

Looking at some TE families have indeed returned contradictory results (Charlesworth and Charlesworth, 1995; Lockton and Gaut, 2010; Tam et al., 2007; Wright and Finnegan, 2001; Young et al., 1994). However, more recently, a clearer view has started to emerge from whole genome comparisons between selfers and outcrossers. *A. thaliana* has less TEs than *A. lyrata* (24% and 30% of the genome respectively, Hu et al., 2011). Interestingly, TEs are less frequent nearby genes in selfers than in outcrossers (Agren et al., 2014; Slotte et al., 2013). Also, deletions of TEs are selected for in *A. thaliana* (Hu et al., 2011), clearly pointing to a more efficient selection against TEs in selfers than in outcrossers. In *A. thaliana*, most TEs are old, older than the *A. thaliana-A. lyrata* speciation. In *A. lyrata*, on the contrary, recent bursts of TEs have occurred (Chaux et al., 2012; Hu et al., 2011; Maumus and Quesneville, 2014; Tsukahara et al., 2012; Wright and Agren, 2011). Similar patterns were found in the selfers *Capsella rubella* and *Capsella orientalis* compared to the outcrosser *Capsella grandiflora* (Agren et al., 2014; Slotte et al., 2013). These observations suggest that reduced transposition and efficient selection against TE insertions are dominating TE dynamics in selfers.

Dioecy can be determined by sex chromosomes, which are known to accumulate TEs because of reduced selection triggered by the absence of recombination for the Y chromosome and a reduced N_e for the X chromosome. In papaya (trioecious), the male-specific and female-specific regions of the sex chromosomes have accumulated retrotransposons as predicted (Gschwend et al., 2012; Wang et al., 2012). In the dioecious plant *Silene latifolia*, the X and the Y chromosomes are much larger than other chromosomes, probably because of sex chromosome specific TE accumulation (Cegan et al., 2012). Interestingly, *S. latifolia* has a much larger genome than *Silene vulgaris*, a non-dioecious close relative. However, whether dioecious species tend to accumulate TEs on the sex chromosomes only or on the whole genome as would predict the dead-end hypothesis is not clear.

1.9.4 Breeding systems and B chromosomes

B chromosomes are also considered parasitic DNA elements, which are mainly deleterious to their hosts. Their dynamics may be similar to TEs with outcrossing favoring their spread; they are therefore expected to be associated with outcrossing or to evolve into

mutualists when associated with selfing (Burt and Trivers, 1998). A significant positive correlation was found between the presence of B chromosomes and outcrossing, when taking phylogenetic inertia into account (Burt and Trivers, 1998; Trivers et al., 2004). However, B chromosomes do not seem to contribute significantly to the evolution of genome size (Trivers et al., 2004).

1.9.5 Global effect of breeding systems on genome size

An analysis of C-values of 14 pairs of closely related outcrossing and selfing species revealed that selfers tend to have smaller genome sizes than outcrossers, and this is not due to differences in ploidy levels (Wright et al., 2008). A detailed analysis of genome size in the *Veronica* genus suggests that selfing, not annuality, is associated with genome size reduction after taking phylogenetic inertia into account (Albach and Greilhuber, 2004). The analysis of a dataset of 205 species showed a significant association between outcrossing and large genomes after taking phylogenetic inertia into account (Whitney et al., 2010). However, these two studies used the phylogenic independent contrast method (Felsenstein, 1985) to correct for phylogenetic inertia. This simple approach (in which sister branches are compared along the tree and ancestral states of characters are inferred in internal nodes by averaging the values of the next lower nodes) has been criticized for the study of genome size as it neglects the possibility of parallel evolution (Lynch, 2011; Whitney et al., 2011). The question of the global effect of selfing on genome size remains thus open. Studies on the effects of asexuality and dioecy on genome size are lacking.

1.10 Conclusion

Current data suggest that changes in breeding system can have a strong impact on genome evolution (Table 1.4). An important point is that the transitions to new breeding systems have to be recent for the data to fit with the theory. In the future, it will be important to date those transitions much more precisely than they are today (when such dating is available) to confirm this. Another important point is about the forces to explain the difference in genomic features between two species with different breeding systems. Breeding system is usually not the only difference, even among closely related species, demography for instance can be a confounding factor. This is a very difficult point; increasing the number of studied systems is probably the only way to address it. We are starting to be in this situation for the transition from outcrossing to selfing, which is the best studied. The transitions to asexuality and dioecy have been less studied. Only a handful of studies are available on the effects of dioecy on genome evolution. However, dioecy is expected to drive substantial changes in evolutionary forces acting on genomes (Figure 1.1C). In particular, sexual selection should be much stronger than in non-dioecious species, probably driving the evolution of sex-biased expression for many genes in the genome. Future studies on dioecious species could provide data on sex-biased expres-

sion (for which we currently know nothing in plants) and other aspects. It could help to understand why dioecy is surprisingly rare in angiosperms.

Table 1.4: Summary of theoretical expectations and observations concerning breeding systems effects on genome evolution in plants. Selfing species are compared to outcrossing species, asexual to sexual and dioecious to non-dioecious. Observation: yes if observation fits theory, no otherwise, - if observation is unavailable.

		Selfing	Asexuality	Dioecy
neutral diversity	Theory	lower	lower	lower?
	Observation	yes	yes	-
selection (K_a/K_s)	Theory	reduced	reduced	reduced?
	Observation	no	no	yes
selection (π_N/π_S)	Theory	reduced	reduced	reduced?
	Observation	yes	-	-
positive selection	Theory	reduced	reduced	reduced?
	Observation	yes	yes	-
codon usage bias	Theory	weaker	weaker	lower?
	Observation	yes	-	-
GC content	Theory	lower	opposing effects	opposing effects
	Observation	yes	-	-
linkage disequilibrium	Theory	stronger	stronger	-
	Observation	yes	-	-
genome size	Theory	opposing effects	opposing effects	larger
	Observation	smaller	larger	yes

1.11 References

Agren, J. A., Wang, W., Koenig, D., Neuffer, B., Weigel, D., and Wright, S. I. (2014). Mating system shifts and transposable element evolution in the plant genus *Capsella*. eng. *BMC genomics* 15, 602. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-602.

Albach, D. C. and Greilhuber, J. (2004). Genome size variation and evolution in *Veronica*. eng. *Annals of Botany* 94(6), 897–911. ISSN: 0305-7364. DOI: 10.1093/aob/mch219.

Ashman, T.-L., Kwok, A., and Husband, B. (2013). Revisiting the Dioecy-Polyploidy Association: Alternate Pathways and Research Opportunities. en. *Cytogenetic and Genome Research* 140(2-4), 241–255. ISSN: 1424-859X, 1424-8581. DOI: 10.1159/000353306.

Asker, S. and Jerling, L. (1992). *Apomixis in Plants*. Anglais. Boca Raton: CRC Press. ISBN: 9780849345456.

Bachtrog, D., Kirkpatrick, M., Mank, J. E., McDaniel, S. F., Pires, J. C., Rice, W., and Valenzuela, N. (2011). Are all sex chromosomes created equal? eng. *Trends in genetics: TIG* 27(9), 350–357. ISSN: 0168-9525. DOI: 10.1016/j.tig.2011.05.005.

Bainard, J. D., Forrest, L. L., Goffinet, B., and Newmaster, S. G. (2013). Nuclear DNA content variation and evolution in liverworts. eng. *Molecular Phylogenetics and Evolution* 68(3), 619–627. ISSN: 1095-9513. DOI: 10.1016/j.ympev.2013.04.008.

Baker, H. G. (1955). Self-Compatibility and Establishment After 'Long-Distance' Dispersal. *Evolution* 9(3), 347–349. ISSN: 00143820. DOI: 10.2307/2405656.

Barrett, S. and Husband, B. (1990). Variation in outcrossing rate in *Eichhornia paniculata* : the role of demographic and reproductive factors. *Plant Species Biology* 5, 41–56.

Barrett, S. C. H. (2002). The evolution of plant sexual diversity. eng. *Nature Reviews. Genetics* 3(4), 274–284. ISSN: 1471-0056. DOI: 10.1038/nrg776.

Baudry, E., Kerdelhué, C., Innan, H., and Stephan, W. (2001). Species and recombination effects on DNA variability in the tomato genus. eng. *Genetics* 158(4), 1725–1735. ISSN: 0016-6731.

Bestor, T. H. (1999). Sex brings transposons and genomes into conflict. eng. *Genetica* 107(1-3), 289–295. ISSN: 0016-6707.

Bhat, V., Dwivedi, K. K., Khurana, J. P., and Sopory, S. K. (2005). Apomixis: an enigma with potential applications. *Current Science* 89(11), 1879–1893. ISSN: 0011-3891.

Boutin, T. S., Le Rouzic, A., and Capy, P. (2012). How does selfing affect the dynamics of selfish transposable elements? eng. *Mobile DNA* 3, 5. ISSN: 1759-8753. DOI: 10.1186/1759-8753-3-5.

Brandvain, Y., Slotte, T., Hazzouri, K. M., Wright, S. I., and Coop, G. (2013). Genomic Identification of Founding Haplotypes Reveals the History of the Selfing Species *Capsella rubella*. en. *PLoS Genetics* 9(9). Ed. by Glemin, S., e1003754. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003754.

Burt, A. and Trivers, R. (1998). Selfish DNA and breeding system in flowering plants. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265(1391), 141–146.

Busch, J. W., Joly, S., and Schoen, D. J. (2011). Demographic Signatures Accompanying the Evolution of Selfing in *Leavenworthia alabamica*. en. *Molecular Biology and Evolution* 28(5), 1717–1729. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msq352.

Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., and Hartl, D. L. (2002). The cost of inbreeding in *Arabidopsis*. eng. *Nature* 416(6880), 531–534. ISSN: 0028-0836. DOI: 10.1038/416531a.

Castric, V. and Vekemans, X. (2004). Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. eng. *Molecular Ecology* 13(10), 2873–2889. ISSN: 0962-1083. DOI: 10.1111/j.1365-294X.2004.02267.x.

Cegan, R., Vyskot, B., Kejnovsky, E., Kubat, Z., Blavet, H., Šafář, J., Doležel, J., Blavet, N., and Hobza, R. (2012). Genomic Diversity in Two Related Plant Species with and without Sex Chromosomes - *Silene latifolia* and *S. vulgaris*. en. *PLoS ONE* 7(2). Ed. by Freitag, M., e31898. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0031898.

Charlesworth, B. (1992). Evolutionary rates in partially self-fertilizing species. eng. *The American Naturalist* 140(1), 126–148. ISSN: 0003-0147. DOI: 10.1086/285406.

Charlesworth, B. and Langley, C. H. (1989). The population genetics of Drosophila transposable elements. eng. *Annual Review of Genetics* 23, 251–287. ISSN: 0066-4197. DOI: 10.1146/annurev.ge.23.120189.001343.

Charlesworth, B., Nordborg, M., and Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. eng. *Genetical Research* 70(2), 155–174.

Charlesworth, B. (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10(3), 195–205. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg2526.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4), 1289–1303.

Charlesworth, D. and Charlesworth, B. (1995). Transposable elements in inbreeding and outbreeding populations. eng. *Genetics* 140(1), 415–417. ISSN: 0016-6731.

Charlesworth, D. and Pannell, J. (2001). Mating systems and population genetic structure in the light of coalescent theory. en. In: *Integrating Ecology and Evolution in a Spatial Context: 14th Special Symposium of the British Ecological Society*. Jonathan Silvertown, Janis Antonovics. Cambridge University Press, 73–96. ISBN: 9780521549332.

Charlesworth, D. and Yang, Z. (1998). Allozyme diversity in Leavenworthia populations with different inbreeding levels. eng. *Heredity* 81 (Pt 4), 453–461. ISSN: 0018-067X.

Charlesworth, D. (2003). Effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358(1434), 1051–1070.

Charlesworth, D. and Wright, S. I. (2001). Breeding systems and genome evolution. *Current opinion in genetics & development* 11(6), 685–690.

Chaux, N. de la, Tsuchimatsu, T., Shimizu, K. K., and Wagner, A. (2012). The predominantly selfing plant Arabidopsis thaliana experienced a recent reduction in transposable element abundance compared to its outcrossing relative Arabidopsis lyrata. eng. *Mobile DNA* 3(1), 2. ISSN: 1759-8753. DOI: 10.1186/1759-8753-3-2.

Chiang, Y.-C., Schaal, B. A., Chou, C.-H., Huang, S., and Chiang, T.-Y. (2003). Contrasting selection modes at the Adh1 locus in outcrossing Miscanthus sinensis vs. inbreeding Miscanthus condensatus (Poaceae). eng. *American Journal of Botany* 90(4), 561–570. ISSN: 0002-9122. DOI: 10.3732/ajb.90.4.561.

Cosendai, A.-C., Wagner, J., Ladinig, U., Rosche, C., and Hörandl, E. (2013). Geographical parthenogenesis and population genetic structure in the alpine species Ranunculus kuepferi (Ranunculaceae). eng. *Heredity* 110(6), 560–569. ISSN: 1365-2540. DOI: 10.1038/hdy.2013.1.

Crawford, D. J., Ruiz, E., Stuessy, T. F., Tepe, E., Aqeveque, P., Gonzalez, F., Jensen, R. J., Anderson, G. J., Bernardello, G., Baeza, C. M., Swenson, U., and Silva O, M. (2001). Allozyme diversity in endemic flowering plant species of the Juan Fernandez Archipelago, Chile: ecological and historical factors with implications for conservation. eng. *American Journal of Botany* 88(12), 2195–2203. ISSN: 0002-9122.

Docking, T. R., Saadé, F. E., Elliott, M. C., and Schoen, D. J. (2006). Retrotransposon sequence variation in four asexual plant species. eng. *Journal of Molecular Evolution* 62(4), 375–387. ISSN: 0022-2844. DOI: 10.1007/s00239-004-0350-y.

Dvorák, J., Luo, M. C., and Yang, Z. L. (1998). Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. eng. *Genetics* 148(1), 423–434. ISSN: 0016-6731.

Eckert, C. and Barrett, S. (1993). Clonal reproduction and patterns of genotypic diversity in *Decodon verticillatus* (Lythraceae). *American Jourrnal of Botany* 80, 1175–1182.

Ellstrand, N. C. and Roose, M. L. (1987). Patterns of Genotypic Diversity in Clonal Plant Species. *American Journal of Botany* 74(1), 123. ISSN: 00029122. DOI: 10 . 2307 / 2444338.

Escobar, J. S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J., and Glémin, S. (2010). An integrative test of the dead-end hypothesis of selfing evolution in Triticeae(Poaceae). eng. *Evolution; International Journal of Organic Evolution* 64(10), 2855–2872. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2010.01045.x.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist* 125(1), 1–15.

Foxe, J. P., Dar, V.-u.-N., Zheng, H., Nordborg, M., Gaut, B. S., and Wright, S. I. (2008). Selection on Amino Acid Substitutions in *Arabidopsis*. en. *Molecular Biology and Evolution* 25(7), 1375–1383. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msn079.

Foxe, J. P., Slotte, T., Stahl, E. A., Neuffer, B., Hurka, H., and Wright, S. I. (2009). Recent speciation associated with the evolution of selfing in *Capsella*. *Proceedings of the National Academy of Sciences* 106(13), 5241–5245.

Glémin, S. (2007). Mating Systems and the Efficacy of Selection at the Molecular Level. en. *Genetics* 177(2), 905–916. ISSN: 0016-6731. DOI: 10.1534/genetics.107.073601.

Glémin, S., Bazin, E., and Charlesworth, D. (2006). Impact of mating systems on patterns of sequence polymorphism in flowering plants. en. *Proceedings of the Royal Society B: Biological Sciences* 273(1604), 3011–3019. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2006.3657.

Glémin, S. and Muyle, A. (2014). Mating systems and selection efficacy: a test using chloroplastic sequence data in Angiosperms. en. *Journal of Evolutionary Biology* 27(7), 1386–1399. ISSN: 1010061X. DOI: 10.1111/jeb.12356.

Glémin, S., Clément, Y., David, J., and Ressayre, A. (2014). GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. en. *Trends in Genetics* 30(7), 263–270. ISSN: 01689525. DOI: 10.1016/j.tig.2014.05.002.

Glémin, S. and Galtier, N. (2012). Genome Evolution in Outcrossing Versus Selfing Versus Asexual Species. In: *Evolutionary Genomics*. Ed. by Anisimova, M. Vol. 855. Totowa, NJ: Humana Press, 311–335. ISBN: 978-1-61779-581-7, 978-1-61779-582-4.

Glémin, S. and Ronfort, J. (2013). Adaptation and maladaptation in selfing and outcrossing species: new mutations versus standing variation. eng. *Evolution; International Journal of Organic Evolution* 67(1), 225–240. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2012.01778.x.

Gordo, I. and Charlesworth, B. (2001). Genetic linkage and molecular evolution. eng. *Current biology: CB* 11(17), R684–686. ISSN: 0960-9822.

Gschwend, A. R., Yu, Q., Tong, E. J., Zeng, F., Han, J., VanBuren, R., Aryal, R., Charlesworth, D., Moore, P. H., Paterson, A. H., and Ming, R. (2012). Rapid divergence and expansion of the X chromosome in papaya. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), 13716–13721. ISSN: 1091-6490. DOI: 10.1073/pnas.1121096109.

Guo, Y.-L., Bechsgaard, J. S., Slotte, T., Neuffer, B., Lascoux, M., Weigel, D., and Schierup, M. H. (2009). Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. *Proceedings of the National Academy of Sciences* 106(13), 5246–5251.

Haag, C. R. and Roze, D. (2007). Genetic Load in Sexual and Asexual Diploids: Segregation, Dominance and Genetic Drift. en. *Genetics* 176(3), 1663–1678. ISSN: 0016-6731. DOI: 10.1534/genetics.107.073080.

Hamrick, J. L. and Godt, M. J. W. (1990). Allozyme diversity in plant species. English. In: *Plant population genetics, breeding, and genetic resources*. Brown, A. H. D.;Clegg, M. T.;Kahler, A. L.;Weir, B. S. Sinauer, Sunderland, MA., 43–63.

Hamrick, J. L. and Godt, M. J. W. (1996). Effects of Life History Traits on Genetic Diversity in Plant Species. *Royal Society of London Philosophical Transactions Series B* 351(1345), 1291–1298. ISSN: 0962-8436. DOI: 10.1098/rstb.1996.0112.

Haudry, A., Cenci, A., Guilhaumon, C., Paux, E., Poirier, S., Santoni, S., David, J., and GléMin, S. (2008). Mating system and recombination affect molecular evolution in four Triticeae species. en. *Genetics Research* 90(01). ISSN: 0016-6723, 1469-5073. DOI: 10.1017/S0016672307009032.

Hazzouri, K. M., Escobar, J. S., Ness, R. W., Killian Newman, L., Randle, A. M., Kalisz, S., and Wright, S. I. (2013). Comparative population genomics in *Collinsia* sister species reveals evidence for reduced effective population size, relaxed selection, and evolution of biased gene conversion with an ongoing mating system shift. eng. *Evolution; International Journal of Organic Evolution* 67(5), 1263–1278. ISSN: 1558-5646. DOI: 10.1111/evo.12027.

Heilbuth, J. C. (2000). Lower species richness in dioecious clades. *The American Naturalist* 156(3), 221–241.

Heilbuth, J. C., Ilves, K. L., and Otto, S. P. (2001). The consequences of dioecy for seed dispersal: modeling the seed-shadow handicap. *Evolution* 55(5), 880–888.

Heller, R. and Maynard Smith, J. (1978). Does Muller's ratchet work with selfing? *Genetics Research* 32(03), 289–293. ISSN: 1469-5073. DOI: 10.1017/S0016672300018784.

Hersch-Green, E. I., Myburg, H., and Johnson, M. T. J. (2012). Adaptive molecular evolution of a defence gene in sexual but not functionally asexual evening primroses: Plant sex and molecular evolution in defence. en. *Journal of Evolutionary Biology* 25(8), 1576–1586. ISSN: 1010061X. DOI: 10.1111/j.1420-9101.2012.02542.x.

Hickey, D. A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101(3-4), 519–531.

Hojsgaard, D., Klatt, S., Baier, R., Carman, J. G., and Hörandl, E. (2014). Taxonomy and Biogeography of Apomixis in Angiosperms and Associated Biodiversity Characteristics. *Critical Reviews in Plant Sciences* 33(5), 414–427. ISSN: 0735-2689. DOI: 10.1080/07352689.2014.898488.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F. X., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y.-L. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* 43(5), 476–481. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.807.

Igic, B. and Kohn, J. R. (2006). The distribution of plant mating systems: study bias against obligately outcrossing species. eng. *Evolution; International Journal of Organic Evolution* 60(5), 1098–1103. ISSN: 0014-3820.

Ingvarsson, P. K. (2002). A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. eng. *Evolution; International Journal of Organic Evolution* 56(12), 2368–2373. ISSN: 0014-3820.

James, E. A. and McDougall, K. L. (2014). Spatial genetic structure reflects extensive clonality, low genotypic diversity and habitat fragmentation in *Grevillea renwickiana* (Proteaceae), a rare, sterile shrub from south-eastern Australia. en. *Annals of Botany* 114(2), 413–423. ISSN: 0305-7364, 1095-8290. DOI: 10.1093/aob/mcu049.

Käfer, J., Boer, H. J. de, Mousset, S., Kool, A., Dufay, M., and Marais, G. A. B. (2014). Dioecy is associated with higher diversification rates in flowering plants. en. *Journal of Evolutionary Biology* 27(7), 1478–1490. ISSN: 1010061X. DOI: 10.1111/jeb.12385.

Käfer, J., Talianová, M., Bigot, T., Michu, E., Guéguen, L., Widmer, A., žlůvová, J., Glémin, S., and Marais, G. A. B. (2013). Patterns of molecular evolution in dioecious and non-dioecious *Silene*. en. *Journal of Evolutionary Biology* 26(2), 335–346. ISSN: 1010061X. DOI: 10.1111/jeb.12052.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* 39(9), 1151–1155. ISSN: 1061-4036. DOI: 10.1038/ng2115.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. eng. *Genetics* 47, 713–719. ISSN: 0016-6731.

Koltunow, A. M. and Grossniklaus, U. (2003). Apomixis: a developmental perspective. eng. *Annual Review of Plant Biology* 54, 547–574. ISSN: 1543-5008. DOI: 10.1146/annurev.arplant.54.110901.160842.

Lambertini, C., Riis, T., Olesen, B., Clayton, J. S., Sorrell, B. K., and Brix, H. (2010). Genetic diversity in three invasive clonal aquatic species in New Zealand. *BMC genetics* 11(1), 52.

Lande, R. (1985). The fixation of chromosomal rearrangements in a subdivided population with local extinction and colonization. eng. *Heredity* 54 (Pt 3), 323–332. ISSN: 0018-067X.

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., and Duret, L. (2014). The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. eng. *PLoS genetics* 10(11), e1004790. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004790.

Lesecque, Y., Mouchiroud, D., and Duret, L. (2013). GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. eng. *Molecular Biology and Evolution* 30(6), 1409–1419. ISSN: 1537-1719. DOI: 10.1093/molbev/mst056.

Liu, F., Charlesworth, D., and Kreitman, M. (1999). The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus Leavenworthia. eng. *Genetics* 151(1), 343–357. ISSN: 0016-6731.

Liu, F., Zhang, L., and Charlesworth, D. (1998). Genetic diversity in Leavenworthia populations with different inbreeding levels. eng. *Proceedings. Biological Sciences / The Royal Society* 265(1393), 293–301. ISSN: 0962-8452. DOI: 10.1098/rspb.1998.0295.

Lockton, S. and Gaut, B. S. (2010). The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. *BMC evolutionary biology* 10(1), 10.

Lovell, J. T., Grogan, K., Sharbel, T. F., and McKay, J. K. (2014). Mating system and environmental variation drive patterns of adaptation in *Boechera spatifolia* (Brassicaceae). en. *Molecular Ecology* 23(18), 4486–4497. ISSN: 09621083. DOI: 10.1111/mec.12879.

Lynch, M. (2011). Statistical inference on the mechanisms of genome evolution. eng. *PLoS genetics* 7(6), e1001389. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001389.

Marais, G., Charlesworth, B., and Wright, S. I. (2004). Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. *Genome biology* 5(7), R45.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. en. *Trends in Genetics* 19(6), 330–338. ISSN: 01689525. DOI: 10.1016/S0168-9525(03)00116-1.

Matzk, F., Hammer, K., and Schubert, I. (2003). Coevolution of apomixis and genome size within the genus *Hypericum*. *Sexual Plant Reproduction* 16(2), 51–58. ISSN: 0934-0882, 1432-2145. DOI: 10.1007/s00497-003-0174-8.

Maumus, F. and Quesneville, H. (2014). Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. eng. *Nature Communications* 5, 4104. ISSN: 2041-1723. DOI: 10.1038/ncomms5104.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* 23, 23–35. ISSN: 1469-5073. DOI: 10.1017/S0016672308009579.

Miller, J. S. (2000). Polyploidy and the Evolution of Gender Dimorphism in Plants. *Science* 289(5488), 2335–2338. ISSN: 00368075, 10959203. DOI: 10.1126/science.289.5488.2335.

Ming, R., Bendahmane, A., and Renner, S. S. (2011). Sex Chromosomes in Land Plants. en. *Annual Review of Plant Biology* 62(1), 485–514. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-042110-103914.

Miyashita, N. T., Kawabe, A., and Innan, H. (1999). DNA variation in the wild plant Arabidopsis thaliana revealed by amplified fragment length polymorphism analysis. eng. *Genetics* 152(4), 1723–1731. ISSN: 0016-6731.

Montgomery, E. A., Huang, S. M., Langley, C. H., and Judd, B. H. (1991). Chromosome rearrangement by ectopic recombination in Drosophila melanogaster: genome structure and evolution. eng. *Genetics* 129(4), 1085–1098. ISSN: 0016-6731.

Morgan, M. T. (2001). Transposable element number in mixed mating populations. eng. *Genetical Research* 77(3), 261–275.

Muller, H. J. (1964). The relation of recombination to mutational advance. eng. *Mutation Research* 106, 2–9. ISSN: 0027-5107.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glémin, S. (2011). GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). eng. *Molecular Biology and Evolution* 28(9), 2695–2706. ISSN: 1537-1719. DOI: 10.1093/molbev/msr104.

Neiman, M., Sharbel, T. F., and Schwander, T. (2014). Genetic causes of transitions from sexual reproduction to asexuality in plants and animals. en. *Journal of Evolutionary Biology* 27(7), 1346–1359. ISSN: 1010061X. DOI: 10.1111/jeb.12357.

Ness, R. W., Wright, S. I., and Barrett, S. C. H. (2010). Mating-System Variation, Demographic History and Patterns of Nucleotide Diversity in the Tristylous Plant *Eichhornia paniculata*. en. *Genetics* 184(2), 381–392. ISSN: 0016-6731. DOI: 10.1534/genetics.109.110130.

Ness, R. W., Siol, M., and Barrett, S. C. (2012). Genomic consequences of transitions from cross-to self-fertilization on the efficacy of selection in three independently derived selfing plants. *BMC genomics* 13(1), 611.

Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. eng. *Genetics* 154(2), 923–929. ISSN: 0016-6731.

Nordborg, M. and Donnelly, P. (1997). The coalescent process with selfing. eng. *Genetics* 146(3), 1185–1195. ISSN: 0016-6731.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., and Bergelson, J. (2005). The Pattern of Polymorphism in *Arabidopsis thaliana*. en. *PLoS Biology* 3(7), e196. ISSN: 1544-9173, 1545-7885. DOI: 10.1371/journal.pbio.0030196.

Nybom, H. (2004). Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. eng. *Molecular Ecology* 13(5), 1143–1155. ISSN: 0962-1083. DOI: 10.1111/j.1365-294X.2004.02141.x.

Pannell, J. R. and Charlesworth, B. (2000). Effects of metapopulation processes on measures of genetic diversity. eng. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 355(1404), 1851–1864. ISSN: 0962-8436. DOI: 10.1098/rstb.2000.0740.

Peck, J. R. (1994). A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. eng. *Genetics* 137(2), 597–606. ISSN: 0016-6731.

Pellino, M., Hojsgaard, D., Schmutzer, T., Scholz, U., Hörandl, E., Vogel, H., and Sharbel, T. F. (2013). Asexual genome evolution in the apomictic *Ranunculus auricomus* complex: examining the effects of hybridization and mutation accumulation. en. *Molecular Ecology* 22(23), 5908–5921. ISSN: 09621083. DOI: 10.1111/mec.12533.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. B. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. eng. *Genome Biology and Evolution* 4(7), 675–682. ISSN: 1759-6653. DOI: 10.1093/gbe/evs052.

Pettengill, J. B. and Moeller, D. A. (2012). Tempo and mode of mating system evolution between incipient *Clarkia* species. eng. *Evolution; International Journal of Organic Evolution* 66(4), 1210–1225. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2011.01521.x.

Pollak, E. (1987). On the theory of partially inbreeding finite populations. I. Partial selfing. eng. *Genetics* 117(2), 353–360. ISSN: 0016-6731.

Qiu, S., Zeng, K., Slotte, T., Wright, S., and Charlesworth, D. (2011). Reduced Efficacy of Natural Selection on Codon Usage Bias in Selfing *Arabidopsis* and *Capsella* Species. en. *Genome Biology and Evolution* 3, 868–880. ISSN: 1759-6653. DOI: 10.1093/gbe/evr085.

Renner, S. S. (2014). The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. eng. *American Journal of Botany* 101(10), 1588–1596. ISSN: 1537-2197. DOI: 10.3732/ajb.1400196.

Renner, S. S. and Ricklefs, R. E. (1995). Dioecy and Its Correlates in the Flowering Plants. *American Journal of Botany* 82(5), 596–606. ISSN: 00029122. DOI: 10.2307/2445418.

Richards, A. J. (1997). *Plant Breeding Systems*. en. Garland Science. ISBN: 9780412574504.

Rick, C. M., Fobes, J. F., and Holle, M. (1977). Genetic variation inLycopersicon pimpinellifolium: Evidence of evolutionary change in mating systems. en. *Plant Systematics and Evolution* 127(2-3), 139–170. ISSN: 0378-2697, 1615-6110. DOI: 10.1007/BF00984147.

Robertson, K., Goldberg, E. E., and Igić, B. (2011). Comparative evidence for the correlated evolution of polyploidy and self-compatibility in Solanaceae. eng. *Evolution; International Journal of Organic Evolution* 65(1), 139–155. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2010.01099.x.

Schaack, S., Gilbert, C., and Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in ecology & evolution* 25(9), 537–546. ISSN: 0169-5347. DOI: 10.1016/j.tree.2010.06.001.

Schoen, D. J. and Brown, A. H. (1991). Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proceedings of the National Academy of Sciences* 88(10), 4494–4497.

Serres-Giardi, L., Belkhir, K., David, J., and Glemin, S. (2012). Patterns and Evolution of Nucleotide Landscapes in Seed Plants. en. *The Plant Cell* 24(4), 1379–1397. ISSN: 1040-4651, 1532-298X. DOI: 10.1105/tpc.111.093674.

Slotte, T., Foxe, J. P., Hazzouri, K. M., and Wright, S. I. (2010). Genome-Wide Evidence for Efficient Positive and Purifying Selection in *Capsella grandiflora*, a Plant Species with a Large Effective Population Size. en. *Molecular Biology and Evolution* 27(8), 1813–1821. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msq062.

Slotte, T. et al. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics* 45(7), 831–835. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.2669.

Stephan, W. and Langley, C. H. (1998). DNA polymorphism in lycopersicon and crossing-over per physical length. eng. *Genetics* 150(4), 1585–1593. ISSN: 0016-6731.

Sun, M. and Wong, K. C. (2001). Genetic structure of three orchid species with contrasting breeding systems using RAPD and allozyme markers. eng. *American Journal of Botany* 88(12), 2180–2188. ISSN: 0002-9122.

Sweigart, A. L. and Willis, J. H. (2003). Patterns of nucleotide diversity in two species of *Mimulus* are affected by mating system and asymmetric introgression. eng. *Evolution; International Journal of Organic Evolution* 57(11), 2490–2506. ISSN: 0014-3820.

Tam, S. M., Causse, M., Garchery, C., Burck, H., Mhiri, C., and Grandbastien, M.-A. (2007). The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species. eng. *Journal of Evolutionary Biology* 20(3), 1056–1072. ISSN: 1010-061X. DOI: 10.1111/j.1420-9101.2007.01293.x.

Thomas, F., Lefevre, T., and Raymond, M. (2015). *Biologie évolutive*. Français. Édition : Deuxième édition. Bruxelles: DE BOECK UNIVERSITE. ISBN: 9782804101619.

Trivers, R., Burt, A., and Palestis, B. G. (2004). B chromosomes and genome size in flowering plants. eng. *Genome / National Research Council Canada = Génome / Conseil National De Recherches Canada* 47(1), 1–8. ISSN: 0831-2796. DOI: 10.1139/g03-088.

Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-i, T., Toyoda, A., Fujiyama, A., Tarutani, Y., and Kakutani, T. (2012). Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. eng. *Genes & Development* 26(7), 705–713. ISSN: 1549-5477. DOI: 10.1101/gad.183871.111.

Vamosi, J. C. and Otto, S. P. (2002). When looks can kill: the evolution of sexually dimorphic floral display and the extinction of dioecious plants. en. *Proceedings of the Royal Society B: Biological Sciences* 269(1496), 1187–1194. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2002.2004.

Wang, J., Na, J.-K., Yu, Q., Gschwend, A. R., Han, J., Zeng, F., Aryal, R., VanBuren, R., Murray, J. E., Zhang, W., Navajas-Pérez, R., Feltus, F. A., Lemke, C., Tong, E. J., Chen, C., Wai, C. M., Singh, R., Wang, M.-L., Min, X. J., Alam, M., Charlesworth, D., Moore, P. H., Jiang, J., Paterson, A. H., and Ming, R. (2012). Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), 13710–13715. ISSN: 1091-6490. DOI: 10.1073/pnas.1207833109.

Whitney, K. D., Baack, E. J., Hamrick, J. L., Godt, M. J. W., Barringer, B. C., Bennett, M. D., Eckert, C. G., Goodwillie, C., Kalisz, S., Leitch, I. J., and Ross-Ibarra, J. (2010). A role for nonadaptive processes in plant genome size evolution? eng. *Evolution; International Journal of Organic Evolution* 64(7), 2097–2109. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2010.00967.x.

Whitney, K. D., Boussau, B., Baack, E. J., and Garland, T. (2011). Drift and genome complexity revisited. eng. *PLoS genetics* 7(6), e1002092. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002092.

Wright, S. I. and Agren, J. A. (2011). Sizing up Arabidopsis genome evolution. eng. *Heredity* 107(6), 509–510. ISSN: 1365-2540. DOI: 10.1038/hdy.2011.47.

Wright, S. and Finnegan, D. (2001). Genome evolution: sex and the transposable element. eng. *Current biology: CB* 11(8), R296–299. ISSN: 0960-9822.

Wright, S. I., Iorgovan, G., Misra, S., and Mokhtari, M. (2007). Neutral Evolution of Synonymous Base Composition in the Brassicaceae. en. *Journal of Molecular Evolution* 64(1), 136–141. ISSN: 0022-2844, 1432-1432. DOI: 10.1007/s00239-005-0274-1.

Wright, S. I., Lauga, B., and Charlesworth, D. (2002). Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Molecular Biology and Evolution* 19(9), 1407–1420.

Wright, S. I., Lauga, B., and Charlesworth, D. (2003). Subdivision and haplotype structure in natural populations of Arabidopsis lyrata. eng. *Molecular Ecology* 12(5), 1247–1263. ISSN: 0962-1083.

Wright, S. I., Ness, R. W., Foxe, J. P., and Barrett, S. C. H. (2008). Genomic Consequences of Outcrossing and Selfing in Plants. en. *International Journal of Plant Sciences* 169(1), 105–118. ISSN: 1058-5893, 1537-5315. DOI: 10.1086/523366.

Wright, S. I. and Schoen, D. J. (1999). Transposon dynamics and the breeding system. In: *Transposable Elements and Genome Evolution*. Springer, 139–148.

Young, R. J., Francis, D. M., St Clair, D. A., and Taylor, B. H. (1994). A dispersed family of repetitive DNA sequences exhibits characteristics of a transposable element in the genus Lycopersicon. eng. *Genetics* 137(2), 581–588. ISSN: 0016-6731.

Chapter 2

The evolution of sex chromosomes and dosage compensation in plants and algae

This second chapter of the introductory part focuses more on the topic of my PhD – compared to the previous one that illustrated my broad interest in plant molecular evolution – and reviews the literature on plant and algal sex chromosome evolution (i.e. non-animal systems). I wrote this manuscript with the help of my supervisor, and Rylan Shearn helped with the figures. I intend to submit it to a journal as a review paper.

2.1 Introduction

Sex chromosomes are a special pair of chromosomes with genetic material that differs partially from one another, allowing them to determine the sex of individuals. They derive from an ordinary pair of autosomes, as shown by their pairing during cell division and by sequence comparison to related species.

In this review, we will focus on the evolution of sex chromosomes in non-animal species. Indeed, most of our knowledge on sex chromosome evolution comes from animal models and it is interesting to take a broader phylogenic point of view to test if it can be generalised to other eukaryotes. Outside animals, sex chromosomes have mainly been studied in plants, and to a lesser extent in green and brown algae. Fungi have mating type loci that are sometimes wrongfully called sex chromosomes because of the features they share (see for example Fontanillas et al., 2015). However, there is an absence of mating type dimorphism, unlike sexual dimorphism where males and females typically produce gametes of different sizes and in different numbers (Charlesworth, 2013). Consequently, this review will mainly focus on plant sex chromosomes and will mention data on algae whenever available.

In female heterogametic systems, females are ZW and males ZZ, as in willow (*Salix suchowensis*). In male heterogametic systems, males are XY and females XX, as in white campion (*Silene latifolia*). In species with an independent haploid phase in their life cycle, sex can be determined in the haploid gametophytes by a UV system (females U, males V) and the diploid sporophyte is then heterogametic (UV), as in *Marchantia polymorpha* (Figure 2.1).

For sex chromosomes to evolve, a species must have separate sexes (dioecy). This requirement is fulfilled in 75% of liverworts (6000 species), 50% of leafy mosses (~7250 species), 36% of gymnosperms (381 species) and 5–6% of angiosperms (15 600 species), providing a high number of plant species possibly carrying sex chromosomes (Ming et al., 2011; Renner, 2014). Data on percentage of dioecy is currently missing in brown and green algae. In angiosperms, dioecy evolved independently from the ancestral hermaphrodite state somewhere between 871 to 5000 times (Renner, 2014), see Box 1.3 for possible evolutionary pathways. In angiosperms, ZW systems are less frequent than XY systems because transitions to gynodioecy are more frequent than transitions to androdioecy and evolution of dioecy through gynodioecy is expected to lead more easily to an XY system (Ming et al., 2011).

We will summarise here how the study of plant and algal sex chromosomes contributed to our current understanding of sex chromosome evolution, ask what their particularities are and suggest further research directions.

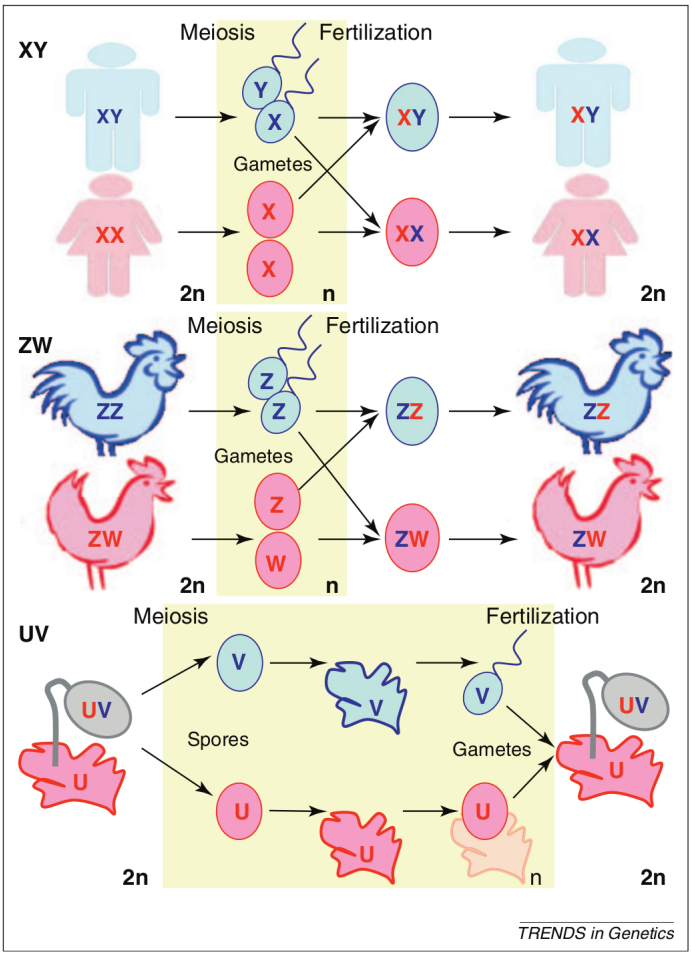


Figure 2.1: Sex chromosome systems: (from Bachtrog et al., 2011) female heterogamety (ZW), male heterogamety (XY) and haplo-diploid system (UV).

2.2 What are the sex determining genes in plants?

Once sex chromosomes have been identified with cytology (if they differ in size) or with genetic markers associated with sex, it is possible to test how they determine sex thanks to mutants. For instance XXY males indicate a dominant Y chromosome that determines sex, whereas XXY females rather show an X to autosome dose sex determining system, where the number of X chromosomes relative to autosomes matters (Table 2.1).
Theory predicts that two closely linked sex determining genes are necessary for the birth of sex chromosomes (Charlesworth and Charlesworth, 1978, and see Figure 2.2): a male sterile mutation (recessive in X/Y systems, dominant in Z/W systems) and a female sterile mutation (dominant in X/Y systems, recessive in Z/W systems). In UV systems the dominance of mutations does not matter as sex is expressed in the haploid phase.

An illustration of this theoretical expectation was provided by the plant *Fragaria virginiana*. In this subdioecious species males and females coexist with hermaphrodites. The male function locus and the female function locus were mapped to the same linkage group, separated by ~6cM and recombination between the two loci was shown to lead to hermaphrodites and neuters (both male and female sterile individuals) in cross pro-

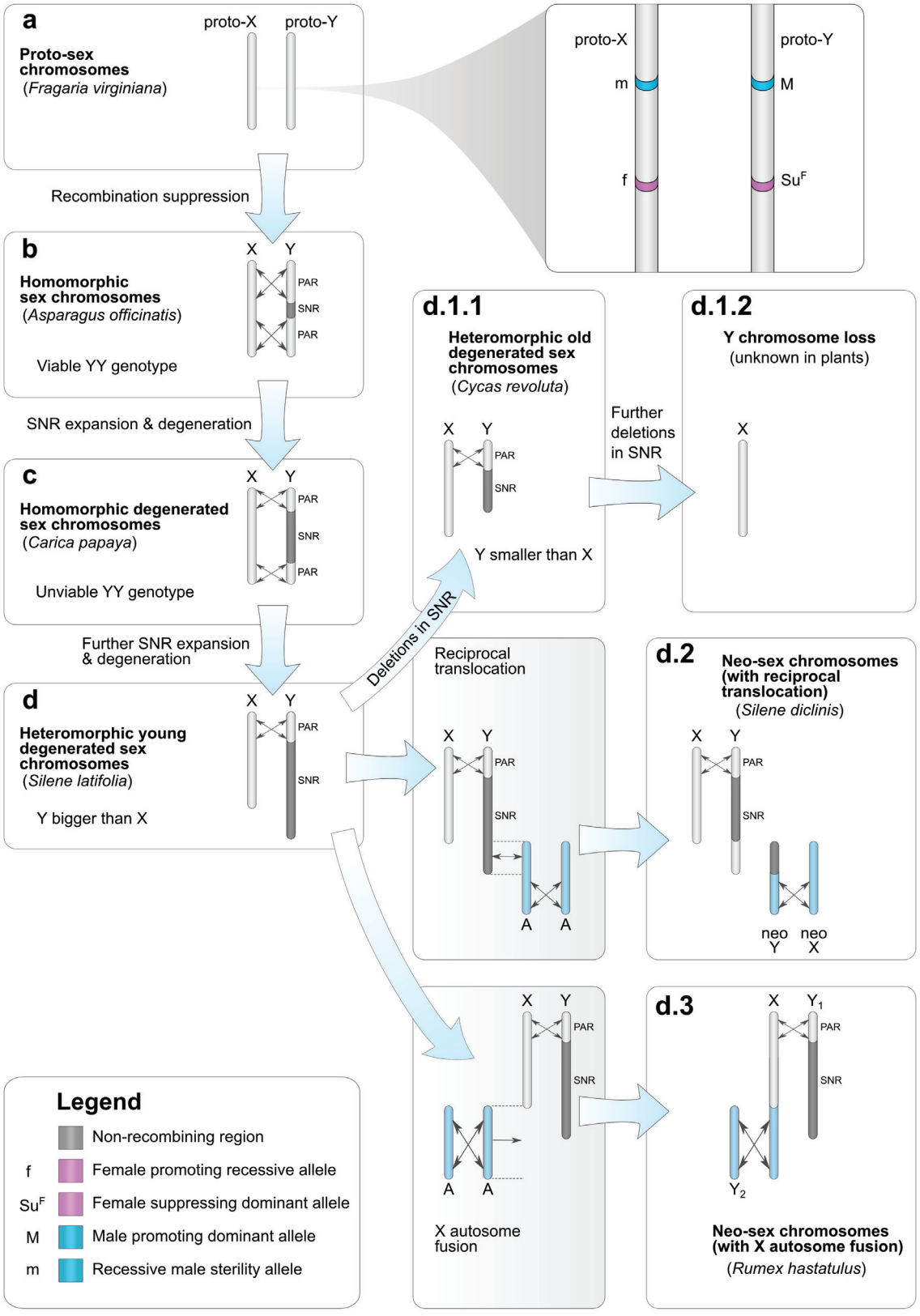


Figure 2.2: Possible stages in XY sex chromosome evolution: note that not all steps are obligatory and that steps are not necessarily correlated with the age of the system. In **a**, the YY genotype is viable and only sex determining genes differ (as shown on the zoom). Neo-sex chromosomes can evolve (as in *Silene diclinis*, **d.2**, Howell et al., 2009 and *Rumex hastatulus*, **d.3**, Smith, 1964). The Y chromosome can become smaller than the X chromosome (as in *Cycas revoluta*, **d.1.1**, Segawa et al., 1971).

genies (Spigler et al., 2008). This represents the first stage of sex chromosome evolution and is called proto-sex chromosomes (Figure 2.2). In *Fragaria virginiana*, because the mutation causing female sterility is dominant, the system consists of proto-ZW sex chromosomes.

In papaya, hermaphrodites are determined by a Y^h specific region that recently evolved (~ 4000y, VanBuren et al., 2015) from the much older Y specific region (~9My), possibly after loss of the female sterility dominant mutation, but further genetic analyses will be necessary to confirm the two gene model in this species.

So far, the only known sex determining genes located on sex chromosomes in plants are in the crop species *Diospyros lotus* (persimmon). The authors identified *OGI*, a small RNA encoded on the Y specific region with male specific expression, that downregulates the expression of the *MeGi* gene which represses male function and promotes female function (Akagi et al., 2014). The *MeGi* gene is autosomal and the interplay between *MeGi* and *OGI* does not fit well with the theoretical two gene model as *OGI* acts both as male promoting and female suppressing.

In *Asparagus officinalis*, pollen development aborts late in females: at the microspore maturation stage. Harkess et al. (2015) showed that the *AMS* (Aborted Microspores) gene has male-biased expression, making it a good candidate for the male sterility gene as it is known to be involved in microspore maturation in *A. thaliana*. But further analyses are required to check whether this gene is located on the Y chromosome of *Asparagus officinalis*.

Other sex-determining genes have been identified in the monoecious species melon (reviewed in Ming et al., 2011), but it remains unknown whether these genes have also been recruited as sex-determining genes in closely related dioecious species with sex chromosomes.

In green algae, the gene *MID* is present only in the minus haplotype and governs gametic differentiation (Ferris and Goodenough, 1997).

2.3 How do sex chromosomes become differentiated?

Hermaphrodites and neuters that can be observed in *Fragaria virginiana* are due to recombination events between the two sex determining loci of the proto-Z and the proto-W. Neuters are selected against as they cannot reproduce, and hermaphrodites are expected to be selected against if they have a lower fitness compared to males and females. This would select for recombination suppression between the two sex determining loci and transform the proto-sex chromosomes into actual sex chromosomes. As a consequence, the male specific region of the Y chromosome stops recombining, it becomes a sex-specific non-recombining region (hereafter SNR). The homologous region on the X still recombines in females and the regions surrounding the SNR are called pseudo-autosomal regions (PAR), and recombine in both males and females (Figure 2.2). Sim-

ilarly in ZW systems the female specific region on the W does not recombine. In UV systems both male and female specific regions do not recombine as the sporophyte is always heterogametic. An example of this early stage of sex chromosome evolution after recombination suppression between the two sex determining genes is provided by the plant *Asparagus officinalis* (Figure 2.2).

Box 2.1 : Sex chromosome sequence data

The amount of repetitive sequences present in SNRs (see Section 2.5) makes their sequencing virtually impossible with whole shot-gun approaches, especially when using short read NGS technologies. The few fully sequenced SNRs in plants are rather small regions (Table 2.1), that were obtained through sequencing of BAC clones organised in a physical map, a strategy that is far too costly to be applied to large SNRs.
Producing high-quality assembly is not always necessary and alternative, less expensive strategies have been recently developed for identifying sex chromosome sequences based on NGS data. A first category of approaches relies on the comparison of one male and one female genome. Identifying X-linked scaffolds can be done by studying the genomic male over female read coverage ratio along the genome: autosomal contigs will have a ratio of 1 while X-linked ones will have a ratio of 0.5 (Figure 2.3A, Vicoso and Bachtrog, 2011; Vicoso et al., 2013a,b). The Y scaffolds are simply those that are exclusively present in the male genome. A more sophisticated analysis can be done by a prior exclusion of the repeats shared by the Y and the female genome (Figure 2.3B, Carvalho and Clark, 2013). Also, a combination of RNA-seq and genome data of male and female individuals has been used to increase the number of known Y-linked genes in well-studied systems (Figure 2.3D, Cortez et al., 2014). A method relying on DNA-seq data on many individuals sampled from different populations exists (Figure 2.3C, Gautier, 2014), but is only adapted to old and diverged sex chromosomes. A second category of approaches relies on studying how markers segregate among sexes in genetic maps. Using these markers, scaffolds from sex chromosomes can be ascertained (Figure 2.3F, Al-Dous et al., 2011; Hou et al., 2015).
These approaches, however, are suitable only if reasonably well assembled reference genomes are available, in the studied species or in a close relative, possibly explaining why they have not often been applied to plants nor algae. If such genomic resources are lacking, as in many species with large genomes, complexity and size can be reduced by using transcriptomes instead of complete genomes. A new strategy relying on the use of a cross (parents and their progeny of each sex) sequenced by RNA-seq is promising as it is cheap and gives direct access to sex-linked gene (genes located on SNRs and their recombining homologs) and their expression levels. This approach was applied to *Silene latifolia* and *Rumex hastatulus* (Table 2.1). More recently, a model-based method was developped for this strategy (Figure 2.3E, Chapter 4).

However, unlike *Asparagus officinalis*, many species with sex chromosomes have much larger SNRs than the expected small region containing the two closely linked sex determining genes, which suggests that sex chromosomes undergo further recombination suppression events. This is confirmed at the sequence level whenever sequences are available (Box 2.1) by the study of X-Y, Z-W or U-V divergence. Indeed, after recombination suppression, SNRs accumulate substitutions separately and, using the molecular clock, it is possible to estimate the time when recombination stopped. Such analyses

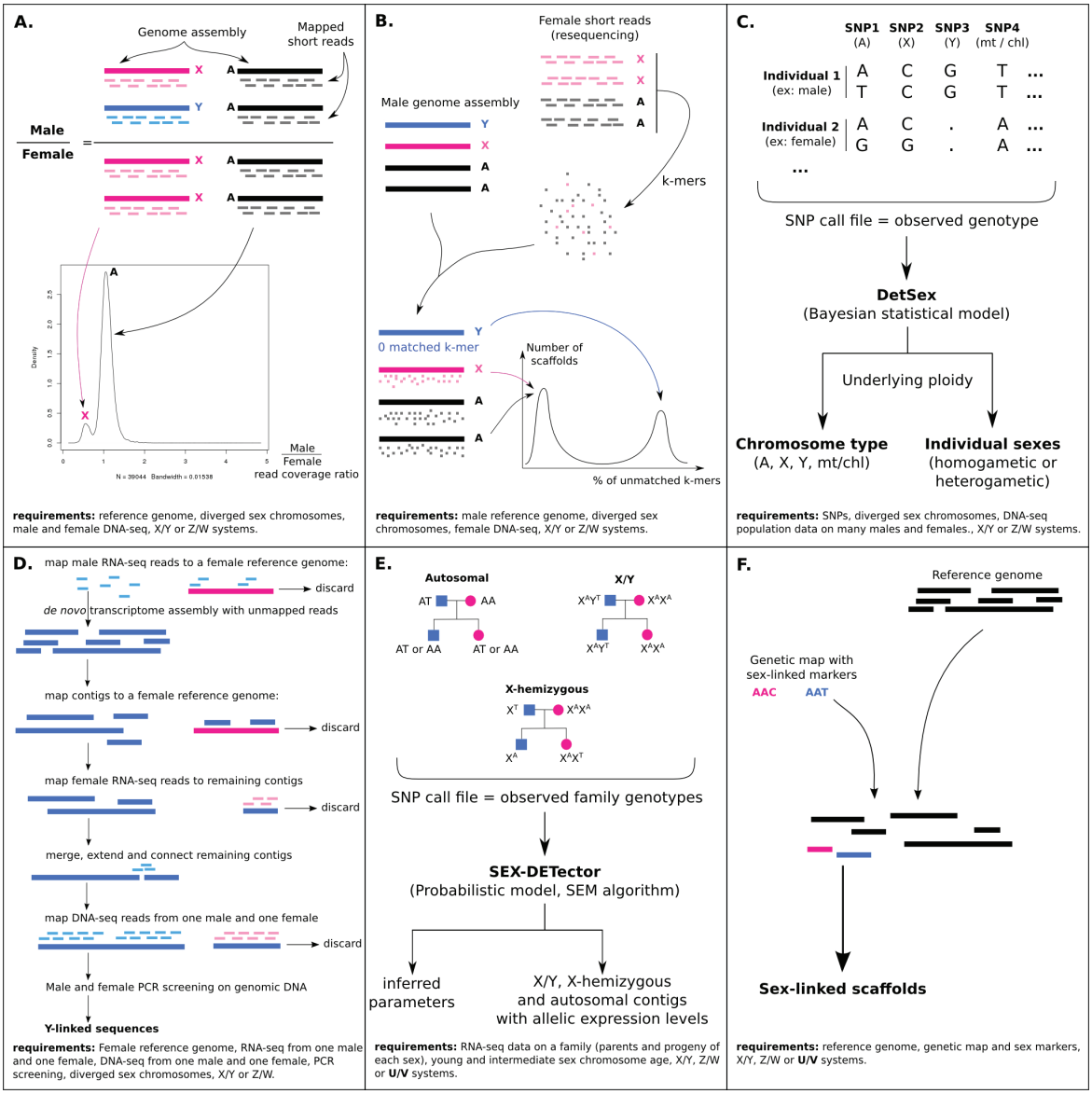


Figure 2.3: Available methods for sequencing sex chromosomes: see Box 2.1 for more details. **A** Vicoso and Bachtrog (2011) and Vicoso et al. (2013a,b). **B** Carvalho and Clark (2013). **C** Gautier (2014). **D** Cortez et al. (2014). **E** Chapter 4. **F** Al-Dous et al. (2011) and Hou et al. (2015).

showed that, in *Silene latifolia*, there has been at least two recombination suppression events (Bergero et al., 2007, 2013) as genes fall into two categories of X-Y divergence. Sex-linked genes with similar X-Y, Z-W or U-V divergence levels are located in a same genomic region, called stratum, where recombination between the sex chromosomes stopped around the same time, probably during a single evolutionary event. Recombination suppression seems to be associated with chromosomal rearrangements, in particular inversions. In papaya two large inversion events appear to define the two strata present in this species (Wang et al., 2012, see Table 2.1 for information on known strata in plant species).

But what is the evolutionary force driving these further events of recombination suppression on sex chromosomes? Theory predicts that sex antagonistic mutations (advantageous for one sex and deleterious for the other) will be selected to be linked to the SNR,

this way providing advantage to one sex without damaging the other (Charlesworth and Charlesworth, 1980; Jordan and Charlesworth, 2012; Rice, 1984, 1987). This would lead to addition of genetic material on the SNR, either through spread onto the flanking PAR (if the PAR contains sex antagonistic genes, see Figure 2.4), or through chromosomal rearrangements with autosomes (if sex antagonistic genes are located on autosomes), generating neo-sex chromosomes (Figure 2.2).

However, no firm evidence definitively connects sex antagonistic alleles to the evolution of reduced recombination of Y, W, U or V sex chromosomes. Qiu et al. (2013) observed a global high genetic diversity in *Silene latifolia* PAR (an evidence of balancing selection maintaining alleles polymorphic for a long evolutionary time), but observed evidence of linkage disequilibrium between the PAR and the SNR only for one gene. The high diversity could be due to sexually antagonistic polymorphism maintained through partial linkage with the SNR, but it is not clear yet whether a neutral model of partial linkage would be sufficient to explain the high diversity level.

Chromosomal rearrangements involving sex chromosomes and autosomes are common and have been observed for example in *S. latifolia* where autosomal regions have been translocated to the sex chromosomes and are now non-recombining (Bergero et al., 2013). Another example comes from the plant *Silene diclinis*, where a reciprocal translocation between an autosome and the ancestral Y chromosome led to two Xs and two Ys with a chain quadrivalent at meiosis metaphase I (Howell et al., 2009, and see Figure 2.2). In *Rumex hastatulus* some individuals have a XY1Y2 system after an X-autosome fusion (Smith, 1964, and see Figure 2.2). These chromosomal rearrangements make it possible for genes that were autosomal to become sex-linked. If these events are selected for because of sex antagonism, they should be more frequent than rearrangements that do not involve sex chromosomes. Data to test this hypothesis is currently missing.

A comparison of the two green algae *Chlamydomonas reinhardtii* and *Volvox carteri* suggests that the UV sex chromosomes of *Volvox carteri* derive from a mating type loci homologous to the one of *Chlamydomonas reinhardtii*, after the addition of genetic material to the SNRs (Ferris et al., 2010). The newly acquired genes of the SNRs show sex-specific patterns of expression and could be linked to the strong differentiation in gamete size (oogamy) observed in *Volvox carteri*, as compared to *Chlamydomonas reinhardtii* where male and female gametes have similar sizes (isogamy). This supports the link between sexual dimorphism and recombination suppression, possibly through sexually antagonistic selection. However a more recent study showed that the sex determining gene *MID* alone is able to change the sex of *Volvox carteri* transgenic individuals (Geng et al., 2014), suggesting sex dimorphism evolved through changes that are intrinsic to the *MID* gene product, rather than from the recruitment of other genes to U and V SNRs.

In *Ectocarpus sp.* the U and V SNRs are surprisingly small (under 1Mb) given the old age of the system (>70My), which could be related to the low level of sexual dimorphism in this species and therefore the low level of sexually antagonistic selection (Ahmed et al., 2014).

Indeed male and female gametes have different behaviour (females do not swim and attract swimming males with pheromones, a process called physiological anisogamy), but differ very little in size (Lipinska et al., 2015).
Recombination suppression leads to the divergence of sex chromosomes and sometimes the formation of strata, but also causes the sex specific regions to degenerate (Section 2.5). Indeed theory predicts that the efficacy of selection will be reduced in nonrecombining regions, because of selective interference among genetically linked loci (called Hill-Robertson interferences, reviewed in Bachtrog, 2006, see Box 2.2).

2.4 What are the forces opposing sex chromosome differentiation?

Theories have been proposed to explain that the SNRs do not necessarily expand and PARs can remain very large in some organisms. In animals, recombination patterns depend on phenotypic, rather than genotypic sex. Therefore SNRs are expected to recombine in sex-reversed XY females (or sex-reversed ZW males) if the SNR is not too diverged, a mechanism called the fountain of youth (Perrin, 2009), as it prevents SNRs degeneration. A similar process could occur in plants where the expression of dioecy can be labile: inconstant ZW females can be bisexual on some years in *Populus trichocarpa* (Stettler, 1971), possibly allowing for Z-W recombination in pollen grains and explaining that sex chromosomes are homomorphic in this species. But whether recombination patterns depend on phenotypic sex rather than genotypic sex (as in animals) still needs to be tested in plants.
Sex chromosome turnover can also prevent the accumulation of recombination suppression events. Theory predicts that a sex chromosome pair can be replaced by a new pair through two mechanisms (Blaser et al., 2014). The first mechanism (Doorn and Kirkpatrick, 2007; Doorn and Kirkpatrick, 2010) involves sex-antagonistic selection: a male-benefiting, female-deleterious mutation appearing on an autosome selects for a masculinising mutation in its vicinity (and conversely for a female-benefiting mutation), possibly overtaking the previous masculinising mutation on the old Y chromosome and leading to the evolution of a new sex chromosome pair (Figure 2.4). In this first mechanism, transitions can change the system (from XY to ZW and conversely). The second mechanism (Blaser et al., 2013) involves the accumulation of deleterious loss-of-function mutations in SNRs that is expected to lower survival in the heterogametic sex. Transitions should occur as soon as this survival cost exceeds the benefits of having sex-antagonistic mutations linked to the sex determining genes. In this second mechanism, transitions cannot change the heterogametic sex as transitions from XY to ZW are expected to fix the Y homologue as an autosomal pair, which is of course detrimental if transitions are precisely triggered by the mutational load accumulating on these Y chromosomes.
A sex chromosome turnover event has been documented in *Silene otites* from XY to ZW

(Slancarova et al., 2013). The whole Salicaceae family is dioecious, which suggests that the ancestor was dioecious and that Salicaceae species could carry old sex chromosomes (~45My, Manchester et al., 2006). However, sex mapped to chromosome XV in both *Salix viminalis* (Pucholt et al., 2015) and *Salix suchowensis* (Hou et al., 2015) with a ZW system in both cases, but, in *Populus*, sex mapped to a different pair of chromosome (different positions on chromosome XIX, depending on the species, reviewed in Tuskan et al., 2012), with some species having an XY system (*Populus nigra*, *Populus tremuloides*, *Populus tremula*) and others a ZW system (*Populus alba*, *Populus deltoides*), in *Populus trichocarpa* contradicting papers reported a ZW (Yin et al., 2008) and an XY system (Geraldes et al., 2015). These observations suggest that there has been high rates of sex chromosome turnover in the Salicaceae family, but the mechanisms that caused these turnovers are yet unknown.

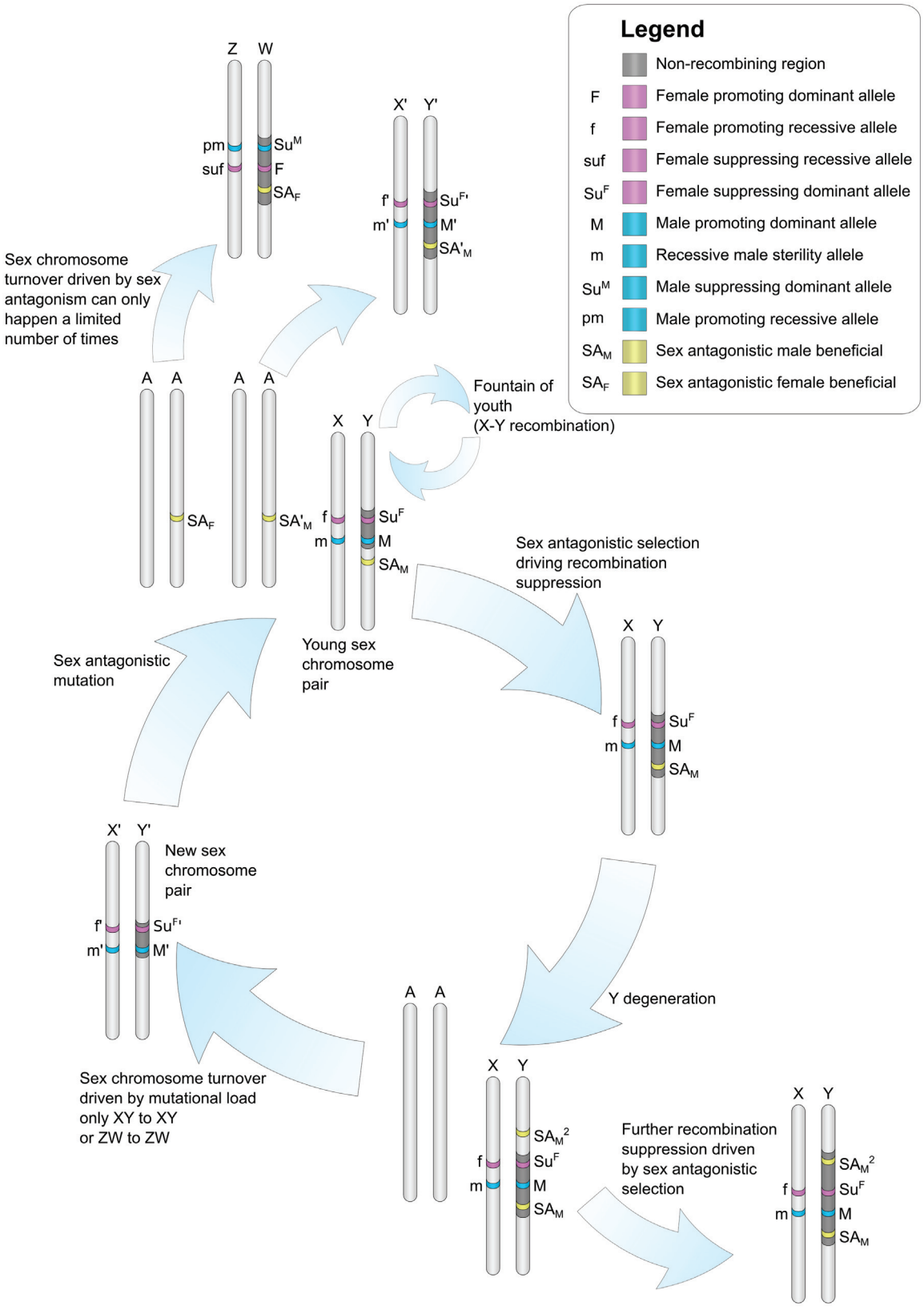


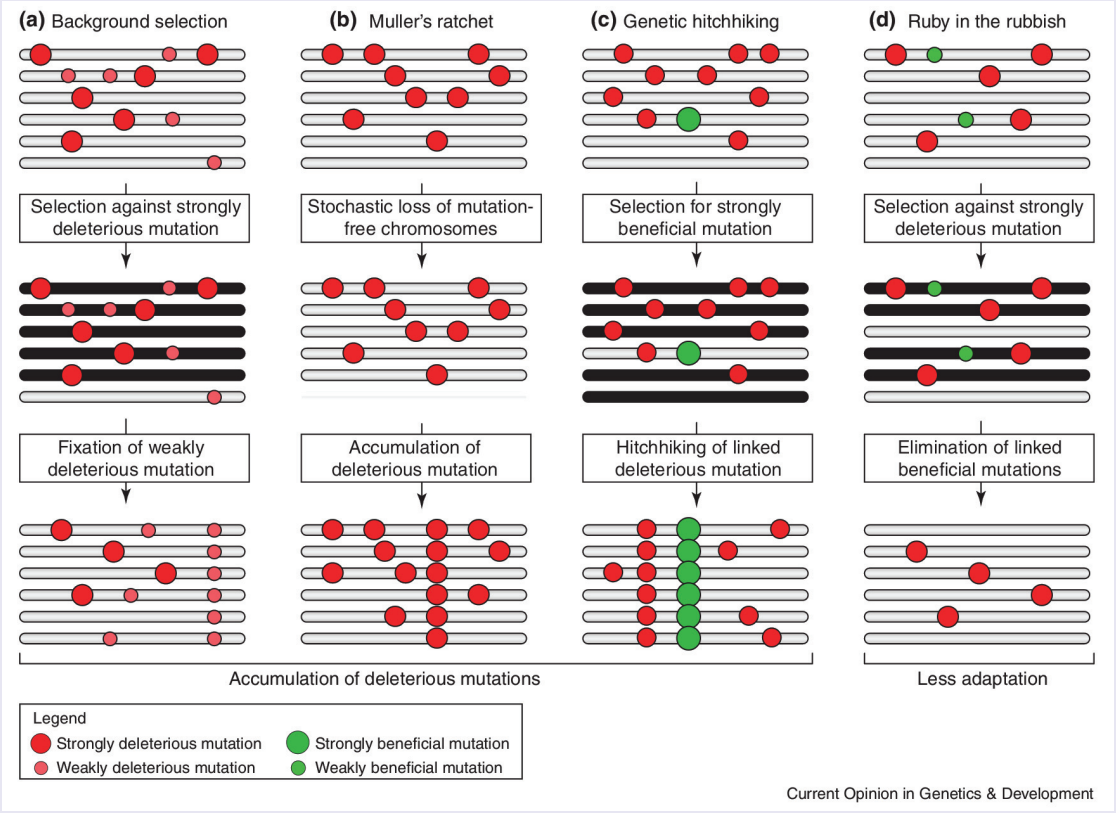
Figure 2.4: Mechanisms of sex chromosome turnover and of absence or presence of recombination suppression: Sex antagonistic genes can induce recombination suppression if located on the PAR, resulting in larger SNRs and heteromorphic sex chromosomes (SA_M here a male beneficial sex antagonitic gene induces further recombination suppresion of the Y SNR). The accumulation of deleterious mutations on SNRs following recombination suppression can induce a sex chromosome turnover, possibly with endless cycles (but never from an XY pair to a ZW pair nor from ZW to XY). Sex chromosomes can also be replaced by a new pair if sex antagonistic genes are located on autosomes (with this mechanism a change of system is possible, from ZW to XY and conversely, but the new pair can only replace a very young sex chromosome pair). The fountain of youth could maintain sex chromosomes at a homomorphic state through X-Y (or Z-W) recombination.

2.5 Are sex-specific non-recombining regions (SNRs) degenerated in plants and algae?

The theories predicting degeneration of SNRs have been verified on various aspects in plants and algae (Table 2.1):

- SNRs accumulate repetitive sequences, for instance papaya hermaphrodite SNR harbours 79.3% repetitive sequences, compared to 52% genome wide (Wang et al., 2012). The Y chromosome of *Silene latifolia* is 570Mb, much larger than the 420Mb X (Matsunaga et al., 1994) and the *Coccinia grandis* Y chromosome has twice the size of any of the other chromosomes (Sousa et al., 2013), suggesting an accumulation of repetitive sequences. Steflova et al. (2013) observed that specific transposable elements and satellites accumulate on the two Ys compared to other chromosomes of *Rumex acetosa*.
- Y copies were found to be less expressed than their X counterparts in *Silene latifolia* (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012), *Rumex hastatulus* (Hough et al., 2014) and in *Ectocarpus sp.* both U and V genes were less expressed than autosomal genes (Ahmed et al., 2014). This could be due to mutations in promoter regions impairing the fixation of transcription factors, or to silencing of nearby repetitive elements.
- Y copies were found to have a higher d_N/d_S than their X counterparts, suggesting relaxed selection and accumulation of deleterious nonsynonymous mutations, in *Silene latifolia* (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Marais et al., 2008) and *Rumex hastatulus* (Hough et al., 2014).
- A lower nucleotide diversity was found in the SNR compared to the rest of the genome in *Silene latifolia* (Qiu et al., 2010) and other plant species (Table 2.1), a consequence of a lower SNR effective population size that in turn decreases the efficacy of selection.
- Codon usage was found to be less optimized in the SNR in *Rumex hastatulus* (Hough et al., 2014) and other plant species (Table 2.1), again suggesting relaxed selection.
- Intron lengths were found to be longer for Y copies compared to their X counterparts in *Silene latifolia* (Marais et al., 2008), *Phoenix dactylifera* (Al-Dous et al., 2011) and in *Ectocarpus sp.* both U and V genes had longer introns than autosomal genes (Ahmed et al., 2014).
- The SNR showed a high number of gene loss in *Silene latifolia* (Bergero and Charlesworth, 2011; Bergero et al., 2015; Chibalina and Filatov, 2011), but X and Y had similar gene loss numbers in papaya (Gschwend et al., 2012; Wang et al., 2012).
- SNRs showed lower gene density (Table 2.1) compared to their recombining counterparts, for instance there is a four fold decrease in gene density on Y compared to X in *S. latifolia* (Chapter 3).

Box 2.2 : Hill-Robertson interferences (adapted from Bachtrog, 2006)



In the absence of recombination, the action of directional selection at one locus can interfere with the action of directional selection at a linked locus. Thus, the net efficacy of natural selection on a non-recombining chromosome, such as the Y, is impaired; deleterious mutations are less efficiently purged, and beneficial mutations are less efficiently incorporated.

(a) Accumulation of weakly deleterious mutations by background selection. A strongly deleterious mutation will be removed by selection along with the mutations genetically linked to it, even if these are slightly advantageous (Charlesworth et al., 1993).

(b) Muller's ratchet. This process involves the stochastic loss of the class of Y chromosomes carrying the fewest number of deleterious mutations from a finite population (Muller, 1964). In the absence of recombination and back mutation, this class of chromosomes cannot be restored. The next best class then replaces it (i.e. the class of chromosomes with the next fewest number of deleterious mutations). This class can in turn be lost, in a succession of irreversible steps. Each such loss of a class of chromosomes is quickly followed by the fixation of a specific deleterious mutation on the Y.

(c) Genetic hitchhiking by favorable mutations. The spread of a favorable mutation in a population of non-recombining Y-chromosomes can drag to fixation any deleterious mutant alleles initially associated with it, as long as the chromosome still has a net fitness advantage (Rice, 1987). Thus, the hitchhiking model requires that selection coefficients for beneficial mutations be larger than for deleterious alleles. Successive adaptive substitutions on an evolving Y chromosome can lead to the fixation of deleterious mutations at many loci. Under this model, Y-chromosome degeneration reflects adaptation at a few loci, at the cost of most other genes on this chromosome.

(d) Lack of adaptation on the non-recombining Y chromosome. The rate of adaptation on a non-recombining chromosome can be greatly reduced, owing to interference of positive mutations with linked deleterious alleles (Peck, 1994). If selection coefficients for beneficial mutations are of the same magnitude or smaller than those for deleterious mutations, only beneficial mutations on Y-chromosomes free of deleterious alleles can contribute to adaptation.

Models **(a–c)** assume that purifying selection against deleterious mutations is reduced on the Y, whereas model **(d)** assumes that positive selection for beneficial mutations is less efficient on the Y. Any accumulation of deleterious alleles on the proto-Y will reduce the fitness of their carriers (models **(a–c)**).

The degeneration of SNRs causes YY and WW genotypes to be lethal in old and moderately old systems (Table 2.1), probably because of the loss of essential gene. However, it was proposed that plants would loose less genes on their SNRs compared to animals because of haploid selection acting in pollen (Bachtrog, 2011; Charlesworth, 2008; Chibalina and Filatov, 2011). Indeed, many genes are expressed in pollen grains (Honys and Twell, 2004), unlike spermatozoa in animals. It is expected that loss of genes expressed in pollen grains from SNRs will be prevented by selection in haploid Y or W pollen grains. However, not all genes carried by SNRs are expressed in pollen and more data (both in animals and plants) will be needed to confirm this hypothesis. The same idea applies to UV systems where gametophytes are haploid individuals, however data is currently lacking in UV systems to test whether gene loss is lower than in ZW or XY systems (outgroups without sex chromosomes are required in order to differentiate gene loss from gene translocation on only one chromosome of the sex chromosome pair). Gene densities in *Ectocarpus sp* and *Marchantia polymorpha* SNRs compared to the rest of the genome indirectly suggest few gene losses from the SNRs (Ahmed et al., 2014; Yamato et al., 2007).

The accumulation of repetitive sequences is responsible for the increase in SNR size observed in some plants (Figure 2.2). It is currently unknown if this is an obligatory step in degeneration, and whether it later evolves into smaller SNRs, after extensive gene loss makes deletions possible (Figure 2.2). TEs are hypothesized to favor chromosomal rearrangements through ectopic recombination and could also promote recombination suppression if heterozygous in the PAR (Dooner and He, 2008).

In papaya, the X region homologous to the SNR also shows a high repetitive content (65.4%) compared to the genome-wide average (52%), although to a lesser extent than the SNR, suggesting that the efficacy of selection is also reduced on the X (Gschwend et al., 2012). This is consistent with the fact that effective population size is also reduced for the X chromosome (Bellott et al., 2010): $N_e(X) = \frac{3}{4} N_e(autosomes)$.

Table 2.1: Review of plant and algal sex chromosome sequence data. Note that the number of plant species with sequence data contrasts dramatically with the number of plants (~50) with known sex chromosomes (Ming et al., 2011), as well as the number of dioecious plants possibly having sex chromosomes (~30,000 see Section 2.1).

Species (Taxon)	reference genome	system	age	strata X-Y/Z-W/U-V divergence	Y/W/U/V degeneration	sequence data
Asparagus officinalis (Asparagaceae, angiosperm)	-	XY (dominant Y)	-	-	homomorphic, viable YY genotype, accumulation of repetitive sequences, low gene density	four BAC clones (Telgmann-Rauber et al., 2007)
Carica papaya (Caricaceae, angiosperm)	Ming et al., 2008	XX females, XY males and XY^h hermaphrodites (dominant Y and Y^h)	~9My	2 strata	homomorphic, lethal YY, Y^hY and Y^hY^h genotypes, high repeat density, 25% Y^h and 14% X gene loss	3.5Mb X and 8.1Mb Y and Y^h sex specific complete regions (Gschwend et al., 2012; VanBuren et al., 2015; Wang et al., 2012)
Diospyros lotus (Ebenaceae, angiosperm)	Akagi et al., 2014	XY	-	-	-	~1Mb male specific region (Akagi et al., 2014)
Phoenix dactylifera (Arecaceae, angiosperm)	Al-Dous et al., 2011; Al-Mssallem et al., 2013	XY (dominant Y)	~50My	-	Heteromorphic (unusual smaller Y than X), low nucleotide diversity, long introns	genetic markers of sex (Cherif et al., 2013), 4 non recombining Y scaffolds (Al-Dous et al., 2011) and X genetic map (Mathew et al., 2014)

Continued on next page

Table 2.1 – continued from previous page

Species (Taxon)	reference genome	system	age	strata X-Y/Z-W/U-V divergence	Y/W/U/V degeneration	sequence data
Rumex hastatulus (Polygonaceae, angiosperm)	-	females XX, males ancestrally XY but derived XY1Y2 (X-autosome dose)	15–16My (Navajas-Pérez et al., 2005)	-	Heteromorphic. 28% gene loss in ancestral Y, 8% in derived Y. High d_N/d_S. Less optimized codon usage. Reduced expression.	RNA-seq data (Hough et al., 2014)
Silene latifolia (Caryophyllaceae, angiosperm)	-	XY (dominant Y)	~5My (Rautenberg et al., 2010)	1 PAR, 2-3 strata 5-25% synonymous divergence (Bergero et al., 2007, 2013)	heteromorphic (giant Y). Lethal YY genotype. Reduced expression. 14.5% Y gene loss (Bergero et al., 2015). High d_N/d_S. Low nucleotide diversity. Increased intron length. High repeat density. Low gene density	RNA-seq data (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012), BAC clones (Chapter 3) and other genes referenced in Bergero et al. (2013)
Vitis vinifera* L. subsp *sylvestris (Vitaceae, angiosperm)	Jaillon et al., 2007	XY	-	-	Homomorphic. Low nucleotide diversity	partial 154.8kb genomic sequence (Picq et al., 2014)
Populus trichocarpa (Salicaceae, angiosperm)	Tuskan et al., 2006	unclear (XY or ZW)	-	-	Homomorphic, high d_N/d_S, reduced nucleotide diversity	~100kbp Y sepcific region (Geraldes et al., 2015) ~ 1Mb W specific region (Tuskan et al., 2012; Yin et al., 2008)
Salix suchowensis (Salicaceae, angiosperm)	Dai et al., 2014	ZW	-	-	Homomorphic	sex linked scaffolds (Hou et al., 2015)

Continued on next page

Table 2.1 – continued from previous page

Species (Taxon)	reference genome	system	age	strata X-Y/Z-W/U-V divergence	Y/W/U/V degeneration	sequence data
Ceratodon purpureus (Ditrichaceae, bryophyte)	-	UV	~6My	possibly three strata	Heteromorphic, reduced nucleotide diversity, premature stop codon	9 genes (McDaniel et al., 2013, 2007)
Ectocarpus sp. (Ectocarpaceae, brown algae)	Cock et al., 2010	UV (dominant V)	>70My	No strata	Low gene density, high repeated DNA density, underrepresentation of optimal codons, lower expression level, longer introns	complete 1Mb genomic sequences (Ahmed et al., 2014)
Marchantia polymorpha (Hepaticeae, bryophyte)	-	UV	-	70.6–93.6% nucleotide identity	Heteromorphic, low gene density	Complete 10Mb V specific region (Yamato et al., 2007)
Volvox carteri (Volvocaceae, green algae)	Prochnik et al., 2010	UV	-	possibly 2 strata	low gene density, high repeat density, reduced codon usage bias	complete genomic sequences (Ferris et al., 2010)

2.6 Has dosage compensation evolved in non-animal taxa?

The availability of sex-linked gene sequences (see Box 2.1) allows functional analyses such as the study of dosage compensation. With SNR degeneration, the heterogametic sex in ZW and XY systems has less expression than the homogametic sex (Section 2.5), most of all after SNR gene loss that make the heterogametic sex partially aneuploid, which can be deleterious. A mechanism called dosage compensation has evolved in some species that allows for similar male and female expression levels (Charlesworth, 1996), but, more importantly, similar expression levels between sex chromosomes and their ancestral autosomal pair (Mank, 2013). Three canonical dosage compensation mechanisms have so far been described in *Drosophila*, mammals and *C. elegans* (reviewed in Disteche, 2012; Ercan, 2015, and see Figure 2.5). In *Drosophila*, the single X in males is hyperexpressed. In mammals, X chromosomes are hyperexpressed in both males and females compared to autosomes but one X is inactivated in females. In *C. elegans*, X chromosomes are also hyperexpressed in both males and hermaphrodites but both Xs are down-regulated in hermaphrodites.

Recent work revealed that these canonical dosage compensation mechanisms only hold

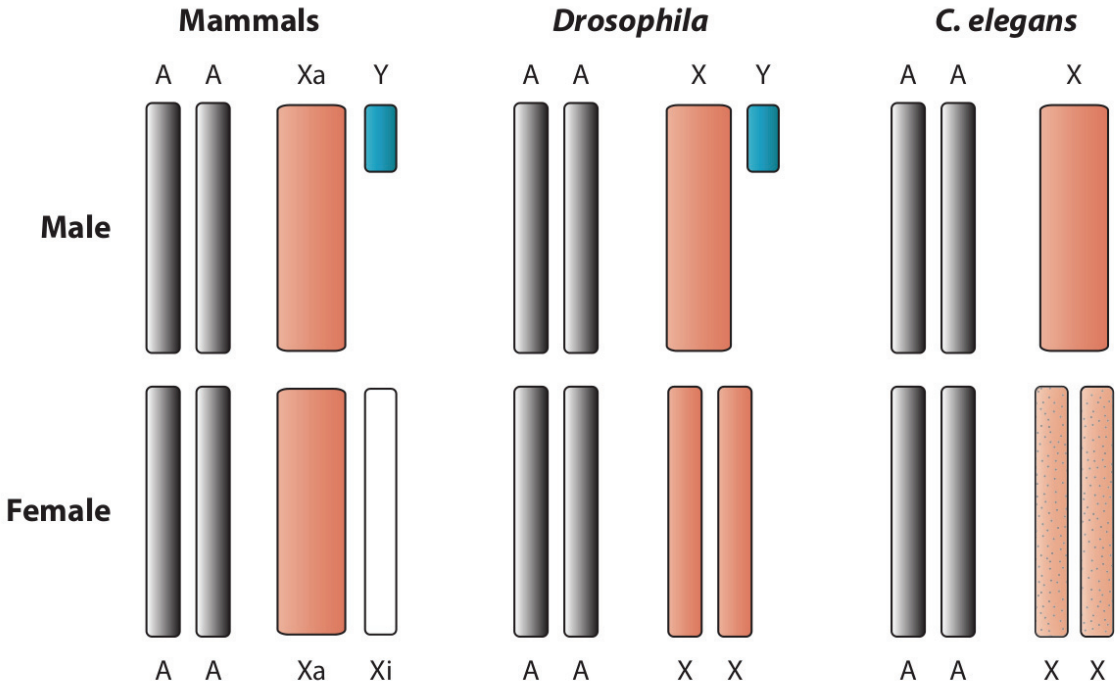


Figure 2.5: Dosage compensation mechanisms (Disteche, 2012): Male and female sexes are indicated. From left to right: Mammals, in which expressed genes on the active X (Xa) are upregulated in both sexes and genes on the inactive X (Xi) are silenced in females; *Drosophila*, in which the X is upregulated in males only; *Caenorhabditis elegans*, in which the X is upregulated in both sexes and downregulated in hermaphrodites (mottled).

for some genes on the X. In mammals genes involved in protein complexes, where stoichiometric balance is important, show the expected two fold increase in gene expression

compared to autosomes (Pessia et al., 2012) but this is not observed when taking all X genes together, at the transcript (Xiong et al., 2010) or protein level (Chen and Zhang, 2015). Similarly, the X chromosome is largely not upregulated in *C. elegans* (Albritton et al., 2014). This suggests only some genes are dosage sensitive and dosage compensation evolves only for this type of genes (Pessia et al., 2014; Veitia et al., 2015). In humans, an alternative strategy involved downregulation of autosomal genes that are within the same protein interaction network as downregulated sex-linked genes (Julien et al., 2012). In *C. elegans*, orhologs of yeast haploinsufficient genes (genes causing deleterious effects when in a haploid state) are depleted from the X chromosome (Albritton et al., 2014), suggesting that another strategy is to move a dosage sensitive gene to an autosome. Therefore, Ohno's hypothesis of X two-fold upregulation is one of many different mechanisms that counteracted the potential haploinsufficiency of individual X-linked genes in males (Ercan, 2015). Also, the study of dosage compensation in other taxa has shown that chromosome-wide dosage compensation has not evolved in all sex chromosomes, in particular it is more common in XY than ZW systems (Mank, 2013).

It was only recently that dosage compensation started to be studied in non-animal systems. Chibalina and Filatov (2011) tested for dosage compensation in *Silene latifolia* by comparing male and female expression levels for X-hemizygous genes (sex-linked genes for which the Y copy is not expressed or was lost). They found that male and female expression significantly differed for these genes and concluded that there was no dosage compensation in this species. However, another study that focused on sex-linked genes with a preserved Y copy observed that males and females maintained similar expression levels even when Y expression was reduced due to degeneration (Muyle et al., 2012). The X-hemizygous gene *SlWUS1* in *Silene latifolia* was shown to have lower expression levels in males than in females and to have both X copies expressed in females using qRT-PCR (Kazama et al., 2012), therefore the authors concluded there was no evidence that this gene was dosage compensated or X-inactivated. But, without outgroup reference, the authors could not rule out that expression in female buds had been increased, perhaps because of a specialisation in the female function. More recently, Bergero et al. (2015) found that male expression was globally halved compared to females for 99 X-hemizygous genes, as expected under no dosage compensation (however a few genes individually showed patterns of dosage compensation). The authors also tested for dosage compensation in 99 XY gene pairs with high expression (over a 100 Y reads), and degenerated Y copy (Y over X expression ratios below 0.75), male expression relative to female expression was globally higher than the value expected without dosage compensation both in bud and leaf transcriptomes. The authors concluded that partial dosage compensation for these XY genes could not be excluded, but that results could also be produced by male-biased expression. Unfortunately the authors did not identify male-biased genes in their dataset (unlike Muyle et al., 2012), which could have made the two hypotheses distinguishable. From these four studies (Bergero et al., 2015; Chibalina and Filatov, 2011;

Kazama et al., 2012; Muyle et al., 2012), it seems clear that X-hemizygous genes are mainly not dosage compensated in *Silene latifolia*, however the status of XY gene pairs remains unclear. It is possible that partial dosage compensation exists in this species (Toups et al., 2015).

Only one other plant has so far been studied for dosage compensation: *Rumex hastatulus* (Hough et al., 2014). The authors focused on 119 X-hemizygous genes and found that 79% were significantly less expressed in males than in females, which led them to conclude there was no dosage compensation in *Rumex hastatulus*. The authors, however, did not comment on the fact that genes with a preserved but degenerated Y copy maintained similar male and female expression levels in their data (Figure 3 in Hough et al., 2014), which strongly suggests dosage compensation.

The reason why most studies focused on X-hemizygous genes to test for dosage compensation is that it was expected that selection for dosage compensation would be strongest for these genes as the complete absence of Y expression would be most deleterious (especially in haploid tissues). However this verbal expectation has not been tested and the opposite could easily be argued: X-hemizygous genes could have lost their Y copy because dosage was not important for them and selection did not slow down the loss of the Y copy nor selected for dosage compensation when inevitable degeneration happened.

In a recent analysis, both XY and X-hemyzygous genes were analysed to test for dosage compensation in *Silene latifolia*, using RNA-seq data from a cross (see Chapter 5). A closely related species without sex chromosomes, *Silene vulgaris*, was also included in the study in order to have an estimate of the ancestral autosomal expression levels of sex-linked genes in *Silene latifolia*. Results showed that X expression levels in males have increased with Y degeneration, allowing similar expression levels between *Silene latifolia* males and the outgroup without sex chromosomes. However, similarly to mammals and *C. elegans*, not all genes were affected by this pattern, in particular X-hemizygous genes show very little dosage compensation, only highly expressed X-hemizygous genes were dosage compensated. This suggests that dosage compensation evolved only for dosage sensitive genes and that X-hemizygous genes are mainly insensitive to changes in dosage. Consistently, X-hemizygous genes were found to be depleted in proteins involved in the ribosomal complex, a striking example for a large protein complex requiring stoichiometric balance. Remarkably, the increase in X expression in males also happened in females (a phenomenon that has already been observed in flour beetle Prince et al., 2010, and in stickleback, Schultheiß et al., 2015). However, in *Silene latifolia*, the X in females is hyperexpressed to a lesser extent than in males, and only for the maternal X. This suggests that some imprinting mechanism is responsible for dosage compensation in *Silene latifolia* which increases expression of the maternal X, both in males and females, and this could possibly be deleterious for females. This shows a conflict between males and females over optimal expression levels. This observation could either be explained by a downregulation in males being more deleterious than an upregulation in females, or by

sexual selection, that is stronger on males than on females, and would select for upregulation of the single X chromosome in males, at the expense of females, until a second correcting mechanism evolves in females to prevent hyperexpression (Mank, 2013). This second mechanism correcting for hyperexpression in females could already be evolving in *Silene latifolia*, as X hyperexpression is lower in females than in males.

2.7 How are sexual conflicts resolved in dioecious plants and algae?

The evolution of sex chromosomes and separate sexes from hermaphroditism should trigger a burst of adaptation throughout the genome as trade-offs between reproductive male and female functions can be resolved through expressing genes differently in male and female individuals (Ellegren and Parsch, 2007, and see Figure 2.6) and also through evolving male and female specific sequences with SNRs. Other mechanisms described in animal species that solve sexual conflicts involve alternative splicing and gene duplication followed by neofunctionalisation (reviewed in Betrán et al., 2004).

In plants, very little is known about sex-biased genes (Barrett and Hough, 2013). Sex

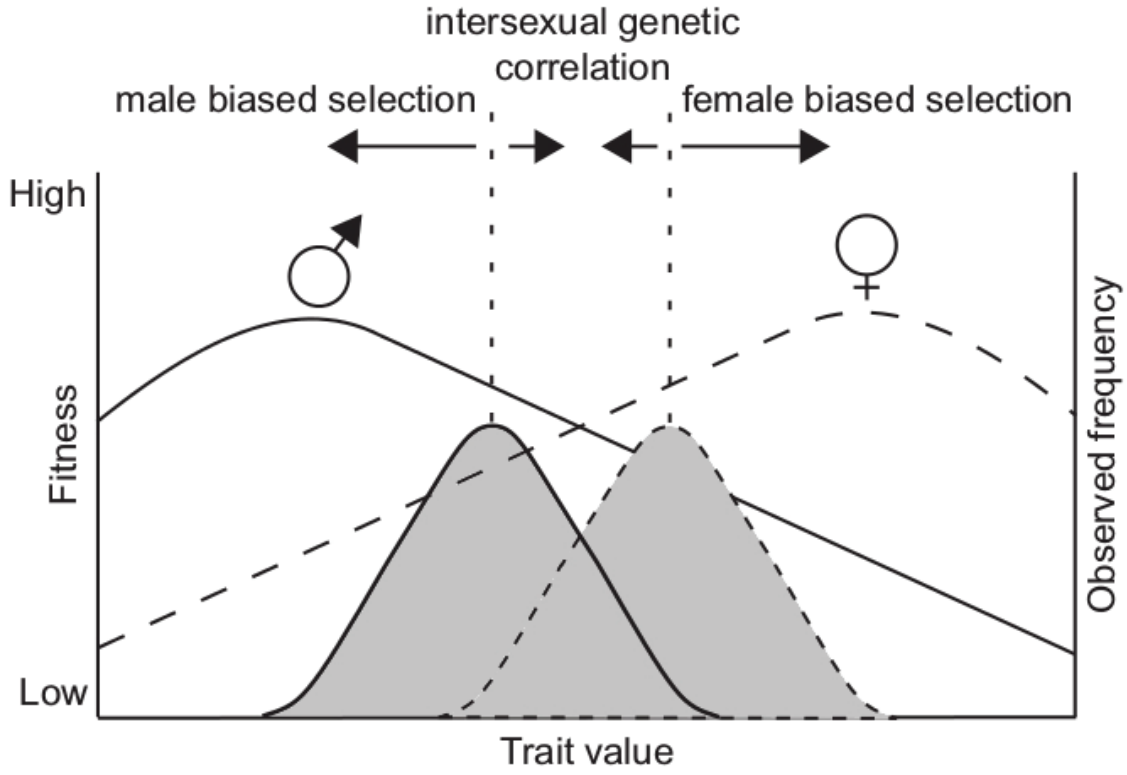


Figure 2.6: Sex antagonism and evolution of gene expression level (from Barrett and Hough, 2013): A hypothetical scenario in which females (dashed lines) and males (solid lines) have different optima for the same trait, causing sex-biased selection (long arrows). A shared genetic architecture can constrain the sexes from evolving toward their respective trait optima. However, sexual dimorphism can still evolve when such trade-offs exist, and this can involve sex-limited gene expression and the breakdown of strong intersexual genetic correlations, possibly facilitated by the evolution of SNRs.

differential gene expression was detected in *Populus yunnanensis*, a species having a ZW system with a small female-specific region, but genes located in autosomes rather than in the female specific region were the major contributors to the sexual differences in the salinity tolerance of poplars (Jiang et al., 2012), suggesting the SNR had little role in the evolution of sexual dimorphism in this species. In *Populus tremula*, only two significantly differentially expressed genes were identified from an RNA-seq data, showing very little sexual dimorphism in this species (Robinson et al., 2014). In the dioecious alga *Fucus vesiculosus*, there are more male-biased than female-biased genes, but the possible contribution of sex chromosomes in that pattern remains unknown, as no sex-determining locus has so far been characterised in this species (Martins et al., 2013). Similarly, in *Asparagus officinalis* more male-biased than female-biased genes were identified, but the contribution of sex chromosomes in that pattern remains unknown (Harkess et al., 2015). In *Ectocarpus sp.*, under 12% of genes were sex-biased, which is consistent with the low level of morphological sexual dimorphism in this species (Lipinska et al., 2015). More male-biased than female-biased genes were identified and the PAR was significantly enriched in female-biased genes compared to autosomes, suggesting a role of partial linkage with the SNR in the evolution of sex-biased expression.

The contribution of sex chromosomes in the evolution of sexual dimorphism in non-animal species was clear in the study of *Silene latifolia* (see Chapter 6). In this species, sex-biased genes are significantly enriched on the sex chromosomes compared to the autosomes. Selection to reduce fitness costs brought about by sexually antagonistic genes were mainly solved through changes in expression levels in females rather than males. In *Drosophila*, the X chromosome is enriched in female-biased genes, selected as the X spends more time in females than in males, a process called feminisation of the X (Ellegren and Parsch, 2007). But no feminisation of the X chromosome was found in *Silene latifolia* when looking at numbers of female-biased genes, maybe due to the young age of the system that has not allowed for changes in the structure of the sex chromosomes (i.e. the number of female-biased genes). However, when looking at expression levels, the X chromosome has undergone feminisation and demasculinisation in *Silene latifolia* since its expression level increased in females for female-biased genes and decreased in females for male-biased genes. Similarly, the Y has undergone masculinisation as its expression has increased in male-biased genes and decreased in female-biased genes.

In *Marchantia polymorpha*, a Y specific repeated gene was found to be expressed specifically in male sexual organ, similarly to human Y palindromes (Okada et al., 2001), which suggests Y masculinisation in this species. In *Volvox*, the retinoblastoma tumor suppressor homolog MAT3 is a mating type gene that displays sexually regulated alternative splicing and evidence of gender-specific selection, again showing SNRs specialise in sexual functions, probably by solving sexual conflicts (Ferris et al., 2010).

2.8 Conclusions and Perspectives

Plants and algae are unique systems to study the evolution of sex chromosomes as their sex chromosomes are often of recent origin (Table 2.1) and closely related species without sex chromosomes allow for comparative analyses that reveal the direction of the evolution of sex chromosomes and make it possible to precisely infer gene loss on SNRs and to study how expression levels evolved for sex-linked genes. The various pathways that can lead to dioecy in plants and algae promise a wide diversity of genetic mechanisms of sex determination. But this diversity has been understudied, with only three sex-determining genes described so far on plant and algal sex chromosomes (Section 2.2). More studies are needed to describe the diversity of sex determination in non-model organisms (Bachtrog et al., 2014), which will allow to test for the two gene model of sex chromosome origin (Charlesworth and Charlesworth, 1978).

In plants, different species exhibit various stages of sex chromosome degeneration levels (Figure 2.2), it is however unknown whether large Ys (compared to Xs), are of younger origin than small Ys (compared to Xs). More studies on sex chromosome degeneration levels are required, along with dating of their origin. This will be facilitated by the use of new methods to obtain sex-linked sequences in non-model organisms (see Box 2.1). The availability of sex-linked sequences will also allow to study strata and their age, which will make it possible to test whether small SNRs are due to a recent origin (as in *Fragaria virginiana*) or to the absence of further events of recombination suppression (as in *Ectocarpus sp*). One mechanism preventing the degeneration of SNRs could be X-Y (or Z-W) recombination in sex reverted individuals, as in animals. But it is unknown whether X-Y (or Z-W) recombination happens in sex reverted individuals in dioecious species. Also, the role of sex-antagonistic alleles in recombination suppression was not formerly demonstrated but only suggested by Qiu et al. (2013). High sex chromosome turnover has been suggested in Salicaceae, but the forces driving them are unknown.

It will be interesting to test for dosage compensation in other plant and algal species. It should however be kept in mind that dosage compensation mechanisms are rarely total, even in the canonical models such as those of placentals or *C. elegans*. From the study of *Silene latifolia* and *Rumex hastatulus*, it appears that sex-linked genes without Y copies are rarely dosage compensated. Outgroups without sex chromosomes should be used to orientate expression changes in sex-linked genes since the study of male over female expression ratio can be biased if both sexes evolved in the same direction (as is the case in *Silene latifolia*).

Finally, estimates of SNRs genes loss are currently lacking in order to test whether UV systems lose less genes compared to XY and ZW systems due to haploid selection opposing degeneration in gametophytes. Similarly, data in plants are lacking to test whether Y SNRs loose less genes in plants than in animals due to haploid selection in pollen grains.

2.9 References

Ahmed, S., Cock, J. M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A. F., Dittami, S. M., Corre, E., Valero, M., Aury, J.-M., Roze, D., Van de Peer, Y., Bothwell, J., Marais, G. A. B., and Coelho, S. M. (2014). A haploid system of sex determination in the brown alga Ectocarpus sp. eng. *Current biology: CB* 24(17), 1945–1957. ISSN: 1879-0445. DOI: 10.1016/j.cub.2014.07.042.

Akagi, T., Henry, I. M., Tao, R., and Comai, L. (2014). Plant genetics. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. eng. *Science (New York, N.Y.)* 346(6209), 646–650. ISSN: 1095-9203. DOI: 10.1126/science.1257225.

Albritton, S. E., Kranz, A.-L., Rao, P., Kramer, M., Dieterich, C., and Ercan, S. (2014). Sex-biased gene expression and evolution of the x chromosome in nematodes. eng. *Genetics* 197(3), 865–883. ISSN: 1943-2631. DOI: 10.1534/genetics.114.163311.

Bachtrog, D. (2006). A dynamic view of sex chromosome evolution. eng. *Current Opinion in Genetics & Development* 16(6), 578–585. ISSN: 0959-437X. DOI: 10.1016/j.gde.2006.10.007.

Bachtrog, D. (2011). Plant Sex Chromosomes: A Non-Degenerated Y? English. *Current Biology* 21(18), R685–R688. ISSN: 0960-9822. DOI: 10.1016/j.cub.2011.08.027.

Bachtrog, D., Kirkpatrick, M., Mank, J. E., McDaniel, S. F., Pires, J. C., Rice, W., and Valenzuela, N. (2011). Are all sex chromosomes created equal? eng. *Trends in genetics: TIG* 27(9), 350–357. ISSN: 0168-9525. DOI: 10.1016/j.tig.2011.05.005.

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., Hahn, M. W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., Vamosi, J. C., and Tree of Sex Consortium (2014). Sex determination: why so many ways of doing it? eng. *PLoS biology* 12(7), e1001899. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001899.

Barrett, S. C. H. and Hough, J. (2013). Sexual dimorphism in flowering plants. eng. *Journal of Experimental Botany* 64(1), 67–82. ISSN: 1460-2431. DOI: 10.1093/jxb/ers308.

Bellott, D. W., Skaletsky, H., Pyntikova, T., Mardis, E. R., Graves, T., Kremitzki, C., Brown, L. G., Rozen, S., Warren, W. C., Wilson, R. K., and Page, D. C. (2010). Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. eng. *Nature* 466(7306), 612–616. ISSN: 1476-4687. DOI: 10.1038/nature09172.

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Forrest, A., Kamau, E., and Charlesworth, D. (2007). Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: evidence from new sex-linked genes. eng. *Genetics* 175(4), 1945–1954. ISSN: 0016-6731. DOI: 10.1534/genetics.106.070110.

Bergero, R., Qiu, S., and Charlesworth, D. (2015). Gene Loss from a Plant Sex Chromosome System. ENG. *Current biology: CB*. ISSN: 1879-0445. DOI: 10.1016/j.cub.2015.03.015.

Bergero, R., Qiu, S., Forrest, A., Borthwick, H., and Charlesworth, D. (2013). Expansion of the pseudo-autosomal region and ongoing recombination suppression in the Silene latifolia sex chromosomes. eng. *Genetics* 194(3), 673–686. ISSN: 1943-2631. DOI: 10.1534/genetics.113.150755.

Betrán, E., Emerson, J. J., Kaessmann, H., and Long, M. (2004). Sex chromosomes and male functions: where do new genes go? eng. *Cell Cycle (Georgetown, Tex.)* 3(7), 873–875. ISSN: 1551-4005.

Blaser, O., Grossen, C., Neuenschwander, S., and Perrin, N. (2013). Sex-chromosome turnovers induced by deleterious mutation load. eng. *Evolution; International Journal of Organic Evolution* 67(3), 635–645. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2012.01810.x.

Blaser, O., Neuenschwander, S., and Perrin, N. (2014). Sex-chromosome turnovers: the hot-potato model. eng. *The American Naturalist* 183(1), 140–146. ISSN: 1537-5323. DOI: 10.1086/674026.

Carvalho, A. B. and Clark, A. G. (2013). Efficient identification of Y chromosome sequences in the human and Drosophila genomes. eng. *Genome Research* 23(11), 1894–1907. ISSN: 1549-5469. DOI: 10.1101/gr.156034.113.

Charlesworth, B. (1996). The evolution of chromosomal sex determination and dosage compensation. eng. *Current biology: CB* 6(2), 149–162. ISSN: 0960-9822.

Charlesworth, B. and Charlesworth, D. (1978). Model for Evolution of Dioecy and Gynodioecy. English. *American Naturalist* 112(988). WOS:A1978FX45000001, 975–997. ISSN: 0003-0147. DOI: 10.1086/283342.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4), 1289–1303.

Charlesworth, D. (2008). Plant sex chromosomes. eng. *Genome Dynamics* 4, 83–94. ISSN: 1660-9263. DOI: 10.1159/000126008.

Charlesworth, D. and Charlesworth, B. (1980). Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. eng. *Genetical Research* 35(2), 205–214.

Charlesworth, D. (2013). Plant sex chromosome evolution. eng. *Journal of Experimental Botany* 64(2), 405–420. ISSN: 1460-2431. DOI: 10.1093/jxb/ers322.

Chen, X. and Zhang, J. (2015). No X-Chromosome Dosage Compensation in Human Proteomes. ENG. *Molecular Biology and Evolution*. ISSN: 1537-1719. DOI: 10.1093/molbev/msv036.

Cherif, E., Zehdi, S., Castillo, K., Chabrillange, N., Abdoulkader, S., Pintaud, J.-C., Santoni, S., Salhi-Hannachi, A., Glémin, S., and Aberlenc-Bertossi, F. (2013). Male-specific DNA markers provide genetic evidence of an XY chromosome system, a recombination ar-

rest and allow the tracing of paternal lineages in date palm. eng. *The New Phytologist* 197(2), 409–415. ISSN: 1469-8137. DOI: 10.1111/nph.12069.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Cock, J. M. et al. (2010). The Ectocarpus genome and the independent evolution of multicellularity in brown algae. eng. *Nature* 465(7298), 617–621. ISSN: 1476-4687. DOI: 10.1038/nature09016.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. eng. *Nature* 508(7497), 488–493. ISSN: 1476-4687. DOI: 10.1038/nature13151.

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G. A., Milne, R., Chen, Y., Wan, Z., Wang, Z., Luo, W., Wang, K., Wan, D., Wang, M., Wang, J., Liu, J., and Yin, T. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. eng. *Cell Research* 24(10), 1274–1277. ISSN: 1748-7838. DOI: 10.1038/cr.2014.83.

Disteche, C. M. (2012). Dosage compensation of the sex chromosomes. eng. *Annual Review of Genetics* 46, 537–560. ISSN: 1545-2948. DOI: 10.1146/annurev-genet-110711-155454.

Dooner, H. K. and He, L. (2008). Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. eng. *The Plant Cell* 20(2), 249–258. ISSN: 1040-4651. DOI: 10.1105/tpc.107.057596.

Doorn, G. S. van and Kirkpatrick, M. (2007). Turnover of sex chromosomes induced by sexual conflict. eng. *Nature* 449(7164), 909–912. ISSN: 1476-4687. DOI: 10.1038/nature06178.

Doorn, G. S. van and Kirkpatrick, M. (2010). Transitions between male and female heterogamety caused by sex-antagonistic selection. eng. *Genetics* 186(2), 629–645. ISSN: 1943-2631. DOI: 10.1534/genetics.110.118596.

Al-Dous, E. K., George, B., Al-Mahmoud, M. E., Al-Jaber, M. Y., Wang, H., Salameh, Y. M., Al-Azwani, E. K., Chaluvadi, S., Pontaroli, A. C., DeBarry, J., Arondel, V., Ohlrogge, J., Saie, I. J., Suliman-Elmeer, K. M., Bennetzen, J. L., Kruegger, R. R., and Malek, J. A. (2011). De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). eng. *Nature Biotechnology* 29(6), 521–527. ISSN: 1546-1696. DOI: 10.1038/nbt.1860.

Ellegren, H. and Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. eng. *Nature Reviews. Genetics* 8(9), 689–698. ISSN: 1471-0056. DOI: 10.1038/nrg2167.

Ercan, S. (2015). Mechanisms of x chromosome dosage compensation. eng. *Journal of Genomics* 3, 1–19. ISSN: 1839-9940. DOI: 10.7150/jgen.10404.

Ferris, P. J. and Goodenough, U. W. (1997). Mating type in Chlamydomonas is specified by mid, the minus-dominance gene. eng. *Genetics* 146(3), 859–869. ISSN: 0016-6731.

Ferris, P., Olson, B. J. S. C., De Hoff, P. L., Douglass, S., Casero, D., Prochnik, S., Geng, S., Rai, R., Grimwood, J., Schmutz, J., Nishii, I., Hamaji, T., Nozaki, H., Pellegrini, M., and Umen, J. G. (2010). Evolution of an expanded sex-determining locus in Volvox. eng. *Science (New York, N.Y.)* 328(5976), 351–354. ISSN: 1095-9203. DOI: 10.1126/science.1186222.

Fontanillas, E., Hood, M. E., Badouin, H., Petit, E., Barbe, V., Gouzy, J., Vienne, D. M. de, Aguileta, G., Poulain, J., Wincker, P., Chen, Z., Toh, S. S., Cuomo, C. A., Perlin, M. H., Gladieux, P., and Giraud, T. (2015). Degeneration of the nonrecombining regions in the mating-type chromosomes of the anther-smut fungi. eng. *Molecular Biology and Evolution* 32(4), 928–943. ISSN: 1537-1719. DOI: 10.1093/molbev/msu396.

Gautier, M. (2014). Using genotyping data to assign markers to their chromosome type and to infer the sex of individuals: a Bayesian model-based classifier. eng. *Molecular Ecology Resources* 14(6), 1141–1159. ISSN: 1755-0998. DOI: 10.1111/1755-0998.12264.

Geng, S., De Hoff, P., and Umen, J. G. (2014). Evolution of sexes from an ancestral mating-type specification pathway. eng. *PLoS biology* 12(7), e1001904. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001904.

Geraldes, A., Hefer, C. A., Capron, A., Kolosova, N., Martinez-Nuñez, F., Soolanayakanahally, R. Y., Stanton, B., Guy, R. D., Mansfield, S. D., Douglas, C. J., and Cronk, Q. C. B. (2015). Recent Y chromosome divergence despite ancient origin of dioecy in poplars (Populus). ENG. *Molecular Ecology*. ISSN: 1365-294X. DOI: 10.1111/mec.13126.

Gschwend, A. R., Yu, Q., Tong, E. J., Zeng, F., Han, J., VanBuren, R., Aryal, R., Charlesworth, D., Moore, P. H., Paterson, A. H., and Ming, R. (2012). Rapid divergence and expansion of the X chromosome in papaya. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), 13716–13721. ISSN: 1091-6490. DOI: 10.1073/pnas.1121096109.

Harkess, A., Mercati, F., Shan, H.-Y., Sunseri, F., Falavigna, A., and Leebens-Mack, J. (2015). Sex-biased gene expression in dioecious garden asparagus (Asparagus officinalis). ENG. *The New Phytologist*. ISSN: 1469-8137. DOI: 10.1111/nph.13389.

Honys, D. and Twell, D. (2004). Transcriptome analysis of haploid male gametophyte development in Arabidopsis. eng. *Genome Biology* 5(11), R85. ISSN: 1465-6914. DOI: 10.1186/gb-2004-5-11-r85.

Hou, J., Ye, N., Zhang, D., Chen, Y., Fang, L., Dai, X., and Yin, T. (2015). Different autosomes evolved into sex chromosomes in the sister genera of Salix and Populus. eng. *Scientific Reports* 5, 9076. ISSN: 2045-2322. DOI: 10.1038/srep09076.

Hough, J., Hollister, J. D., Wang, W., Barrett, S. C. H., and Wright, S. I. (2014). Genetic degeneration of old and young Y chromosomes in the flowering plant Rumex hastatulus.

eng. *Proceedings of the National Academy of Sciences of the United States of America* 111(21), 7713–7718. ISSN: 1091-6490. DOI: 10.1073/pnas.1319227111.

Howell, E. C., Armstrong, S. J., and Filatov, D. A. (2009). Evolution of neo-sex chromosomes in Silene diclinis. eng. *Genetics* 182(4), 1109–1115. ISSN: 1943-2631. DOI: 10.1534/genetics.109.103580.

Jaillon, O. et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. eng. *Nature* 449(7161), 463–467. ISSN: 1476-4687. DOI: 10.1038/nature06148.

Jiang, H., Peng, S., Zhang, S., Li, X., Korpelainen, H., and Li, C. (2012). Transcriptional profiling analysis in Populus yunnanensis provides insights into molecular mechanisms of sexual differences in salinity tolerance. eng. *Journal of Experimental Botany* 63(10), 3709–3726. ISSN: 1460-2431. DOI: 10.1093/jxb/ers064.

Jordan, C. Y. and Charlesworth, D. (2012). The potential for sexually antagonistic polymorphism in different genome regions. eng. *Evolution; International Journal of Organic Evolution* 66(2), 505–516. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2011.01448.x.

Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., Daish, T., Grützner, F., and Kaessmann, H. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. eng. *PLoS biology* 10(5), e1001328. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001328.

Kazama, Y., Nishihara, K., Bergero, R., Fujiwara, M. T., Abe, T., Charlesworth, D., and Kawano, S. (2012). SlWUS1; an X-linked gene having no homologous Y-linked copy in Silene latifolia. eng. *G3 (Bethesda, Md.)* 2(10), 1269–1278. ISSN: 2160-1836. DOI: 10.1534/g3.112.003749.

Lipinska, A., Cormier, A., Luthringer, R., Peters, A. F., Corre, E., Gachon, C. M. M., Cock, J. M., and Coelho, S. M. (2015). Sexual Dimorphism and the Evolution of Sex-Biased Gene Expression in the Brown Alga Ectocarpus. ENG. *Molecular Biology and Evolution*. ISSN: 1537-1719. DOI: 10.1093/molbev/msv049.

Manchester, S. R., Judd, W. S., and Handley, B. (2006). Foliage and Fruits of Early Poplars (Salicaceae: Populus) from the Eocene of Utah, Colorado, and Wyoming. *International Journal of Plant Sciences* 167(4), 897–908. ISSN: 1058-5893. DOI: 10.1086/503918.

Mank, J. E. (2013). Sex chromosome dosage compensation: definitely not for everyone. eng. *Trends in genetics: TIG* 29(12), 677–683. ISSN: 0168-9525. DOI: 10.1016/j.tig.2013.07.005.

Marais, G. A. B., Nicolas, M., Bergero, R., Chambrier, P., Kejnovsky, E., Monéger, F., Hobza, R., Widmer, A., and Charlesworth, D. (2008). Evidence for degeneration of the Y chromosome in the dioecious plant Silene latifolia. eng. *Current biology: CB* 18(7), 545–549. ISSN: 0960-9822. DOI: 10.1016/j.cub.2008.03.023.

Martins, M. J. F., Mota, C. F., and Pearson, G. A. (2013). Sex-biased gene expression in the brown alga Fucus vesiculosus. eng. *BMC genomics* 14, 294. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-294.

Mathew, L. S., Spannagl, M., Al-Malki, A., George, B., Torres, M. F., Al-Dous, E. K., Al-Azwani, E. K., Hussein, E., Mathew, S., Mayer, K. F. X., Mohamoud, Y. A., Suhre, K., and Malek, J. A. (2014). A first genetic map of date palm (Phoenix dactylifera) reveals long-range genome structure conservation in the palms. eng. *BMC genomics* 15, 285. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-285.

Matsunaga, S., Hizume, M., Kawano, S., and Kuroiwa, T. (1994). Cytological Analyses in *Melandrium album*: Genome Size, Chromosome Size and Fluorescence *in situ* Hybridization. *Cytologia* 59(1), 135–141. DOI: 10.1508/cytologia.59.135.

McDaniel, S. F., Neubig, K. M., Payton, A. C., Quatrano, R. S., and Cove, D. J. (2013). Recent gene-capture on the UV sex chromosomes of the moss Ceratodon purpureus. eng. *Evolution; International Journal of Organic Evolution* 67(10), 2811–2822. ISSN: 1558-5646. DOI: 10.1111/evo.12165.

McDaniel, S. F., Willis, J. H., and Shaw, A. J. (2007). A linkage map reveals a complex basis for segregation distortion in an interpopulation cross in the moss Ceratodon purpureus. eng. *Genetics* 176(4), 2489–2500. ISSN: 0016-6731. DOI: 10.1534/genetics.107.075424.

Ming, R., Bendahmane, A., and Renner, S. S. (2011). Sex Chromosomes in Land Plants. en. *Annual Review of Plant Biology* 62(1), 485–514. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-042110-103914.

Ming, R. et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). eng. *Nature* 452(7190), 991–996. ISSN: 1476-4687. DOI: 10.1038/nature06856.

Al-Mssallem, I. S. et al. (2013). Genome sequence of the date palm Phoenix dactylifera L. eng. *Nature Communications* 4, 2274. ISSN: 2041-1723. DOI: 10.1038/ncomms3274.

Muller, H. J. (1964). The relation of recombination to mutational advance. eng. *Mutation Research* 106, 2–9. ISSN: 0027-5107.

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., and Marais, G. A. B. (2012). Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. eng. *PLoS biology* 10(4), e1001308. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001308.

Navajas-Pérez, R., Herrán, R. de la, López González, G., Jamilena, M., Lozano, R., Ruiz Rejón, C., Ruiz Rejón, M., and Garrido-Ramos, M. A. (2005). The evolution of reproductive systems and sex-determining mechanisms within rumex (polygonaceae) inferred from nuclear and chloroplastidial sequence data. eng. *Molecular Biology and Evolution* 22(9), 1929–1939. ISSN: 0737-4038. DOI: 10.1093/molbev/msi186.

Okada, S., Sone, T., Fujisawa, M., Nakayama, S., Takenaka, M., Ishizaki, K., Kono, K., Shimizu-Ueda, Y., Hanajiri, T., Yamato, K. T., Fukuzawa, H., Brennicke, A., and

Ohyama, K. (2001). The Y chromosome in the liverwort Marchantia polymorpha has accumulated unique repeat sequences harboring a male-specific gene. eng. *Proceedings of the National Academy of Sciences of the United States of America* 98(16), 9454–9459. ISSN: 0027-8424. DOI: 10.1073/pnas.171304798.

Peck, J. R. (1994). A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. eng. *Genetics* 137(2), 597–606. ISSN: 0016-6731.

Perrin, N. (2009). Sex reversal: a fountain of youth for sex chromosomes? eng. *Evolution; International Journal of Organic Evolution* 63(12), 3043–3049. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2009.00837.x.

Pessia, E., Engelstädter, J., and Marais, G. A. B. (2014). The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? eng. *Cellular and molecular life sciences: CMLS* 71(8), 1383–1394. ISSN: 1420-9071. DOI: 10.1007/s00018-013-1499-6.

Pessia, E., Makino, T., Bailly-Bechet, M., McLysaght, A., and Marais, G. A. B. (2012). Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(14), 5346–5351. ISSN: 1091-6490. DOI: 10.1073/pnas.1116763109.

Picq, S., Santoni, S., Lacombe, T., Latreille, M., Weber, A., Ardisson, M., Ivorra, S., Maghradze, D., Arroyo-Garcia, R., Chatelet, P., This, P., Terral, J.-F., and Bacilieri, R. (2014). A small XY chromosomal region explains sex determination in wild dioecious V. vinifera and the reversal to hermaphroditism in domesticated grapevines. eng. *BMC plant biology* 14, 229. ISSN: 1471-2229. DOI: 10.1186/s12870-014-0229-z.

Prince, E. G., Kirkland, D., and Demuth, J. P. (2010). Hyperexpression of the X chromosome in both sexes results in extensive female bias of X-linked genes in the flour beetle. eng. *Genome Biology and Evolution* 2, 336–346. ISSN: 1759-6653. DOI: 10.1093/gbe/evq024.

Prochnik, S. E., Umen, J., Nedelcu, A. M., Hallmann, A., Miller, S. M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L. K., Hellsten, U., Chapman, J., Simakov, O., Rensing, S. A., Terry, A., Pangilinan, J., Kapitonov, V., Jurka, J., Salamov, A., Shapiro, H., Schmutz, J., Grimwood, J., Lindquist, E., Lucas, S., Grigoriev, I. V., Schmitt, R., Kirk, D., and Rokhsar, D. S. (2010). Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. eng. *Science (New York, N.Y.)* 329(5988), 223–226. ISSN: 1095-9203. DOI: 10.1126/science.1188800.

Pucholt, P., Rönnberg-Wästljung, A.-C., and Berlin, S. (2015). Single locus sex determination and female heterogamety in the basket willow (Salix viminalis L.) ENG. *Heredity*. ISSN: 1365-2540. DOI: 10.1038/hdy.2014.125.

Qiu, S., Bergero, R., and Charlesworth, D. (2013). Testing for the footprint of sexually antagonistic polymorphisms in the pseudoautosomal region of a plant sex chromosome

pair. eng. *Genetics* 194(3), 663–672. ISSN: 1943-2631. DOI: 10.1534/genetics.113.152397.

Qiu, S., Bergero, R., Forrest, A., Kaiser, V. B., and Charlesworth, D. (2010). Nucleotide diversity in Silene latifolia autosomal and sex-linked genes. eng. *Proceedings. Biological Sciences / The Royal Society* 277(1698), 3283–3290. ISSN: 1471-2954. DOI: 10.1098/rspb.2010.0606.

Rautenberg, A., Hathaway, L., Oxelman, B., and Prentice, H. C. (2010). Geographic and phylogenetic patterns in Silene section Melandrium (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. eng. *Molecular Phylogenetics and Evolution* 57(3), 978–991. ISSN: 1095-9513. DOI: 10.1016/j.ympev.2010.08.003.

Renner, S. S. (2014). The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. eng. *American Journal of Botany* 101(10), 1588–1596. ISSN: 1537-2197. DOI: 10.3732/ajb.1400196.

Rice, W. (1984). Sex-Chromosomes and the Evolution of Sexual Dimorphism. English. *Evolution* 38(4). WOS:A1984TJ95300003, 735–742. ISSN: 0014-3820. DOI: 10.2307/2408385.

Rice, W. (1987). The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination Between Primitive Sex-Chromosomes. English. *Evolution* 41(4). WOS:A1987J007200019, 911–914. ISSN: 0014-3820. DOI: 10.2307/2408899.

Robinson, K. M., Delhomme, N., Mähler, N., Schiffthaler, B., Onskog, J., Albrectsen, B. R., Ingvarsson, P. K., Hvidsten, T. R., Jansson, S., and Street, N. R. (2014). Populus tremula (European aspen) shows no evidence of sexual dimorphism. eng. *BMC plant biology* 14, 276. ISSN: 1471-2229. DOI: 10.1186/s12870-014-0276-5.

Schultheiß, R., Viitaniemi, H. M., and Leder, E. H. (2015). Spatial dynamics of evolving dosage compensation in a young sex chromosome system. ENG. *Genome Biology and Evolution*. ISSN: 1759-6653. DOI: 10.1093/gbe/evv013.

Segawa, M., Kishi, S., and Tatuno, S. (1971). Sex Chromosomes of Cycas-Revoluta. English. *Japanese Journal of Genetics* 46(1). WOS:A1971J337300004, 33–&. ISSN: 0021-504X. DOI: 10.1266/jjg.46.33.

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. eng. *Evolution; International Journal of Organic Evolution* 67(12), 3669–3677. ISSN: 1558-5646. DOI: 10.1111/evo.12223.

Smith, B. W. (1964). The evolving karyotype of *Rumex hastatulus*. *Evolution*, 93–104.

Sousa, A., Fuchs, J., and Renner, S. S. (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. eng. *Cytogenetic and Genome Research* 139(2), 107–118. ISSN: 1424-859X. DOI: 10.1159/000345370.

Spigler, R. B., Lewers, K. S., Main, D. S., and Ashman, T.-L. (2008). Genetic mapping of sex determination in a wild strawberry, Fragaria virginiana, reveals earliest form of sex chromosome. eng. *Heredity* 101(6), 507–517. ISSN: 1365-2540. DOI: 10.1038/hdy.2008.100.

Steflova, P., Tokan, V., Vogel, I., Lexa, M., Macas, J., Novak, P., Hobza, R., Vyskot, B., and Kejnovsky, E. (2013). Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY1Y2) in the dioecious plant Rumex acetosa. eng. *Genome Biology and Evolution* 5(4), 769–782. ISSN: 1759-6653. DOI: 10.1093/gbe/evt049.

Stettler, R. (1971). Variation in sex expression of black cottonwood and related hybrids. *Silvae Genet*, 42–46.

Telgmann-Rauber, A., Jamsari, A., Kinney, M. S., Pires, J. C., and Jung, C. (2007). Genetic and physical maps around the sex-determining M-locus of the dioecious plant asparagus. eng. *Molecular genetics and genomics: MGG* 278(3), 221–234. ISSN: 1617-4615. DOI: 10.1007/s00438-007-0235-z.

Toups, M., Veltsos, P., and Pannell, J. (2015). Plant sex chromosomes: lost genes with little compensation. *in press*.

Tuskan, G. A. et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). eng. *Science (New York, N.Y.)* 313(5793), 1596–1604. ISSN: 1095-9203. DOI: 10.1126/science.1128691.

Tuskan, G. A., DiFazio, S., Faivre-Rampant, P., Gaudet, M., Harfouche, A., Jorge, V., Labbé, J. L., Ranjan, P., Sabatti, M., Slavov, G., Street, N., Tschaplinski, T. J., and Yin, T. (2012). The obscure events contributing to the evolution of an incipient sex chromosome in Populus: a retrospective working hypothesis. en. *Tree Genetics & Genomes* 8(3), 559–571. ISSN: 1614-2942, 1614-2950. DOI: 10.1007/s11295-012-0495-6.

VanBuren, R., Zeng, F., Chen, C., Zhang, J., Wai, C. M., Han, J., Aryal, R., Gschwend, A. R., Wang, J., Na, J.-K., Huang, L., Zhang, L., Miao, W., Gou, J., Arro, J., Guyot, R., Moore, R. C., Wang, M.-L., Zee, F., Charlesworth, D., Moore, P. H., Yu, Q., and Ming, R. (2015). Origin and domestication of papaya Yh chromosome. eng. *Genome Research* 25(4), 524–533. ISSN: 1549-5469. DOI: 10.1101/gr.183905.114.

Veitia, R. A., Veyrunes, F., Bottani, S., and Birchler, J. A. (2015). X chromosome inactivation and active X upregulation in therian mammals: facts, questions, and hypotheses. eng. *Journal of Molecular Cell Biology* 7(1), 2–11. ISSN: 1759-4685. DOI: 10.1093/jmcb/mjv001.

Vicoso, B. and Bachtrog, D. (2011). Lack of global dosage compensation in Schistosoma mansoni, a female-heterogametic parasite. eng. *Genome Biology and Evolution* 3, 230–235. ISSN: 1759-6653. DOI: 10.1093/gbe/evr010.

Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S., and Bachtrog, D. (2013a). Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of

global dosage compensation. eng. *PLoS biology* 11(8), e1001643. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001643.

Vicoso, B., Kaiser, V. B., and Bachtrog, D. (2013b). Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. eng. *Proceedings of the National Academy of Sciences of the United States of America* 110(16), 6453–6458. ISSN: 1091-6490. DOI: 10.1073/pnas.1217027110.

Wang, J., Na, J.-K., Yu, Q., Gschwend, A. R., Han, J., Zeng, F., Aryal, R., VanBuren, R., Murray, J. E., Zhang, W., Navajas-Pérez, R., Feltus, F. A., Lemke, C., Tong, E. J., Chen, C., Wai, C. M., Singh, R., Wang, M.-L., Min, X. J., Alam, M., Charlesworth, D., Moore, P. H., Jiang, J., Paterson, A. H., and Ming, R. (2012). Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), 13710–13715. ISSN: 1091-6490. DOI: 10.1073/pnas.1207833109.

Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Wang, X., Zhang, J., and He, X. (2010). RNA sequencing shows no dosage compensation of the active X-chromosome. eng. *Nature Genetics* 42(12), 1043–1047. ISSN: 1546-1718. DOI: 10.1038/ng.711.

Yamato, K. T. et al. (2007). Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. eng. *Proceedings of the National Academy of Sciences of the United States of America* 104(15), 6472–6477. ISSN: 0027-8424. DOI: 10.1073/pnas.0609054104.

Yin, T., Difazio, S. P., Gunter, L. E., Zhang, X., Sewell, M. M., Woolbright, S. A., Allan, G. J., Kelleher, C. T., Douglas, C. J., Wang, M., and Tuskan, G. A. (2008). Genome structure and emerging evidence of an incipient sex chromosome in Populus. eng. *Genome Research* 18(3), 422–430. ISSN: 1088-9051. DOI: 10.1101/gr.7076308.

Part II

Results

Chapter 3

Identifying new sex-linked genes through BAC sequencing in the dioecious plant *Silene latifolia*

Chapter 2 showed the need for more sex-linked sequences in plants to allow one to study sex chromosomes. The current chapter reports a work aiming at obtaining more sex-linked genes in the plant *Silene latifolia* through the sequencing of BAC clones. The obtained sequences were used to study gene densities and gene loss in X and Y chromosomes of *Silene latifolia*.

This project mainly involved three labs: Roman Hobza and Alex Widmer's labs generated the sequences. Alex Widmer's lab assembled the data and annotated the genes. Gabriel Marais' lab coordinated the project, analysed the data and prepared the manuscript with inputs from all authors. I was in charge of comparing the BAC clone genes and the RNA-seq contigs from previous studies, which was used to assess the potential bias in inferring sex-linked genes from RNA-seq data and to estimate (roughly) gene loss (see authors contributions in Section 3.7).

The manuscript was accepted in *BMC genomics* pending minor revisions in April 2015. The revised version was resubmitted in May 2015.

Identifying new sex-linked genes through BAC sequencing in the dioecious plant *Silene latifolia*

Blavet N[1,2,8], Blavet H[2,3,8], Muyle A[4,8], Käfer J[4], Cegan R[3], Deschamps C[5], Zemp N[1], Mousset S[4], Aubourg S[6], Bergero R[7], Charlesworth D[7], Hobza R[2,3,9], Widmer A[1,9], Marais GAB[4,9]

3.1 Keywords

Sex chromosomes, sex-linked genes, plant, BAC, RNA-seq, gene loss, Y degeneration, *Silene latifolia, Silene vulgaris.*

Box 3.1 : List of abbreviations

- BAC = Bacterial Artificial Chromosome
- EST = Expressed Sequence Tag
- NUMT = Nuclear Mitochondrial DNA
- NUPT = Nuclear plastid DNA
- PAR = pseudoautosomal region
- RPKM = Reads per Kilobases per Million
- SNP = Single Nucleotide Polymorphism

3.2 Abstract

Background: *Silene latifolia* represents one of the best-studied plant sex chromosome systems. A new approach using RNA-seq data has recently identified hundreds of new sex-linked genes in this species. However, this approach is expected to miss genes that are either not expressed or are expressed at low levels in the tissue(s) used for RNA-seq. Therefore other independent approaches are needed to discover such sex-linked genes.

[1]Institute of Integrative Biology (IBZ), ETH Zurich, Zurich, Switzerland
[2]Institute of Experimental Botany AS CR, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc - Holice, Czech Republic
[3]Department of Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic
[4]Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France
[5]Pole Rhone-Alpes de Bioinformatique (PRABI), Villeurbanne, France
[6]Unité de Recherche en Gńomique Vǵale (UMR 1165), INRA/Universitd'Evry-Val-d'Essonne – ERL CNRS 8196, Evry, France
[7]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
[8]equal contribution as first authors
[9]equal contribution as senior authors

Results: Here we used 10 well-characterized *S. latifolia* sex-linked genes and their homologs in *Silene vulgaris*, a species without sex chromosomes, to screen BAC libraries of both species. We isolated and sequenced 4 Mb of BAC clones of *S. latifolia* X and Y and *S. vulgaris* genomic regions, which yielded to 59 new sex-linked genes (with *S. vulgaris* homologs for some of them).We assembled sequences that we believe represent the tip of the Xq arm. These sequences are clearly not pseudoautosomal, so we infer that the *S. latifolia* X has a single pseudoautosomal region (PAR) on the Xp arm. The estimated mean gene density in X BACs is 2.2 times lower than that in *S. vulgaris* BACs, agreeing with the genome size difference between these species. Gene density was found extremely low in the Y BAC clones. We compared our BAC-located genes to the sex-linked genes identified in previous RNA-seq studies, and found that about half of them (those with low expression in flower buds) were not identified as sex-linked in previous RNA-seq studies. We compiled a set of validated ~70 X/Y genes and X-hemizygous genes (without Y copies) from the literature, and used these genes to show that X-hemizygous genes have a higher probability of being undetected by the RNA-seq approach, compared with X/Y genes; we used this to estimate that about 30% of our BAC-located genes must be X-hemizygous. The estimate is similar when we use BAC-located genes that have *S. vulgaris* homologs, which excludes genes that were gained by the X chromosome.
Conclusions: Our BAC sequencing identified 59 new sex-linked genes, and our analysis of these BAC-located genes, in combination with RNA-seq data suggests that gene losses from the *S. latifolia* Y chromosome could be as high as 30%, higher than previous estimates of 10-20%.

3.3 Background

Of only a handful of plant sex chromosome systems that have been investigated at the molecular level, the XY chromosome system of *Silene latifolia* is one of the best-studied (Bernasconi et al., 2009; Ming et al., 2011). However, finding sex-linked genes in this species has been a slow process and is still ongoing. Approaches such as screening cDNA libraries with probes from microdissected *S. latifolia* Y chromosomes identified only a few sex-linked genes (reviewed in Filatov, 2005b). Segregation analysis of intron variants and SNPs within plant families revealed more sex-linked genes (e.g. Bergero et al., 2007, 2013). Combined, these approaches yielded about 30 validated *S. latifolia* sex-linked genes.

Recently, however, three studies used RNA-seq to identify hundreds of *S. latifolia* sex-linked genes, either using segregation patterns within families (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011) or male and female full siblings from an inbred population (Muyle et al., 2012). Sex-linked genes were identified either by following allele transmission from parents to their progeny (in the two studies using families, Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011), or by searching for SNPs homozy-

gous in females and heterozygous in males, indicating Y-linkage (Muyle et al., 2012). As no *S. latifolia* reference genome is available, these searches started with either a *de novo* assembled reference transcriptome using the *S. latifolia* RNA-seq data (Chibalina and Filatov, 2011; Muyle et al., 2012) or using 454 EST data from *S. vulgaris*, a close relative without sex chromosomes Bergero and Charlesworth, 2011; Sloan et al., 2012, to map the *S. latifolia* reads and perform SNP-calling. Both approaches are subject to errors, especially when sex-linkage of a contig is inferred from the segregation pattern of only a single SNP, so the inferences were assessed by checking for complete sex-linkage of some of the inferred sex-linked genes, using PCR on sets of unrelated males and females (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011). Further tests were done to check whether "tester sets" of well-validated sex-linked and autosomal genes (see above) were correctly assigned (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012). The results were encouraging, with most genes tested being correctly assigned. However, only a few newly inferred genes (~10 in each study) were checked experimentally, and the tester sets included only 10-20 sex-linked and ~10 autosomal genes. Moreover, the RNA-seq studies focused on RNA from only one tissue (flower buds) and any sex-linked genes not expressed in flower buds, or expressed at low levels, must be missed (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012).

The number of sex-linked genes in *S. latifolia* is therefore not yet accurately known. An alternative approach to discovering new sex-linked genes is to sequence BAC clones from the sex chromosome. A handful of BACs from the *S. latifolia* X and Y chromosomes have already been sequenced (e.g. Blavet et al., 2012; Ishii et al., 2010), but they yielded few new sex-linked genes. To improve the yield, we screened a BAC library with probes from validated X-linked or Y-linked genes of *S. latifolia*, which establishes sex-linkage of all genes found in the BAC sequences. Identifying both X-linked and Y-linked genes is important for estimating the proportion of X-linked genes that have lost their Y counterparts, indicating genetic degeneration of this plant sex chromosome system. Sequencing BACs should help identify genes with low expression levels, some of which were probably missed by previous studies, because most sex-linked genes identified so far in *S. latifolia* come from cDNA, ESTs or RNA-seq data, which will be enriched for highly expressed genes. Sequencing the complete *S. latifolia* sex chromosomes using BACs would be extremely costly as the X is 400 Mb, and the Y 550 Mb. However, BAC sequencing to obtain sequences of portions of the sex chromosomes is very useful. In particular, it can provide larger tester set to compare with results from RNA-seq studies (see above), as well as for analyses (explained below) for estimating changes in gene densities during the evolution of the X and Y chromosomes, and gene losses from the Y chromosome.

We obtained ~4 Mb of BAC sequences from the *S. latifolia* sex chromosomes and from *Silene vulgaris*, a closely related non-dioecious plant without sex chromosomes, in order to identify both new sex-linked genes and their *S. vulgaris* homologs, which can serve as outgroup sequences for comparing the evolution of *S. latifolia* X-linked and Y-linked

genes. A BAC library from an *S. latifolia* male was screened using probes specific for X-linked and Y-linked alleles of 10 previously validated X/Y gene pairs (see Section 3.6 and Table S1). Orthologs of all 10 genes have been identified in *S. vulgaris*, all mapping to a single linkage group (Bergero et al., 2013; Filatov, 2005a), indicating that they were all on the ancestral chromosome, and not gained during the evolution of the *S. latifolia* sex chromosomes. Their map locations in *S. latifolia* indicate that they represent both evolutionary strata (chromosomal regions with different levels of X-Y divergence, previously described for this species (Bergero et al., 2008a; Bergero et al., 2013), see also Figure 3.S1.A). Annotation of the BAC sequences yielded 49 new X-linked genes and 10 new Y-linked genes. We analysed the gene densities of the X-linked, Y-linked and *S. vulgaris* BACs. We also searched by Blast the previously published RNA-seq data with the sequences of the new sex-linked genes in the BACs, and used it to develop a new, combined approach to estimate Y gene loss. The results of our re-evaluation suggest that gene loss may have been underestimated based on RNA-seq alone, although more work is still needed to get a precise estimate of Y gene loss in *S. latifolia*.

3.4 Results and Discussion

3.4.1 Obtaining *S. latifolia* X and Y genomic sequences and identifying genes

A total of 25 positive BAC clones were selected and sequenced (see Section 3.6, Tables S1, S2). After further validation (see Section 3.6), 24 clones were retained for analysis. These included 6 triplets of X/Y/*vulgaris* sequences, one X/*vulgaris* pair, one Y/*vulgaris* pair, and two single X BAC clones without Y chromosome or *S. vulgaris* homologs (Table S1). The 16 sex-linked chromosomal fragments sequenced total ~2.5 Mb, the largest set of *S. latifolia* sex-linked genomic sequences so far obtained. These BAC sequences were assembled and annotated (see Section 3.6, Tables S1, S2), revealing a total of 153 genes, 78 of which are from *S. vulgaris*. Including the probe genes, the *S. latifolia* genes total 58 X-linked and 17 Y-linked genes (Table 3.1 and Tables S1, S2). 59 of them are newly identified in *S. latifolia*, tripling the number of *S. latifolia* fully sex-linked genes with complete genomic sequences; 49 of these 59 new sex-linked genes are X-linked, and 10 are Y-linked.

Table 3.1: Gene number and density in *S. latifolia* X and Y and in *S. vulgaris* BAC clones. Gene density was computed using all available BAC data. When only triplets are used, the results are similar.

	S. latifolia X	S. latifolia Y	S. vulgaris
Total of all genes, including genes used as "probes"	58	17	78
Total number of new genes	49	10	70
Total number of S. latifolia / S. vulgaris homologous gene pairs	13	2	-
Total physical size (Mb)	1.7	1.09	1.05
Gene density per Mb	34	16	74

An all-against-all Blast search among the BAC-located genes revealed conserved blocks

of several tens of Kb around each probe gene in the *S. latifolia* X and *S. vulgaris* BAC sequences (Figure 3.S2). These blocks include only 13 new X-*vulgaris* homologous gene pairs (Table 3.1 and Table S2). When aligned using MAUVE (Section 3.6), we found conserved gene orders in the blocks around the probe genes, and sequence similarities in the intergenic regions. In contrast, Blast searching found only two new Y-*vulgaris* gene pairs (Table 3.1 and Table S2), and MAUVE alignments found similarity between Y and *S. vulgaris* sequences mostly restricted to the probe gene itself (Figure 3.S2). This suggests the occurrence of insertions, deletions and other chromosomal rearrangements of the *S. latifolia* Y chromosome at a small (within BAC) scale, in addition to the large-scale rearrangements previously found (Bergero et al., 2008a,b; Cermak et al., 2008; Hobza et al., 2007, 2006; Kejnovsky et al., 2006; Macas et al., 2011; Pritham et al., 2003; Steflova et al., 2014).

To directly evaluate the extent of gene losses from the *S. latifolia* Y chromosome, we first searched for X/Y gene pairs (often called "gametologous pairs", in which X and Y genes are alleles that diverged since X-Y recombination became suppressed), where one is clearly recognizable as a pseudogene. We found no such pairs. All pseudogenes found in the BAC sequences were duplicates of other genes in the same BAC clone. The only X/Y gene pairs in our BAC sequences are the "probe" genes, which were already known (Table S2); none of the new X-linked genes have gametologs in the corresponding Y chromosome BAC sequence (Table S2).

3.4.2 Assembling BACs from the *X4*, *X7* and *X6a* regions and implications for the number of pseudoautosomal regions in *S. latifolia* sex chromosomes

We found overlaps between the X BAC sequences from three probes, genes *X4*, *X7* and *X6a*. These BAC sequences were therefore assembled into a scaffold (Figure 3.S1.B). The end of this scaffold (BAC clone BAC65P13) consists of X43.1 repeats typical of *Silene* telomeres (Sýkorová et al., 2003). These X43.1 repeats probably represent the X telomere, based on the following reasoning. BAC assembly and sequencing statistics indicate that 7% of reads in BAC65P13 are from X43.1, yielding an estimate that the X43.1 repeat forms a ~6 kb region of this BAC. No interstitial X.43.1 signal was detected on the X chromosome in previous work using FISH (Cermak et al., 2008), but a 6 kb sequence composed of units arranged in tandem should yield a clear fluorescent signal with the X43.1 probe. A non-telomeric location is therefore unlikely. Our results therefore suggest that we have reached the end of the Xq arm in *S. latifolia*. In turn, this implies that only the Xp end is pseudoautosomal. Our results are therefore consistent with the *S. latifolia* sex chromosomes having only a single pseudoautosomal region, and not two as AFLP mapping suggested (Scotti and Delph, 2006); a single pseudoautosomal region (PAR) is consistent with the latest genetic mapping (Bergero et al., 2013).

3.4.3 Gene densities in *S. latifolia* X, Y and *S. vulgaris* BAC clones

We found an average of 34 genes/Mb in the *S. latifolia* X BAC sequences and 74 genes/Mb in those from *S. vulgaris* (Table 3.1). The gene densities we observed in both species' BAC sequences are quite high, which suggests that we have sequenced gene-dense regions. The 2.2-fold lower gene density in the *S. latifolia* X is, however, consistent with the expectation based purely on the genome sizes of the two species (2.7 Gb for *S. latifolia* and the 1 Gb for *S. vulgaris*; see the Plant DNA C-value Database, `http://data.kew.org/cvalues/`). Assuming the same total number of genes in both species (which is likely as they are closely related species with an identical chromosome number of 2n = 24), and neglecting possible inter-chromosomal translocations in *S. latifolia* or *S. vulgaris* (Bergero et al., 2013), the relative total genome sizes predict a 2.7-fold lower gene density in *S. latifolia*.

In contrast, the *S. latifolia* Y BACs have an estimated average gene density of only 16 genes/Mb (Table 3.1), 2.1 times lower than the X. The *S. latifolia* Y chromosome is 550 Mb, considerably larger than the X (400 Mb; see Matsunaga et al., 1994). If the number of genes were the same on both sex chromosomes (that is, if their size difference is due solely to the accumulation on the Y of sequences not present on the X, including transposable elements, NUMTs and NUPTs Bergero et al., 2008b; Cermak et al., 2008; Kejnovsky et al., 2006; Pritham et al., 2003; Steflova et al., 2013), the ratio of gene densities for Y versus X should be the same as the ratio of Y/X chromosome sizes, 550/400, predicting a mean Y density 1.4 times lower than that of the X. The observed value in the *S. latifolia* Y BAC sequences is nevertheless considerably lower than the expectation, and suggests losses of as much as 34% of genes from the Y.

3.4.4 Searching for the BAC-located genes in RNA-seq data

We blasted our BAC-located genes to the RNA-seq contigs found in previous studies (see Section 3.6), which produced significant matches for 54 out of 63 genes (Table 3.2 and Table 3.S3), showing that most of our BAC-located genes (~85%) are expressed in flower buds. Only half of these genes were identified as sex-linked by any of the previous studies (Table 3.2). As predicted (see Section 3.3) the genes not detected as sex-linked in any of the RNA-seq studies have much lower expression levels (as estimated by Muyle et al., 2012) than those where sex-linkage was detected (RPKM values: 3008.3 versus 11251.2, respectively; the difference is significant by a one-tailed Student's *t* test, p-value = 0.004). This suggests that failure to ascertain genes as sex-linked when they have low expression affects inferences using RNA-seq, in addition to absence of expression of some genes in flower buds.

Table 3.2: Comparison of BAC and RNA-seq data All BAC-located genes are included except the 6 probe genes for which both X and Y copies were available in the BAC data.

	Total number	BAC-located genes matching RNA-seq contigs	BAC-located genes matching X/Y RNA-seq contigs	BAC-located genes matching X-hemizygous RNA-seq contigs	Sources of RNA-seq data
X-linked BAC-located gene	52	44	12	-	Muyle et al., 2012
		36	5	2	Bergero and Charlesworth, 2011
		31	13	5	Chibalina and Filatov, 2011
		46	19	5	3 combined
Y-linked BAC-located gene	11	7	3	-	Muyle et al., 2012
		6	1	1	Bergero and Charlesworth, 2011
		6	1	3[1]	Chibalina and Filatov, 2011
		8	3	1	3 combined
All BAC-located genes	63	54	22	6	3 combined

3.4.5 Re-evaluating Y gene loss using both BAC and RNA-seq data

Two RNA-seq studies have used X-linked genes to estimate Y gene loss in *S. latifolia*. Only 10 to 20% of X-linked genes were estimated to have no Y transcripts, suggesting that Y degeneration and male hemizygosity may be modest in *S. latifolia* (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011). Correct inference of X-hemizygous genes is critical for reliably estimating Y gene loss. If the Y copy of a XY gene pair is not expressed, or is expressed at low levels in the tissue(s) used for RNA-seq analysis, hemizygosity will be incorrectly inferred and gene losses from the Y will be overestimated. We found some examples of this when comparing the BAC and RNA-seq data (using stringent Blast criteria, see Section 3.6). Two BAC-located genes matched contigs inferred as X/Y gene pairs from one study but with contigs inferred as X-hemizygous in others, and one Y-linked gene matched a contig inferred as X-hemizygous (Table 3.2).

Among our X-linked BAC-located genes, five matched contigs inferred to be X-hemizygous (Table 3.2). Using our BAC-located genes that match RNA-seq contigs detected as sex-linked, this yields an estimate of 20% of Y gene loss, the same as in the published RNA-seq studies (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011). However, if coverage is low due to a low expression level, SNPs may not be identified; individuals cannot then be genotyped and no inferences about sex-linkage are possible. Recent data

[1]Among those 3, two genes were found to be X-hemizygous in Chibalina and Filatov (2011), and XY in Bergero and Charlesworth (2011) and Muyle et al. (2012). In the combined data (see details in Section 3.6), we considered these genes to be XY.

from animals suggests that average expression levels are lower for X-hemizygous genes than for X/Y gene pairs (Bellott et al., 2014; Cortez et al., 2014), and therefore the RNA-seq approach may fail to detect X-hemizygous genes more often than X/Y gene pairs, resulting in an underestimation of gene losses from the Y. If this bias occurs, the BAC-located genes not matching contigs inferred as sex-linked should include more X-hemizygous genes than the ~20% estimate above.

To evaluate this possibility, it would be helpful to have an estimate of the proportion of X-hemizygous genes that were undetected by the RNA-seq studies. When these studies were done, very few validated X-hemizygous genes were available in *S. latifolia*. Only two fully degenerated Y-linked genes in *S. latifolia* have so far been documented (Guttman and Charlesworth, 1998; Kazama et al., 2012). Two recent studies used segregation analysis in large families and inferred further X-hemizygous genes, one being an unpublished segregation analysis using RadSeq data (Bergero et al., 2013, Qiu S, Charlesworth D, unpublished); however a comparison with the sex-linked contigs from RNA-seq studies reveals that ~57% may be X/Y gene pairs and not truly X-hemizygous (see the list of genes with X-hemizygous segregation patterns in Table S4). X-hemizygous genes from (Bergero et al., 2013, Qiu S, Charlesworth D, unpublished) cannot therefore be used to estimate the proportion of X-hemizygous genes that failed to be detected by the RNA-seq studies.

We therefore used an indirect approach. Many well-validated X/Y gene pairs are now available, and can be used to estimate the probability that the combined RNA-seq studies fail to detect such a gene pair. Given this estimate, one can infer how many of the BAC-located genes that do not match sex-linked RNA-seq contigs could represent such missed X/Y gene pairs, and thus how many are probably truly X-hemizygous genes (schematized in Figure 3.S3). For the required estimate, we used all published well-validated X/Y gene pairs: the 17 experimentally validated ones (see references in Table S4), 20 sex-linked contigs from RNA-seq studies that were validated by PCR (Bergero and Charlesworth, 2011), and 12 more from a recent segregation analysis (Bergero et al., 2013). All these are probably highly expressed genes. We added 21 more X/Y gene pairs from the unpublished RadSeq study (Qiu S, Charlesworth D, unpublished), which uses genomic DNA, and can therefore ascertain genes even if their expression levels are low, for a total of 70 tester genes that were previously inferred as sex-linked. 78% of these genes had significant matches with contigs from at least one of the three RNA-seq studies, implying that they are expressed in flower buds. Genes matching contigs not assigned as sex-linked in one study often matched sex-linked ones in another, so that only around 25% of true X/Y gene pairs remained undetected in the three RNA-seq studies combined (Table S4).

This estimated proportion suggests that, out of our total number of 46 X-linked BAC-located genes expressed in flower buds, 0.25*46 = 11.5 are probably X/Y gene pairs undetected in the combined RNA-seq data. Thus, 11.5 of the 22 BAC-located genes not matching sex-linked RNA-seq contigs (category (iii) in Table 3.3) are accounted for. This leaves 22 – 11.5 = 10.5 genes that are probably X-hemizygous, but failed to be detected

by the RNA-seq studies. Only X-linked genes newly ascertained by our BAC sequencing are "ancestral" genes relevant for estimating gene losses (the probe genes were ascertained through detecting Y-linked variants, and were therefore previously known to have Y copies); there were probably 50 "ancestral" genes in our BAC sequences, 43 X BAC-located genes that lack copies in our Y BACs but have RNA-seq matches, plus the 7 Y-only BAC-located genes with RNA-seq matches (the total is 60 including the probe genes). The estimated number of Y gene losses is then as follows: 5 genes detected as X-hemizygous (category (ii) in Table 3.3) + 10.5 X-hemizygous genes that failed to be detected by the RNA-seq studies (see above). Dividing by 50 ancestral genes yields 33% (or 26% including the probe genes, Table 3.3). Using a similar approach to estimate gene losses from the X chromosome gives a considerably lower fraction, 5% (or 3% including the probe genes), significantly different from the estimate for the Y (Table 3.3, Fisher's exact test p-values $< 10^{-3}$ in either case). Estimates of ancestral gene numbers are particularly reliable when an outgroup is used to exclude genes that were gained after the sex chromosomes originated, by duplication and/or relocation onto the X. We therefore repeated this analysis, restricting it to genes with homologs on the *S. vulgaris* BAC sequences (which must have been present on the ancestral proto-sex chromosomes). The results are similar; excluding the "probe" genes, we estimate 34% gene loss from the Y, and none from the X (Fisher's exact test p-value = 0.003; see Table 3.S5, or, including the "probe" genes, 22% and 0% Y and X gene loss, respectively; Fisher's exact test p-value < 0.05).

Table 3.3: Analysis of gene loss in X and Y chromosomes combining BAC and RNA-seq data

Categories of genes	X-linked genes	Y-linked genes
All new genes in BAC sequences	52	11
No match to RNA-seq contigs	6	3
Genes retained for analysis	46	8
Category (i): XY results in RNA-seq analysis	19	3
Category (ii): X-hemizygous results in RNA-seq analysis	5	1
Category (iii): Not ascertained as sex-linked by RNA-seq analysis	22	4
Estimated X/Y false negative rate for gene pairs for RNA-seq analysis[1]	25%	25%
Expected number of XY pairs undetected in RNA-seq analysis	11.5	2
Potential number of X-hemizygous (X0) or Y0 genes undetected in RNA-seq analysis	10.5	2
Potential total number of X-hemizygous (X0) or Y0 genes (sum of detected + undetected in RNA-seq analysis numbers above)	15.5	2
Potential proportion of X-hemizygous (X0) or Y0 genes[2]	26-33%	3-5%

Correct estimation of the proportion of X-hemizygous genes among the BAC-located genes depends of the representativeness of the X/Y gene pairs used as tester set. To further check our set of inferred X-hemizygous genes, we searched for genes that were

[1] Based on 39 genes previously known to have X- and Y-linked copies, see Table S4.
[2] Potential total number of genes absent/number of ancestral genes (including or not the probe genes), see text for details.

wrongly classified as X-hemizygous, but which were actually X/Y gene pairs whose sequences are so diverged that they assembled into different contigs, one of which (the Y contig) was not detected. RNA-seq contigs representing the Y copies of these X-hemizygous genes should be found only in males. To test for such sequences among the RNA-seq contigs, we examined the BAC-located genes that the published RNA-seq analyses did not ascertain as sex-linked by blasting them against a set of RNA-seq contigs that were found only in males (from Muyle et al., 2012). This yielded only between 3 and 5 significant matches (depending on the filtering of the RNA-seq data, see Section 3.6). Thus, very few potentially divergent Y copies are present among the RNA-seq contigs; moreover, some of the male-specific contigs may not represent divergent Y copies but may simply be autosomal paralogs specifically expressed in males. The lack of evidence for the existence of many undetected X/Y gene pairs with diverged Y-linked copies agrees with our estimate that no more than 10 of the genes not ascertained as sex-linked by RNA-seq analysis are actually X/Y gene pairs (Table 3.3).

3.5 Conclusions

Our BAC sequencing effort resulted in 59 new validated sex-linked genes in *S. latifolia*, adding to the 43 published ones already available (listed in Table S4). Comparing our new genes to sex-linked genes identified by RNA-seq studies shows that failure to ascertain genes as sex-linked when they have low expression is an important limitation of RNA-seq, in addition to non-expression in the flower bud tissues that have been used, illustrating the difficulty of reliably inferring sex-linkage, X-hemizygosity and gene loss from the Y chromosome without a reference genome. Analyses to take this ascertainment bias into account suggest that gene losses from the *S. latifolia* Y could be higher than previously thought, perhaps around 30%, consistent with the gene densities in X/Y and *S. vulgaris* BACs. However, further work is needed to estimate Y gene loss in this species more precisely.

3.6 Methods

3.6.1 Isolation and sequencing of BAC clones

The BAC library was screened following (Cegan et al., 2010). Clones were gridded on nylon membrane filters and hybridized. The *S. latifolia* BAC library includes a total of 119,808 clones, with an average insert-size of 128 kb, which equates to 5.3 times the male haploid genome. The *S. vulgaris* BAC library (total of 55,296 clones), with an average insert-size of 110 kb, represents 6.8 haploid genomes of this species. We screened these libraries using probes designed from 10 published sex-linked genes and their homologs in *S. vulgaris* (shown in Figure 3.S1.A, plus the triplet SlAP3X/Y-SvAP3). For each “probe”

gene, the X-linked copy was used to screen the *S. latifolia* BAC library, and the Y copy to identify Y-linked BAC clones in the *S. latifolia* BAC library, while the *S. vulgaris* homolog was used to identify *S. vulgaris* BAC clones. For each probe, we found 1 to >100 positive clones. We selected clones showing strong hybridization with the probe, and only those that were confirmed by PCR with probe-derived primers were used in further analyses. Whenever possible, we sequenced one BAC clone for each probe gene. These clones were sequenced with coverage varying from 5-6 to 8-600 X for Sanger and 454, respectively (some clones with mate-pairs, and some without). The BAC sequences were validated by comparing the sequence of the "probe" gene from the BAC to the published sequence of the "probe" gene; this excluded only one BAC clone. This yielded complete triplets of X, Y and *S. vulgaris* BACs for some probe genes, but not all (Table S1). All the "probe" genes except *SlAP3* have already been mapped on the *S. latifolia* X chromosomes (Bergero et al., 2007, 2013), and their Y copies have been mapped on Y chromosome physical maps, see (Bergero et al., 2008a). All the BAC contigs are available in Genbank (Accession numbers KC978922-KC977838). Table S1 provides more details.

3.6.2 Assembly and annotation of BAC sequences

For each BAC clone, the reads were assembled *de novo* using Newbler v.2.5.3 (2010), except for three BAC clones sequenced using Sanger sequencing (19P24, 93L17 and 78D08), which were assembled with phrap v.16 (2007). The assembly statistics in Table S1 were obtained using QUAST (Gurevich et al., 2013). Annotation (see Table S2) was done using both homology-based and expression-data-based strategies using Uniprot and *S. latifolia* RNA-seq data from (Muyle et al., 2012). Truncated genes and genes with premature stop codons and/or frameshifts were annotated as pseudogenes. DNA repeats (including transposable elements) were annotated using the latest update of the database of DNA repeats in *S. latifolia*, based on an extensive search using genomic library screening and low coverage sequencing of the *S. latifolia* data (Cermak et al., 2008; Macas et al., 2011).

3.6.3 Sequence analysis

Homology among BAC clones from the same X/Y probe gene pair was assessed by aligning the BAC sequences with MAUVE 2.3.1 (Darling et al., 2010) after masking the repeats using RepeatMasker v3.3.0 (http://www.repeatmasker.org/) with the *Silene* DNA repeat database mentioned above. Homology between X/Y BAC pairs was also assessed by performing an all-against-all Blast search (with the default parameters) among the genes found in the X/Y BAC pair. The results are shown in Figure 3.S2, and the *X-vulgaris* and *Y-vulgaris* pairs that we found are listed in Table 3.S6.

To obtain the results shown in Table 2, we performed a Blast search of all coding sequences (CDS, obtained by annotating the BAC sequences, see previous section) against the RNA-seq data from the three previous studies (Bergero and Charlesworth, 2011; Chi-

balina and Filatov, 2011; Muyle et al., 2012) using data available in Genbank (Chibalina and Filatov, 2011) and our own data (Bergero and Charlesworth, 2011; Muyle et al., 2012). We retained only manually checked Blast hits with e-values < 10^{-5}, % identities > 90%, and alignment lengths > 50 bp. Multiple corresponding RNA-seq contigs were allowed for a single BAC CDS to allow for assembly problems in the RNA-seq data. The three RNA-seq studies were then combined to infer each CDS gene as being X/Y, X-hemizygous, or not detected as sex-linked in the RNA-seq data (Table S2). A gene was classified as X/Y in RNA-seq data if any one of the matching RNA-seq contigs was classified as X/Y, and as X-hemizygous if it satisfied two criteria: (i) at least one matching RNA-seq contig was classified as X-hemizygous, and (ii) all other matching RNA-seq contigs were not classified as XY gene pairs. Finally, the gene was classified as not having been detected as sex-linked in RNA-seq data whenever all matching RNA-seq contigs failed to be detected as sex-linked. Expression level estimates were obtained from (Muyle et al., 2012).
To check our X-hemizygous genes, we blasted them all (including those detected as X-hemizygous in the RNA-seq studies) against a set of RNA-seq contigs expressed only in males (using data from Muyle et al., 2012). Some of these genes might correspond to sex-linked genes with highly diverged X and Y copies that assemble in separate RNA-seq contigs and might therefore be wrongly classified as X-hemizygous, or not be detected as sex-linked at all. To test for potentially Y-linked sequences, we used a set of male-specific contigs from the RNA-seq results. We required these contigs to be expressed in all males and none of the females, using (i) all male-specific contigs, N = 5,504 (ii) male-specific contigs without matches to any transposable element sequence (using the *S. latifolia* TE database mentioned above) and with more than 10 mapped reads in one of the libraries (to remove noisy expression), N = 3,400. Only sequences with Blast hits of length > 100 bp, e-values < 10^{-4}, scores > 80 and identities > 80% were retained.
Fisher's exact tests and Student's *t* tests were done using the relevant statistical functions in R (http://www.r-project.org/).

3.7 Authors' contributions

NB assembled and annotated the BACs and identified the orthologous genes between X, Y and *S. vulgaris* sequences, HB selected and sequenced the BACs, AM did the comparison between the BACs and RNA-seq data and all associated analyses, JK analyzed the CDS of the BAC-located genes, RD contributed to the annotation of the BACs (in particular, the transposable elements), CD performed the alignments of the X/Y/*vulgaris* BACs, NZ performed some validation of the X-hemizygous genes inferred in this study, SA contributed to the annotation (in particular, finding coding regions and pseudogenes), SM set the statistical framework of the study and performed some of the tests, RB provided one of the RNA-seq data, DC provided one of the RNA-seq data and contributed to the writing, RH contributed to the design of the study, supervised the sequencing and assembly of BACs

and contributed to the writing, AW contributed to the design of the study, supervised the sequencing and assembly of BACs and contributed to the writing, GM contributed to the design of the study, supervised the data analysis, and coordinated the writing. All authors read and approved the final manuscript.

3.8 Acknowledgements

This work was supported by the French National Research Agency (grant numbers ANR-08-JCJC-0109, ANR-11-BSV7-013-03 to G.A.B.M.); the Czech Science Foundation (grant numbers P501/12/2220, 522/09/0083 to R.H.); ETH Zurich (grant number TH-07 06-3 to A.W) and the Swiss National Foundation (grant number 31003A-116455 to A.W.); Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07-/2.3.00/20.0165 to N.B.).

3.9 References

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T.-J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., Kremitzki, C., Fulton, R. S., Dugan, S., Ding, Y., Morton, D., Khan, Z., Lewis, L., Buhay, C., Wang, Q., Watt, J., Holder, M., Lee, S., Nazareth, L., Alföldi, J., Rozen, S., Muzny, D. M., Warren, W. C., Gibbs, R. A., Wilson, R. K., and Page, D. C. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. eng. *Nature* 508(7497), 494–499. ISSN: 1476-4687. DOI: 10.1038/nature13206.

Bergero, R., Charlesworth, D., Filatov, D. A., and Moore, R. C. (2008a). Defining regions and rearrangements of the Silene latifolia Y chromosome. eng. *Genetics* 178(4), 2045–2053. ISSN: 0016-6731. DOI: 10.1534/genetics.107.084566.

Bergero, R., Forrest, A., and Charlesworth, D. (2008b). Active miniature transposons from a plant genome and its nonrecombining Y chromosome. eng. *Genetics* 178(2), 1085–1092. ISSN: 0016-6731. DOI: 10.1534/genetics.107.081745.

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Forrest, A., Kamau, E., and Charlesworth, D. (2007). Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: evidence from new sex-linked genes. eng. *Genetics* 175(4), 1945–1954. ISSN: 0016-6731. DOI: 10.1534/genetics.106.070110.

Bergero, R., Qiu, S., Forrest, A., Borthwick, H., and Charlesworth, D. (2013). Expansion of the pseudo-autosomal region and ongoing recombination suppression in the Silene latifolia sex chromosomes. eng. *Genetics* 194(3), 673–686. ISSN: 1943-2631. DOI: 10.1534/genetics.113.150755.

Bernasconi, G., Antonovics, J., Biere, A., Charlesworth, D., Delph, L. F., Filatov, D., Giraud, T., Hood, M. E., Marais, G. a. B., McCauley, D., Pannell, J. R., Shykoff, J. A., Vyskot, B., Wolfe, L. M., and Widmer, A. (2009). Silene as a model system in ecology and evolution. eng. *Heredity* 103(1), 5–14. ISSN: 1365-2540. DOI: 10.1038/hdy.2009.34.

Blavet, N., Blavet, H., Cegan, R., Zemp, N., Zdanska, J., Janoušek, B., Hobza, R., and Widmer, A. (2012). Comparative analysis of a plant pseudoautosomal region (PAR) in Silene latifolia with the corresponding S. vulgaris autosome. eng. *BMC genomics* 13, 226. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-226.

Cegan, R., Marais, G. A., Kubekova, H., Blavet, N., Widmer, A., Vyskot, B., Dolezel, J., Safár, J., and Hobza, R. (2010). Structure and evolution of Apetala3, a sex-linked gene in Silene latifolia. eng. *BMC plant biology* 10, 180. ISSN: 1471-2229. DOI: 10.1186/1471-2229-10-180.

Cermak, T., Kubat, Z., Hobza, R., Koblizkova, A., Widmer, A., Macas, J., Vyskot, B., and Kejnovsky, E. (2008). Survey of repetitive sequences in Silene latifolia with respect to their distribution on sex chromosomes. eng. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 16(7), 961–976. ISSN: 1573-6849. DOI: 10.1007/s10577-008-1254-2.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. eng. *Nature* 508(7497), 488–493. ISSN: 1476-4687. DOI: 10.1038/nature13151.

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. eng. *PloS One* 5(6), e11147. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0011147.

Filatov, D. A. (2005a). Evolutionary history of Silene latifolia sex chromosomes revealed by genetic mapping of four genes. eng. *Genetics* 170(2), 975–979. ISSN: 0016-6731. DOI: 10.1534/genetics.104.037069.

Filatov, D. A. (2005b). Isolation of genes from plant Y chromosomes. eng. *Methods in Enzymology* 395, 418–442. ISSN: 0076-6879. DOI: 10.1016/S0076-6879(05)95023-4.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. eng. *Bioinformatics (Oxford, England)* 29(8), 1072–1075. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt086.

Guttman, D. S. and Charlesworth, D. (1998). An X-linked gene with a degenerate Y-linked homologue in a dioecious plant. eng. *Nature* 393(6682), 263–266. ISSN: 0028-0836. DOI: 10.1038/30492.

Hobza, R., Kejnovsky, E., Vyskot, B., and Widmer, A. (2007). The role of chromosomal rearrangements in the evolution of Silene latifolia sex chromosomes. eng. *Molecular genetics and genomics: MGG* 278(6), 633–638. ISSN: 1617-4615. DOI: 10.1007/s00438-007-0279-0.

Hobza, R., Lengerova, M., Svoboda, J., Kubekova, H., Kejnovsky, E., and Vyskot, B. (2006). An accumulation of tandem DNA repeats on the Y chromosome in Silene latifolia during early stages of sex chromosome evolution. eng. *Chromosoma* 115(5), 376–382. ISSN: 0009-5915. DOI: 10.1007/s00412-006-0065-5.

Ishii, K., Amanai, Y., Kazama, Y., Ikeda, M., Kamada, H., and Kawano, S. (2010). Analysis of BAC clones containing homologous sequences on the end of the Xq arm and on chromosome 7 in the dioecious plant Silene latifolia. eng. *Genome / National Research Council Canada = Génome / Conseil National De Recherches Canada* 53(4), 311–320. ISSN: 1480-3321. DOI: 10.1139/g10-008.

Kazama, Y., Nishihara, K., Bergero, R., Fujiwara, M. T., Abe, T., Charlesworth, D., and Kawano, S. (2012). SlWUS1; an X-linked gene having no homologous Y-linked copy in Silene latifolia. eng. *G3 (Bethesda, Md.)* 2(10), 1269–1278. ISSN: 2160-1836. DOI: 10.1534/g3.112.003749.

Kejnovsky, E., Kubat, Z., Hobza, R., Lengerova, M., Sato, S., Tabata, S., Fukui, K., Matsunaga, S., and Vyskot, B. (2006). Accumulation of chloroplast DNA sequences on the Y chromosome of Silene latifolia. eng. *Genetica* 128(1-3), 167–175. ISSN: 0016-6707. DOI: 10.1007/s10709-005-5701-0.

Macas, J., Kejnovský, E., Neumann, P., Novák, P., Koblížková, A., and Vyskot, B. (2011). Next generation sequencing-based analysis of repetitive DNA in the model dioecious [corrected] plant Silene latifolia. eng. *PloS One* 6(11), e27335. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027335.

Matsunaga, S., Hizume, M., Kawano, S., and Kuroiwa, T. (1994). Cytological Analyses in *Melandrium album*: Genome Size, Chromosome Size and Fluorescence *in situ* Hybridization. *Cytologia* 59(1), 135–141. DOI: 10.1508/cytologia.59.135.

Ming, R., Bendahmane, A., and Renner, S. S. (2011). Sex Chromosomes in Land Plants. en. *Annual Review of Plant Biology* 62(1), 485–514. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-042110-103914.

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., and Marais, G. A. B. (2012). Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. eng. *PLoS biology* 10(4), e1001308. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001308.

Pritham, E. J., Zhang, Y. H., Feschotte, C., and Kesseli, R. V. (2003). An Ac-like transposable element family with transcriptionally active Y-linked copies in the white campion, Silene latifolia. eng. *Genetics* 165(2), 799–807. ISSN: 0016-6731.

Scotti, I. and Delph, L. F. (2006). Selective trade-offs and sex-chromosome evolution in Silene latifolia. eng. *Evolution; International Journal of Organic Evolution* 60(9), 1793–1800. ISSN: 0014-3820.

Sloan, D. B., Keller, S. R., Berardi, A. E., Sanderson, B. J., Karpovich, J. F., and Taylor, D. R. (2012). De novo transcriptome assembly and polymorphism detection in the flowering plant Silene vulgaris (Caryophyllaceae). eng. *Molecular Ecology Resources* 12(2), 333–343. ISSN: 1755-0998. DOI: 10.1111/j.1755-0998.2011.03079.x.

Steflova, P., Hobza, R., Vyskot, B., and Kejnovsky, E. (2014). Strong accumulation of chloroplast DNA in the Y chromosomes of Rumex acetosa and Silene latifolia. eng. *Cytogenetic and Genome Research* 142(1), 59–65. ISSN: 1424-859X. DOI: 10.1159/000355212.

Steflova, P., Tokan, V., Vogel, I., Lexa, M., Macas, J., Novak, P., Hobza, R., Vyskot, B., and Kejnovsky, E. (2013). Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY1Y2) in the dioecious plant Rumex acetosa. eng. *Genome Biology and Evolution* 5(4), 769–782. ISSN: 1759-6653. DOI: 10.1093/gbe/evt049.

Sýkorová, E., Cartagena, J., Horáková, M., Fukui, K., and Fajkus, J. (2003). Characterization of telomere-subtelomere junctions in Silene latifolia. eng. *Molecular genetics and genomics: MGG* 269(1), 13–20. ISSN: 1617-4615. DOI: 10.1007/s00438-003-0811-9.

3.10 Supplementary information

Large Tables available online:

Table S1: https://drive.google.com/open?id=0B7KhiXOz6_OLM2Q5MHdDTWdJcU0&authuser=

Table S2: https://drive.google.com/open?id=0B7KhiXOz6_OLVjI4cDdYSzhtaTg&authuser=0

Table S4: https://drive.google.com/open?id=0B7KhiXOz6_OLUjM4cVUwNmJ3OWc&authuser=0

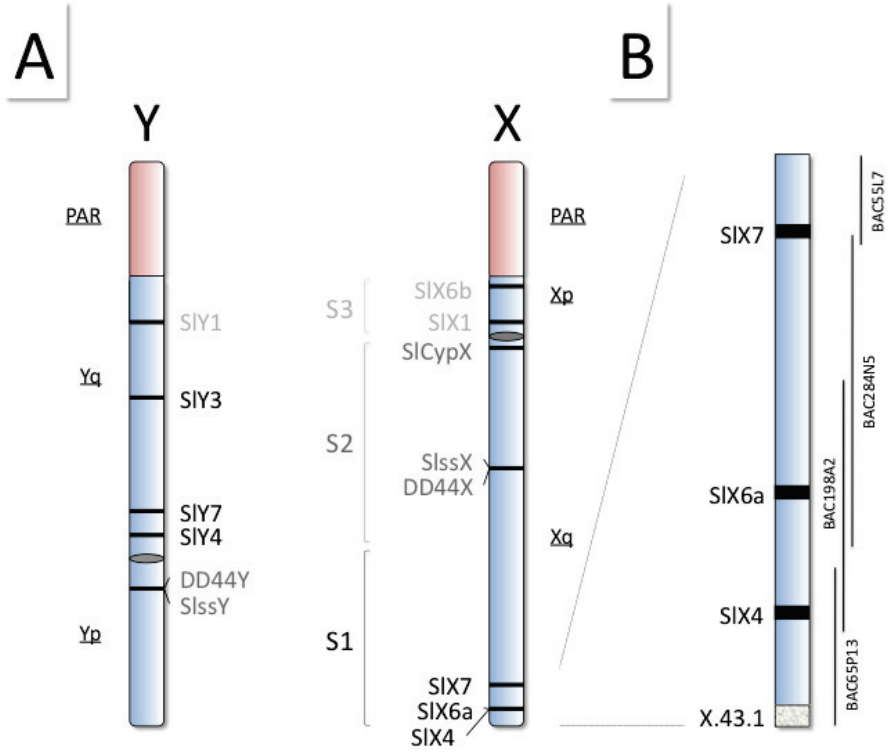


Figure 3.S1: BAC clones A) Localization of the BAC clones: on the *S. latifolia* X and Y chromosomes. Schematic view of the X chromosome genetic map (adapted from Bergero et al., 2007, 2013) and of the Y chromosome deletion map (adapted from Bergero et al., 2008a). The three strata (S1, S2, S3) as defined in Bergero et al. (2008a) are shown on the X chromosome. The positions of the X-linked and Y-linked genes used as probes to screen the *S. latifolia* male BAC library are shown. *SlAP3X* was also used as a probe but mapping data are not available for this gene. Selection and validation of several BAC clones was unsuccessful, which explains why we do not have just BAC triplets (Section 3.6). See Table S1 for a complete list of the sequenced BACs. **B)** Assembly of the Xq arm. The BACs including *SlX7, SlX6a* and *SlX4* are overlapping and we used this to assemble of the end of the Xq arm, where the typical X.43.1 telomeric repeats were found.

[1]Among those 3, two genes were found to be X-hemizygous in Chibalina and Filatov (2011), and XY in Bergero and Charlesworth (2011) and Muyle et al. (2012). In the combined data (see details in Section 3.6), we considered these genes to be XY.
[1]Based on 39 genes previously known to have X- and Y-linked copies, see Table S4.
[2]Potential total number of genes absent/number of ancestral genes (including or not the probe genes), see text for details.

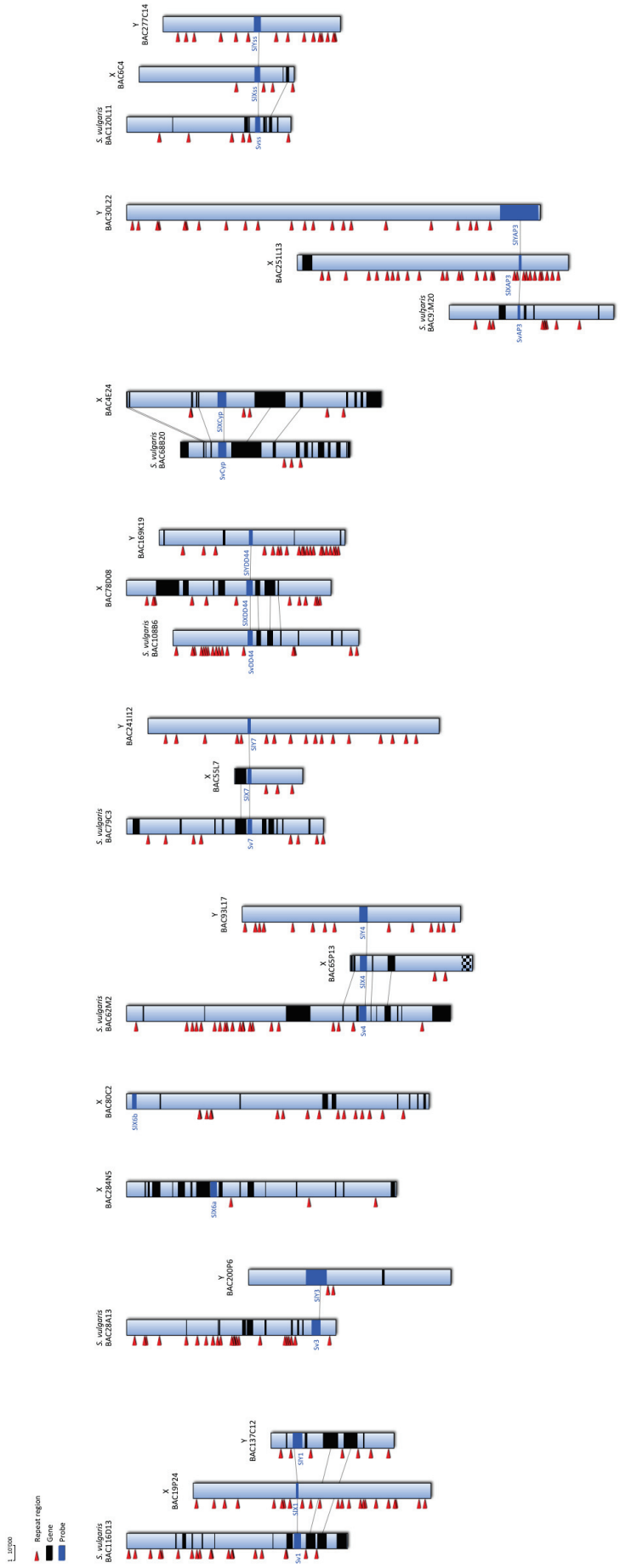


Figure 3.S2: Annotation of all BAC clones: Blue bars = "probe" genes, black bars = new genes, red triangles = transposable elements. Homology relationships are depicted by grey lines connecting genes.

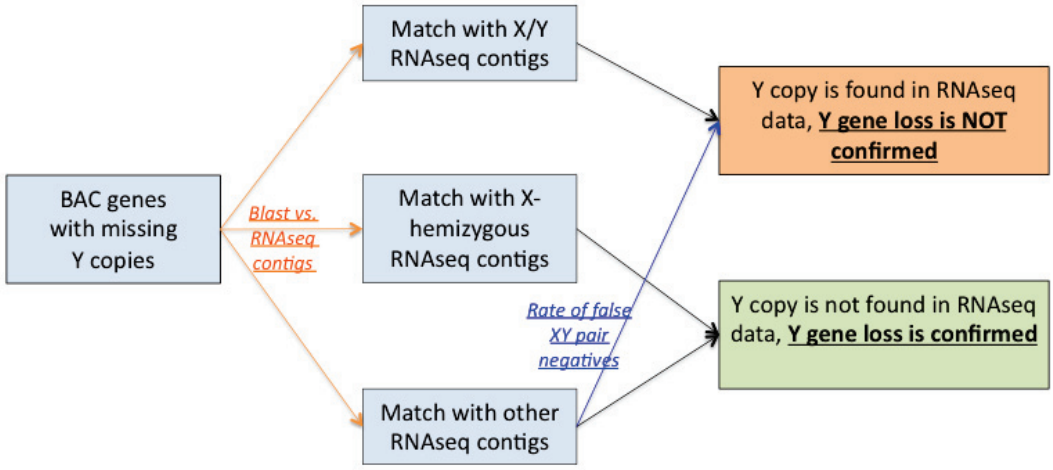


Figure 3.S3: Pipeline for inferring Y gene loss: X-linked BAC-located genes are blasted against the RNAseq contigs. This gives three categories of BAC-located genes: (i) match with a X/Y RNAseq contig, (ii) match with a X-hemizygous contig, (iii) match with an RNAseq contig not detected as sex-linked (see Section 3.6 for details about the blast search and how we combined results using different RNAseq datasets). Using the rate of false negatives for X/Y gene pairs, the numbers of undetected X/Y gene pairs and X-hemizygous genes are obtained. This then gives the total number of X-hemizygous genes, which is used to compute the percentage of gene loss on the Y chromosome. Gene loss on the X chromosome is computed similarly using Y-linked BAC-located genes but is less precise, as the Y0 genes are not inferred in RNAseq studies.

Table 3.S3: Comparison of BAC and RNA-seq data (detailed table). All BAC-located genes are included except the 6 probe genes for which both X and Y copies were available in the BAC data. Numbers of genes with *S. vulgaris* orthologs are indicated in parentheses

	Total number	BAC-located genes matching RNA-seq contigs	BAC-located genes matching X/Y RNA-seq contigs	BAC-located genes matching X-hemizygous RNA-seq contigs	Sources of RNA-seq data
X-linked BAC-located gene	52 (15)	44 (15)	12 (4)	-	Muyle et al., 2012
		36 (11)	5 (4)	2 (1)	Bergero and Charlesworth, 2011
		31 (8)	13 (5)	5 (2)	Chibalina and Filatov, 2011
		46 (15)	19 (6)	5 (2)	3 combined
Y-linked BAC-located gene	11 (3)	7 (3)	3 (3)	-	Muyle et al., 2012
		6 (3)	1 (1)	1 (0)	Bergero and Charlesworth, 2011
		6 (3)	1 (1)	3 (2)[1]	Chibalina and Filatov, 2011
		8 (3)	3 (3)	1 (0)	3 combined

Table 3.S5: Analysis of gene loss in X and Y chromosomes combining BAC and RNA-seq data (for X-vulgaris and Y-vulgaris pairs only)

Categories of genes	X-linked genes	Y-linked genes
All new genes in BAC sequences	15	3
No match to RNA-seq contigs	0	0
Genes retained for analysis	15	3
Category (i): XY results in RNA-seq analysis	6	3
Category (ii): X-hemizygous results in RNA-seq analysis	2	0
Category (iii): Not ascertained as sex-linked by RNA-seq analysis	7	0
Estimated X/Y false negative rate for gene pairs for RNA-seq analysis[1]	25%	25%
Expected number of XY pairs undetected in RNA-seq analysis	3.75	0.75
Potential number of X-hemizygous (X0) or Y0 genes undetected in RNA-seq analysis	3.25	0
Potential total number of X-hemizygous (X0) or Y0 genes (sum of detected + undetected in RNA-seq analysis numbers above)	5.25	0
Potential proportion of X-hemizygous (X0) or Y0 genes[2]	22-29%	0%

Table 3.S6: List of new genes with *S. vulgaris* homolog

***S. latifolia* sex-linked BAC-located gene**	***S. vulgaris* homolog**
BAC137C12_SlY1_CDS03_AT3G13870.1	BAC116D13_Sv1_CDS02_AT3G13870.1
BAC137C12_SlY1_CDS05_AT3G13750.1	BAC116D13_Sv1_CDS07_AT3G13750.1
BAC65P13_SlX4_CDS02_AT2G31410.1	BAC62M2_Sv4_CDS10_AT2G31410.1
BAC65P13_SlX4_CDS03_AT4G12970.1	BAC62M2_Sv4_CDS03_AT4G12970.1
BAC65P13_SlX4_CDS04_AT3G26790.1	BAC62M2_Sv4_CDS01_AT3G26790.1
BAC55L7_SlX7_CDS01_AT3G02050.1	BAC79C3_Sv7_CDS04_AT3G02050.1
BAC78D08_SlDD44X_CDS07_AT3G63530.1	BAC108B6_SvDD44_CDS03_AT3G63530.1
BAC78D08_SlDD44X_CDS05_AT3G08505.1	BAC108B6_SvDD44_CDS05_AT3G08505.1
BAC78D08_SlDD44X_CDS06_AT3G12650.1	BAC108B6_SvDD44_CDS04_AT3G12650.1
BAC4E24_SlCypX_CDS05_AT3G11590.1	BAC68B20_SvCyp_CDS09_AT3G11590.1
BAC4E24_SlCypX_CDS06_AT1G08380.1	BAC68B20_SvCyp_CDS10_AT1G08380.1
BAC4E24_SlCypX_CDS10_AT2G36100.1	BAC68B20_SvCyp_CDS11_AT2G36100.1
BAC4E24_SlCypX_CDS11_AT1G08260.1	BAC68B20_SvCyp_CDS01_AT1G08260.1
BAC4E24_SlCypX_CDS07_AT2G27690.1	BAC68B20_SvCyp_CDS12_AT2G27690.1
BAC6C4_SlssX_CDS01_tr|A5B1K4|A5B1K4_VITVI	BAC120L11_Svss_CDS05_tr|A5B1K4|A5B1K4_VITVI

Chapter 4

A probabilistic method for identifying sex-linked genes using RNA-seq-derived genotyping data

This project aimed at designing a method that would reliably identify sex-linked genes in non-model organisms at low cost. It has followed the project using BAC clones to obtain sex-linked genes in *Silene latifolia* (Chapter 3) as this first project provided reliable sex-linked sequences but at a high cost and only few new genes were obtained. The genes obtained through BAC clone sequencing allowed to test the new method on real data, along with other previously known sex-linked genes in *Silene latifolia*. The approach relies on RNA-seq data on a cross (parents and progeny of each sex), as it was previously proved to be efficient (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Hough et al., 2014), but, thanks to a probabilistic model, this project potentially makes this strategy applicable to any kind of sex chromosomes (XY, ZW and UV) and any dataset (this strategy had previously been applied with empirical filters that limited its broad use). I was particularly keen on developing a method with its own probabilistic model in order to provide a reliable set of sex-linked genes for further biological analyses.

This project was led in Gabriel Marais' lab, Niklaus Zemp in Alex Widmer's lab in Zurich generated the RNA-seq datasets. I developed the method, tested it on real and simulated data, prepared the first draft of the manuscript (including most figures and tables) which was modified by my supervisor and other authors, see authors contributions in Section 4.8 for more details.

The manuscript was submitted to *Genome Research* in December 2014 and sent to the referees but rejected after one round of revision. It was submitted to *MBE* in May 2015 and is currently under review.

A probabilistic method for identifying sex-linked genes using RNA-seq-derived genotyping data

Aline Muyle[1], Jos Käfer[1], Niklaus Zemp[2], Sylvain Mousset[1], Franck Picard[1,3], Gabriel AB Marais[1,3]

4.1 Keywords

Sex chromosomes, XY, ZW, UV, sex-linked genes, non-model organisms, RNA-seq, Galaxy workflow.

4.2 Abstract

The genetic basis of sex determination remains unknown for the vast majority of organisms with separate sexes. A key question is whether a species has sex chromosomes. Sex chromosome presence indicates genetic sex determination, and their sequencing may help identifying the sex-determining genes and understanding the molecular mechanisms of sex determination. Identifying sex chromosomes can be difficult, especially when they are homomorphic. Sequencing them is also very challenging, in particular the non-recombining regions as they are usually repeat-rich. A novel approach for identifying sex-linked genes and sex chromosomes consisting of using RNA-seq to genotype male and female individuals and study sex-linkage has recently been proposed. This approach entails a modest sequencing effort and does not require prior genomic or genetic resources, and is thus particularly suited to study non-model organisms. Applying this approach to many organisms is, however, difficult due to the lack of an appropriate statistically-grounded pipeline to analyse the data. Here we propose a model-based method to infer sex-linkage using a maximum likelihood framework and genotyping data from a full-sib family, which can be obtained for most organisms that can be grown in the lab and for economically important animals/plants. Our method works on any type of sex chromosomes (XY, ZW, UV) and has been embedded in a pipeline that includes a genotyper specifically developed for RNA-seq data. Validation on empirical and simulated data indicates that our pipeline is particularly relevant to study sex chromosomes of recent or intermediate age but can return useful information in old systems as well; it is available as a Galaxy workflow.

[1]Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France
[2]ETH Zurich, Institute of Integrative Biology, Universitätstrasse 16, 8092 Zürich, Switzerland
[3]equal contribution as senior authors

4.3 Introduction

Species with separate sexes (males and females) are common. They represent ~95% of animal species (Weeks, 2012) and as much as 50-75% of the species in some land plant lineages (Ming et al., 2011). They are rarer in angiosperms; yet ~15,000 species with separate sexes (dioecious) have been found (Renner, 2014). Many crops (e.g. papaya, strawberries, kiwi, spinach) are dioecious or derive from a dioecious progenitor (Ming et al., 2011). However, the mechanisms for sex determination remain unknown for most plant species and a number of animal species; in many cases it is not known whether sex chromosomes are present. Sex chromosomes of similar size and morphology (homomorphic) are particularly difficult to identify with cytology. Homomorphic sex chromosomes are probably frequent in groups such as angiosperms where many dioecious species have evolved recently from hermaphroditic ancestors and sex chromosomes are expected to be young and weakly diverged (Ming et al., 2011), in groups such as fish where sex determination mechanisms evolve quickly and the replacement of a pair of sex chromosomes by another (sex chromosome turn over) is high (Mank and Avise, 2009), or in groups such as amphibians where occasional recombination limits sex chromosome divergence (Stöck et al., 2013). Nevertheless, the exact frequency of homomorphic sex chromosomes remains unknown in those groups. In angiosperms, for example, dioecy has evolved probably >800 times independently (Renner, 2014), but less than 40 sex chromosome pairs (including 20 homomorphic pairs) have been reported so far (Ming et al., 2011).

Obtaining sequences of sex chromosomes is also very difficult since they have non-recombining regions. Those regions can be very large and comprise most (or all) of the Y chromosome in the heteromorphic systems. The human Y chromosome, for instance, is largely non-recombining, i.e. does not make cross-over with the X during meiosis. Only two small regions of the human Y, called the pseudoautosomal regions can do so. Non-recombining regions are found in all sex chromosome types: the Y and W in diploid XY and ZW systems and also in both sex chromosomes in the UV haploid systems found in some mosses and algae (Bachtrog et al., 2011). The non-recombining regions of the genome are known to accumulate large amounts of repeats, including transposable elements (Charlesworth et al., 1994; Gaut et al., 2007). In some species, the X and Z chromosomes also accumulate repeats, although at a lesser extent than their Y/W counterparts, because selection against repeats is reduced on these chromosomes due to a smaller effective population size compared to the rest of the genome (Bellott et al., 2010; Gschwend et al., 2012).

This makes the sex chromosomes (particularly the non-recombining portions of the Y, W and U/V chromosomes) difficult to assemble, especially when using short-read sequencing technologies. Consequently, many genome projects have focused on individuals of the homogametic sex (XX or ZZ) to avoid this difficulty and also the problem of

reduced coverage of the X and the Z when sequencing individuals of the heterogametic sex (discussed in Hughes and Rozen, 2012). Y chromosomes have been sequenced using strategies relying on establishing a BAC map prior to sequencing. The single-haplotype iterative mapping and sequencing (SHIMS, described in Hughes and Rozen, 2012) in particular has provided high-quality assemblies of the mammalian Y chromosome (Bellott et al., 2014; Hughes et al., 2012, 2010; Skaletsky et al., 2003). These strategies are, however, labour-demanding and costly, which explains why only a handful of Y chromosomes have been fully sequenced to date (<15), many of which have small non-recombining regions: the livevort *Marchantia* (Yamato et al., 2007), the fish medaka (Kondo et al., 2006), the green alga *Volvox* (Ferris et al., 2010), the tree papaya (Wang et al., 2012) and the brown alga *Ectocarpus* (Ahmed et al., 2014).

Producing high-quality assembly is not always necessary and alternative, less expensive strategies have been recently developed for identifying sex chromosome sequences based on next-generation sequencing (NGS) data. A first category of approaches relies on the comparison of one male and one female genome. Identifying X-linked scaffolds can be done by studying the genomic male over female read coverage ratio along the genome: autosomal contigs will have a ratio of 1 while X-linked ones will have a ratio of 0.5 (Vicoso and Bachtrog, 2011; Vicoso et al., 2013a,b). The Y scaffolds are simply those that are exclusively present in the male genome. A more sophisticated analysis can be done by a prior exclusion of the repeats shared by the Y and the female genome (Akagi et al., 2014; Carvalho and Clark, 2013). Also, a combination of RNA-seq and genome data of male and female individuals has been used to increase the number of known Y-linked genes in well-studied systems (Cortez et al., 2014). Similar analyses were done for ZW systems (Ayers et al., 2013; Moghadam et al., 2012; Vicoso and Bachtrog, 2011; Vicoso et al., 2013a,b). This approach, however, is suitable only if reasonably well-assembled reference genome is available, in the studied species or in a close relative.

A second category of approaches relies on studying how SNPs segregate among sexes. Full genome sequencing data of several male and female individuals can be used to genotype individuals of different sexes and study sex-linkage of single-nucleotide polymorphisms (SNPs) and scaffold from sex chromosomes can be ascertained (Al-Dous et al., 2011). If such genomic resources are lacking, as in many species with large genomes, complexity and size can be reduced by using transcriptomes instead of complete genomes. RNA-seq can be used instead of DNA-seq data to genotype individuals of different sexes and identify sex-linked SNPs and sex-linked genes. Sex chromosomes can thus be investigated in species with hitherto unknown genomes. One possibility is to sequence several male and female individuals of an inbred line and identify X/Y gene pairs by looking for SNPs showing Y-linkage (see Muyle et al., 2012; and current Supplementary Figure 4.S1). On the other hand, sequencing the parents and a few offspring individuals of known sex from a specific cross allows the identification of sex-linked genes using both Y-linkage and X-linkage information (see Bergero and Charlesworth, 2011; Chibalina and Filatov,

2011, and current Figure 4.1.B). The approach based on RNA-seq derived genotypes has identified hundreds of new sex-linked genes in species where only a few were known before (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012) or in species without previously known sex-linked genes or genomic resources available (Hough et al., 2014). Those are promising results, but genotyping and inference of sex-linkage were done without a statistical framework. The genotyping was performed (i) with tools that were developed for DNA-seq data, which do not account for uneven allele expression as may be the case for X/Y gene pairs or (ii) using reads number thresholds to distinguish true SNPs from sequencing errors, which were determined empirically for a given dataset, but may not be correct for another dataset as they depend on the sequencing coverage and the number of offspring that were used for sequencing. Arbitrary numbers of sex-linked SNPs per gene were used to classify genes as sex-linked or not. The lack of appropriate and statistically-grounded methods and pipelines clearly limits the application of the RNA-seq-based genotyping approach to more organisms despite its high potential.

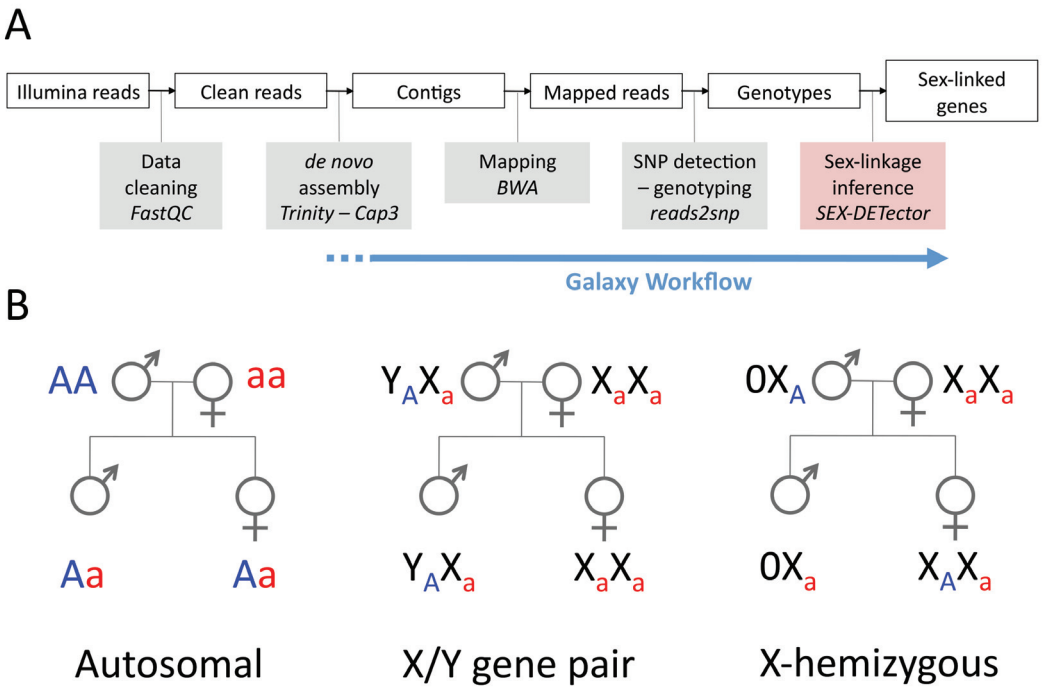


Figure 4.1: Pipeline: Schematic steps of our pipeline (**A**) and examples of family genotypes for the three types of segregation types in an X/Y system (**B**).

Here we propose a model-based method (called SEX-DETector) for inferring sex-linkage from genotyping data from a full-sib family. Our model describes the genotypes of the parents and the F1 offspring for autosomal and sex-linked genes and accounts for genotyping errors. A likelihood-based approach is used to compute the posterior probabilities of being autosomal, X/Y and X-hemizygous (X-linked copy only) for each RNA-seq con-

tig. The method is developed for any chromosome type (XY, ZW, UV), and the likelihood framework additionally offers the possibility to test for the presence and type of sex chromosomes in the data based on model selection. SEX-DETector is embedded in a pipeline including different steps from assembly to sex-linkage inference (Figure 4.1.A). Genotyping is done using a genotyper specifically designed for RNA-seq data (Gayral et al., 2013; Tsagkogeorga et al., 2012) that takes unequal allelic expression into account, which is relevant here as the Y copy of a X/Y gene pair tends to be less expressed than the X copy (reviewed in Bachtrog, 2013).

We tested our pipeline on RNA-seq data of a family from *Silene latifolia*, a dioecious plant with relatively recent but heteromorphic sex chromosomes, for which some sex-linked and autosomal genes have been experimentally characterized. Our method could detect 83% of known sex-linked genes expressed in the tissue used to obtain our RNA-seq data (i.e. flower bud). To compare our pipeline to previous ones that rely on arbitrary thresholds, we computed sensitivity and specificity values on all known *S. latifolia* genes and our pipeline showed a much higher sensitivity (0.63 compared to 0.25-0.43) while specificity remained close to 1, which suggests that the number of sex-linked genes has been underestimated in previous work. Applying our pipeline to a comparable RNA-seq data from *Silene vulgaris*, a plant without sex chromosomes yielded no sex-linked genes as expected. We further tested the SEX-DETector method using simulations, which indicated good performance with a modest experimental effort and on different sex chromosome systems. The advantages, limits and potential solutions to those limitations of our method/pipeline and the approach based on RNA-seq derived genotyping data in general are discussed. Our method/pipeline is particularly relevant to study sex chromosomes of recent or intermediate age, for which the approaches relying on male vs. female genome comparisons are not adapted (as they require X and Y reads not to co-assemble). It can also provide useful information about old systems, which may complement that given by the approaches relying on male vs. female genome comparisons specifically devoted to study those systems.

For an easy use of the method, we developed a Galaxy workflow (including extra assembly, mapping, genotyping and sex-linked gene detection) that takes assembled contigs and the raw reads from any system (XY, ZW, UV) as input and returns a set of sex-linked gene sequences and allele-specific expression level estimates.

4.4 Results

4.4.1 A probabilistic method for inferring sex-linkage from family genotyping data

SEX-DETector is based on the genotypes of two parents and their progeny from which we infer the segregation type of each contig. The model considers that SNPs can be trans-

mitted to the progeny by three segregation modes: (i) autosomal, (ii) sex-linked with both X and Y (or Z and W) alleles present and (iii) X (or Z) hemizygous, i.e. sex-linked with only the X (or Z) allele present (the Y or W allele being inactivated, lost, too weakly expressed or in a different contig due to X/Y or Z/W divergence, see Section 4.5). We also considered the case of UV sex chromosomes, similar to the one for XY and ZW, without hemizygous segregation and with only one parent (the sporophyte). Our method relies on the segregation of SNPs (see Supplementary Table S1) that are fully described in the progeny when both parent genotypes and segregation type are known (see Figure 4.1.B for an example). We use a likelihood-based framework to assess the *posterior* probability of each segregation type for each informative SNP given the observed genotype data.

A strong advantage of our model-based approach compared to empirical methods is the amount of information captured from the data thanks to a hierarchical probabilistic model. Observed genotypes of parents and progeny were incorporated in a model using genotype probabilities (see Section 4.6). We accounted for discrepancies that may exist between observed and true genotypes (because of genotyping errors) by introducing two genotyping error parameters: one for any type of genotyping error, and one specific to the Y (or W) as genotyping errors are more frequent on the Y (or W) allele due to reduced expression and low RNA-seq read coverage. The UV model does not contain any specific genotyping error for the non-recombining sex chromosome as both U and V are non-recombining. Our model accounts for genotyping errors that are likely to be present on parental genotypes as well. These steps (i.e. estimating genotype probabilities and genotyping errors) are essential as each true parental genotype has a different probability to occur in the dataset due to the level of heterozygosity and the base composition of a given species, and they will ensure that the method can apply to different species. Then, for each SNP and individual, we compute the *posterior* probabilities of genotyping errors, which allows us to compute the *posterior* probabilities of observing the true parental genotypes and then the segregation types *posterior* probabilities for each SNP. The segregation type of each contig is finally inferred by averaging informative SNP posteriors. All X-hemizygous SNPs are informative whereas informative X/Y SNPs are positions for which the heterogametic parent is heterozygous and different from the homogametic parent (otherwise it is not possible to distinguish X/Y and autosomal segregation). Each SNP posterior is weighted by its posterior probability of genotyping error (so that SNPs with higher genotyping error posteriors have less effect on the final inference about a contig segregation type). The contig inferred segregation type corresponds to the one with the highest *posterior* (which corresponds to the maximum *a posteriori* rule).

4.4.2 A pipeline for analysing RNA-seq data from a family

An important step in the pipeline is read assembly. To be able to detect Y-linked SNPs as well as XY gene pairs, the reads from X and Y transcripts must co-assemble into a single RNA-seq contig. This is achieved using Trinity and Cap3 for further assembly: Trinity

will produce groups of contigs including alternative transcripts and alleles; joining the alleles into one contig is done with Cap3 (see Section 4.6). Note that the detection of the X-linked SNPs will not depend on the efficiency of the X and Y read co-assembly, and it will still be possible to identify sex-linked genes in case of low or no X and Y read co-assembly (see Section 4.5). Most genotypers have been developed for analysing genomic data, not transcriptomic data. A major difference between these two types of data is that coverage can significantly differ among alleles in transcriptomic data because of differences in expression level among alleles. For X/Y gene pairs, such differences are frequent, with the Y copy being less expressed than the X one (reviewed in Bachtrog, 2013). Standard genotypers will typically consider the less expressed alleles as sequencing errors as we have experienced and corrected manually in previous work (Muyle et al., 2012). To solve this problem, we used a genotyper specifically developed for RNA-seq data, called reads2snp, which allows differences in expression level among alleles (Gayral et al., 2013; Tsagkogeorga et al., 2012). Our genotype inferences were different compared to standard genotypers when X and Y copies had different expression levels (data not shown). We developed Galaxy wrappers for SEX-DETector and used available wrappers for other tools (including reads2snp) to prepare a Galaxy workflow.

4.4.3 Testing our pipeline's performance using a *Silene latifolia* dataset

The SEX-DETector pipeline (Figure 4.1.A) was run on a *Silene latifolia* dataset. *S. latifoli*a is a dioecious plant species with well-studied XY chromosomes that had several interesting characteristics for benchmarking our method/pipeline: (1) *S. latifolia* genome and sex chromosomes are quite large (the genome is 3Gb, the X is 400 Mb and the Y is 550 Mb); (2) no reference genome is available in this species; (3) *S. latifolia* sex chromosomes are relatively recent (~5 MY old; Rautenberg et al., 2010) but clearly heteromorphic; X-Y synonymous divergence ranges from 5 to 25% (Bergero et al., 2007); *S. latifolia* thus represents a system of intermediate age; (4) a tester set of 209 genes for which segregation type has been established is available in this species (Supplementary Table S2). The dataset consists of a cross (two parents and four offspring of each sex). RNA-seq data was obtained for each of these individuals tagged separately and the reads were assembled using Trinity and then Cap3, the final assembly included 46,178 ORFs (Table 4.1). RNA-seq reads were mapped onto this assembly (see Supplementary Table 4.S3 for library sizes and mapping statistics) and genotyping was done for each individual using reads2snp. SEX-DETector was run on the genotyping data to infer autosomal and sex-linked genes (Table 4.1). For further analysis, only contigs having at least one SNP without genotyping error and showing a *posterior* probability ≥ 0.8 (of being autosomal or sex-linked) were retained. Figure 4.2.A-D shows examples from the tester set. For some genes, all SNPs show clearly the same correct segregation type (Figure 4.2.A-C), whereas in some genes mixed segregation patterns were inferred, which we attribute to co-assembly of recent

paralogs or other assembly/mapping problems (see Figure 4.2.D and Section 4.5).

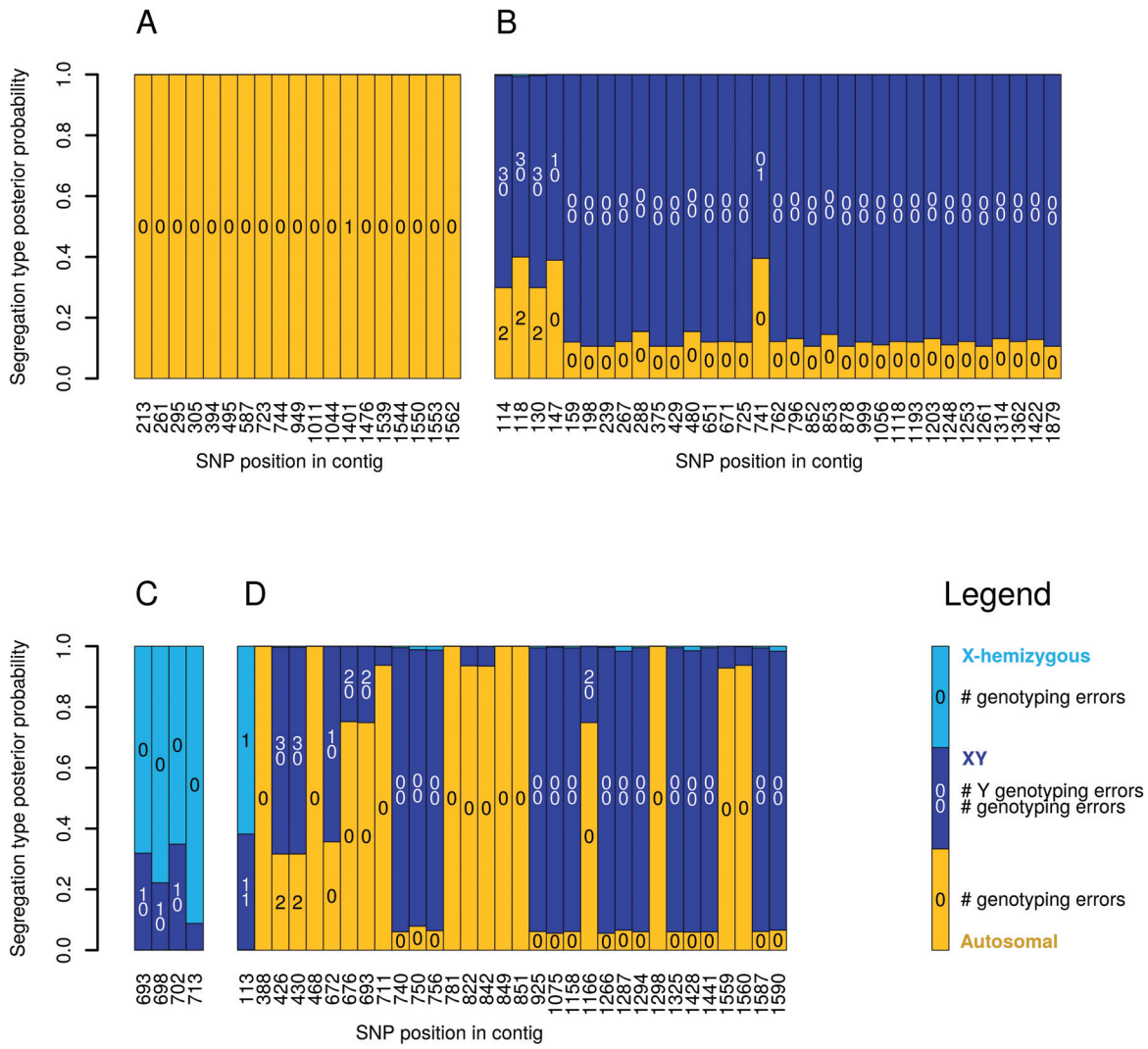


Figure 4.2: Results of the SEX-DETector pipeline for known *S. latifolia* genes. Only informative SNPs are shown: positions that are inferred as polymorphic, and for which, in case of autosomal or X/Y segregation, the heterogametic parent is heterozygous and different from the homogametic parent (otherwise it is not possible to differentiate between X/Y and autosomal segregation). Segregation type posterior probabilities are shown for each informative SNP (see legend on figure for colour code) and inferred number of genotyping errors for each segregation type are shown inside the bars (a genotyping error is inferred when its posterior probability is higher than 0.5). **A)** SlE72 (this gene is known to be autosomal), its weighted autosomal mean probability is 0.99. **B)** SlCypX (this gene is known to be X/Y), its weighted sex-linked mean probability is 0.96. **C)** WUS1 (this gene is known to be X-hemizygous), its weighted sex-linked mean probability is 0.99. **D)** BAC284N5-CDS13_SlX6a (this gene is known to be sex-linked), its weighted sex-linked mean probability is 0.82.

We used our tester set to test the performance of our pipeline, i.e. estimate its sensitivity (the capacity to detect true sex-linked genes) and specificity (the capacity not to assign autosomal genes as sex-linked). 83% of the known sex-linked genes expressed in the RNA-seq data used here (i.e. flower bud) were detected, indicating a high sensitivity. We obtained a specificity of 99% for this dataset as one gene, OxRZn, was supposedly wrongly assigned as a sex-linked gene by SEX-DETector. However, this gene was ear-

Table 4.1: Results of our pipeline on the *S. latifolia* dataset.

ORF Types	Numbers
ORFs in final assembly	46178
ORFs with enough coverage to be studied	43901
ORFs with enough informative SNPs to compute a segregation probability	17189
ORFs with posterior segregation probability over 0.8	15164
ORFs assigned to an autosomal segregation type	13807 (91%)
ORFs assigned to a X/Y segregation type	1025 (7%)
ORFs assigned to a X hemizygous segregation type	332 (2%)

lier assessed as autosomal on the basis of the absence of male specific alleles (Marais et al., 2011) and SEX-DETector assigned it to a sex-linked category because of two clear X-hemizygous SNPs, without genotyping error. It is therefore likely that OxRZn is in fact a true positive and more research on that gene is required.

4.4.4 Comparing our pipeline to others using a *S. latifolia* dataset

We compared the performance of our pipeline to those used in previous work on inferring sex-linkage with RNA-seq data in *S. latifolia* (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012). Those pipelines differ in many ways, and the data themselves can be different. In previous work, offspring individuals of the same sex were sometimes pooled before sequencing (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011). We used again the tester set of 209 *S. latifolia* genes with known segregation types, which we blasted onto each dataset to find the corresponding contigs and their inferred segregation type (for details see Supplementary Table S2). Because the different pipelines require different types of data (pooled progeny versus individually-tagged offspring) and with different read coverages, we computed sensitivity on all known genes (expressed or not). Our pipeline outperformed other pipelines in terms of sensitivity (Table 4.2), while specificity was comparable (Table 4.2), which indicates that the number of sex linked genes in *S. latifolia* has previously been underestimated.

Table 4.2: Comparison with other methods: sensitivity and specificity values obtained with different methods using 209 known *S. latifolia* genes.

	Sensitivity	Specificity
SEX-DETector	0.625 (0.509 – 0.731)	0.99 (0.958 – 0.999)
Muyle et al., 2012	0.275 (0.181 – 0.386)	0.984 (0.945 – 0.998)
Bergero and Charlesworth, 2011	0.25 (0.159 – 0.359)	0.992 (0.958 – 0.999)
Chibalina and Filatov, 2011	0.425 (0.315 – 0.541)	1 (0.972 – 1)

As further analysis showed, this under-estimation was due to overly conservative filtering in previous work. To exclude false positives, genes with at least 5 sex-linked SNPs were retained in previous studies. More filtering was done by excluding contigs with autosomal SNPs (Bergero and Charlesworth, 2011; Hough et al., 2014). As shown in Fig-

ure 4.3, keeping only contigs with at least 5 sex-linked SNPs removes nearly half of the contigs inferred as sex-linked by SEX-DETector, many of which have a high posterior probability. Excluding further those with autosomal SNPs (keeping those with sex-linked SNPs only) removes 74% of the contigs (Figure 4.3.B). Comparatively, SEX-DETector removes 12% of contigs when filtering for a posterior probability higher than 0.8 (Table 4.1), as most genes have a very high *posterior* segregation type probability which indicates a strong signal in the data and illustrates the benefits of using a model-based approach.

4.4.5 Simulations show that SEX-DETector requires a modest experimental effort and works on different sex chromosome systems

We simulated genotypes for a cross (parents and progeny) by generating a coalescent tree with autosomal or sex-linked history (Supplementary Figure 4.S2) and generated the parental sequences using that tree and molecular evolution parameters. Progeny genotypes were obtained by random segregation of alleles from the parents and a genotyping error layer was added (see Section 4.6). 10,000 contigs were simulated for each dataset.

In order to know how many offspring of each sex should be sequenced to achieve the best sensitivity and specificity trade-off using SEX-DETector, we varied the number of progeny individuals in the simulations. For an X/Y or Z/W system, optimal results were obtained when sequencing five progeny individuals of each sex (Figure 4.4.A); sequencing more progeny individuals did not improve the results further. This suggests that sequencing 12 individuals (two parents and five progeny individuals of each sex) may be sufficient to achieve optimal performances with SEX-DETector on an X/Y or Z/W system. For a U/V system, two progeny individuals of each sex seems sufficient to obtain optimal SEX-DETector performance (Figure 4.4.B), which suggests that sequencing five individuals (the sporophyte parent and two progeny of each sex) may be enough in the case of a U/V system. Our simulations thus suggest that SEX-DETector requires a modest experimental effort to reliably identify expressed sex-linked genes.
In order to assess the applicability of SEX-DETector to different types of sex chromosomes (old versus young, homomorphic versus heteromorphic) and species (highly versus weakly polymorphic), we used the same simulation procedure and tested the effect of one parameter at a time on SEX-DETector sensitivity and specificity. In our simulations, the degree of polymorphism within species had no influence on the performance of our method (Supplementary Figure 4.S3.A). As for the influence of the size of the non-recombining region (homomorphic or heteromorphic sex chromosomes), it was tested using different % of sex-linked genes in a genome with no effect on the performance of SEX-DETector (Supplementary Figure 4.S3.B). The limit of detection of a sex-linked contig was reached only when 1 sex-linked contig out of 10,000 contigs was present. Finally,

the simulations indicated that our method is robust to the X-Y divergence, as young and old sex chromosomes were evenly detected (Supplementary Figure 4.S3.C).

4.4.6 SEX-DETector identifies unknown sex chromosomes using model selection

It is common that species with separated sexes have unknown sex determination system, i.e. it is unknown whether they have sex chromosomes and if they have, of what type (Z/W versus X/Y). The likelihood-based framework of SEX-DETector allows us to test for these assumptions by comparing models' fit to the data using the Bayesian Information Criterion (BIC, see Section 4.6). In species for which sex determination is unknown, it is possible to compare models with and without sex chromosomes, and then if sex chromosomes are detected, it is possible to compare models with X/Y or Z/W system. This was tested on real data and simulated data.

In the *S. latifolia* dataset, the best model inferred by SEX-DETector was a model with sex chromosomes as expected, with 1357 sex-linked contigs (which represents 9% of the contigs with a posterior probability higher than 0.8). In the *Silene vulgaris* dataset (a species without sex chromosomes), no sex-linked contigs were inferred, the best model fit to the data was thus a model without sex chromosomes as expected (see Section 4.6).

In order to know from which proportion of sex-linked genes sex chromosomes can be detected, we compared models on simulated data with varying numbers of sex-linked contigs out of 10,000 simulated contigs (Table 4.3 and Supplementary Table S4). When no sex-linked contigs were simulated, as expected the best model was the one without sex chromosomes. This was also the case when a single sex-linked contig was simulated. In this case, SEX-DETector could not detect it due to lack of information in the dataset. When ten or more sex-linked contigs were simulated, the best model was the one with sex chromosomes as expected. Thus, ten sex-linked contigs out of 10,000 provide sufficient information for SEX-DETector (i.e. 1 sex-linked gene out of 1000 genes can be detected).

Once the presence of sex chromosomes has been inferred, it can be tested whether the system is X/Y or Z/W. The model comparison between X/Y and Z/W systems worked on both real and simulated data: the best model for *S. latifolia* was, as expected, the X/Y system (Table 4.3 and Supplementary Table S4).

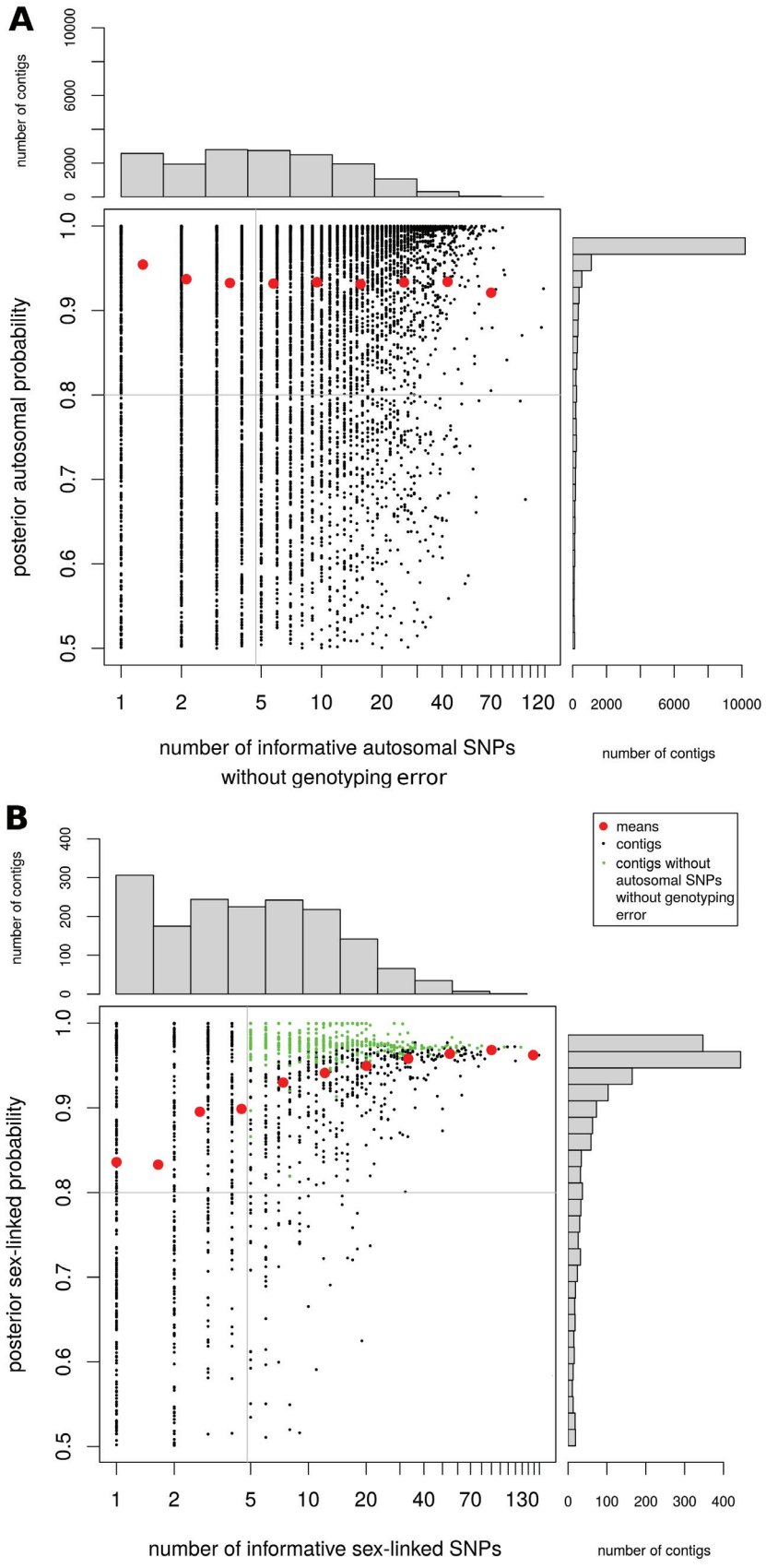


Figure 4.3: Performance of the method. The number of SNPs without genotyping error was plotted against the posterior segregation type probability for each autosomal (**A**) and sex-linked (**B**) contigs of the *S. latifolia* dataset. The distributions of both variables are shown, and means for each category on the histograms are indicated by red dots. Sex-linked genes kept after the filter commonly used in empirical methods are shown in green (at least five sex-linked SNPs and no autosomal SNPs).

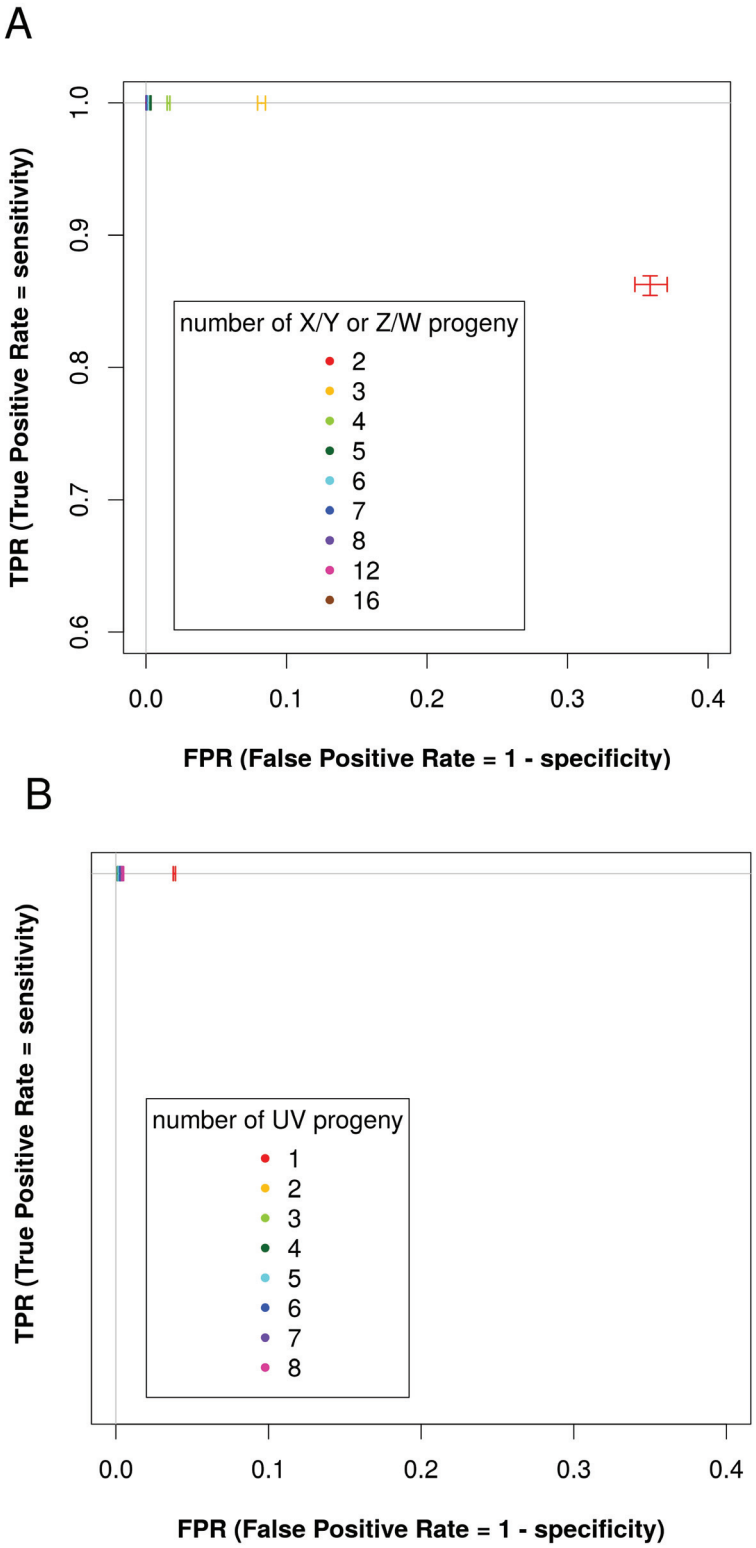


Figure 4.4: Results on simulations. ROC curve showing the effect of the number of progeny sequenced on sensitivity (TPR, true positive rate) and specificity (1-FPR, false positive rate) in simulated data. A perfect classification of contigs would lead to a point having TPR equal to one and FPR equal to zero (top left corner of the graph). **A)** X/Y or Z/W sex determination system (all points overlap in the top left corner when over five progeny of each sex are used). **B)** U/V system (all points overlap in the top left corner when over two progeny of each sex are used).

Table 4.3: Model comparison using SEX-DETector on real datasets (*S. latifolia* (with sex chromosomes) and *S. vulgaris* (without sex chromosomes) and simulated X/Y datasets with varying number of sex-linked contigs out of 10,000 simulated contigs. The best model is chosen as the one having the lowest BIC value (see Supplementary Table S4 for details).

		models with sex chromosomes				**BIC model without sex chromosomes**
		XY model		**ZW model**		
		BIC	**number of sex-linked genes**	**BIC**	**number of sex-linked genes**	
Real datasets	***Silene latifolia* (XY system)**	best	15164	-	9	-
	***Silene vulgaris* (no sex chromosomes)**	-	0	best	0	-
Simulated datasets of 10,000 genes with different numbers of sex-linked genes (XY system)	**0 sex-linked genes**	-	0–1	-	0	best
	1 sex-linked genes	-	0	-	0	best
	10 sex-linked genes	best	16–57	-	0–1	-
	100 sex-linked genes	best	156–181	-	0–10	-
	500 sex-linked genes	best	592–624	-	23–40	-
	3000 sex-linked genes	best	3159–3200	-	1688–1807	-

4.5 Discussion

4.5.1 Strengths and limits of the SEX-DETector pipeline

Our pipeline offers a number of interesting features. Both real *S. latifolia* data and simulations suggest that very few individuals need to be sequenced: twelve individuals (two parents plus five male and five female offprings) in the case of a X/Y or Z/W system, and five individuals (the sporophytic parent plus two male and two female gametophytic offprings) for a U/V system appears to be enough to get very good performance from our pipeline. This makes the strategy very accessible given the cost of RNA-seq. Our pipeline is user-friendly thanks to a Galaxy workflow that goes from further assembly of contigs using Cap3 to sex-linked genes and allelic expression levels (Figure 4.1.A).

New sex chromosomes can thus be characterised in species that have separated sexes and for which sex-determination has remained unknown. Indeed, the method allows one to test for the presence of sex chromosomes in the data, and then test for an X/Y versus a Z/W system, using the BIC.

Sensitivity results (Table 4.2) showed that the SEX-DETector pipeline is more powerful for detecting sex-linked genes compared to previous pipelines relying on empirical methods (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012). The use of a cross allows identification of both X/X and X/Y SNPs unlike in (Muyle et al., 2012). Individually tagged progeny individuals give more information than the pools (that were used in Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011). The reads2snp genotyper suited for RNA-seq data with allelic expression biases prevents the weakly expressed Y alleles to be wrongly classified as sequencing errors. Further, the probabilistic framework allows to filter contigs on posterior segregation type probability rather than the number of sex-linked SNPs as done in other studies (Bergero and Charlesworth, 2011; Hough et al., 2014), allowing to preserve more true sex-linked genes without increasing false positive inferences (Table 4.2, Figure 4.3).

SEX-DETector infers both sex-linked and autosomal genes. In earlier approaches, contigs were inferred as either sex-linked or not sex-linked, and the latter consisted of a mix of autosomal and undetected sex-linked genes. In SEX-DETector, contigs for which no segregation type inference was possible, for example because of a lack of informative SNPs, are classified as undetermined and are not merged with autosomal genes. Having reliable inferences on autosomal genes is highly useful for comparative studies between autosomal and sex-linked genes.

RNA-seq derived genotyping data approaches, including ours, have limitations. The use of RNA-seq suffers from the limitation that some genes may not be expressed, or may be too weakly expressed in the data for these approaches to infer segregation types, which implies that using approaches based on RNA-seq genotyping data necessarily underestimates the number of truly sex-linked genes. Also, no expression of a Y copy in the studied tissue of a X/Y gene pair will result in the X copy being ascertained as X-hemizygous.

Extracting RNA from tissues with complex transcriptomes (many expressed genes) will attenuate these problems, for example flower buds in plants or reproductive organs (e.g. testis) in animals. Using not a single but several tissues/organs/development stages may also help solving these problems, but will increase the cost. Pooling the tissues for each individual before sequencing may be a way to avoid such extra-cost.

Our pipeline includes a *de novo* assembly, which may lead to co-assembly of very recent paralogs into chimeric contigs. This problem is common to all available approaches to obtain sex chromosome sequences except the SHIMS approach (Hughes and Rozen, 2012). However, and most importantly, our method will be able to identify the genes where such problems might have occurred when paralogs are on different chromosomes (as these genes will show a mixture of sex-linked and autosomal SNPs). These genes can be excluded from further analysis by the user. Paralogs from the same chromosome will be more difficult to be handled by our method (and nearly all others, except the SHIMS).

Our simulations suggest that SEX-DETector should work on different systems; its performance was indeed excellent in a wide range of situations even when introducing genotyping errors (Supplementary Figure 4.S3). The simulations were manned to assess the SEX-DETector method performance and not that of the whole pipeline. We therefore directly simulated genotypes and did not include all the possible errors occurring upstream the genotyping steps (assembly and mapping). In particular, the failure to co-assemble X and Y reads and assembly errors were not simulated, which may have implications on the applicability of the pipeline to old systems.

Highly divergent X/Y genes are expected to assemble into separate X and Y contigs (Muyle et al., 2012). The X-Y divergence threshold beyond which co-assembly will not be possible is not known. *S. latifolia* does not have a particularly low X-Y synonymous divergence since it ranges from 5% to 25% (Bergero et al., 2007). By comparison, most X/Y gene pairs in humans exhibit a divergence lower than 30% (mean X-Y synonymous divergence for strata 3, 4 and 5 is respectively 30%, 10% and 5%, Skaletsky et al., 2003). When running SEX-DETector on our tester set, none of the known X/Y gene pairs was inferred as X-hemizygous, whereas several are the oldest *S. latifolia* stratum (with X-Y synonymous divergence of 20-25%, Bergero et al., 2007). The good performance of our methods on *S. latifolia* suggests that the failure to co-assemble X and Y reads will not be an issue for systems with moderate X-Y divergence. A recent study using the RNA-seq-based segregation approach on *Rumex hastatulus*, an older system (~15 MY), returned hundreds of sex-linked genes and suggests this approach works on even more divergent systems than *S. latifolia* (Hough et al., 2014).

Moreover, as already explained, in the case of failure to co-assemble X and Y reads, the X contigs will still be identified (through X-linked SNPs) and the SEX-DETector analysis will return X-hemizygous genes. The Y contig will not be identified as SEX-DETector does not detect Y contigs alone (they cannot be distinguished from autosomal genes that are exclusively expressed in males, i.e. that have male-limited expression). To identify

the Y contigs that have been missed, other strategies need to be investigated, for example testing whether the X-hemizygous genes match to male-specific contigs, which may represent the divergent Y contigs. Note that when this was done for our 332 inferred X-hemizygous genes, only 5 of them had a significant match with a male-specific contigs, suggesting that our set of inferred X-hemizygous genes includes only a few wrongly inferred X/Y gene pairs with a divergent Y expressed in flower bud.

Our simulations also suggested that SEX-DETector might work similarly on homomorphic (few sex-linked genes) and heteromorphic (many sex-linked genes) systems. In homomorphic systems, the sex chromosomes typically include one or two large pseudoautosomal regions (PARs). The genes close to the pseudoautosomal boundary may exhibit partial linkage. It is likely that 10 offsprings or so will not be enough to tell apart the partially sex-linked from the fully sex-linked genes (Supplementary Figure 4.S4). For some analyses, it will not be a limitation, it can even be useful to have as many genes as possible from the sex chromosomes including those in the PARs. But for others, it may be important to distinguish genes in the sex-specific regions form genes in the PARs. In this case, the number of sequenced offsprings should be increased as it is expected that partial sex-linkage should vanish when analysing many individuals.

A difficulty for the identification of X-linked SNPs may be the presence of X chromosome inactivation, or any instance of dominance effects in which only one X-linked transcript is present in RNA-seq data. X chromosome inactivation is the inactivation of one of the two X chromosomes in females. If in the studied tissue, the same X chromosome is consistently inactivated, heterozygous mothers will appear homozygous while different alleles are found in their sons, and this will make the detection of sex-linked SNPs using X-linkage information more difficult. The method will not give erroneous results but basically loose power: X-hemizygous genes may not be detected and even X/Y gene pairs may be less easy to detect (with only Y-linkage information). The consequences should not be too serious for recent / intermediate systems. In old systems where the method will identify mostly the X-linked genes, what can we expect? At present X chromosome inactivation is only known in mammals and it is not known whether it exists in other taxa. In the case of random X chromosome inactivation (as in placentals) tissues will include a mixture of cells with one X or the other inactivated, both transcripts will be present in the RNA-seq data and the problem will vanish. In marsupials, the paternal X is inactivated but X-inactivation is incomplete (Al Nadaf et al., 2010), not all of the genes are fully inactivated, which leaves some room to find some X-linked genes, even in this extreme case.

Importantly, SEX-DETector will also work on genotyping data derived from genome sequencing. In some cases, it might be more efficient to use DNA-seq data instead of RNA-seq data to genotype individuals. This alternative approach will however scale with genome size, and may be costly for species with large genome. For example, if (1) the sex-linked genes have very different expression patterns so that RNA-seq of many tissues would be required to identify many sex-linked genes, (2) the sex-specific region of the sex

chromosomes is expected to be very small and all genes might not be expressed in all tissues so that the number of expressed sex-linked genes might be outside the range of SEX-DETector power (<10 genes), (3) X chromosomes inactivation (especially non-random) is suspected, one could collect DNA instead of RNA sequences from a family for solving these potential problems.

4.5.2 What strategy is best for detecting sex-linked genes?

Strategies relying on sequencing male and a female genome/transcriptome have successfully been used in several organisms (Ayers et al., 2013; Carvalho and Clark, 2013; Cortez et al., 2014; Moghadam et al., 2012; Vicoso and Bachtrog, 2011; Vicoso et al., 2013a,b). They are expected to work better for small genomes and require the sex chromosomes to be divergent enough. In species with very large and complex genomes with many repeats, the assembly of NGS genomic data will be challenging (if at all possible), which may harden the male/female genome comparison and may result in obtaining highly fragmented and incomplete sex-linked gene catalogues. In young and weakly diverged sex chromosome systems, X and Y (or Z and W, or V and U) sequences will assemble together, which will prevent their identification by these strategies.
Strategies relying on getting RNA-seq data to perform a segregation analysis are expected to be insensitive to genome size. This has some practical aspects as the sequencing costs will not scale with genome size. The approach will thus remain affordable also for species with large genomes. It concentrates on the transcribed part of the sex chromosomes and will directly give the sequences of many sex-linked genes, the primary material for studies of sex chromosomes. It also provides expression data that can be used to address various questions about the evolution of gene expression on sex chromosomes. As discussed in the previous section, these approaches are ideal for weakly to moderately diverged X/Y gene pairs whose reads will co-assemble. In more diverged systems, X and Y copy co-assembly may be more difficult, and these approaches will mainly return X-hemizygous genes (even if a Y copy exists). Both types of strategies are thus complementary and may be used hand-in-hand in some systems (e.g. sex chromosome systems showing different levels of divergence).

4.5.3 Conclusions - Perspectives

Our SEX-DETector method/pipeline requires family data and will work optimally on young or intermediate systems (although returning useful information for old systems) as discussed above. Family data can be obtained in many organisms (e.g. all those that have genetic map), typically all the organisms that can be grown in the lab. Such data are also available in many agronomically important animals/crops. As mentioned in the introduction, it is likely that many sex chromosome systems that remain to be characterized are homomorphic. The applicability of SEX-DETector is thus broad.

SEX-DETector along with the other methods recently developed for obtaining sex-linked sequences using NGS at low cost will render feasible the large-scale comparative analysis of different sex chromosome systems. In particular, the comparison of systems from closely related species becomes possible, which will give much more information than previously performed comparisons of phylogenetically highly distantly related systems.
To study sex chromosomes, in some organisms for which family cannot be obtained easily, a method to infer sex-linkage from population data (males and females sampled from a population) would be very useful. Such a method has recently been proposed (Gautier, 2014) but it has mainly been developed to sort markers (autosomal, X, Y, organelles) before performing population genetics analysis on NGS genomic data, and relies on ploidy levels to detect sex-linked contigs. This method is thus designed for old X/Y or Z/W sex chromosomes only. Our pipeline currently includes the empirical method used in (Muyle et al., 2012) for population data (Supplementary Figure 4.S1), but the extension of a SEX-DETector version for population data (based on a population genetics model) is currently under development.
Finally, our pipeline could also be used, pending such adjustments, on other systems than sex chromosomes, such as mating type loci, B chromosomes, incompatibility loci, supergenes and any other type of dominant loci associated with a phenotype, for which a cross between a heterozygous and a homozygous individual is possible and for which both alleles are expressed at the transcript level in heterozygous individuals.

4.6 Material and Methods

4.6.1 Description of the probabilistic Model

The observed data consists of contigs containing offspring and parental genotypes (OG, PG respectively). From these genotyping data we want to infer the unknown segregation type (S) of each contig. We first suppose that S is a Multinomial random variable $\mathcal{M}(1,\pi_1,\pi_2,\pi_3)$ such that π_1, π_2, π_3 are the probabilities for one contig of being autosomal, X/Y (or Z/W), or X (or Z) hemizygous, respectively. Our strategy relies on the introduction of genotyping errors that may concern offspring as well as parent genotypes, either on the Y ($YGE \sim \mathcal{B}(p)$) contigs, or on the other alleles ($GE \sim \mathcal{B}(\varepsilon)$). We further introduce true homogametic and heterogametic parental genotypes (TMG for true mother genotype and TFG for true father genotype respectively), which are also unknown. Variable TMG is supposed to be Multinomial with parameters $(\alpha_1, \ldots, \alpha_M)$, α_m standing for the probability for genotype m, and TFG is also multinomial, with parameters depending on the segregation type $TFG|S_j \sim \mathcal{M}(1,\beta_1...,\beta_{N_j})$. Finally, when there is no genotyping error, if the segregation type and true parental genotypes were known, the conditional distribution of variables (OG, PG) is Multinomial, with parameters fully determined by segregation tables (see Supplementary Table S1). Parameters are estimated by maximum likelihood using an Expectation-Maximisation (EM) algorithm that includes a Stochastic

step (SEM algorithm) to deal with initialization issues. The outputs of the method are maximum likelihood estimates of parameters π, p, ε, α, β, and the posterior probabilities for hidden variables (TMG, TFG, YGE, GE, and S) given the observed data (OG, PG). The maximum *a posteriori* rule is used to infer parental genotypes, genotyping errors, and most importantly segregation types. Every notation and computation details are provided in Section 4.10.1 and Section 4.10.2, see also Supplementary Figure 4.S5.
To infer contig status, we defined what we call informative SNPs, which are autosomal or X/Y positions for which the heterogametic parent is heterozygous and different from the homogametic parent (otherwise it is not possible to differentiate between X/Y and autosomal segregation). Only informative SNPs are considered for computing a contig average segregation type, where SNPs are weighted by their posterior genotyping error probability (lower weight for contigs with higher posterior genotyping error probability). Contigs are assigned as sex-linked if they have at least one informative sex-linked (X/Y or X-hemizygous) SNP without genotyping error, and if the average sex-linked posterior probability is higher than the autosomal one and higher than a chosen threshold. Accordingly, a contig is inferred as autosomal if it has at least one autosomal SNP without genotyping error and if the autosomal posterior probability is higher than the sex-linked one and higher than the given threshold. A posterior segregation type probability threshold of 0.8 was chosen here. This parameter can be changed by the user.
The code of SEX-DETector was written in Perl.

4.6.2 Data analysis

4.6.2.1 Plant material and sequencing

RNA-seq data were generated from a cross in the dioecious plant *S. latifolia*, which has sex chromosomes and from a cross in the gynodioecious plant *S. vulgaris*, which does not have sex chromosomes. We used the following RNAseq libraries that were used in previous studies: Leuk144-3_father, a male from a wild population; U10_37_mother, a female from a ten-generation inbred line (Muyle et al., 2012); and their progeny (C1_01_male, C1_3_male, C1_04_male, C1_05_male, C1_26_female, C1_27_female, C1_29_female, C1-_34_female). For *S. vulgaris* the hermaphrodite father came from a wild population (Guarda_1), the female mother from another wild population (Seebach_2) and their hermaphrodite (V1_1, V1_2, V1_4) and female (V1_5, V1_8, V1_9) progeny.
Individuals were grown in a temperature-controlled greenhouse. The QiagenRNeasy Mini Plant extraction kit was used to extract total RNA two times separately from four flower buds at developmental stages B1–B2 after removing the calyx. Samples were treated additionally with QiagenDNase. RNA quality was assessed with an Aligent Bioanalyzer (RIN.9) and quantity with an Invitrogen Qubit. An intron-spanning PCR product was checked on an agarose gel to exclude the possibility of genomic DNA contamination. Then, the two extractions of the same individual were pooled. Individuals were tagged and then pooled

for sequencing. Samples were sequenced by FASTERIS SA on an Illumina HiSeq2000 following an Illumina paired-end protocol (fragment lengths 150–250bp, 100 bp sequenced from each end).
A normalized 454 library was generated for *S. latifolia* using bud extracts from 4 different developmental stages.

4.6.2.2 Assembly

Adaptors, low quality and identical reads were removed. The transcriptome was then assembled using TRINITY (Haas et al., 2013) on the combined 10 individuals described previously as well as the 6 individuals from (Muyle et al., 2012) and the normalized 454 sequencing that was transformed to illumina using 454-to-illumina-transformed-reads. Then, isoforms were collapsed using /trinity-plugins/rsem-1.2.0/rsem-prepare-reference. PolyA tails, bacterial RNAs and ribosomal RNAs were removed using ribopicker. ORFs were predicted with trinity transcripts_to_best_scoring_ORFs.pl.
In order to increase the probability of X and Y sequences to be assembled in the same contig, ORFs were further assembled using CAP3 (cap3 -p 70, Version Date: 10/15/07, Huang and Madan, 1999) inside of TRINITY components.

4.6.2.3 Mapping, genotyping and segregation inference

Illumina reads from the 10 individuals of the cross were mapped onto the assembly using BWA (version 0.6.2, bwa aln -n 5 and bwa sampe, Li and Durbin, 2009). The libraries were then merged using SAMTOOLS (Version 0.1.18, Li et al., 2009). The obtained alignments were locally realigned using IndelRealigner (GATK, DePristo et al., 2011; McKenna et al., 2010) and were analysed using reads2snp (Version 3.0, -fis 0 -model M2 -output_genotype best -multi_alleles acc -min_coverage 3 -par false, Tsagkogeorga et al., 2012) in order to genotype individuals at each loci while allowing for biases in allele expression and not cleaning for paralogous SNPs as X/Y SNPs tend to be filtered out by paraclean (the program that removes paralogous positions, Gayral et al., 2013). SEX-DETector was then used to infer contigs segregation types after estimation of parameters using an EM algorithm. Posterior segregation types probabilities were filtered to be higher than 0.8. See pipeline in Figure 4.1.A.

4.6.2.4 The tester set in *S. latifolia*

For various tests, we used 209 genes with previously known segregation type : 129 experimentally known autosomal genes, 31 experimentally known sex-linked genes (X/Y or X-hemizygous) and 49 X CDS from BAC sequences (Supplementary Table S2).
The sequences of these 209 genes were blasted (blast -e 1E-5, Altschul et al., 1990) onto the *de novo* assembly in order to find the corresponding ORF of each gene. Blasts were filtered for having a percentage of identity over 90% and an alignment length over 100bp

and manually checked. Multiple RNA-seq contigs were accepted for a single gene if they matched different regions of the gene. If multiple contigs matched the same region of a gene, only the contig with the best identity percentage was kept. The gene was considered inferred as sex-linked if at least one of his matching contig was sex-linked. The inferred status of the genes by SEX-DETector was then used to compute specificity and sensitivity values.
The same approach was used to compute sensitivity and specificity values for three previous studies that inferred *S. latifolia* RNA-seq contigs segregation patterns (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012).

4.6.2.5 BIC test for the presence of sex chromosomes in *S. latifolia* and *S. vulgaris*

The ML framework of the method allows the use of statistical tests, for instance testing for the actual presence of sex-linked genes in the dataset. A model with the three possible segregation types can be compared to a model with only autosomal segregation type using BIC.
$BIC(M) = -2log\mathscr{L} + klogn$
Where $BIC(M)$ is the BIC value of model M, $\mathscr{L}$ is the likelihood of the model, k is the number of free parameters and n is the sample size (number of polymorphic positions used to estimate parameters in the EM algorithm). The model with the lower BIC value is chosen. It is also possible to test for a X/Y versus a Z/W system by comparing both BIC values. In case a model with sex chromosomes fits best the data but no sex-linked genes are inferred, then it means there are no sex chromosomes in the dataset.

4.6.3 Simulations

Sequences were simulated for two parents (or a single parent in the case of a UV system) using ms to generate a coalescent tree (Hudson, 2002, see Supplementary Figure 4.S2) and then seq-gen to generate sequences using the ms tree and molecular evolution parameters (version 1.3.2x, seq-gen -mHKY -l contig_length -f 0.26 0.21 0.23 0.3 -t 2 -s theta, Rambaut and Grassly, 1997). Different types of sequences were generated: either autosomal (ms 4 1 -T) or X/Y (ms 4 1 -T -I 2 3 1 -n 2 0.25 -n 1 0.75 -ej XY_divergence_time 2 1 -eN XY_divergence_time 1) or X-hemizygous (same parameters as X/Y but no Y sequence drawn) or U/V (ms 2 1 -T -I 2 1 1 -n 2 0.5 -n 1 0.5 -ej UV_divergence 2 1 -eN UV_divergence 1). Then, allele segregation was randomly carried on for a given number of progeny of each sex, using the segregation pattern determined when generating sequences with ms and seq-gen (see Supplementary Table S1 for segregation tables).
$\theta = 4N_e\mu$ was set to 0.0275 as estimated in *S. latifolia* by (Qiu et al., 2010). μ was set to 10^{-7}, which implies that $4N_e$ was equal to ~70,000. Contig lengths were randomly attributed from the observed distribution of contigs lengths of the *S. latifolia* assembly presented previously. Equilibrium frequencies used for seq-gen were retrieved from SEX-DETector inferences on the observed *S. latifolia* data. The transition to transversion ratio

was set to 2 as inferred by PAML (Yang, 2007) on *S. latifolia* data (Käfer et al., 2013). The rate of genotyping error (ε) was set to 0.01 and the rate of Y genotyping error (p) was set to 0.13 as inferred by SEX-DETector on the observed *S. latifolia* data.
Five types of datasets were simulated, with ten repetitions for each set of parameters and 10,000 contigs were simulated for each dataset:

- Effect of X-Y divergence: Five different X-Y divergence times in units of $4N_e$ generations were tested, either *S. latifolia* X-Y divergence time (4.5My) or 10 times or 100 times older or younger. The proportion of X-hemizygous contigs among sex-linked contigs was set accordingly to X-Y divergence time: 0.002, 0.02, 0.2, 0.6 and 1 for respectively 45,000 years, 450,000 years, 4.5 My, 45 My and 450 My divergence time. As well as the proportion of Y genotyping error (since Y expression is known to decrease with X-Y divergence): 0, 0.01, 0.13, 0.2 and 1 respectively. Four offspring of each sex were simulated. The proportion of sex-linked contigs was set to 10%.
- Effect of the number of sex-linked contigs: Five different proportions of sex-linked contigs (X/Y pairs or X hemizygous) were tested: 30% (3000 sex-linked contigs out of 10,000), 5%, 1%, 0,1% and 0,01%. Four offspring of each sex were simulated and X-Y divergence was set to 4.5 My.
- Effect of theta: Three different $\theta = 4N_e\mu$ (polymorphism) were tested: 0.000275, 0.00275 and 0.0275. Five offspring of each sex were simulated and X-Y divergence was set to 4.5 My, the X-Y divergence time in unit of $4N_e$ generations varied accordingly to the value of theta. The proportion of sex-linked contigs was set to 10%.
- Effect of the number of individuals in Z/W and X/Y systems: Nine different numbers of offspring individuals of each sex were tested for the X/Y system: 2, 3, 4, 5, 6, 7, 8, 12 or 16 individuals of each sex. Sex chromosome size was set to 10% and X-Y/Z-W divergence to 4.5 My.
- Effect of the number of individuals in U/V systems: Eight different numbers of offspring individuals of each sex were tested for the U/V system: 1, 2, 3, 4, 5, 6, 7 or 8 individuals of each sex. Sex chromosome size was set to 10% and U-V divergence to 4.5 My.

For each simulated dataset, segregation types were inferred using SEX-DETector and were compared to the true segregation types in order to compute sensitivity and specificity values.

4.6.4 Galaxy workflow

A Galaxy workflow has been developed (see user guide and source codes at `http://lbbe.univ-lyon1.fr/-SEX-DETector-.html`).

4.6.5 Empirical method without a cross

The method for inbred brothers and sisters (or males and females sampled from the same population) is not model-based and relies on empirical filtering of SNPs: individuals are first genotyped using read counts, an allele is retained if it represents over 2% of the total read count at the position, a position is considered if it has more than three reads (the thresholds can be changed). Another possibility is to genotype individuals using reads2snp. Then SNPs are filtered to retrieve cases where males are all heterozygous and females all homozygous in the case of an XY system (as in (Muyle et al., 2012), and see current Supplementary Figure 4.S1). A contig is considered sex-linked if it shows at least one such SNP.

4.7 Acknowledgements

We thank Alex Widmer for access to RNA-seq datasets and comments on the manuscript, Nicolas Galtier and Sylvain Glémin (ISEM-Montpellier) for useful discussions about reads2snp, Vincent Miele (LBBE) for SEX-DETector profiling and advice on code performance, Khalid Belkhir (ISEM-montpellier) for providing and adapting Galaxy wrappers for analyses used upstream of SEX-DETector (BWA, reads2snp) and Philippe Veber (LBBE) for help with Galaxy. This work was financially supported by Agence Nationale de la Recherche grants to GABM (grants numbers: ANR-11-BSV7-013, ANR-11-BSV7-024; ANR-14-CE19-0021) and SNF project to Alex Widmer (SNF 31003A_141260).

4.8 Author contributions

AM and GABM conceived the method. FP, AM and SM designed the model. SM designed the simulations. NZ generated the RNA-seq data and built the reference transcriptome. AM implemented the method, ran the simulations and analysed the data. AM, FP and GABM interpreted the results. JK developed the Galaxy wrapper for SEX-DETector and built the Galaxy workflow. AM, FP and GABM wrote the paper with input from all authors.

4.9 References

Ahmed, S., Cock, J. M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A. F., Dittami, S. M., Corre, E., Valero, M., Aury, J.-M., Roze, D., Van de Peer, Y., Bothwell, J., Marais, G. A. B., and Coelho, S. M. (2014). A haploid system of sex determination in the brown alga Ectocarpus sp. eng. *Current biology: CB* 24(17), 1945–1957. ISSN: 1879-0445. DOI: 10.1016/j.cub.2014.07.042.

Akagi, T., Henry, I. M., Tao, R., and Comai, L. (2014). Plant genetics. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. eng. *Science (New York, N.Y.)* 346(6209), 646–650. ISSN: 1095-9203. DOI: 10.1126/science.1257225.

Al Nadaf, S., Waters, P. D., Koina, E., Deakin, J. E., Jordan, K. S., and Graves, J. A. (2010). Activity map of the tammar X chromosome shows that marsupial X inactivation is incomplete and escape is stochastic. eng. *Genome Biology* 11(12), R122. ISSN: 1465-6914. DOI: 10.1186/gb-2010-11-12-r122.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. eng. *Journal of Molecular Biology* 215(3), 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

Ayers, K. L., Davidson, N. M., Demiyah, D., Roeszler, K. N., Grützner, F., Sinclair, A. H., Oshlack, A., and Smith, C. A. (2013). RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome. eng. *Genome Biology* 14(3), R26. ISSN: 1465-6914. DOI: 10.1186/gb-2013-14-3-r26.

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. eng. *Nature Reviews. Genetics* 14(2), 113–124. ISSN: 1471-0064. DOI: 10.1038/nrg3366.

Bachtrog, D., Kirkpatrick, M., Mank, J. E., McDaniel, S. F., Pires, J. C., Rice, W., and Valenzuela, N. (2011). Are all sex chromosomes created equal? eng. *Trends in genetics: TIG* 27(9), 350–357. ISSN: 0168-9525. DOI: 10.1016/j.tig.2011.05.005.

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T.-J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., Kremitzki, C., Fulton, R. S., Dugan, S., Ding, Y., Morton, D., Khan, Z., Lewis, L., Buhay, C., Wang, Q., Watt, J., Holder, M., Lee, S., Nazareth, L., Alföldi, J., Rozen, S., Muzny, D. M., Warren, W. C., Gibbs, R. A., Wilson, R. K., and Page, D. C. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. eng. *Nature* 508(7497), 494–499. ISSN: 1476-4687. DOI: 10.1038/nature13206.

Bellott, D. W., Skaletsky, H., Pyntikova, T., Mardis, E. R., Graves, T., Kremitzki, C., Brown, L. G., Rozen, S., Warren, W. C., Wilson, R. K., and Page, D. C. (2010). Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. eng. *Nature* 466(7306), 612–616. ISSN: 1476-4687. DOI: 10.1038/nature09172.

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Forrest, A., Kamau, E., and Charlesworth, D. (2007). Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: evidence from new sex-linked genes. eng. *Genetics* 175(4), 1945–1954. ISSN: 0016-6731. DOI: 10.1534/genetics.106.070110.

Carvalho, A. B. and Clark, A. G. (2013). Efficient identification of Y chromosome sequences in the human and Drosophila genomes. eng. *Genome Research* 23(11), 1894–1907. ISSN: 1549-5469. DOI: 10.1101/gr.156034.113.

Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. eng. *Nature* 371(6494), 215–220. ISSN: 0028-0836. DOI: 10.1038/371215a0.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. eng. *Nature* 508(7497), 488–493. ISSN: 1476-4687. DOI: 10.1038/nature13151.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Angel, G. del, Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. eng. *Nature Genetics* 43(5), 491–498. ISSN: 1546-1718. DOI: 10.1038/ng.806.

Al-Dous, E. K., George, B., Al-Mahmoud, M. E., Al-Jaber, M. Y., Wang, H., Salameh, Y. M., Al-Azwani, E. K., Chaluvadi, S., Pontaroli, A. C., DeBarry, J., Arondel, V., Ohlrogge, J., Saie, I. J., Suliman-Elmeer, K. M., Bennetzen, J. L., Kruegger, R. R., and Malek, J. A. (2011). De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). eng. *Nature Biotechnology* 29(6), 521–527. ISSN: 1546-1696. DOI: 10.1038/nbt.1860.

Ferris, P., Olson, B. J. S. C., De Hoff, P. L., Douglass, S., Casero, D., Prochnik, S., Geng, S., Rai, R., Grimwood, J., Schmutz, J., Nishii, I., Hamaji, T., Nozaki, H., Pellegrini, M., and Umen, J. G. (2010). Evolution of an expanded sex-determining locus in Volvox. eng. *Science (New York, N.Y.)* 328(5976), 351–354. ISSN: 1095-9203. DOI: 10.1126/science.1186222.

Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J., and Anderson, L. K. (2007). Recombination: an underappreciated factor in the evolution of plant genomes. eng. *Nature Reviews. Genetics* 8(1), 77–84. ISSN: 1471-0056. DOI: 10.1038/nrg1970.

Gautier, M. (2014). Using genotyping data to assign markers to their chromosome type and to infer the sex of individuals: a Bayesian model-based classifier. eng. *Molecular Ecology Resources* 14(6), 1141–1159. ISSN: 1755-0998. DOI: 10.1111/1755-0998.12264.

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. eng. *PLoS genetics* 9(4), e1003457. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003457.

Gschwend, A. R., Yu, Q., Tong, E. J., Zeng, F., Han, J., VanBuren, R., Aryal, R., Charlesworth, D., Moore, P. H., Paterson, A. H., and Ming, R. (2012). Rapid divergence and expansion of the X chromosome in papaya. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), 13716–13721. ISSN: 1091-6490. DOI: 10.1073/pnas.1121096109.

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. eng. *Nature Protocols* 8(8), 1494–1512. ISSN: 1750-2799. DOI: 10.1038/nprot.2013.084.

Hough, J., Hollister, J. D., Wang, W., Barrett, S. C. H., and Wright, S. I. (2014). Genetic degeneration of old and young Y chromosomes in the flowering plant Rumex hastatulus. eng. *Proceedings of the National Academy of Sciences of the United States of America* 111(21), 7713–7718. ISSN: 1091-6490. DOI: 10.1073/pnas.1319227111.

Huang, X. and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. en. *Genome Research* 9(9), 868–877. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.9.9.868.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. eng. *Bioinformatics (Oxford, England)* 18(2), 337–338. ISSN: 1367-4803.

Hughes, J. F. and Rozen, S. (2012). Genomics and genetics of human and primate y chromosomes. eng. *Annual Review of Genomics and Human Genetics* 13, 83–108. ISSN: 1545-293X. DOI: 10.1146/annurev-genom-090711-163855.

Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Graves, T., Fulton, R. S., Dugan, S., Ding, Y., Buhay, C. J., Kremitzki, C., Wang, Q., Shen, H., Holder, M., Villasana, D., Nazareth, L. V., Cree, A., Courtney, L., Veizer, J., Kotkiewicz, H., Cho, T.-J., Koutseva, N., Rozen, S., Muzny, D. M., Warren, W. C., Gibbs, R. A., Wilson, R. K., and Page, D. C. (2012). Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. eng. *Nature* 483(7387), 82–86. ISSN: 1476-4687. DOI: 10.1038/nature10843.

Hughes, J. F., Skaletsky, H., Pyntikova, T., Graves, T. A., Daalen, S. K. M. van, Minx, P. J., Fulton, R. S., McGrath, S. D., Locke, D. P., Friedman, C., Trask, B. J., Mardis, E. R., Warren, W. C., Repping, S., Rozen, S., Wilson, R. K., and Page, D. C. (2010). Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. eng. *Nature* 463(7280), 536–539. ISSN: 1476-4687. DOI: 10.1038/nature08700.

Käfer, J., Talianová, M., Bigot, T., Michu, E., Guéguen, L., Widmer, A., žlůvová, J., Glémin, S., and Marais, G. A. B. (2013). Patterns of molecular evolution in dioecious and non-dioecious *Silene*. en. *Journal of Evolutionary Biology* 26(2), 335–346. ISSN: 1010061X. DOI: 10.1111/jeb.12052.

Kondo, M., Hornung, U., Nanda, I., Imai, S., Sasaki, T., Shimizu, A., Asakawa, S., Hori, H., Schmid, M., Shimizu, N., and Schartl, M. (2006). Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. eng. *Genome Research* 16(7), 815–826. ISSN: 1088-9051. DOI: 10.1101/gr.5016106.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. eng. *Bioinformatics (Oxford, England)* 25(14), 1754–1760. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp324.

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. eng. *Bioinformatics (Oxford, England)* 25(16), 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

Mank, J. E. and Avise, J. C. (2009). Evolutionary diversity and turn-over of sex determination in teleost fishes. eng. *Sexual Development: Genetics, Molecular Biology, Evolution, Endocrinology, Embryology, and Pathology of Sex Determination and Differentiation* 3(2-3), 60–67. ISSN: 1661-5433. DOI: 10.1159/000223071.

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. eng. *PloS One* 6(8), e21915. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021915.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. eng. *Genome Research* 20(9), 1297–1303. ISSN: 1549-5469. DOI: 10.1101/gr.107524.110.

Ming, R., Bendahmane, A., and Renner, S. S. (2011). Sex Chromosomes in Land Plants. en. *Annual Review of Plant Biology* 62(1), 485–514. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-042110-103914.

Moghadam, H. K., Pointer, M. A., Wright, A. E., Berlin, S., and Mank, J. E. (2012). W chromosome expression responds to female-specific selection. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(21), 8207–8211. ISSN: 1091-6490. DOI: 10.1073/pnas.1202721109.

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., and Marais, G. A. B. (2012). Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. eng. *PLoS biology* 10(4), e1001308. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001308.

Qiu, S., Bergero, R., Forrest, A., Kaiser, V. B., and Charlesworth, D. (2010). Nucleotide diversity in Silene latifolia autosomal and sex-linked genes. eng. *Proceedings. Biological Sciences / The Royal Society* 277(1698), 3283–3290. ISSN: 1471-2954. DOI: 10.1098/rspb.2010.0606.

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. eng. *Computer applications in the biosciences: CABIOS* 13(3), 235–238. ISSN: 0266-7061.

Rautenberg, A., Hathaway, L., Oxelman, B., and Prentice, H. C. (2010). Geographic and phylogenetic patterns in Silene section Melandrium (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. eng. *Molecular Phylogenetics and Evolution* 57(3), 978–991. ISSN: 1095-9513. DOI: 10.1016/j.ympev.2010.08.003.

Renner, S. S. (2014). The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. eng. *American Journal of Botany* 101(10), 1588–1596. ISSN: 1537-2197. DOI: 10.3732/ajb.1400196.

Skaletsky, H. et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. eng. *Nature* 423(6942), 825–837. ISSN: 0028-0836. DOI: 10.1038/nature01722.

Stöck, M., Savary, R., Betto-Colliard, C., Biollay, S., Jourdan-Pineau, H., and Perrin, N. (2013). Low rates of X-Y recombination, not turnovers, account for homomorphic sex chromosomes in several diploid species of Palearctic green toads (Bufo viridis subgroup). eng. *Journal of Evolutionary Biology* 26(3), 674–682. ISSN: 1420-9101. DOI: 10.1111/jeb.12086.

Tsagkogeorga, G., Cahais, V., and Galtier, N. (2012). The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate Ciona intestinalis. eng. *Genome Biology and Evolution* 4(8), 740–749. ISSN: 1759-6653. DOI: 10.1093/gbe/evs054.

Vicoso, B. and Bachtrog, D. (2011). Lack of global dosage compensation in Schistosoma mansoni, a female-heterogametic parasite. eng. *Genome Biology and Evolution* 3, 230–235. ISSN: 1759-6653. DOI: 10.1093/gbe/evr010.

Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S., and Bachtrog, D. (2013a). Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. eng. *PLoS biology* 11(8), e1001643. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001643.

Vicoso, B., Kaiser, V. B., and Bachtrog, D. (2013b). Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. eng. *Proceedings of the National Academy of Sciences of the United States of America* 110(16), 6453–6458. ISSN: 1091-6490. DOI: 10.1073/pnas.1217027110.

Wang, J., Na, J.-K., Yu, Q., Gschwend, A. R., Han, J., Zeng, F., Aryal, R., VanBuren, R., Murray, J. E., Zhang, W., Navajas-Pérez, R., Feltus, F. A., Lemke, C., Tong, E. J., Chen, C., Wai, C. M., Singh, R., Wang, M.-L., Min, X. J., Alam, M., Charlesworth, D., Moore, P. H., Jiang, J., Paterson, A. H., and Ming, R. (2012). Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), 13710–13715. ISSN: 1091-6490. DOI: 10.1073/pnas.1207833109.

Weeks, S. C. (2012). The role of androdioecy and gynodioecy in mediating evolutionary transitions between dioecy and hermaphroditism in the animalia. eng. *Evolution; International Journal of Organic Evolution* 66(12), 3670–3686. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2012.01714.x.

Yamato, K. T. et al. (2007). Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. eng. *Proceedings of the National Academy of Sciences of the United States of America* 104(15), 6472–6477. ISSN: 0027-8424. DOI: 10.1073/pnas.0609054104.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. eng. *Molecular Biology and Evolution* 24(8), 1586–1591. ISSN: 0737-4038. DOI: 10.1093/molbev/msm088.

4.10 Supplementary information

Large Tables available online:

Supplementary Table S1: Segregation tables, observed genotypes probabilities given true parent genotypes, segregation types and genotyping errors. https://drive.google.com/open?id=0B7KhiXOz6_OLRlVua2JRcHhzT1k&authuser=0

Supplementary Table S2: Known genes used to test SEX-DETector and compare it with other methods in *S. latifolia*.

https://drive.google.com/open?id=0B7KhiXOz6_OLTnA1UElpTU9XUzQ&authuser=0

Supplementary Table S4: Details of model comparison using SEX-DETector on real datasets (*Silene latifolia* which has sex chromosomes and *Silene vulgaris* which does not have sex chromosomes) and simulated X/Y datasets with varying number of sex-linked contigs out of 10,000 simulated contigs. The best model is chosen as the one having the lowest BIC value (bold and stressed). https://drive.google.com/open?id=0B7KhiXOz6_OLQ0lMVVphb1Zucm8&authuser=0

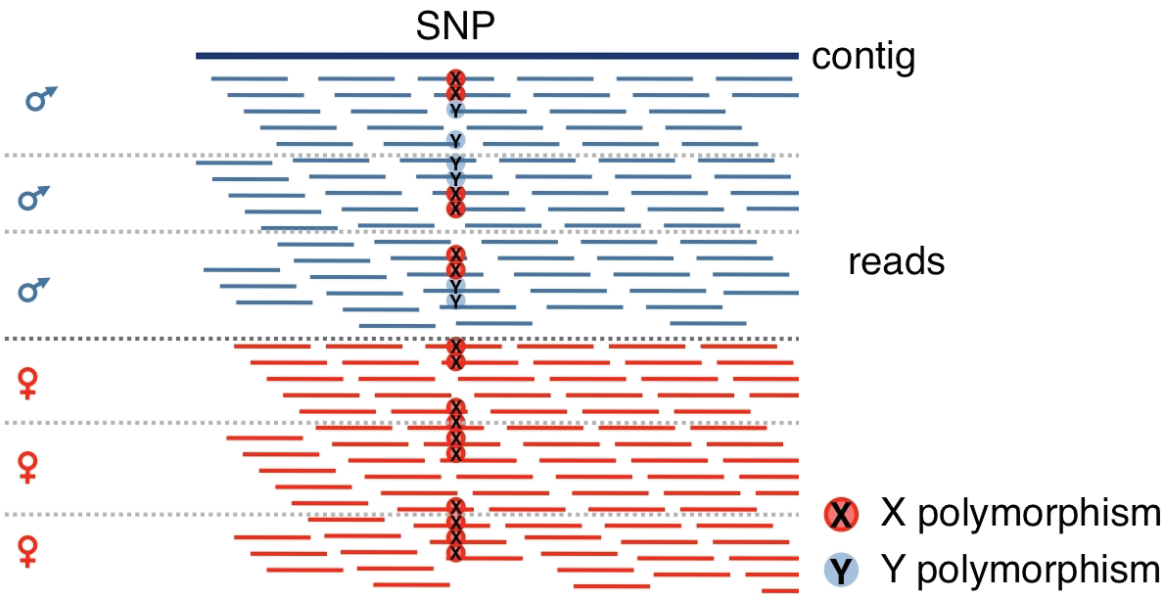


Figure 4.S1: Method without a cross: male and female genotypes are studied, a SNP is considered sex-linked if all males are heterozygous (XY) and all females homozygous (XX), the pattern is reversed in the case of a ZW system. Genotypes are either inferred using the genotyper reads2snp or from read counts.

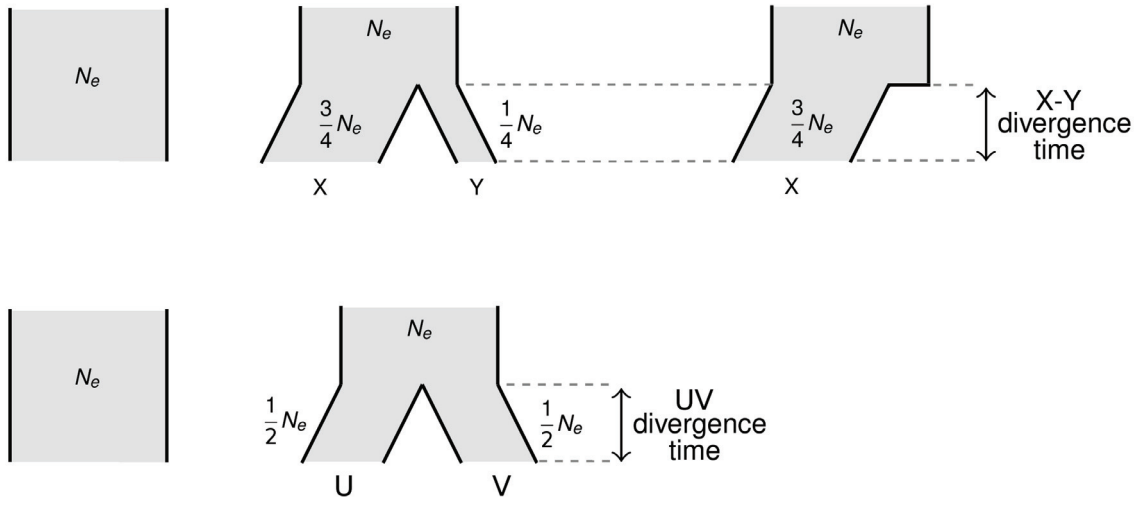


Figure 4.S2: simulations design with the program ms (Hudson, 2002), for X/Y system (upper part) and U/V system (lower part).

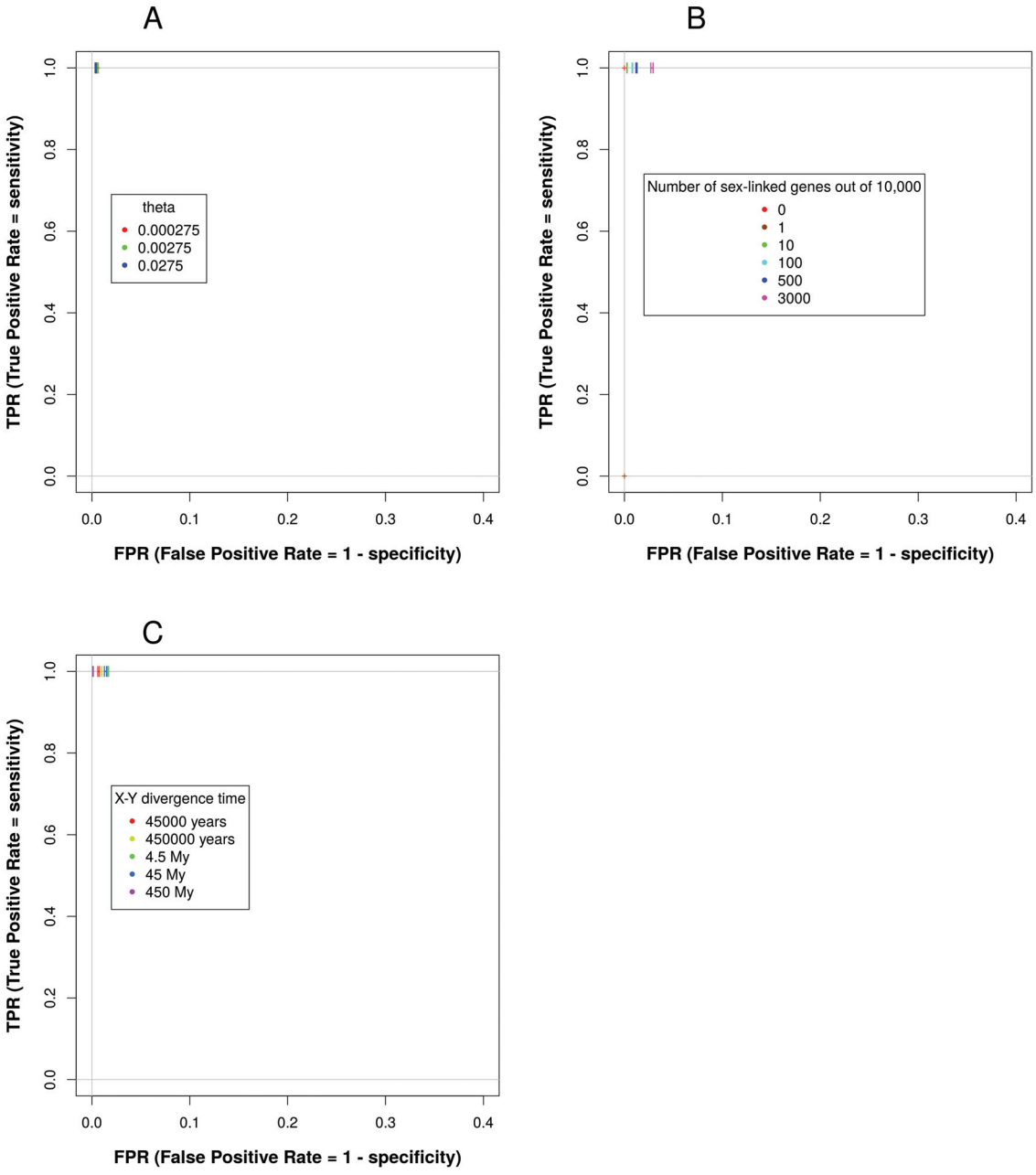


Figure 4.S3: simulations results: ROC curve showing the effect of different parameters on sensitivity (TPR, true positive rate) and specificity (1-FPR, false positive rate) in simulated data. A perfect classification of contigs would lead to a point having TPR equal to one and FPR equal to zero (top left corner of the graph). **A)** effect of X-Y divergence time. **B)** effect of the number of sex-linked contigs out of 10,000 simulated contigs. **C)** effect of theta (polymorphism).

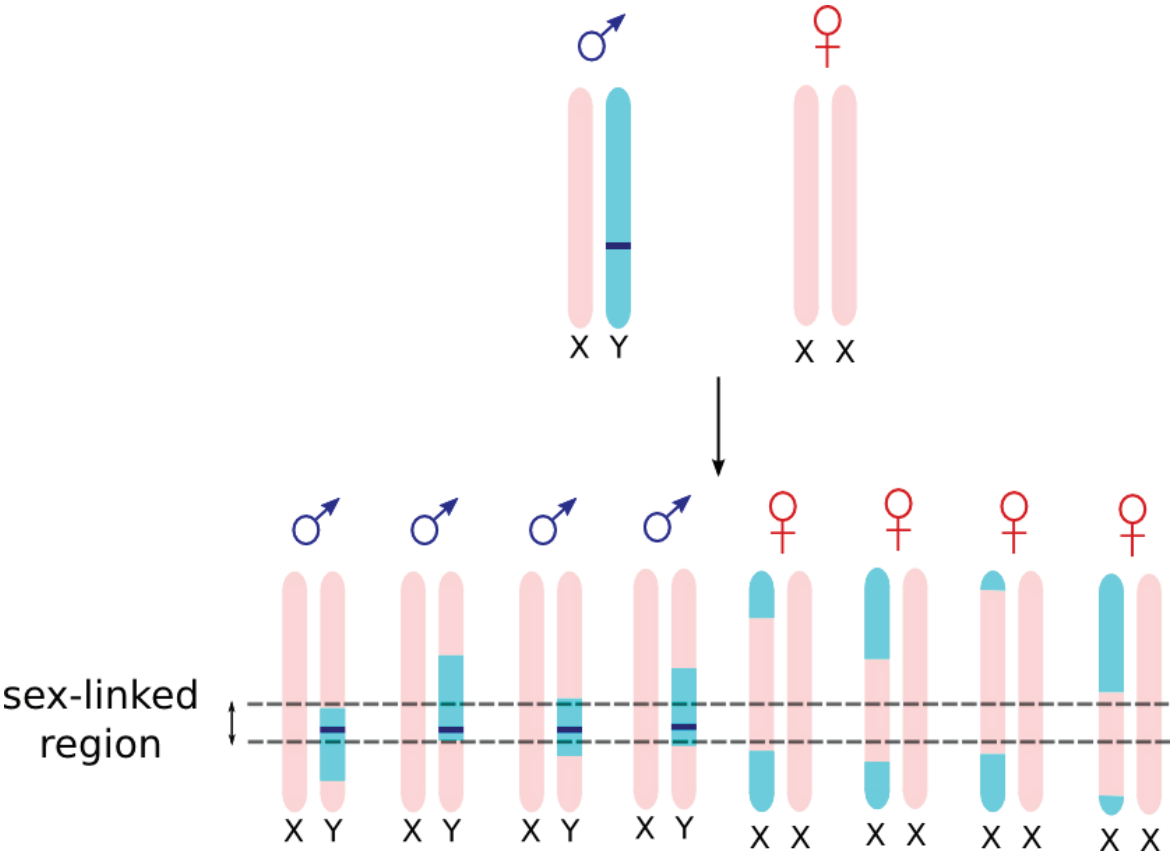


Figure 4.S4: Effect of linkage on the method: expected segregation result in a family when only one gene is sex-linked, closely linked genes will also seem sex-linked so that there is never a single sex-linked gene in a dataset that could not be detected by SEX-DETector (unlike what was observed in simulations where recombination events are not simulated).

Table 4.S3: Library sizes (in number of reads) and mapping statistics.

Species	Individual	Library size (# reads)	# mapped reads	proportion mapped reads
Silene latifolia	Leuk144-3_father	115377548	64840769	0.56
	U10_37_mother	75800468	23262670	0.31
	C1_01_male	71711204	41812869	0.58
	C1_3_male	82186876	49305677	0.6
	C1_04_male	69957770	41597899	0.59
	C1_05_male	181295644	106427848	0.59
	C1_26_female	63717900	35910978	0.56
	C1_27_female	145384384	88301626	0.61
	C1_29_female	161456150	94931546	0.59
	C1_34_female	190950550	116240597	0.61
	Total	1157838494	662632479	0.57
Silene vulgaris	V1_1	32836858	12299198	0.37
	V1_2	59903112	27797320	0.46
	V1_4	56879642	25633403	0.45
	V1_5	54786992	23011360	0.42
	V1_8	57192532	25685569	0.45
	V1_9	47704414	21347225	0.45
	Guarda1	64185890	30885401	0.48
	Seebach1	115453516	49841201	0.43
	Seebach2	98602086	45304549	0.46
	Total	587545042	261805226	0.45

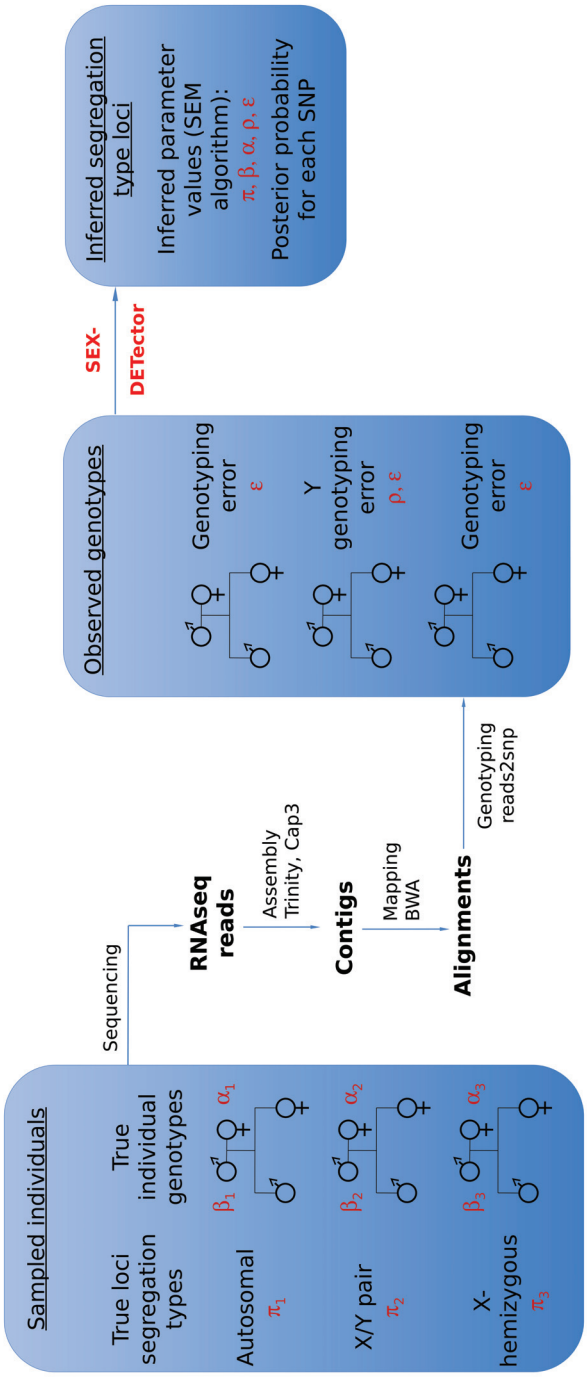


Figure 4.S5: Schematic steps of the SEX-DETector pipeline with parameters of the model. Π_j are the segregation types probabilities in the dataset ($j = 1$ for autosomal, $j = 2$ for X/Y and $j = 3$ for X-hemizygous). α_j are the proportions of real homogametic parent genotypes, in segregation type j, in the whole dataset. β_j are the proportions of real heterogametic parent genotypes, in segregation type j, in the whole dataset. ε is the probability of a genotyping error happening on any allele. p is the probability of a genotyping error happening on the Y allele, which is higher than ε due to low Y expression level in RNA-seq data. Observed genotypes can differ from real genotypes due to genotyping errors. The different parameters are estimated using an EM algorithm which then allows to compute a posterior segregation type probability for each SNP and then each contig.

4.10.1 Supplementary Text S1: SEX-DETector for XY and ZW systems

4.10.1.1 Model

4.10.1.1.1 Data description

The data consists in observed genotypes of offspring (OG) and parents (PG). The aim of this model is to fully describe the probability of each genotype for given parental genotypes (TMG, TPG) and segregation types (S).

i_r **:** "the individual i of sex r, either heterogametic (i_{het}) or homogametic (i_{hom}). $i \in [1, I]$."

k **:** "the contig $k \in [1, K]$."

t **:** "the polymorphic position t in contig k, $t \in [1, T_k]$."

j **:** "the segregation type $j \in [1, J]$.

$$j = \begin{cases} 1 & \text{if the segregation type is autosomal} \\ 2 & \text{if the segregation type is X/Y (or Z/W)} \\ 3 & \text{if the segregation type is X (or Z) hemizygous (Y or W not expressed)} \end{cases}$$

m **:** "the true homogametic parent genotype $m \in [1, M]$ with $M = 10$: (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)."

n **:** "the true heterogametic parent genotype n under segregation type j, $n \in [1, N_j]$.

$$N_j = \begin{cases} 10 & \text{if } j = 1 \text{ (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)} \\ 12 & \text{if } j = 2 \ (X^A Y^C, X^C Y^A, X^A Y^G, X^G Y^A, X^A Y^T, X^T Y^A, \\ & \quad X^C Y^G, X^G Y^C, X^C Y^T, X^T Y^C, X^G Y^T, X^T Y^G) \\ 4 & \text{if } j = 3 \ (X^A, X^C, X^G, X^T) \end{cases}$$

ℓ **:** "the diploid genotype $\ell \in [1, L]$ with $L = 10$: (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)."

Individuals, contigs and positions inside a contig are assumed independent.

4.10.1.1.2 Unobserved data

Segregation Type. $S_{ktj} = 1$ if the position t in contig k has a segregation type j (0 otherwise). It is supposed to follow a Multinomial distribution such that

$$\mathbf{S_{kt}} = (S_{kt1}, \ldots, S_{ktJ}) \sim \mathcal{M}(1; \pi_1, ..., \pi_J),$$

in the following, we denote by $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)$ and $\sum_{j=1}^{J} \pi_j = 1$.

$$\log \mathcal{L}(\mathbf{S}; \boldsymbol{\pi}) = \sum_{k,t} \sum_{j} [S_{ktj}] \log \pi_j,$$

True Homogametic Parent Genotype. $\mathrm{TMG}^m_{ktj} = 1$ if the assumed true (T) homogametic (M) parent genotype (G) is m, common to all individuals i at position t in contig k of segregation type j (0 otherwise). This is introduced to account for genotyping errors. Given the segregation type, it is supposed to follow a multinomial distribution such that:

$$\mathbf{TMG_{ktj}} = (\mathrm{TMG}^1_{ktj}, \ldots, \mathrm{TMG}^M_{ktj}) | \{S_{ktj} = 1\} \sim \mathscr{M}(1; \alpha_1, \ldots, \alpha_M)$$

In the following, we denote by $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_M)$ and $\sum_m \alpha_m = 1$.

$$\log \mathscr{L}(\mathbf{TMG}|\mathbf{S}; \boldsymbol{\alpha}) \quad = \quad \sum_{k,t} \sum_j \left[S_{ktj}\right] \sum_m \left[\mathrm{TMG}^m_{ktj}\right] \log \alpha_m$$

True Heterogametic Parent Genotype. $\mathrm{TFG}^n_{ktj} = 1$ if the assumed true (T) heterogametic (F) parent genotype (G) is n at position t in contig k of segregation type j, 0 otherwise. Given the segregation type, it is supposed to follow a Multinomial distribution such that:

$$\mathbf{TFG_{ktj}} = (\mathrm{TFG}^1_{ktj}, \ldots, \mathrm{TFG}^N_{ktj}) | \{S_{ktj} = 1\} \sim \mathscr{M}(1; \beta_1, \ldots, \beta_{N_j})$$

In the following, we denote by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)$, where $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jn}, ..., \beta_{jN_j})$ and $\sum_n \beta_{jn} = 1$. In an autosomal segregation, there is no difference between the homogametic and the heterogametic sex so that $\boldsymbol{\beta}_1 = \boldsymbol{\alpha}$.

$$\log \mathscr{L}(\mathbf{TFG}|\mathbf{S}; \boldsymbol{\beta}) \quad = \quad \sum_{k,t} \sum_j \left[S_{ktj}\right] \sum_n \left[\mathrm{TFG}^n_{ktj}\right] \log \beta_{jn}$$

Genotyping error. $\mathrm{GE}^i_{ktjmn} = 1$ if there is a genotyping error for individual i at position t in contig k of segregation type j, given parental genotypes m and n, 0 otherwise. We assume a Bernoulli distribution for this variable such that

$$\mathbb{P}(\mathrm{GE}^i_{ktjmn} = 1 | S_{ktj} = 1, \mathrm{TMG}^m_{ktj} = 1, \mathrm{TFG}^n_{ktj} = 1) = \varepsilon$$

$$\begin{aligned} &\log \mathscr{L}(\mathbf{GE}|\mathbf{S}, \mathbf{TMG}, \mathbf{TFG}; \varepsilon) \\ &= \sum_{k,t} \sum_j \left[S_{ktj}\right] \sum_{m,n} \left[\mathrm{TMG}^m_{ktj}\right] \left[\mathrm{TFG}^n_{ktj}\right] \sum_i \Big(\left[\mathrm{GE}^i_{ktjmn}\right] \log \varepsilon + (1 - \left[\mathrm{GE}^i_{ktjmn}\right]) \log(1 - \varepsilon) \Big) \end{aligned}$$

Y Genotyping Error. $\mathrm{YGE}^{i_r}_{ktjmn} = 1$ if there is a genotyping error on Y (or W) at position t in contig k for individual i (sexe r), given parental genotypes m and n, 0 otherwise. We assume a Bernoulli distribution for this variable, with p the probability of a Y (or W) genotyping error in segregation type $j = 2$ for the heterogametic sex ($r = het$) :

$$\mathbb{P}(\mathrm{YGE}^{i_{\mathrm{het}}}_{kt2mn} = 1 | S_{kt2} = 1, \mathrm{TMG}^m_{kt2} = 1, \mathrm{TFG}^n_{kt2} = 1) = p$$

For other cases ($j \neq 2$, or $j = 2$ and $r = hom$) :

$$\mathbb{P}(\mathrm{YGE}^{i_r}_{ktjmn} = 1 | S_{ktj} = 1, \mathrm{TMG}^m_{ktj} = 1, \mathrm{TFG}^n_{ktj} = 1) = 0$$

$$\begin{aligned} &\log \mathscr{L}(\mathbf{YGE}|\mathbf{S}, \mathbf{TMG}, \mathbf{TFG}; p) \\ &= \sum_{k,t} \left[S_{kt2}\right] \sum_{m,n} \left[\mathrm{TMG}^m_{kt2}\right] \left[\mathrm{TFG}^n_{kt2}\right] \sum_{i_{\mathrm{het}}} \Big(\left[\mathrm{YGE}^{i_{\mathrm{het}}}_{kt2mn}\right] \log p + (1 - \left[\mathrm{YGE}^{i_{\mathrm{het}}}_{kt2mn}\right]) \log(1 - p) \Big) \end{aligned}$$

All terms but those in ($j = 2, r = het$) vanish as they are non informative regarding parameter p.

4.10.1.1.3 Observed data

The data consists in the observed genotypes :

$\mathbf{G^{i_r}_{kt}} = (\mathbf{OG^{i_r}_{kt}}, \mathbf{PG^{i_r}_{kt}})$ **:** The observed genotype (G) for individual i, Parent (P) or Offspring (O) of sex r in contig k at position t, with $\mathbf{OG^{i_r}_{kt}} = (\mathrm{OG}^{i_r\ell}_{kt})_{\ell\in[1,L]}$, and $\mathrm{OG}^{i_r\ell}_{kt} = 1$ if offspring i of sex r has genotype ℓ in contig k at position t, 0 otherwise (same for $\mathbf{PG^{i_r}_{kt}}$).

When conditioning by every hidden variable the distribution of the observed genotype is multinomial, and fully determined by the parameters contained in the segregation tables:

- $\lambda^{hd,r}_{jmn\ell}$ is the probability of observing the genotype ℓ in the offspring for an individual of sex r (heterogametic or homogametic) given the segregation type j, the homogametic parent true genotype m, the heterogametic parent true genotype n, the genotyping error h (either with an error $h = \varepsilon$ or without error $h = (1-\varepsilon)$) and the Y (or W) genotyping error d (either with an error $d = p$ or without error $d = (1-p)$). $\lambda^{hd,r}_{jmn\ell}$ is fully determined by the transmission probability of each genotype (see segregation tables).

- $\mu^{hd,r}_{jw\ell}$ is the probability of observing the genotype ℓ in the parent of sex r (heterogametic or homogametic) given the segregation type j, the true genotype w (either m or n), the genotyping error h (either with an error $h = \varepsilon$ or without error $h = (1-\varepsilon)$) and the Y (or W) genotyping error d (either with an error $d = p$ or without error $d = (1-p)$). $\mu^{hd,r}_{jw\ell}$ is fully determined by the equiprobability of observing any genotype given that a genotyping error occured (see segregation tables).

This leads to the following conditional likelihood that can be computed separately for different contigs, positions, and segregation types, such that:

$$\begin{aligned}
&\log \mathscr{L}(\mathbf{G}|\mathbf{S},\mathbf{TMG},\mathbf{TFG},\mathbf{GE},\mathbf{YGE};\boldsymbol{\theta}) \\
&= \sum_{k,t,j} \log \mathscr{L}(\mathbf{G_{kt}}|S_{ktj},\mathbf{TMG_{kt}},\mathbf{TFG_{kt}},\mathbf{GE_{kt}},\mathbf{YGE_{kt}};\boldsymbol{\theta}) \\
&= \sum_{k,t,j} \left[S_{ktj}\right] \sum_{m,n} \left[\mathrm{TMG}^m_{ktj}\right]\left[\mathrm{TFG}^n_{ktj}\right] \sum_i \log \mathbb{P}(\mathbf{G}^{i_r}_{kt}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^n_{ktj}=1;\boldsymbol{\theta})
\end{aligned}$$

Parameter $\boldsymbol{\theta}$ stands for the complete parameters of the model, that are $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \varepsilon, p)$ and are estimated by maximum likelihood using SEM algorithm (Stochastic Expectation Maximization).

4.10.1.2 Expectation step

$\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(g)})$ $\boldsymbol{\theta}^{(g)}$ is the current estimation of parameters.

$$\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(g)}) = \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\Big(\log \mathscr{L}(\mathbf{G},\mathbf{S},\mathbf{TMG},\mathbf{TFG},\mathbf{GE},\mathbf{YGE}|\boldsymbol{\theta})\Big|\mathbf{G}\Big)$$

Where

$$\begin{aligned}\log\mathscr{L}(\mathbf{G},\mathbf{S},\mathbf{TMG},\mathbf{TFG},\mathbf{GE},\mathbf{YGE};\boldsymbol{\theta}) &= \log\mathscr{L}(\mathbf{G}|\mathbf{S},\mathbf{TMG},\mathbf{TFG},\mathbf{GE},\mathbf{YGE};\boldsymbol{\theta})\\ &+ \log\mathscr{L}(\mathbf{S};\boldsymbol{\pi})\\ &+ \log\mathscr{L}(\mathbf{TMG}|\mathbf{S},\boldsymbol{\alpha})\\ &+ \log\mathscr{L}(\mathbf{TFG}|\mathbf{S};\boldsymbol{\beta})\\ &+ \log\mathscr{L}(\mathbf{GE}|\mathbf{S},\mathbf{TMG},\mathbf{TFG};\varepsilon)\\ &+ \log\mathscr{L}(\mathbf{YGE}|\mathbf{S},\mathbf{TMG},\mathbf{TFG};p)\end{aligned}$$

As explained previously, this likelihood can be computed separately for contigs, positions, and segregation types,

$$\mathscr{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(g)}) = \sum_{k,t,j}\mathscr{Q}_{ktj}(\boldsymbol{\theta},\boldsymbol{\theta}^{(g)})$$

Posterior Segregation type. Using Bayes rule we get the posterior segregation type:

$$\begin{aligned}\widehat{S}_{ktj} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(S_{ktj}|\mathbf{G_{kt}}\right) = \frac{\pi_j\mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1)}{\sum_{j'}\pi_{j'}\mathbb{P}(\mathbf{G_{kt}}|S_{ktj'}=1)}\\ &= \frac{\pi_j\sum_{m,n}\alpha_m\beta_{jn}\prod_i\mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^n_{ktj}=1)}{\sum_{j'}\pi_{j'}\sum_{m,n}\alpha_m\beta_{j'n}\prod_i\mathbb{P}(G^i_{kt}|S_{ktj'}=1,\mathrm{TMG}^m_{ktj'}=1,\mathrm{TFG}^n_{ktj'}=1)}\end{aligned}$$

Posterior true parent genotypes.

$$\begin{aligned}\widehat{\mathrm{TMG}}^m_{ktj} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(\mathrm{TMG}^m_{ktj}|S_{ktj}=1,\mathbf{G_{kt}}\right)\\ &= \frac{\alpha_m\mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1)}{\sum_{m'}\alpha_{m'}\mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1,\mathrm{TMG}^{m'}_{ktj}=1)}\\ &= \frac{\alpha_m\sum_n\beta_{jn}\prod_i\mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^n_{ktj}=1)}{\sum_{m'}\alpha_{m'}\sum_n\beta_{jn}\prod_i\mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TMG}^{m'}_{ktj}=1,\mathrm{TFG}^n_{ktj}=1)}\end{aligned}$$

$$\begin{aligned}\widehat{\mathrm{TFG}}^n_{ktj} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(\mathrm{TFG}^n_{ktj}|S_{ktj}=1,\mathbf{G_{kt}}\right)\\ &= \frac{\beta_{jn}\mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1,\mathrm{TFG}^n_{ktj}=1)}{\sum_{n'}\beta_{jn'}\mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1,\mathrm{TFG}^{n'}_{ktj}=1)}\\ &= \frac{\beta_{jn}\sum_m\alpha_m\prod_i\mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^n_{ktj}=1)}{\sum_{n'}\beta_{jn'}\sum_m\alpha_m\prod_i\mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^{n'}_{ktj}=1)}\end{aligned}$$

Posterior Genotyping Errors

$$\begin{aligned}\widehat{\mathrm{GE}}^i_{ktjmn} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(\mathrm{GE}^i_{ktjmn}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^n_{ktj}=1,\mathbf{G_{kt}}\right)\\ &= \frac{\varepsilon\times\mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^n_{ktj}=1,\mathrm{GE}^i_{ktjmn}=1)}{\mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TMG}^m_{ktj}=1,\mathrm{TFG}^n_{ktj}=1)}\end{aligned}$$

If $j = 2$ and $r = het$ then:

$$\begin{aligned}\widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(\mathrm{YGE}^{i_{\mathrm{het}}}_{kt2mn} | S_{kt2}=1, \mathrm{TMG}^{m}_{ktj}=1, \mathrm{TFG}^{n}_{ktj}=1, \mathbf{G_{kt}}\right) \\ &= \frac{p \times \mathbb{P}(G^{i_{\mathrm{het}}}_{kt} | S_{kt2}=1, \mathrm{TMG}^{m}_{ktj}=1, \mathrm{TFG}^{n}_{ktj}=1, \mathrm{YGE}^{i_{\mathrm{het}}}_{kt2mn}=1)}{\mathbb{P}(G^{i_{\mathrm{het}}}_{kt} | S_{kt2}=1, \mathrm{TMG}^{m}_{kt2}=1, \mathrm{TFG}^{n}_{kt2}=1)}\end{aligned}$$

Otherwise $\widehat{\mathrm{YGE}}^{i_r}_{ktjmn} = 0$.

Constraints on $\widehat{\mathbf{GE}}^{i_{\mathbf{het}}}_{kt2mn}$ and $\widehat{\mathbf{YGE}}^{i_{\mathbf{het}}}_{kt2mn}$. For X/Y segregation type ($j = 2$) and heterogametic individuals ($r = het$) in some cases the observed genotype can be observed due to many different combinations of genotyping errors (Y or other). In these cases we apply a parsimonious approach and infer there was always the least possible number of genotyping errors. We also infer that when a Y genotyping error is possible, it is more likely than another genotyping error.

For $j = 2$ and heterogametic progeny ($r = het$) :

- if $\mathrm{OG}^{i_{\mathrm{het}}\ell}_{kt} = 1$ and $\lambda^{(1-\varepsilon)(1-p),het}_{2mn\ell} = 0$ and $\lambda^{(1-\varepsilon)p,het}_{2mn\ell} \neq 0$ and $\lambda^{\varepsilon(1-p),het}_{2mn\ell} \neq 0$ then $\widehat{\mathrm{GE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ and $\widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} = 1$ (the observed genotype cannot be obtained without a genotyping error, it is more likely that a Y genotyping error occured).
- if $\mathrm{OG}^{i_{\mathrm{het}}\ell}_{kt} = 1$ and $\lambda^{(1-\varepsilon)(1-p),het}_{2mn\ell} \neq 0$ and $\lambda^{(1-\varepsilon)p,het}_{2mn\ell} = 0$ and $\lambda^{\varepsilon(1-p),het}_{2mn\ell} \neq 0$ and $\lambda^{\varepsilon p,het}_{2mn\ell} \neq 0$ then $\widehat{\mathrm{GE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ and $\widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ (it is more likely that the observed genotype was obtained without any genotyping error, Y or other).
- if $\mathrm{OG}^{i_{\mathrm{het}}\ell}_{kt} = 1$ and $\lambda^{(1-\varepsilon)(1-p),het}_{2mn\ell} \neq 0$ and $\lambda^{(1-\varepsilon)p,het}_{2mn\ell} = 0$ and $\lambda^{\varepsilon(1-p),het}_{2mn\ell} = 0$ and $\lambda^{\varepsilon p,het}_{2mn\ell} \neq 0$ then $\widehat{\mathrm{GE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ and $\widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ (the observed genotype can be obtained either without any genotyping error, or with both a Y and another genotyping error, we infer there was no genotyping error).

For $j = 2$ and heterogametic parent ($r = het$) :

- if $\mathrm{PG}^{i_{\mathrm{het}}\ell}_{kt} = 1$ and $\mu^{(1-\varepsilon)(1-p),het}_{2n\ell} = 0$ and $\mu^{(1-\varepsilon)p,het}_{2n\ell} \neq 0$ and $\mu^{\varepsilon(1-p),het}_{2n\ell} \neq 0$ then $\widehat{\mathrm{GE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ and $\widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} = 1$ (the observed genotype cannot be obtained without a genotyping error, it is more likely that a Y genotyping error occured).
- if $\mathrm{PG}^{i_{\mathrm{het}}\ell}_{kt} = 1$ and $\mu^{(1-\varepsilon)(1-p),het}_{2n\ell} \neq 0$ and $\mu^{(1-\varepsilon)p,het}_{2n\ell} = 0$ and $\mu^{\varepsilon(1-p),het}_{2n\ell} = 0$ and $\mu^{\varepsilon p,het}_{2n\ell} \neq 0$ then $\widehat{\mathrm{GE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ and $\widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} = 0$ (the observed genotype can be obtained either without any genotyping error, or with both a Y and another genotyping error, we infer there was no genotyping error).

4.10.1.3 Maximization step

$$\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\mathbf{g})})}{\partial \boldsymbol{\theta}} = 0$$

Using Lagrange multipliers, new estimations of the parameters can be computed and then used for the next iteration.

$$
\begin{aligned}
\widehat{\pi}_j^{(g+1)} &= \frac{\sum_{k,t} \widehat{S}_{ktj}^{(g)}}{\sum_{k,t} 1} \\
\widehat{\alpha}_m^{(g+1)} = \widehat{\beta}_{1m}^{(g+1)} &= \frac{\sum_{k,t,j} \widehat{S}_{ktj}^{(g)} \widehat{\mathrm{TMG}}_{ktj}^{m(g)} + \sum_{k,t} \widehat{S}_{kt1}^{(g)} \widehat{\mathrm{TFG}}_{kt1}^{m(g)}}{\sum_{k,t} \sum_j \widehat{S}_{ktj}^{(g)} + \sum_{k,t} \widehat{S}_{kt1}^{(g)}} \\
\text{for } j \neq 1 : \widehat{\beta}_{jn}^{(g+1)} &= \frac{\sum_{k,t} \widehat{S}_{ktj}^{(g)} \widehat{\mathrm{TFG}}_{ktj}^{n(g)}}{\sum_{k,t} \widehat{S}_{ktj}^{(g)}} \\
\widehat{p}^{(g+1)} &= \frac{\sum_{k,t} \widehat{S}_{kt2}^{(g)} \sum_{m,n} \widehat{\mathrm{TMG}}_{kt2}^{m(g)} \widehat{\mathrm{TFG}}_{kt2}^{n(g)} \left(\sum_{i_{\mathrm{het}}} \widehat{\mathrm{YGE}}_{kt2mn}^{i_{\mathrm{het}}(g)} \right)}{\sum_{k,t} \widehat{S}_{kt2}^{(g)} \sum_{m,n} \widehat{\mathrm{TMG}}_{kt2}^{m(g)} \widehat{\mathrm{TFG}}_{kt2}^{n(g)} \left(\sum_{i_{\mathrm{het}}} 1 \right)} \\
\widehat{\varepsilon}^{(g+1)} &= \frac{\sum_{k,t} \sum_j \widehat{S}_{ktj}^{(g)} \sum_{m,n} \widehat{\mathrm{TMG}}_{ktj}^{m(g)} \widehat{\mathrm{TFG}}_{ktj}^{n(g)} \left(\sum_i \widehat{\mathrm{GE}}_{ktjmn}^{i(g)} \right)}{\sum_{k,t} \sum_j \widehat{S}_{ktj}^{(g)} \sum_{m,n} \widehat{\mathrm{TMG}}_{ktj}^{m(g)} \widehat{\mathrm{TFG}}_{ktj}^{n(g)} \left(\sum_i 1 \right)}
\end{aligned}
$$

Parameters are estimated by maximum likelihood using a Stochastic Expectation Maximisation (SEM) algorithm : starting from given parameter values we disturb them (S step) in order to randomise initial parameter values. Then we compute the expectation of each hidden variable given the observed data (E step), then these values are disturbed (S step) and are used to compute new parameter values that maximise the expectation of the likelihood (M step). After ten iterations only E and M steps are done until parameter values converge.

4.10.1.4 Attribution of contigs to segregation types

Once parameters have been estimated by the EM algorithm above, contigs are attributed to a segregation category using positions that are polymorphic and informative. A position that was inferred as hemizygous and that is not monomorphic is always informative. A position that was inferred as autosomal or X/Y is considered informative only if the heterogametic parent is heterozygous and has a genotype that is different from the homogametic parent (otherwise it is not possible to differentiate between X/Y and autosomal segregation). A weighted mean segregation type is computed as follow, with different weights attributed to the positions according to the probability of genotyping errors (if a site has a high probability of genotyping error then it will be given less weight in the final

decision for the contig segregation type):

$$
\begin{aligned}
\widehat{S}_{k1} &= \frac{\sum_t \widehat{S}_{kt1} \sum_{m,n} \widehat{\mathrm{TMG}}^m_{kt1} \widehat{\mathrm{TFG}}^n_{kt1} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{kt1mn} \right)}{D} \\
\widehat{S}_{k2} &= \frac{\sum_t \widehat{S}_{kt2} \sum_{m,n} \widehat{\mathrm{TMG}}^m_{kt2} \widehat{\mathrm{TFG}}^n_{kt2} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{kt2mn} \right) \left(\sum_{i_{\mathrm{het}}} 1 - \widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} \right)}{D} \\
\widehat{S}_{k3} &= \frac{\sum_t \widehat{S}_{kt3} \sum_{m,n} \widehat{\mathrm{TMG}}^m_{kt3} \widehat{\mathrm{TFG}}^n_{kt3} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{kt3mn} \right)}{D} \\
\textit{where} & \\
D &= \sum_t \sum_{j \neq 2} \widehat{S}_{ktj} \sum_{m,n} \widehat{\mathrm{TMG}}^m_{ktj} \widehat{\mathrm{TFG}}^n_{ktj} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{ktjmn} \right) \\
&+ \sum_t \widehat{S}_{kt2} \sum_{m,n} \widehat{\mathrm{TMG}}^m_{kt2} \widehat{\mathrm{TFG}}^n_{kt2} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{kt2mn} \right) \left(\sum_{i_{\mathrm{het}}} 1 - \widehat{\mathrm{YGE}}^{i_{\mathrm{het}}}_{kt2mn} \right)
\end{aligned}
$$

A contig is attributed to a sex-linked segregation if its posterior probability to be autosomal is lower than the sum of being X/Y and X hemizygous and if the contig has at least one X/Y or X hemizygous SNP without error. Similarly, a contig is attributed to an autosomal segregation if its posterior probability to be autosomal is higher than the sum of being X/Y and X hemizygous and if the contig has at least one autosomal SNP without error.

For each SNP the true parental genotypes are inferred as the ones that have the highest $\widehat{\mathrm{TMG}}^m_{ktj}$ and $\widehat{\mathrm{TFG}}^n_{ktj}$ probabilities. Expression levels are retrieved using the X and Y (or Z and W) alleles predicted by this method.

4.10.2 Supplementary Text S2: SEX-DETector for UV systems

4.10.2.1 Model

4.10.2.1.1 Data description

The data consists in observed genotypes of offspring (OG) and parent (PG). The aim of this model is to fully describe the probability of each genotype for given parental genotype (TPG) and segregation types (S).

i_r **:** "the individual i of sex r, either male (i_{het}) or female (i_{hom}). $i \in [1, I]$."

k **:** "the contig $k \in [1, K]$."

t **:** "the polymorphic position t in contig k, $t \in [1, T_k]$."

j **:** "the segregation type $j \in [1, J]$.

$$j = \begin{cases} 1 & \text{if the segregation type is autosomal} \\ 2 & \text{if the segregation type is U/V} \end{cases}$$

n **:** "the true parent genotype n under segregation type j, $n \in [1, N_j]$.

$$N_j = \begin{cases} 10 & \text{if } j = 1 \text{ (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)} \\ 12 & \text{if } j = 2 \ (U^A V^C, U^C V^A, U^A V^G, U^G V^A, U^A V^T, U^T V^A, \\ & \quad U^C V^G, U^G V^C, U^C V^T, U^T V^C, U^G V^T, U^T V^G) \end{cases}$$

ℓ **:** "the diploid genotype $\ell \in [1, L]$ with $L = 10$: (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)."

Individuals, contigs and positions inside a contig are assumed independent.

4.10.2.1.2 Unobserved data

Segregation Type. $S_{ktj} = 1$ if the position t in contig k has a segregation type j (0 otherwise). It is supposed to follow a Multinomial distribution such that

$$\mathbf{S_{kt}} = (S_{kt1}, \ldots, S_{ktJ}) \sim \mathcal{M}(1; \pi_1, ..., \pi_J),$$

in the following, we denote by $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)$ and $\sum_{j=1}^{J} \pi_j = 1$.

$$\log \mathcal{L}(\mathbf{S}; \boldsymbol{\pi}) \quad = \quad \sum_{k,t} \sum_{j} \left[S_{ktj} \right] \log \pi_j,$$

True Parent Genotype. $\text{TPG}^n_{ktj} = 1$ if the assumed true (T) parent (P) genotype (G) is n at position t in contig k of segregation type j, 0 otherwise. Given the segregation type, it is supposed to follow a Multinomial distribution such that:

$$\mathbf{TPG_{ktj}} = (\text{TPG}^1_{ktj}, \ldots, \text{TPG}^N_{ktj}) | \{S_{ktj} = 1\} \sim \mathcal{M}(1; \beta_1, \ldots, \beta_{N_j})$$

In the following, we denote by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)$, where $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jn}, ..., \beta_{jN_j})$ and $\sum_n \beta_{jn} = 1$. In an autosomal segregation, there is no difference between the female and the male sex so that $\boldsymbol{\beta}_1 = \boldsymbol{\alpha}$.

$$\log \mathscr{L}(\mathbf{TPG}|\mathbf{S};\boldsymbol{\beta}) = \sum_{k,t}\sum_{j}[S_{ktj}]\sum_{n}[\mathrm{TPG}^n_{ktj}]\log\beta_{jn}$$

Genotyping error. $\mathrm{GE}^i_{ktjn} = 1$ if there is a genotyping error for individual i at position t in contig k of segregation type j, given parental genotype n, 0 otherwise. We assume a Bernoulli distribution for this variable such that

$$\mathbb{P}(\mathrm{GE}^i_{ktjn} = 1 | S_{ktj} = 1, \mathrm{TPG}^n_{ktj} = 1) = \varepsilon$$

$$\begin{aligned}&\log \mathscr{L}(\mathbf{GE}|\mathbf{S},\mathbf{TPG};\varepsilon)\\ &= \sum_{k,t}\sum_{j}[S_{ktj}]\sum_{n}[\mathrm{TPG}^n_{ktj}]\sum_{i}\Big([\mathrm{GE}^i_{ktjn}]\log\varepsilon + (1-[\mathrm{GE}^i_{ktjn}])\log(1-\varepsilon)\Big)\end{aligned}$$

4.10.2.1.3 Observed data

The data consists in the observed genotypes :

$\mathbf{G^{i_r}_{kt}} = (\mathbf{OG^{i_r}_{kt}}, \mathbf{PG^{i}_{kt}})$ **:** The observed genotype (G) for individual i, Parent (P) or Offspring (O) of sex r in contig k at position t, with $\mathbf{OG^{i_r}_{kt}} = (\mathrm{OG}^{i_r\ell}_{kt})_{\ell\in[1,L]}$, and $\mathrm{OG}^{i_r\ell}_{kt} = 1$ if offspring i of sex r has genotype ℓ in contig k at position t, 0 otherwise (same for $\mathbf{PG^{i}_{kt}}$).

When conditioning by every hidden variable the distribution of the observed genotype is multinomial, and fully determined by the parameters contained in the segregation tables:

- $\lambda^{h,r}_{jn\ell}$ is the probability of observing the genotype ℓ in the offspring for an individual of sex r (male or female) given the segregation type j, the parent true genotype n, the genotyping error h (either with an error $h = \varepsilon$ or without error $h = (1-\varepsilon)$). $\lambda^{h,r}_{jn\ell}$ is fully determined by the transmission probability of each genotype (see segregation tables).

- $\mu^{h}_{jn\ell}$ is the probability of observing the genotype ℓ in the parent given the segregation type j, the true genotype n, the genotyping error h (either with an error $h = \varepsilon$ or without error $h = (1-\varepsilon)$). $\mu^{h}_{jn\ell}$ is fully determined by the equiprobability of observing any genotype given that a genotyping error occured (see segregation tables).

This leads to the following conditional likelihood that can be computed separately for different contigs, positions, and segregation types, such that:

$$\begin{aligned}
&\log \mathscr{L}(\mathbf{G}|\mathbf{S},\mathbf{TPG},\mathbf{GE};\boldsymbol{\theta}) \\
&= \sum_{k,t,j} \log \mathscr{L}(\mathbf{G_{kt}}|S_{ktj},\mathbf{TPG_{kt}},\mathbf{GE_{kt}};\boldsymbol{\theta}) \\
&= \sum_{k,t,j} \left[S_{ktj}\right] \sum_{n} \left[\mathrm{TPG}^n_{ktj}\right] \sum_{i} \log \mathbb{P}(\mathbf{G}^{i_r}_{kt}|S_{ktj}=1,\mathrm{TPG}^n_{ktj}=1;\boldsymbol{\theta})
\end{aligned}$$

Parameter $\boldsymbol{\theta}$ stands for the complete parameters of the model, that are $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \varepsilon, p)$ and are estimated by maximum likelihood using SEM algorithm (Stochastic Expectation Maximization).

4.10.2.2 Expectation step

$\mathscr{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(g)})$ $\boldsymbol{\theta}^{(g)}$ is the current estimation of parameters.

$$\mathscr{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(g)}) = \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\Big(\log \mathscr{L}(\mathbf{G},\mathbf{S},\mathbf{TPG},\mathbf{GE}|\boldsymbol{\theta})\Big|\mathbf{G}\Big)$$

Where

$$\begin{aligned}
\log \mathscr{L}(\mathbf{G},\mathbf{S},\mathbf{TPG},\mathbf{GE};\boldsymbol{\theta}) &= \log \mathscr{L}(\mathbf{G}|\mathbf{S},\mathbf{TPG},\mathbf{GE};\boldsymbol{\theta}) \\
&+ \log \mathscr{L}(\mathbf{S};\boldsymbol{\pi}) \\
&+ \log \mathscr{L}(\mathbf{TPG}|\mathbf{S};\boldsymbol{\beta}) \\
&+ \log \mathscr{L}(\mathbf{GE}|\mathbf{S},\mathbf{TPG};\varepsilon)
\end{aligned}$$

As explained previously, this likelihood can be computed separately for contigs, positions, and segregation types,

$$\mathscr{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(g)}) = \sum_{k,t,j} \mathscr{Q}_{ktj}(\boldsymbol{\theta},\boldsymbol{\theta}^{(g)})$$

Posterior Segregation type. Using Bayes rule we get the posterior segregation type:

$$\begin{aligned}
\widehat{S}_{ktj} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(S_{ktj}|\mathbf{G_{kt}}\right) = \frac{\pi_j \mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1)}{\sum_{j'} \pi_{j'} \mathbb{P}(\mathbf{G_{kt}}|S_{ktj'}=1)} \\
&= \frac{\pi_j \sum_n \beta_{jn} \prod_i \mathbb{P}(G^i_{kt}|S_{ktj}=1,,\mathrm{TPG}^n_{ktj}=1)}{\sum_{j'} \pi_{j'} \sum_n \beta_{j'n} \prod_i \mathbb{P}(G^i_{kt}|S_{ktj'}=1,\mathrm{TPG}^n_{ktj'}=1)}
\end{aligned}$$

Posterior true parent genotype.

$$\begin{aligned}
\widehat{\mathrm{TPG}}^n_{ktj} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(\mathrm{TPG}^n_{ktj}|S_{ktj}=1,\mathbf{G_{kt}}\right) \\
&= \frac{\beta_{jn}\mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1,\mathrm{TPG}^n_{ktj}=1)}{\sum_{n'} \beta_{jn'}\mathbb{P}(\mathbf{G_{kt}}|S_{ktj}=1,\mathrm{TPG}^{n'}_{ktj}=1)} \\
&= \frac{\beta_{jn}\prod_i \mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TPG}^n_{ktj}=1)}{\sum_{n'} \beta_{jn'}\prod_i \mathbb{P}(G^i_{kt}|S_{ktj}=1,\mathrm{TPG}^{n'}_{ktj}=1)}
\end{aligned}$$

Posterior Genotyping Errors

$$\begin{aligned}\widehat{\mathrm{GE}}^{i}_{ktjn} &= \mathbb{E}_{\boldsymbol{\theta}^{(g)}}\left(\mathrm{GE}^{i}_{ktjn}|S_{ktj}=1,\mathrm{TPG}^{n}_{ktj}=1,\mathbf{G_{kt}}\right)\\ &= \frac{\varepsilon\times\mathbb{P}(G^{i}_{kt}|S_{ktj}=1,\mathrm{TPG}^{n}_{ktj}=1,\mathrm{GE}^{i}_{ktjn}=1)}{\mathbb{P}(G^{i}_{kt}|S_{ktj}=1,\mathrm{TPG}^{n}_{ktj}=1)}\end{aligned}$$

4.10.2.3 Maximization step

$$\frac{\partial\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(g)})}{\partial\boldsymbol{\theta}}=0$$

Using Lagrange multipliers, new estimations of the parameters can be computed and then used for the next iteration.

$$\begin{aligned}\widehat{\pi}_{j}^{(g+1)} &= \frac{\sum_{k,t}\widehat{S}^{(g)}_{ktj}}{\sum_{k,t}1}\\ \widehat{\beta}_{jn}^{(g+1)} &= \frac{\sum_{k,t}\widehat{S}^{(g)}_{ktj}\widehat{\mathrm{TPG}}^{n(g)}_{ktj}}{\sum_{k,t}\widehat{S}^{(g)}_{ktj}}\\ \widehat{\varepsilon}^{(g+1)} &= \frac{\sum_{k,t}\sum_{j}\widehat{S}^{(g)}_{ktj}\sum_{n}\widehat{\mathrm{TPG}}^{n(g)}_{ktj}\left(\sum_{i}\widehat{\mathrm{GE}}^{i(g)}_{ktjn}\right)}{\sum_{k,t}\sum_{j}\widehat{S}^{(g)}_{ktj}\sum_{n}\widehat{\mathrm{TPG}}^{n(g)}_{ktj}\left(\sum_{i}1\right)}\end{aligned}$$

Parameters are estimated by maximum likelihood using a Stochastic Expectation Maximisation (SEM) algorithm : starting from given parameter values we disturb them (S step) in order to randomise initial parameter values. Then we compute the expectation of each hidden variable given the observed data (E step), then these values are disturbed (S step) and are used to compute new parameter values that maximise the expectation of the likelihood (M step). After ten iterations only E and M steps are done until parameter values converge.

4.10.2.4 Attribution of contigs to segregation types

Once parameters have been estimated by the EM algorithm above, contigs are attributed to a segregation category using positions that are polymorphic and informative. A position that was inferred as autosomal or U/V is considered informative only if the parent is heterozygous (otherwise it is not possible to differentiate between U/V and autosomal segregation). A weighted mean segregation type is computed as follow, with different weights attributed to the positions according to the probability of genotyping errors (if a site has a high probability of genotyping error then it will be given less weight in the final

decision for the contig segregation type):

$$\widehat{S}_{k1} = \frac{\sum_t \widehat{S}_{kt1} \sum_n \widehat{\mathrm{TPG}}^n_{kt1} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{kt1mn} \right)}{\sum_t \sum_j \widehat{S}_{ktj} \sum_n \widehat{\mathrm{TPG}}^n_{ktj} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{ktjmn} \right)}$$

$$\widehat{S}_{k2} = \frac{\sum_t \widehat{S}_{kt2} \sum_n \widehat{\mathrm{TPG}}^n_{kt2} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{kt2mn} \right)}{\sum_t \sum_j \widehat{S}_{ktj} \sum_n \widehat{\mathrm{TPG}}^n_{ktj} \left(\sum_i 1 - \widehat{\mathrm{GE}}^i_{ktjmn} \right)}$$

A contig is attributed to a sex-linked segregation if its posterior probability to be autosomal is lower than the probability to be U/V and if the contig has at least one U/V SNP without error. Similarly, a contig is attributed to an autosomal segregation if its posterior probability to be autosomal is higher than the probability to be U/V and if the contig has at least one autosomal SNP without error.

For each SNP the true parental genotype is inferred as the one that has the highest $\widehat{\mathrm{TPG}}^n_{ktj}$ probability. Expression levels are retrieved using the U and V alleles predicted by this method.

Chapter 5

Evolution of dosage compensation in *Silene latifolia*

Previous studies on dosage compensation in the plant *Silene latifolia* produced contradicting results (Bergero et al., 2015; Chibalina and Filatov, 2011; Muyle et al., 2012). The aim of this project was to try and clarify the situation by making use of SEX-DETector and the newly inferred sex-linked genes.

This project involved three labs: Gabriel Marais' lab coordinated the project and did the bioinformatic analyses, Alex Widmer's lab generated the RNA-seq data and the X chromosome genetic map and Roman Hobza's lab generated the sorted Y chromosome sequences. My contribution for this project was to run SEX-DETector on the RNA-seq datasets for the three tissues, analyse the data, prepare Tables and Figures and write the Material and Method section. I also participated in the supervision of Cécile Fruchard's Master internship from September to December 2013 during which we did a preliminary analysis of flower buds allelic expression levels to test for dosage compensation in that tissue. See Section 5.4 for more details.

The manuscript is under preparation to be submitted to *Science* as a report.

Up-regulation of the maternal X chromosome as a dosage compensation mechanism in a dioecious plant

Aline Muyle[1,2], Nicklaus Zemp[3,2], Cécile Fruchard[1], Radim Cegan[4], Jan Vrana[5], Clothilde Deschamps[6], Raquel Tavares[1], Franck Picard[1], Roman Hobza[4,5], Alex Widmer[3], Gabriel AB Marais[1]

5.1 Abstract

Y chromosome degeneration is compensated, in some animals, by adjusting X chromosome expression. How such dosage compensation mechanisms have evolved and whether they exist outside animals remains unclear. We reconstructed the evolution of expression in *Silene latifolia*, a plant with an evolutionary young XY system, and found that for many genes with reduced expression of the Y copy, the expression of the X counterpart has increased in males. This increase in X expression also happened in females, to a lower extent, and only for the maternal X. This up-regulation of the maternal X chromosome both in males and females suggests that dosage compensation, in this plant, relies on an epigenetic imprinting mechanism. Our findings reveal that dosage compensation can evolve in plants and provide insights into the first steps of the evolution of a dosage compensation mechanism.

One Sentence Summary: Our finding that in a plant XY system, up-regulation of the X chromosome has evolved in response to Y degeneration, reveals that dosage compensation can evolve in non-animal taxa.

5.2 Report

Sex chromosomes have repeatedly evolved from ordinary autosomes in different taxa (Bachtrog, 2013). Initially with similar gene content compared to their X counterparts, the Y chromosomes have degenerated and lost genes (Bachtrog, 2013). Such gene decay on the Y has caused reduced sex chromosome dosage in males, with potential negative effects on fitness, and dosage compensation has evolved to correct for this (Charlesworth, 1996). Three canonical mechanisms have been proposed so far from studies in *Drosophila*, mammals and *C. elegans* (Disteche, 2012; Ercan, 2015). They differ in whether the X

[1]Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France
[2]equal contribution as first authors
[3]ETH Zurich, Institute of Integrative Biology, Universitätstrasse 16, 8092 Zürich, Switzerland
[4]Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno
[5]Center for Biotechnological and Agricultural Research in Hana, Olomouc, Czech Republic
[6]Pole Rhone-Alpes de Bioinformatique (PRABI), Villeurbanne, France

exhibits a two-fold up-regulation in males only (Drosophila) or in both sexes, in which case an additional two-fold down-regulation involving either one (mammals) or both X (*C. elegans*) in females/hermaphrodites. Recent work has revealed that in mammals and *C. elegans*, only a minority of genes are affected by the canonical dosage compensation mechanisms described in these taxa (Albritton et al., 2014; Ercan, 2015; Pessia et al., 2012). The other genes are compensated through other mechanisms, including down-regulation of autosomal genes (Julien et al., 2012), gene relocation out of the X (Albritton et al., 2014), and some genes are not compensated at all (Julien et al., 2012; Pessia et al., 2012). Moreover, the study of dosage compensation in other taxa has shown that chromosome-wide dosage compensation has not evolved in all sex chromosomes, in particular it is more common in XY than ZW systems (Mank, 2013). How dosage compensation mechanisms have evolved and whether they exist in non-animal taxa remains unclear.

In plants, separate sexes (dioecy) has evolved in ~6% of the species and about 40 sex chromosome systems have been described (Ming et al., 2011). While the well-studied plant sex chromosomes clearly show evidence for Y degeneration (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Hough et al., 2014; Wang et al., 2012), whether dosage compensation has evolved is unclear. In *Silene latifolia*, a dioecious plant with a relatively young (~5 MY old) but already heteromorphic XY system, RNA-seq data of male and female individuals have been used to perform a high-throughput segregation analysis and identify hundreds of sex-linked genes (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012). Assuming that the deleterious effects of reduced-dosage are the strongest for sex-linked genes that have completely lost their Y copy (X-hemyzygous genes), dosage compensation is expected to evolve in those genes first. A study focusing on the X-hemyzygous genes found no evidence for dosage compensation in *S. latifolia* (Chibalina and Filatov, 2011). More recently and after an effort to experimentally validate the X-hemizygous genes in *S. latifolia*, this test was done again and returned the same result (Bergero et al., 2015). In *Rumex hastatulus*, the same test was done and provided no evidence for dosage compensation in this older system (~15 My old, Hough et al., 2014). However, reduced dosage in males will affect not only X-hemizygous genes but also sex-linked genes with reduced Y copy expression, which are numerous in *S. latifolia* (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Muyle et al., 2012). Focusing on those latter genes, some evidence for dosage compensation was found (Muyle et al., 2012). The discrepancies between all these studies are not understood yet.

We used a family (2 parents, 4 sons, 4 daughters) that we genotyped using RNA-seq data and tracked SNPs showing X-linkage and Y-linkage transmission patterns (Section 5.3). Both X/Y gene pairs and X-hemizygous genes were inferred in a statistical framework using a model-based method that we developed (Chapter 4). Similar numbers were found for RNA-seq data from three different tissues (Table 5.1); some differences are expected

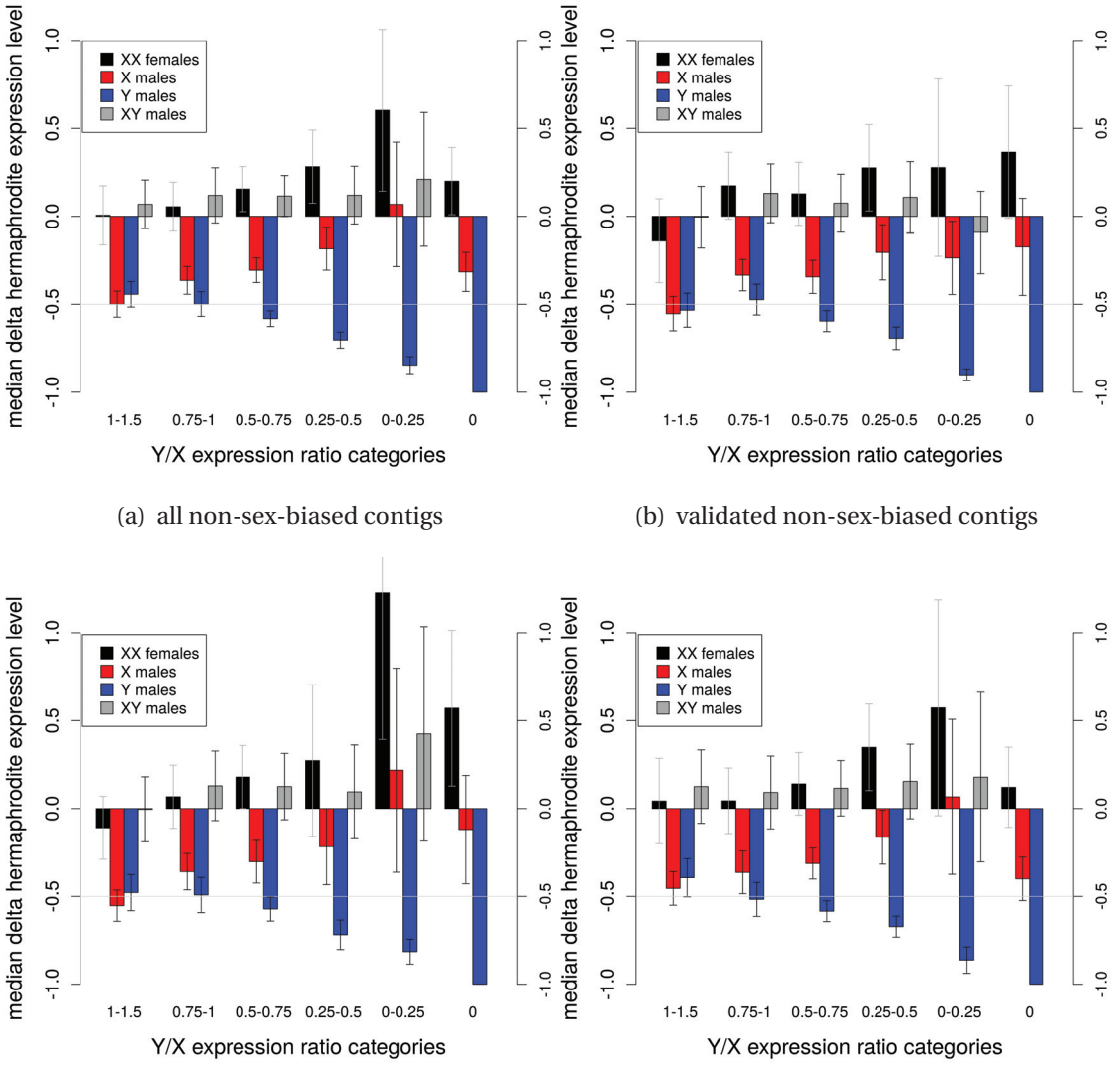
as the statistical power of our method will depend on a gene expression level, which may vary from tissue to tissue (Chapter 4). We tried to validate as many sex-linked genes as possible using different sources of data (Section 5.3): (i) a set of experimentally validated *S. latifolia* sex-linked genes found in the literature plus from a unpublished BAC sequence dataset from the *S. latifolia* sex chromosomes (Chapter 3) totaling ~80 sex-linked genes, (ii) a genetic map of the *S. latifolia* X chromosome that we built using GBS data, (iii) sequence data from isolated *S. latifolia* Y chromosome. Using these data, we validated about half of the inferred genes (Table 5.1), which thus represent a robust set of sex-linked genes. In order to study dosage compensation in *S. latifolia*, we first removed the genes that are differentially expressed among the sexes as shown in Table 1 (see Section 5.3) as no dosage compensation is expected for the sex-biased genes (Ellegren and Parsch, 2007).

Table 5.1: Numbers of inferred autosomal, X/Y, X-hemizygous and unassigned contigs in *S. latifolia* identified by SEX-DETector in three different tissues (Section 5.3). Numbers of X/Y and X-hemizygous contigs are also given after removing sex-biased contigs and after validation using different sources of data (Section 5.3).

Tissue types	total number				without sex-biased genes		without sex-biased genes, validated	
	Unassigned	Autosomal	X/Y	X-hemizygous	X/Y	X-hemizygous	X/Y	X-hemizygous
Flower buds	31072	13807	925	374	712	174	366	89
Leaves	33499	11596	780	303	733	293	374	141
Seedlings	33391	11631	869	287	737	216	395	104

We used RNA-seq read counts to estimate expression levels in *S. latifolia* and *Silene vulgaris*, a close relative without sex chromosomes, which served as an outgroup in our analysis. We used the differences in expression levels between both species to infer the changes in expression that occurred on the sex chromosomes (Section 5.3). Figure 5.1a shows that the reduction of the expression of the Y-linked copy (measured by the Y/X ratio) has been accompanied by an increase of the expression of the X-linked copy in males, so that in that sex the overall expression of the sex-linked genes has remained unchanged since the split between *S. latifolia* and *S. vulgaris* (the inter-species expression difference is close to zero). Importantly, we observed the same results when focusing on the validated sex-linked only (Figure 5.1b). In Figure 5.1 data in leaves were used as this tissue includes the smallest fraction of sex-biased genes (Chapter 6); the same analysis in other tissues (flower buds and seedlings) gave similar results (Figures 5.S1 to 5.S3 a-b). Overall, these results clearly suggest that many X/Y genes are compensated in *S. latifolia*, and that X chromosome dosage compensation occurs in several tissues.

There exists, however, a considerable heterogeneity in dosage compensation among genes on the sex chromosomes. Recent work has revealed that not all the genes are dosage-compensated in the mammalian and *C. elegans* systems; dosage compensation



(a) all non-sex-biased contigs

(b) validated non-sex-biased contigs

(c) highly expressed non-sex-biased contigs

(d) lowly expressed non-sex-biased contigs

Figure 5.1: **Leaf dataset: median delta expression levels between *S. latifolia* and *S. vulgaris* hermaphrodites**, in different Y/X ratio categories. Allelic read numbers were summed at sex-linked SNP locations and normalised for each individual and contig separately, then averaged among individuals for each contig; the median delta for all contigs was then obtained (delta=(*S. latifolia*–*S. vulgaris*)/*S. vulgaris*). If the delta variable is lower than zero then expression in *S. latifolia* is lower compared to *S. vulgaris*. If the delta variable is higher than zero then expression in *S. latifolia* is higher compared to *S. vulgaris*. And if the delta variable is equal to zero then expression in *S. latifolia* and *S. vulgaris* are equal. The Y/X ratio was computed in *S. latifolia* males and averaged between individuals to use as a proxy for Y degeneration. **XX females**: Median delta expression levels of both X-linked alleles in females; **X males**: Median delta expression levels of the single X-linked allele in males; **Y males**: Median delta expression levels of the Y-linked allele in males; **XY males**: Median delta expression levels of the X-linked plus Y-linked alleles in males. In the absence of dosage compensation delta is expected to be -0.5 for the single X in males (expressed half as much as both autosomes from *S. vulgaris*). **(a)** All non sex-biased contigs (977 contigs), with the following sample sizes for the different Y/X ratio categories are: 0, 293; 0–0.25, 89; 0.25–0.5, 151; 0.5–0.75, 160; 0.75–1, 147; 1–1.5, 119. **(b)** Validated non sex-biased contigs (491 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 141; 0–0.25, 41; 0.25–0.5, 65; 0.5–0.75, 93; 0.75–1, 79; 1–1.5, 62. **(c)** Highly expressed non sex-biased contigs (322 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 72; 0–0.25, 34; 0.25–0.5, 62; 0.5–0.75, 58; 0.75–1, 50; 1–1.5, 39. **(d)** Lowly expressed non sex-biased contigs (655 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 221; 0–0.25, 55; 0.25–0.5, 89; 0.5–0.75, 102; 0.75–1, 97; 1–1.5, 80.

seems to have evolved mostly in the dosage-sensitive genes (Albritton et al., 2014; Julien et al., 2012; Pessia et al., 2012; Veitia et al., 2015). Also, dosage-sensitive Y-linked genes tend to degenerate slowly and in some cases never get totally inactivated (Bellott et al., 2014; Cortez et al., 2014; Veitia et al., 2015; White et al., 2015). Precise data on dosage-sensitivity of *S. latifolia* genes are not available but dosage-sensitivity is known to strongly correlate with expression level and breadth (e.g. Gout et al., 2010). We thus studied the expression changes between *S. latifolia* and *S. vulgaris* in lowly and highly expressed genes (Figures 5.1, 5.S1 and 5.S3 c-d), both categories show dosage compensation for X/Y genes, but results differ for the X-hemizygous category. Figure 5.1a clearly indicates that the expression of the X-hemizygous genes has undergone an almost two-fold decrease in *S. latifolia* males compared to *S. latifolia* females and *S. vulgaris*. This suggests that, overall, this category of sex-linked genes does not exhibit dosage compensation as previously reported (Bergero et al., 2015; Chibalina and Filatov, 2011; Hough et al., 2014). Interestingly, highly expressed X-hemizygous genes shows some evidence for dosage compensation (Figures 5.1, 5.S1 and 5.S3 c versus d). This suggests that X-hemizygous genes may include few dosage-sensitive genes, which would explain why they have lost their Y copy so fast and why there is no need to compensate for such loss. In agreement with this idea, using a GO term analysis, we found that the X-hemizygous genes are depleted in protein complex genes (i.e ribosomal protein), which are well-known dosage-sensitive genes (Pessia et al., 2012).

Dosage compensation is clearly partial in *S. latifolia*, not all the sex-linked genes showing Y degeneration are compensated (Figure 5.1, Figures 5.S1 to 5.S3 a). In mammals and *C. elegans*, which are old systems, only some key, dosage-sensitive, genes are compensated (see previous paragraph). In *S. latifolia*, the same may apply but dosage compensation may also be incomplete because the system is young and the evolution of dosage compensation is ongoing. Figure 5.1a reveals another interesting aspect of dosage compensation in *S. latifolia*; the increase in expression of the X copy observed for some sex-linked genes also evolved in female. In this system, up-regulation of the X copy has occurred in both sexes (this was observed in flower buds and seedlings too Figures 5.S1 to 5.S3 a). First, this means that the changes in expression observed in males are not due to a simple buffering mechanism observed, for example, in artificially induced aneuploids, in which auto-adjustment of gene networks partially corrects for reduction in dosage (Malone et al., 2012). Buffering mechanisms are indeed only expected in the sex where dosage has been changed (i.e. in males), not in the other sex. In *S. latifolia*, a specific mechanism has thus evolved. Moreover, this result reveals a conflict between males and females over optimal expression levels as hyperexpression in females is probably deleterious. The fitness costs for females could be overcome by the benefit of increasing expression of these genes in males. Or, because sexual selection is stronger in males than in females, selection could have selected for increased expression in males, at the expense of females, until a second correcting mechanism evolves in females (Mank, 2013). Interestingly, this hy-

perexpression in both sexes corresponds to the first step that Susumu Ohno has hypothesized for the evolution of dosage compensation in mammals (Ohno, 1967), which we may be observing in this young system. The system could also remain as is, as observed in *Tribolium* where the X chromosome is hyperexpressed in both sexes, even though the sex chromosomes are old (Prince et al., 2010). Last, this explains why dosage compensation was not easy to observe in previous work when including only *S. latifolia* RNA-seq data. As expression has changed both in males and females, simply comparing male and female expression of the X-linked copy will make it more difficult to detect a strong effect; only using an outgroup, as done here, allows to clearly uncover this mechanism.

Because we have used family data, we were able to track the expression changes on the X transmitted from the father and from the mother separately (Section 5.3). Figure 5.2 shows the *S. latifolia* and *S. vulgaris* expression differences for maternal and paternal X chromosomes. It appears that the expression of the X from the mother has been up-regulated in the sons, and also, to a lesser extent, in the daughters, while the X from the father has remained unchanged (results were confirmed on flower buds Figure 5.S4 and seedlings Figure 5.S5). This suggests that the dosage compensation mechanism in *S. latifolia* relies on the up-regulation of the X chromosome from the mother. This points towards an epigenetic imprinting mechanism where the X transmitted by the mother would bear epigenetic marks resulting in this chromosome being specifically up-regulated in both sexes, but molecular studies will be needed to confirm this. Some kind of X-inactivation was proposed some time ago (Siroky et al., 1998) because authors observed that one arm of one X chromosome is methylated in *S. latifolia* females. This methylation might be a normal state, corresponding to the paternal X, and the unmethylated X observed for males and for one X in females could be the hyperexpressed maternal X. Interestingly, the up-regulation of the maternal X chromosome is stronger in males than in females (Figure 5.2), which suggests that a mechanism counteracting the up-regulation of one X chromosome might have started to evolve in females.

The finding of dosage compensation in a plant shows that dosage compensation is not restricted to animals and may accompany the evolution of Y degeneration in non-animal taxa. Observing dosage compensation in an evolutionary young system of ~5 MY old shows that dosage compensation can evolve *de novo* relatively fast after the emergence of the sex chromosomes. In the studied plant system, sex-linked reduced-expression in male (due to Y degeneration) clearly seems to be more deleterious than X over-expression in females as assumed by Ohno. Also, many genes are not compensated, possibly because they are not dosage-sensitive or because dosage compensation has not had time to evolve yet. Dosage compensation in this system clearly bears the hallmarks of a recently-evolved and still non-optimal mechanism. How the up-regulation of the maternal X is achieved and how this mechanism was able to evolve so fast remains to be understood.

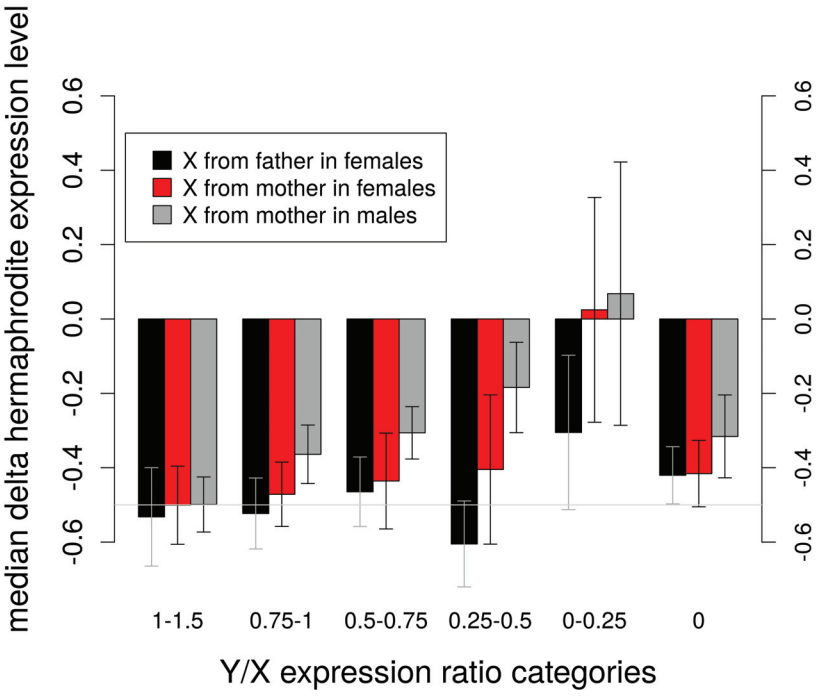


Figure 5.2: Leaf dataset: median delta expression levels between *S. latifolia* maternal and paternal X chromosomes and *S. vulgaris* hermaphrodite autosomes, in different Y/X ratio categories. Same legend as Figure 5.1 except that only SNPs where females were heterozygous were used, reducing contig numbers (626 in total) with the following sample sizes for the different Y/X ratio categories: 0, 273; 0–0.25, 49; 0.25–0.5, 73; 0.5–0.75, 83; 0.75–1, 85; 1–1.5, 63.

5.3 Material and Methods

5.3.1 RNA-seq data

5.3.1.1 Plant material and sequencing

RNA-seq data were generated from a cross in the dioecious plant *S. latifolia*, which has sex chromosomes and from a cross in the gynodioecious plant *S. vulgaris*, which does not have sex chromosomes. The following individuals and tissues were sequenced (See Supplementary Table 5.S1 for library sizes):

- For flower buds for *S. latifolia*: both parents (a male from a wild population : Leuk-144-3_father; and a female from a ten-generation inbred line : U10_37_mother), and their progeny (C1_01_male, C1_3_male, C1_04_male, C1_05_male, C1_26_female, C1_27_female, C1_29_female, C1_34_female) were sequenced. For flower buds for *S. vulgaris* the hermaphrodite father came from a wild population (Guarda_1), the female mother from another wild population (Seebach_2) and their hermaphrodite (V1_1, V1_2, V1_4) and female (V1_5, V1_8, V1_9) progeny were sequenced.
- For leaves for *S. latifolia*: the same cross progeny individuals (C1_01_male, C1-_3_male, C1_04_male, C1_05_male, C1_26_female, C1_27_female, C1_29_female, C1_34_female) were sequenced. For leaves for *S. vulgaris* four hermaphrodites

(V2_10_L, V2_11_L, V2_12_L, V2_13_L) were sequenced.

- For seedlings for *S. latifolia*: cross progeny individuals (Seed_lati_female_1, Seed_lati_female_2, Seed_lati_female_3, Seed_lati_female_4, Seed_lati_male_1, Seed_lati_male_2, Seed_lati_male_3, Seed_lati_male_4) were sequenced. For seedlings for *S. vulgaris* four hermaphrodites (Seed_vulg_herm_1, Seed_vulg_herm_2, Seed_vulg_herm_3, Seed_vulg_herm_4) were sequenced.

For the flower buds and leaves datasets, individuals were grown in a temperature-controlled greenhouse. The QiagenRNeasy Mini Plant extraction kit was used to extract total RNA two times separately from four flower buds at developmental stages B1–B2 after removing the calyx. Samples were treated additionally with QiagenDNase. RNA quality was assessed with an Aligent Bioanalyzer (RIN.9) and quantity with an Invitrogen Qubit. An intron-spanning PCR product was checked on an agarose gel to exclude the possibility of genomic DNA contamination. Then, the two extractions of the same individual were pooled. Individuals were tagged and then pooled for sequencing. Samples were sequenced by FASTERIS SA on an Illumina HiSeq2000 following an Illumina paired-end protocol (fragment lengths 150–250bp, 100 bp sequenced from each end). Seedlings were grown in a temperature controlled climate chamber in Eschikon (Switzerland) using the same conditions as in Zemp et al. (2014). The *S. latifolia* and *S. vulgaris* seedlings were collected without the roots at the four leaf stage. The sexing of the *S. latifolia* seedlings was done using Y specific markers (Hobza and Widmer, 2008) that were amplified with the direct PCAR KAPA3G Plant PCR Kit (note male number 3 was later shown to be a female). High quality RNA (RIN > 7.5) was extracted using the total RNA mini kit of Geneaid. Twelve RNA-seq libraries were produced using the Truseq kit v2 from Illimina. Libraries were tagged individually and sequenced in two Illumna Hiseq 2000 channels at the D-BSSE (ETH Zürich, Switzerland) using 100 bp paired-end read protocol.
A normalised 454 library was generated for each sex of *S. latifolia* using bud extracts from 4 different developmental stages that were pooled.
Plants from an 11 generation inbred line were grown under controlled conditions in a greenhouse in Eschikon (Switzerland). One female (U11_01) and one male (U11_02) were randomly selected. From each plant, high quality RNA (RIN > 7.5) were extracted using the total RNA mini kit of Geneaid from very small flower buds, small and large flower buds, flowers before anthesis without calyces, rosette leaves and seedlings (4 leaves stage). Additionally, pollinated flowers (1 h after pollination) and developing fruits (5 day after pollination) were extracted from the female and pollen from the male. RNA of the different tissues was equally pooled for each sex and cDNA was produced using the Clontech SMARTer Kit. The two cDNA pools were then normalized using a duplex specific endonuclease of the Evrogen TRIMMER kit. For each sex two ranges were selected (1- 1.3 kb and 1.2 -2 kb) using the Pippin Prep (Sage Science). Four SMRTbell libraries were prepared using the C2 Pacific Biosciences (PacBio) chemistry and sequenced on a PacBio RS II at the Functional Genomic Center Zurich (FGCZ).

5.3.1.2 Reference trancriptome

A reference transcriptome was built from the *Silene latifolia* flower bud data and used for all tissue types and species afterwards. Adaptors, low quality and identical reads were removed. The transcriptome was then assembled using TRINITY (Haas et al., 2013) on the combined 10 individuals described previously as well as the 6 individuals from (Muyle et al., 2012) and the normalized 454 sequencing that was transformed to illumina using 454-to-illumina-transformed-reads. Then, isoforms were collapsed using /trinity-plugins/rsem-1.2.0/rsem-prepare-reference. PolyA tails, bacterial RNAs and ribosomal RNAs were removed using ribopicker. ORFs were predicted with trinity transcripts_to_best_scoring_ORFs.pl. In order to increase the probability of X and Y sequences to be assembled in the same contig, ORFs were further assembled using home made perl scripts to run CAP3 (cap3 -p 70, Version Date: 10/15/07, Huang and Madan, 1999) inside of TRINITY components.

5.3.2 Inference of sex-linked genes

Illumina reads from the 10 individuals of the cross were mapped onto the assembly using BWA (version 0.6.2, bwa aln -n 5 and bwa sampe, Li and Durbin, 2009). The libraries were then merged using SAMTOOLS (version 0.1.18, (Li et al., 2009)). The obtained alignments were locally realigned using GATK IndelRealigner (DePristo et al., 2011) and were analysed using reads2snps (Version 3.0, -fis 0 -model M2 -output_genotype best -multi_alleles acc -min_coverage 3 -par false, Tsagkogeorga et al., 2012) in order to genotype individuals at each loci while allowing for biases in allele expression and not cleaning for paralogous SNPs as X/Y SNPs tend to be filtered out by paraclean (the program that removes paralogous positions, Gayral et al., 2013). SEX-DETector (Chapter 4) was then used to infer contigs segregation types after estimation of parameters using an SEM algorithm. Contigs posterior segregation type probabilities were filtered to be higher than 0.8. For the leaf and seedling datasets, as parents were not sequenced for these tissues, SEX-DETector was run using the data from flower buds for the parents.

5.3.3 Validation of sex-linked genes

5.3.3.1 Detection of false X-hemizygous contigs

X-hemizygous contig inference can originate from an X/Y gene which X and Y copies were assembled into different contigs. To detect such cases, X-hemizygous contigs were blasted (blast -e 1E-5, Altschul et al., 1990) against all RNA-seq contigs with male-limited expression (see below for male-limited contig inferences).

5.3.3.2 Validation using data from literature

A few sex-linked genes in *S. latifolia* have already been described in the literature (see Supplementary Table 2).

5.3.3.3 Validation using a genetic map

Plant material and genotyping: A female individual form an interspecific *S. latifolia* cross (C1_37) was back crossed with a single male from an 11 generation inbred line and the offspring (BC1 individuals) were grown under controlled conditions in a greenhouse in Eschikon (Switzerland). High quality RNA form flower buds as descript in Zemp *et al.* (2014) was extracted from 48 individuals (35 females and 13 males) of these BC1 individuals. 48 RNA-seq libraries were produced using the Truseq kit v2 from Illimina with a median insert size of about 200 bp. Individuals were tagged separately and sequenced in four Illumna Hiseq 2000 channels at the D-BSSE (ETH Zürich, Switzerland) using 100 bp paired-end read protocol. The parents used for this cross were already sequenced in another study in a similar way (C1_37, Chapter 6 and U10_49, Muyle et al., 2012).
Linkage group identification: RNA-seq reads were mapped against the *S. latifolia* flower buds reference transcriptome using BWA with a mismatch of 5. Libraries were merged and realigned using GATK (DePristo et al., 2011) and SNPs were analysed using reads2snps (Tsagkogeorga et al., 2012). Using a customized perl script SNP genotypes of the parents and the offspring and the posterior odd probabilities were extracted from the output files of reads2snps. Informative SNPs according to CP and BC1 design and a posterior odd of 0.8 were then converted into JoinMap format using a customized R script. If we had more than one informative SNP per contig we used the SNP with less segregation distortion and less missing values, which led to 8,023 BC1 and 16,243 CP-markers. The data was then imported into JoinMap 4.1 (VAN Ooijen, 2011) and loci with more than 10 % missing values were excluded, which led to 15, 118 CP and 7,951 BC1-markers. Linkage groups were identified using the default setting of JoinMap. Robustness of the assignment of the linkage groups was tested using LepMap (Rastas et al., 2013). Blasting the contigs against known sex-linked genes allowed to identify the linkage group of the X chromosome. Ordering the contigs along the linkage groups was not possible since there were not enough individuals to reach a converged order of the contigs, but contigs could be attributed to the X linkage group.

5.3.3.4 Validation using isolated Y chromosome DNA-seq data

Y chromosome DNA was isolated for sequencing using flow cytometry. The samples for flow cytometric experiments were prepared from root tips according to Vrána et al. (2012) with modifications. Seeds of *S. latifolia* were germinated in a petri dish immersed in water at 25°C for 2 days until optimal length of roots was achieved (1 cm). The root cells were synchronized by treatment with 2mM hydroxyurea at 25°C for 18h. Accumulation

of metaphases was achieved using 2.5μM oryzalin. Approximately 200 root tips were necessary to prepare 1ml of sample. The chromosomes were released from the root tips by mechanical homogenization using a Polytron PT1200 homogenizer (Kinematica AG, Littau, Switzerland) at 18,000rpm for 13 s. The crude suspension was filtered and stained with DAPI (2μg/ml). All flow cytometric experiments were performed on FACSAria II SORP flow cytometer (BD Biosciences, San José, Calif., USA). Isolated Y chromosomes were sequenced with 2x100bp PE Illumina HiSeq.
Reads were filtered for quality and Illumina adapters were removed using the ea-utils FASTQ processing utilities (Aronesty, 2011). The optimal kmer value for assembly was searched using KmerGenie (Chikhi and Medvedev, 2014). Filtered reads were assembled using soapdenovo2 (Luo et al., 2012) with kmer=49, as suggested by KmerGenie.
The obtained assembly was highly fragmented, therefore RNA-seq data was used to join, order and orient the genomic fragments with L_RNA_scaffolder (Xue et al., 2013). The following RNA-seq reads were used (see Section 5.3.1.1): one sample of male flower buds sequenced by 454, 6 samples of male flower buds sequenced by Illumina PE, 4 samples of male leaves sequenced by Illumina PE and one sample of male pooled tissues sequenced by PacBio. The genomic assembly was successively scaffolded with L_RNA_scaffolder using RNA-seq samples one after the other, first 454 samples then Illumina and finally PacBio. The obtained contigs were filtered to be longer than 200pb.

5.3.3.5 Set of validated sex-linked genes

The three sources of data (litterature, genetic map and filtered Y sequence data) were compared to SEX-DETector inferred sex-linked RNA-seq contigs using BLAST (blast -e 1E-5, Altschul et al., 1990). Blasts were filtered for having a percentage of identity over 90%, an alignment length over 100bp and were manually checked. If a sex-linked RNA-seq contig blasted against a sequence from one of the three data sources (literature, X genetic map or filtered Y DNA-seq) it was then considered as validated.

5.3.4 Expression level estimates

5.3.4.1 whole contig expression levels

Whole contig mean expression levels were obtained for each individual using GATK DepthOfCoverage (DePristo et al., 2011) as the sum of every position coverage, divided by the length of the contig. Normalised expression levels (in RPKM, Oshlack et al., 2010) were then computed for each individual by dividing the value by the library size of the individual (total number of mapped reads), allowing different depths of coverage between individuals. Whole contig mean male and female expression levels were then computed by averaging male and female individuals for each contig.

5.3.4.2 Allelic expression levels

In order to study separately X and Y allele expression levels in males and females, expression levels were studied at the SNP level. In *S. latifolia*, for each sex-linked contig expression levels were estimated using read counts from both X/Y and X-hemizygous informative SNPs: SNPs were attributed to an X/Y or X hemizygous segregation type if the according posterior probability was higher than 0.5, and SNPs are informative if the father is heterozygous and has a genotype that is different from the mother (otherwise it is not possible to tell apart the X from the Y allele and therefore it is not possible to compute X and Y expression separately). X/Y SNPs for which at least one female had over two percent of her reads belonging to the Y allele were removed as unlikely to be true X/Y SNPs. Contigwise X, Y, X+X and X+Y normalised expression levels (in RPKM, Oshlack et al., 2010) were computed by summing read numbers for each X-linked or Y-linked alleles for all SNPs and each individual separately and then normalised using the total number of mapped reads per individuals (library size) and the number of studied sex-linked SNPs in the contig:

$$E = \frac{r}{n \times l} \tag{5.1}$$

With E = normalised expression level for a given individual, r = sum of total read counts, n = number of studied SNPs, l = library size of the individual (number of mapped reads). For contigs that only have X/X SNPs (SNPs for which the father's X is different to both Xs of the mother), Y expression level is only computed from the father as all males are homozygous in the progeny. Such contigs were therefore removed when having under 3 X/X SNPs to avoid approximations on the contig mean Y/X expression level (39 contigs removed in the flower buds dataset, 44 in the leaves dataset and 40 in the seedlings dataset).

In order to make *S. latifolia* and *S. vulgaris* expression levels comparable for sex-linked contigs, *S. vulgaris* contig wise expression levels were estimated using only the positions used in *S. latifolia* (informative X/Y or X-hemizygous SNPs). The read count of every position in every contig and for every *S. vulgaris* individual was given by GATK DepthOfCoverage (DePristo et al., 2011). Only positions corresponding to informative X/Y or X-hemizygous SNPs in *S. latifolia* were used to compute the expression level for each contig and each *S. vulgaris* individual as explained in Equation 5.1.

Contigwise *S. latifolia* X, Y, X+X, X+Y and *S. vulgaris* autosomes normalised allelic expression levels were then averaged between individuals.

5.3.5 Expression divergence between *S. latifolia* and *S. vulgaris*

A variable called Δ (Equation 5.2) was computed in order to study how expression levels evolved in *S. latifolia* on the sex chromosomes compared to autosomal expression levels in *S. vulgaris*: Δ is equal to zero if *S. latifolia* and *S. vulgaris* have equal expression levels, Δ

is positive if *S. latifolia* has higher expression levels compared to *S. vulgaris* and Δ is negative otherwise. For the leaf dataset, *S. vulgaris* expression levels were globally lower for the whole transcriptome than *S. latifolia*, even after correcting for individual library sizes. For the flower bud dataset, *S. vulgaris* expression levels were globally higher for the whole transcriptome than *S. latifolia*. Mean expression levels were similar between *S. latifolia* and *S. vulgaris* for the seedling dataset. The mean *S. vulgaris* over *S. latifolia* whole contig expression ratio was computed and averaged for all contigs of the transcriptome. This ratio (0.859 for leaves, 1.093 for flower buds for the *S. vulgaris* hermaphrodite dataset, 1.129 for flower buds for the *S. vulgaris* female dataset and 0.963 for the seedlings) was used to account for the bias of global different expression levels between *S. latifolia* and *S. vulgaris* when computing Δ:

$$\Delta = \frac{\text{S. latifolia expression level} \times \text{correcting ratio} - \text{S. vulgaris expression level}}{\text{S. vulgaris expression level}} \quad (5.2)$$

5.3.6 Identification of sex-biased genes

The analysis was done separately for the three tissues (flower buds, seedling and rosette leaves). Raw reads counts were extracted from the bam files and imported into R. Contigs were filtered to be sufficiently expressed in at least half of the female and/or male libraries using a cutoff of 1 count per million reads. Genes with sex-biased expression were identified using a common dispersion in edgeR (Robinson et al., 2010) including the different replicates for the contigs inferred to sex-linked and X-hemizygous separately. See Table 1 for numbers of sex-biased contigs removed in order to study dosage compensation.
Male-limited expressed contigs were identified by calculating the mean expression values (FPKM) in both sexes and selecting those which were exclusively expressed in males.

5.3.7 GO term analysis

All contigs of the reference transcriptome were blasted against the Uniprot database using blastx with an expectation value cutoff of 0.001, and an upper limit of hit results per sequence of 20. Significant blastx hits were obtained for 35,763 contigs. Gene Ontology terms were associated with the contigs using the Blast2GO PRO version V 2.7.2 (Conesa et al., 2005) and the Data Base version b2g_sep14, with 41,136 GO terms and 4,097 Enzymes available. Blast hits were related to the functional GO annotation by using the "Mapping" function of Blast2GO. The complete annotation of the sequences was performed using Annotation, ANNEX (increasing annotation step), GO-SLIM (using the goslim_plant.obo option to reduce the GO vocabulary) and InterPro (by choosing the whole set of InterPro annotation applications available: BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, HMMPanther, ProfileScan, HAMAP, PatternScan, SuperFamily, Gene3D, Phobius, Coils, SignalIPHMM and TMHMM). Annotation using the default parameters (E-Value-Hit-Filtrer 1.0E-6; Annotation CutOff 55; GO Weight 5; Hsp-Hit Coverage CutOff 0) yielded a total of 27,152 annotated contigs (after merging the InterPro

results with the GO annotation). Tests for enrichment were performed on all GO terms associated with the sequences, using a two-tailed Fisher's exact test with correction for multiple tests by using a false discovery rate of 0.05. Double IDs (both on the test and on the reference set) were removed.

5.4 Author contributions

Aline Muyle, Niklaus Zemp, Alex Widmer and Gabriel Marais conceived the study and experimental design. Niklaus Zemp and Alex Widmer prepared and sequenced the RNA-seq datasets and assembled the reference transcriptome. Aline Muyle ran SEX-DETector on the RNA-seq datasets for the three tissues (including mapping and genotyping steps), produced allelic expression levels, analysed the data, prepared Tables and Figures and wrote the Material and Method section. Niklaus Zemp generated the X chromosome genetic map (with help from Aline Muyle for the mapping and genotyping part). Radim Cegan, Jan Vrana and Roman Hobza did the Y chromosome flow cytometry sorting and sequencing. Clothilde Deschamps did the first assembly of the sorted Y chromosome and improved it with RNA-seq data with the help of Cécile Fruchard. Aline Muyle did the blasts to validate the inferences of SEX-DETector. Raquel Tavares did the GO term analysis. Frank Picard helped with statistical analyses of the data. Gabriel Marais wrote the main text of the manuscript with inputs from other authors.

5.5 References

Albritton, S. E., Kranz, A.-L., Rao, P., Kramer, M., Dieterich, C., and Ercan, S. (2014). Sex-biased gene expression and evolution of the x chromosome in nematodes. eng. *Genetics* 197(3), 865–883. ISSN: 1943-2631. DOI: 10.1534/genetics.114.163311.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. eng. *Journal of Molecular Biology* 215(3), 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

Aronesty, E. (2011). *ea-utils: "Command-line tools for processing biological sequencing data"*.

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. eng. *Nature Reviews. Genetics* 14(2), 113–124. ISSN: 1471-0064. DOI: 10.1038/nrg3366.

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T.-J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., Kremitzki, C., Fulton, R. S., Dugan, S., Ding, Y., Morton, D., Khan, Z., Lewis, L., Buhay, C., Wang, Q., Watt, J., Holder, M., Lee, S., Nazareth, L., Alföldi, J., Rozen, S., Muzny, D. M., Warren, W. C., Gibbs, R. A., Wilson, R. K., and Page, D. C. (2014). Mammalian Y chromosomes retain widely expressed

dosage-sensitive regulators. eng. *Nature* 508(7497), 494–499. ISSN: 1476-4687. DOI: 10.1038/nature13206.

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Qiu, S., and Charlesworth, D. (2015). Gene Loss from a Plant Sex Chromosome System. ENG. *Current biology: CB*. ISSN: 1879-0445. DOI: 10.1016/j.cub.2015.03.015.

Charlesworth, B. (1996). The evolution of chromosomal sex determination and dosage compensation. eng. *Current biology: CB* 6(2), 149–162. ISSN: 0960-9822.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Chikhi, R. and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. eng. *Bioinformatics (Oxford, England)* 30(1), 31–37. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt310.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. eng. *Bioinformatics (Oxford, England)* 21(18), 3674–3676. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti610.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. eng. *Nature* 508(7497), 488–493. ISSN: 1476-4687. DOI: 10.1038/nature13151.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Angel, G. del, Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. eng. *Nature Genetics* 43(5), 491–498. ISSN: 1546-1718. DOI: 10.1038/ng.806.

Disteche, C. M. (2012). Dosage compensation of the sex chromosomes. eng. *Annual Review of Genetics* 46, 537–560. ISSN: 1545-2948. DOI: 10.1146/annurev-genet-110711-155454.

Ellegren, H. and Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. eng. *Nature Reviews. Genetics* 8(9), 689–698. ISSN: 1471-0056. DOI: 10.1038/nrg2167.

Ercan, S. (2015). Mechanisms of x chromosome dosage compensation. eng. *Journal of Genomics* 3, 1–19. ISSN: 1839-9940. DOI: 10.7150/jgen.10404.

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. eng. *PLoS genetics* 9(4), e1003457. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003457.

Gout, J.-F., Kahn, D., Duret, L., and Paramecium Post-Genomics Consortium (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. eng. *PLoS genetics* 6(5), e1000944. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000944.

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. eng. *Nature Protocols* 8(8), 1494–1512. ISSN: 1750-2799. DOI: 10.1038/nprot.2013.084.

Hobza, R. and Widmer, A. (2008). Efficient molecular sexing in dioecious Silene latifolia and S. dioica and paternity analysis in F(1) hybrids. eng. *Molecular Ecology Resources* 8(6), 1274–1276. ISSN: 1755-098X. DOI: 10.1111/j.1755-0998.2008.02344.x.

Hough, J., Hollister, J. D., Wang, W., Barrett, S. C. H., and Wright, S. I. (2014). Genetic degeneration of old and young Y chromosomes in the flowering plant Rumex hastatulus. eng. *Proceedings of the National Academy of Sciences of the United States of America* 111(21), 7713–7718. ISSN: 1091-6490. DOI: 10.1073/pnas.1319227111.

Huang, X. and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. en. *Genome Research* 9(9), 868–877. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.9.9.868.

Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., Daish, T., Grützner, F., and Kaessmann, H. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. eng. *PLoS biology* 10(5), e1001328. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001328.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. eng. *Bioinformatics (Oxford, England)* 25(14), 1754–1760. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp324.

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. eng. *Bioinformatics (Oxford, England)* 25(16), 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., and

Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. eng. *GigaScience* 1(1), 18. ISSN: 2047-217X. DOI: 10.1186/2047-217X-1-18.

Malone, J. H., Cho, D.-Y., Mattiuzzo, N. R., Artieri, C. G., Jiang, L., Dale, R. K., Smith, H. E., McDaniel, J., Munro, S., Salit, M., Andrews, J., Przytycka, T. M., and Oliver, B. (2012). Mediation of Drosophila autosomal dosage effects and compensation by network interactions. eng. *Genome Biology* 13(4), r28. ISSN: 1465-6914. DOI: 10.1186/gb-2012-13-4-r28.

Mank, J. E. (2013). Sex chromosome dosage compensation: definitely not for everyone. eng. *Trends in genetics: TIG* 29(12), 677–683. ISSN: 0168-9525. DOI: 10.1016/j.tig.2013.07.005.

Ming, R., Bendahmane, A., and Renner, S. S. (2011). Sex Chromosomes in Land Plants. en. *Annual Review of Plant Biology* 62(1), 485–514. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-042110-103914.

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., and Marais, G. A. B. (2012). Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. eng. *PLoS biology* 10(4), e1001308. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001308.

Ohno, S. (1967). Sex chromosomes and sex-linked genes. en. *Springer-Verlag, Berlin, Heidelberg, New York*.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. eng. *Genome Biology* 11(12), 220. ISSN: 1465-6914. DOI: 10.1186/gb-2010-11-12-220.

Pessia, E., Makino, T., Bailly-Bechet, M., McLysaght, A., and Marais, G. A. B. (2012). Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(14), 5346–5351. ISSN: 1091-6490. DOI: 10.1073/pnas.1116763109.

Prince, E. G., Kirkland, D., and Demuth, J. P. (2010). Hyperexpression of the X chromosome in both sexes results in extensive female bias of X-linked genes in the flour beetle. eng. *Genome Biology and Evolution* 2, 336–346. ISSN: 1759-6653. DOI: 10.1093/gbe/evq024.

Rastas, P., Paulin, L., Hanski, I., Lehtonen, R., and Auvinen, P. (2013). Lep-MAP: fast and accurate linkage map construction for large SNP datasets. eng. *Bioinformatics (Oxford, England)* 29(24), 3128–3134. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt563.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. eng. *Bioinformatics (Oxford, England)* 26(1), 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

Siroky, J., Castiglione, M. R., and Vyskot, B. (1998). DNA methylation patterns of Melandrium album chromosomes. eng. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 6(6), 441–446. ISSN: 0967-3849.

Tsagkogeorga, G., Cahais, V., and Galtier, N. (2012). The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate Ciona intestinalis. eng. *Genome Biology and Evolution* 4(8), 740–749. ISSN: 1759-6653. DOI: 10.1093/gbe/evs054.

VAN Ooijen, J. W. (2011). Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. eng. *Genetics Research* 93(5), 343–349. ISSN: 1469-5073. DOI: 10.1017/S0016672311000279.

Veitia, R. A., Veyrunes, F., Bottani, S., and Birchler, J. A. (2015). X chromosome inactivation and active X upregulation in therian mammals: facts, questions, and hypotheses. eng. *Journal of Molecular Cell Biology* 7(1), 2–11. ISSN: 1759-4685. DOI: 10.1093/jmcb/mjv001.

Vrána, J., Simková, H., Kubaláková, M., Číhalíková, J., and Doležel, J. (2012). Flow cytometric chromosome sorting in plants: the next generation. eng. *Methods (San Diego, Calif.)* 57(3), 331–337. ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2012.03.006.

Wang, J., Na, J.-K., Yu, Q., Gschwend, A. R., Han, J., Zeng, F., Aryal, R., VanBuren, R., Murray, J. E., Zhang, W., Navajas-Pérez, R., Feltus, F. A., Lemke, C., Tong, E. J., Chen, C., Wai, C. M., Singh, R., Wang, M.-L., Min, X. J., Alam, M., Charlesworth, D., Moore, P. H., Jiang, J., Paterson, A. H., and Ming, R. (2012). Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), 13710–13715. ISSN: 1091-6490. DOI: 10.1073/pnas.1207833109.

White, M. A., Kitano, J., and Peichel, C. L. (2015). Purifying Selection Maintains Dosage-Sensitive Genes during Degeneration of the Threespine Stickleback Y Chromosome. ENG. *Molecular Biology and Evolution*. ISSN: 1537-1719. DOI: 10.1093/molbev/msv078.

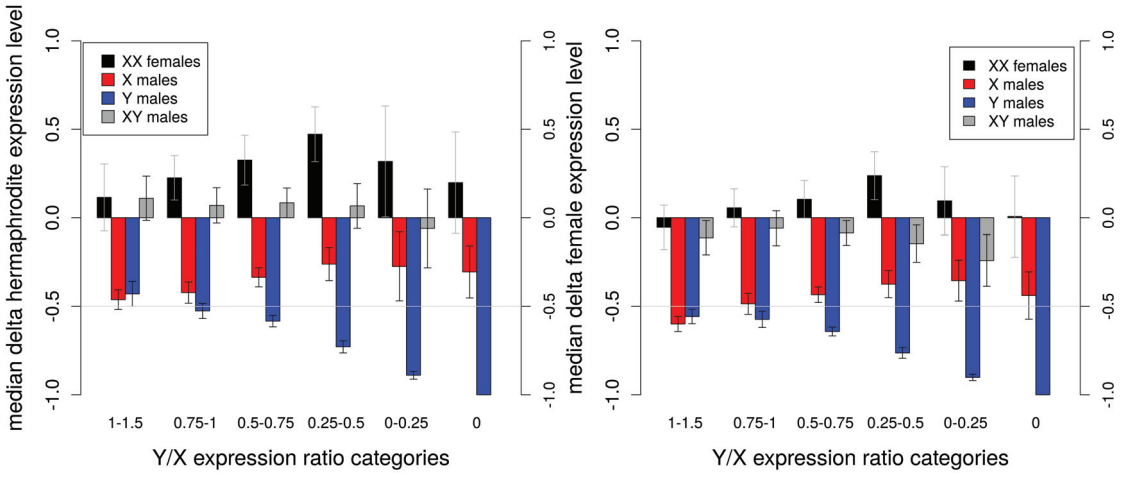
Xue, W., Li, J.-T., Zhu, Y.-P., Hou, G.-Y., Kong, X.-F., Kuang, Y.-Y., and Sun, X.-W. (2013). L_RNA_scaffolder: scaffolding genomes with transcripts. eng. *BMC genomics* 14, 604. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-604.

Zemp, N., Minder, A., and Widmer, A. (2014). Identification of internal reference genes for gene expression normalization between the two sexes in dioecious white Campion. eng. *PloS One* 9(3), e92893. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0092893.

5.6 Supplementary information

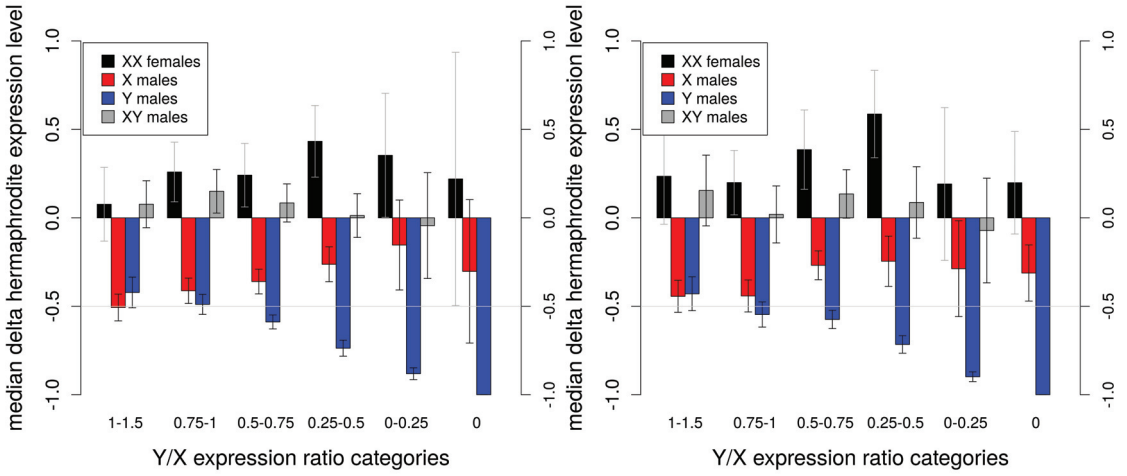
Large Table available online:

Supplementary Table S2: List of known sex-linked genes in *S. latifolia* and associated



(a) all non-sex-biased contigs

(b) validated non-sex-biased contigs



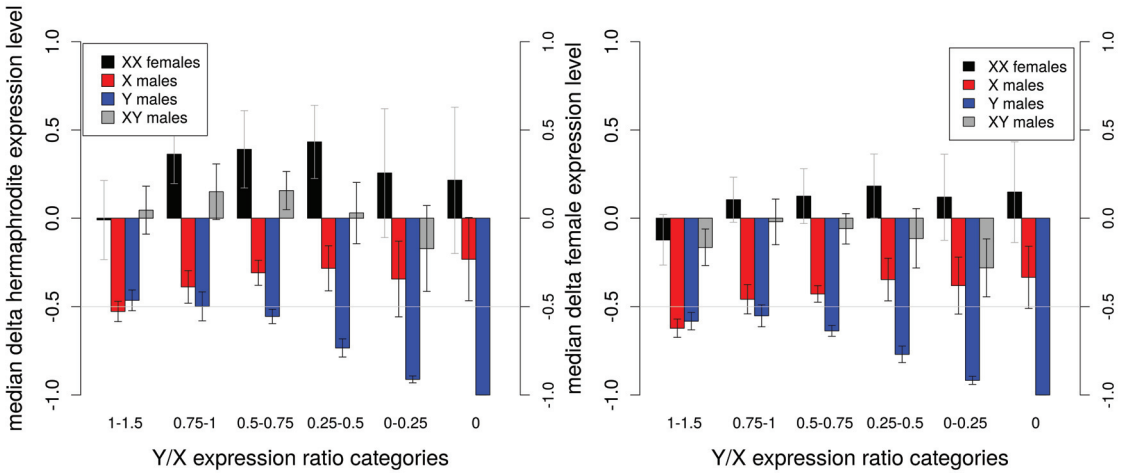
(c) highly expressed non-sex-biased contigs

(d) lowly expressed non-sex-biased contigs

Figure 5.S1: Bud dataset: median delta expression levels between *S. latifolia* and *S. vulgaris* hermaphrodites, in different Y/X ratio categories. Same legend as Figure 5.1. **(a)** All non sex-biased contigs (671 contigs), with the following sample sizes for the different Y/X ratio categories are: 0, 174; 0–0.25, 98; 0.25–0.5, 157; 0.5–0.75, 173; 0.75–1, 136; 1–1.5, 107. **(b)** Validated non sex-biased contigs (435 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 89; 0–0.25, 47; 0.25–0.5, 76; 0.5–0.75, 82; 0.75–1, 84; 1–1.5, 57. **(c)** Highly expressed non sex-biased contigs (327 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 30; 0–0.25, 34; 0.25–0.5, 62; 0.5–0.75, 84; 0.75–1, 71; 1–1.5, 46. **(d)** Lowly expressed non sex-biased contigs (518 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 144; 0–0.25, 64; 0.25–0.5, 95; 0.5–0.75, 89; 0.75–1, 65; 1–1.5, 61.

literature references.

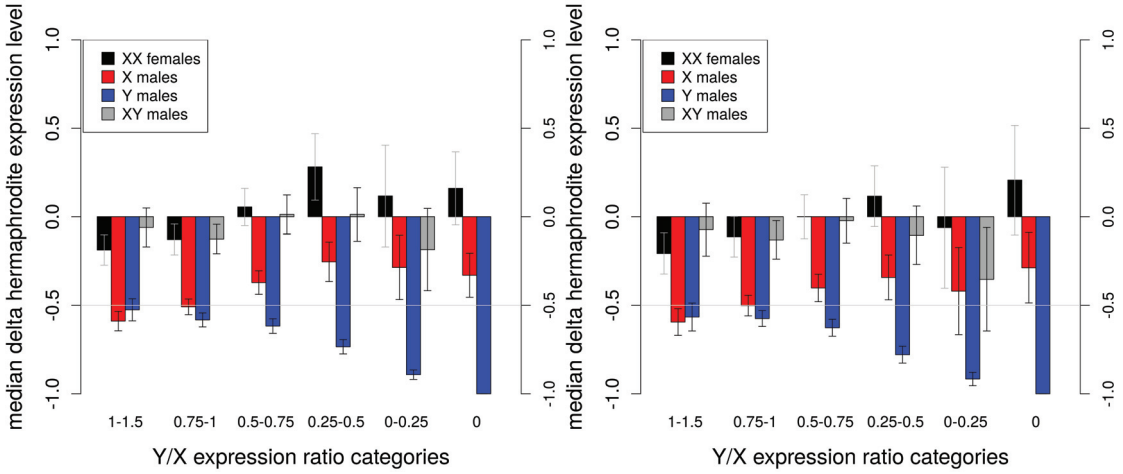
`https://drive.google.com/open?id=0B7KhiXOz6_OLUEMtNVZOdnFsdHM&authuser=0`



(a) all non-sex-biased contigs

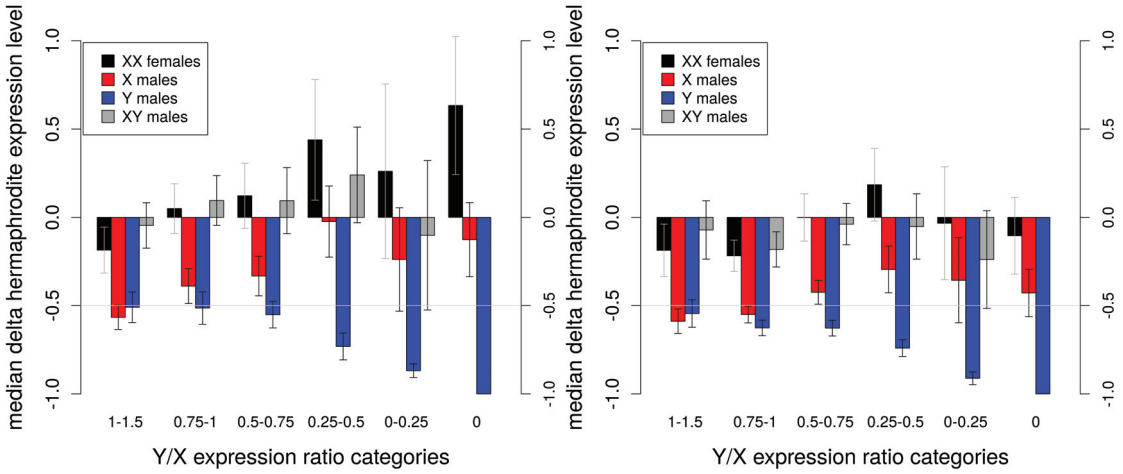
(b) validated non-sex-biased contigs

Figure 5.S2: Bud dataset: median delta expression levels between *S. latifolia* and *S. vulgaris* females, in different Y/X ratio categories. Same legend as Figure 5.S1.



(a) all non-sex-biased contigs

(b) validated non-sex-biased contigs



(c) highly expressed non-sex-biased contigs

(d) lowly expressed non-sex-biased contigs

Figure 5.S3: Seedling dataset: median delta expression levels between *S. latifolia* and *S. vulgaris* hermaphrodites, in different Y/X ratio categories. Same legend as Figure 5.1. **(a)** All non sex-biased contigs (903 contigs), with the following sample sizes for the different Y/X ratio categories are: 0, 216; 0–0.25, 79; 0.25–0.5, 156; 0.5–0.75, 162; 0.75–1, 154; 1–1.5, 121. **(b)** Validated non sex-biased contigs (474 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 104; 0–0.25, 41; 0.25–0.5, 75; 0.5–0.75, 86; 0.75–1, 93; 1–1.5, 66. **(c)** Highly expressed non sex-biased contigs (324 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 69; 0–0.25, 33; 0.25–0.5, 58; 0.5–0.75, 62; 0.75–1, 57; 1–1.5, 39. **(d)** Lowly expressed non sex-biased contigs (579 contigs), with the following sample sizes for the different Y/X ratio categories: 0, 147; 0–0.25, 46; 0.25–0.5, 98; 0.5–0.75, 100; 0.75–1, 97; 1–1.5, 82.

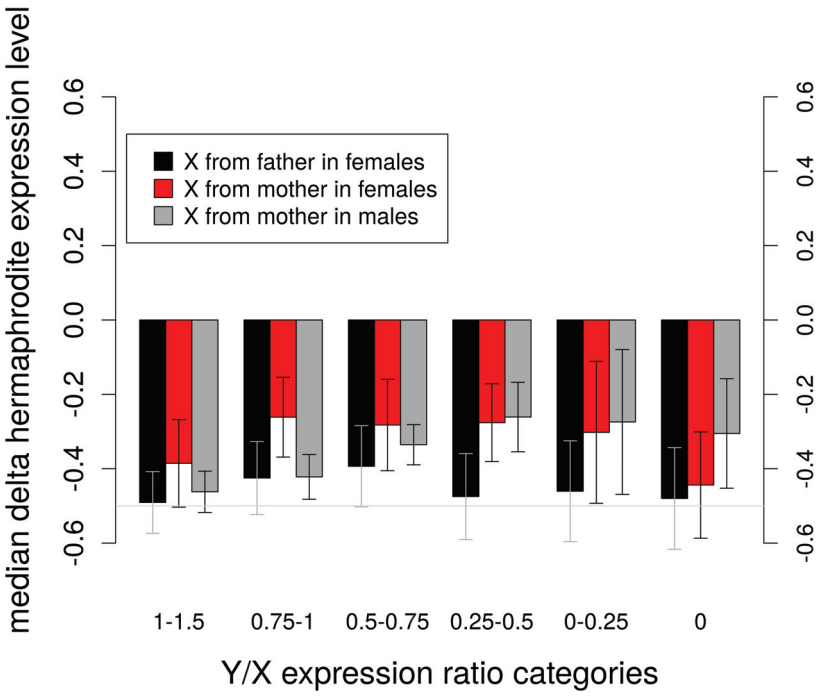


Figure 5.S4: Bud dataset: median delta expression levels between *S. latifolia* maternal and paternal X chromosomes and *S. vulgaris* hermaphrodite autosomes, in different Y/X ratio categories. Same legend as Figure 5.1 except that only SNPs where females were heterozygous were used, reducing contig numbers (562 in total) with the following sample sizes for the different Y/X ratio categories: 0, 163; 0–0.25, 64; 0.25–0.5, 90; 0.5–0.75, 83; 0.75–1, 88; 1–1.5, 74.

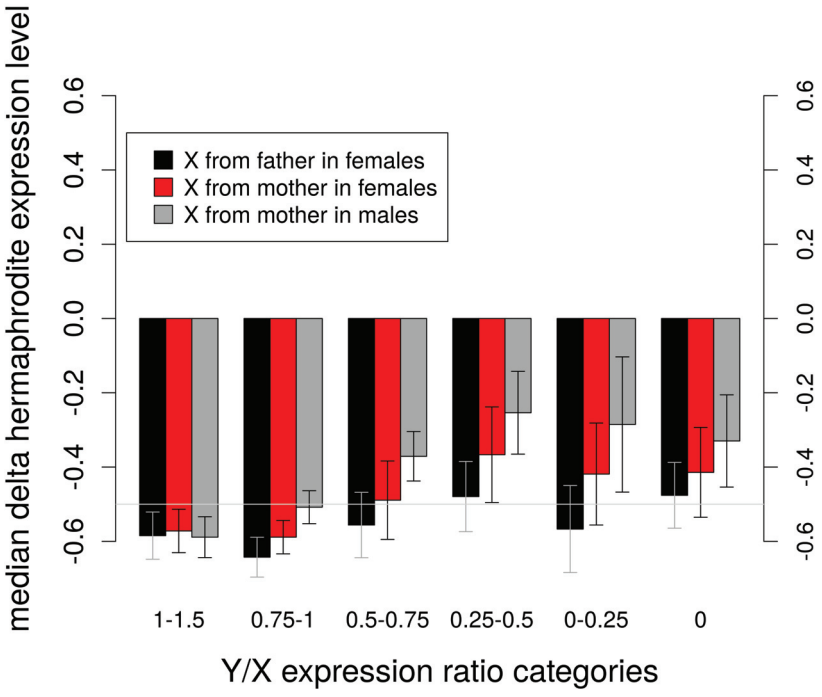


Figure 5.S5: Seedling dataset: median delta expression levels between *S. latifolia* maternal and paternal X chromosomes and *S. vulgaris* hermaphrodite autosomes, in different Y/X ratio categories. Same legend as Figure 5.1 except that only SNPs where females were heterozygous were used, reducing contig numbers (604 in total) with the following sample sizes for the different Y/X ratio categories: 0, 210; 0–0.25, 53; 0.25–0.5, 77; 0.5–0.75, 76; 0.75–1, 106; 1–1.5, 82.

Table 5.S1: library sizes (number of reads) of each individual and mapping statistics.

	individuals	# reads	# mapped reads	% mapped reads
S. latifolia flower buds	U10_37_mother	75800468	23262670	30.6
	Leuk144-3_father	115377548	64840769	56.2
	C1_01_male	71711204	41812869	58.3
	C1_3_male	82186876	49305677	60.1
	C1_04_male	69957770	41597899	59.3
	C1_05_male	181295644	106427848	58.6
	C1_26_female	63717900	35910978	56.6
	C1_27_female	145384384	88301626	61.1
	C1_29_female	161456150	94931546	58.9
	C1_34_female	190950550	116240597	60.9
S. latifolia leaves	C1_1_L_male	67883026	35418817	52.2
	C1_3_L_male	48736220	28632079	58.7
	C1_4_L_male	53539946	30431570	56.8
	C1_5_L_male	89406692	52900963	59.2
	C1_26_L_female	43525226	25632703	58.9
	C1_27_L_female	72164976	41086828	56.9
	C1_29_L_female	132788562	76583589	57.7
	C1_34_L_female	24783678	14640936	59.1
S. latifolia seedlings	Seed_lati_female_1	80090178	44691737	55.8
	Seed_lati_female_2	64993492	37891606	58.3
	Seed_lati_female_3	69893290	38935889	55.7
	Seed_lati_female_4	60155792	29397870	48.9
	Seed_lati_male_1	77527602	28629289	36.9
	Seed_lati_male_2	68485856	38037817	55.5
	Seed_lati_male_3	75155124	43597777	58.0
	Seed_lati_male_4	86674078	49221179	56.8
S. vulgaris flower buds	SV1_1_hermaphrodite	32836858	12299198	37.5
	SV1_2_hermaphrodite	59903112	27797320	46.4
	SV1_4_hermaphrodite	56879642	25633403	45.1
	SV1_5_female	54786992	23011360	42.0
	SV1_8_female	57192532	25685569	44.9
	SV1_9_female	47704414	21347225	44.7
	SV_Guarda1_father_herm	64185890	30885401	48.1
	SV_Seebach1_hermaphrodite	115453516	49841201	43.2
	SV_Seebach2_mother	98602086	45304549	45.9
S. vulgaris leaves	V2_10_L	91373032	37910605	41.5
	V2_11_L	80307234	34598237	43.1
	V2_12_L	55995074	26645813	47.6
	V2_13_L	98768016	35642564	36.1
S. vulgaris seedlings	Seed_vulg_herm_1	82017058	37088232	45.2
	Seed_vulg_herm_2	88125680	41024846	46.6
	Seed_vulg_herm_3	89331776	39853034	44.6
	Seed_vulg_herm_4	79716486	36723084	46.1

Chapter 6

Regulatory changes in females drive the evolution of sex-biased gene expression

This project aimed at understanding how sexual conflicts are resolved with the evolution of dioecy and sex chromosomes. The use of an outgroup allowed for the first time to study in what direction expression changes happened in sex-biased genes. SEX-DETector inferred autosomal and sex-linked genes were used.

This project was led by Niklaus Zemp in Alex Widmer's lab (in Zurich). My contribution was to run SEX-DETector on the data, produce allelic expression levels for sex-linked genes and write the corresponding section in the Material and Method. See Section 6.7 for more details.

The manuscript was submitted in June 2015 to *Nature*.

Changes in females drive the evolution of sex-biased gene expression in a dioecious plant

Niklaus Zemp[1], Raquel Tavares[2], Aline Muyle[2], Deborah Charlesworth[3], Gabriel A. B. Marais[2], Alex Widmer[1]

6.1 Abstract

Separate sexes (dioecy) and sex-biased gene expression have repeatedly evolved in animals and plants, but changes in gene expression associated with the origin of males and females have been little studied. When separate sexes and sex chromosomes have recently evolved in only one of a pair of closely related species, then changes associated with the evolution of males and females can be reconstructed. Here we show in such a pair of plant species that most expression changes of autosomal and sex-linked genes affected females, rather than males, and we infer that sex-biased gene expression has evolved through decreases in the sex experiencing the detrimental effects of sexually antagonistic alleles, although some genes have probably increased expression in the sex in which these alleles are beneficial. We also detect expression changes suggesting masculinization of the Y chromosome, and also feminization and demasculinization of the X in the dioecious species.

6.2 Introduction

Males and females of many plant and animal species differ in morphological, physiological and ecological characteristics, despite their overall genetic similarity (Ellegren and Parsch, 2007; Mank, 2009; Stewart et al., 2010). The evolution of sexual dimorphism can either involve complete sex-linkage, when a gene or allele is restricted to the genome of just one sex (e.g. Y-linked male function genes), or sex-limited or sex-biased expression of genes that are present in both sexes (Bonduriansky and Chenoweth, 2009; Doorn, 2009; Stewart et al., 2010), sometimes involving alternative splicing or duplication of genes followed by sex-specialization of their expression (Stewart et al., 2010). Sex-biased and sex-limited gene expression, and enrichment of such genes on the sex chromosomes, are well documented in animals (Ellegren, 2011; Ellegren and Parsch, 2007; Meisel et al., 2012; Reinius et al., 2012), including humans (Kang et al., 2011), and the evolutionary processes that can produce them are well understood, but the changes that actually led

[1]ETH Zurich, Institute of Integrative Biology, Universitätstrasse 16, 8092 Zürich, Switzerland
[2]Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France
[3]University of Edinburgh, Institute of Evolutionary Biology, Edinburgh EH9 3JT, Midlothian, Scotland

to the observed expression differences have not been studied. Genes with female-biased expression, for example, may have evolved by increased expression in females, or may have undergone a decrease in males, or both (Figure 6.1a, scenarios I-III). In most organisms, these evolutionary changes cannot be studied because separate sexes evolved too long ago. Species in which separate sexes evolved recently, such as some dioecious plants, are therefore of great interest, because gene expression changes can be inferred from comparisons with related species without separate sexes, which should often represent the ancestral state (Figure 6.1b). Only a few studies have yet investigated expression differences between females and males of plants and algae (Harkess et al., 2015; Lipinska et al., 2015; Robinson et al., 2014) and none has investigated the evolutionary changes leading to sex-biased gene expression in plants. Such studies can clarify the roles of sex-linked and sexually antagonistic genes in the evolution of separate sexes (Barrett and Hough, 2013).

In the plant genus *Silene*, gynodioecy (the co-existence of hermaphrodites and females

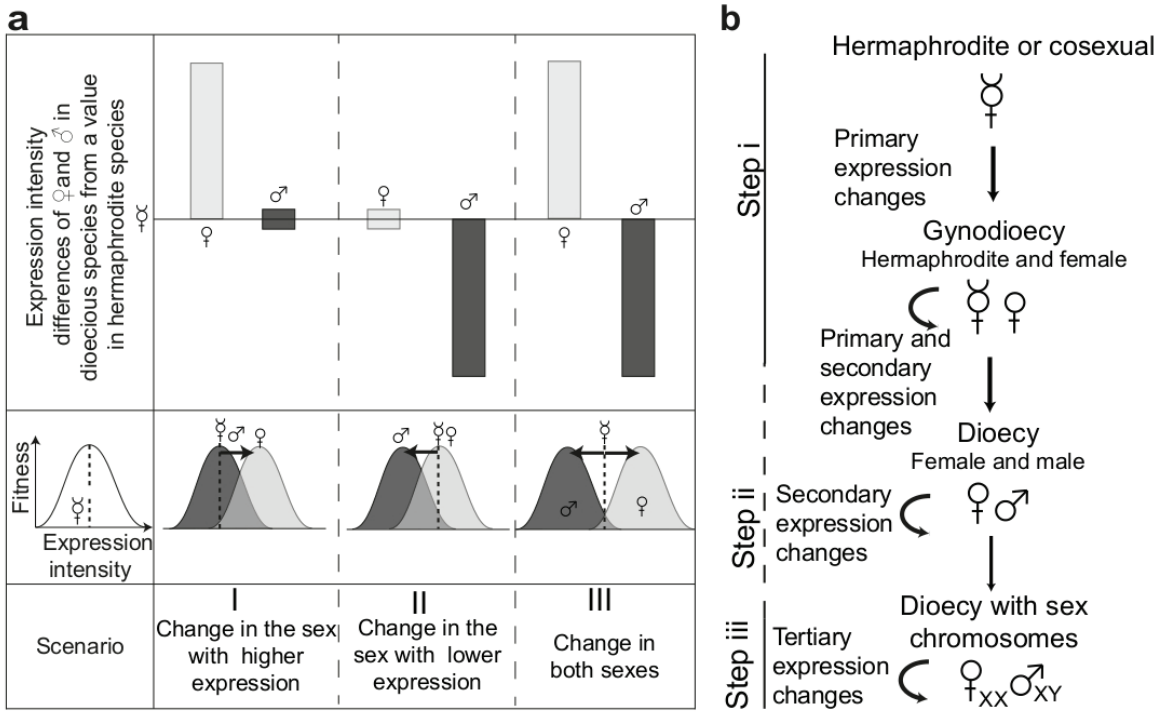


Figure 6.1: Evolution of sex-biased gene expression and transcriptional changes associated with the evolution of separate sexes. a, Hypothetical scenarios for the evolution of a gene with female-biased expression from a co-sexual ancestral state (horizontal line). I: expression increased exclusively in females; scenario II: expression decreased exclusively in males,; scenario III: expression changes in both sexes, with an increase in females and decrease in males. **b,** Model of evolutionary changes in gene expression associated with the transition from hermaphroditism to dioecy, and the evolution of non-recombining sex chromosomes. Step i) Sterility mutations in hermaphrodites cause loss of sexual functions and organs and lead to coexistence of hermaphrodite and unisexual individuals in gynodioecious populations (or androdioecious ones, though this is less likely) and subsequently to separate sexes in dioecious populations (Charlesworth and Charlesworth, 1978). Such primary expression changes can be identified from their fully sex-limited expression. Step ii) Secondary expression changes in gynodioecious and dioecious populations compensate for negative fitness effects of primary expression changes and reduce deleterious effects resulting from intralocus sexual conflict through the evolution of sex-biased gene expression. Step iii) Tertiary expression changes occur on young sex chromosome and lead to feminization or masculinization of the X and Y (or Z and W) chromosomes.

in the same population) represents the ancestral state, and dioecy (i.e. separate sexes) has evolved at least twice independently (Desfeux et al., 1996). *Silene vulgaris* (Moench) Garcke is gynodioecious and closely related to the dioecious species *Silene latifolia* Poiret (White Campion) (Marais et al., 2011), whose sex chromosome evolution is widely studied (Bernasconi et al., 2009). Male and female flowers and inflorescences of *S. latifolia* are sexually dimorphic in terms of flower size, number and physiology (Delph et al., 2002, and see Figure 6.2a) and both fully and partially sex-linked quantitative trait loci affecting sexually dimorphic traits, including some physiological traits, have been inferred (Delph et al., 2010). *S. latifolia* has an XY sex-determination system with heteromorphic sex chromosomes, which evolved within the past 5-10 MY (Marais et al., 2011; Rautenberg et al., 2010), but already displays evidence of Y chromosome degeneration like that in much older animal sex chromosomes (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Marais et al., 2008), and may have started to evolve dosage compensation (Muyle et al., 2012).

We used an mRNA-seq approach to assess gene expression differences between males

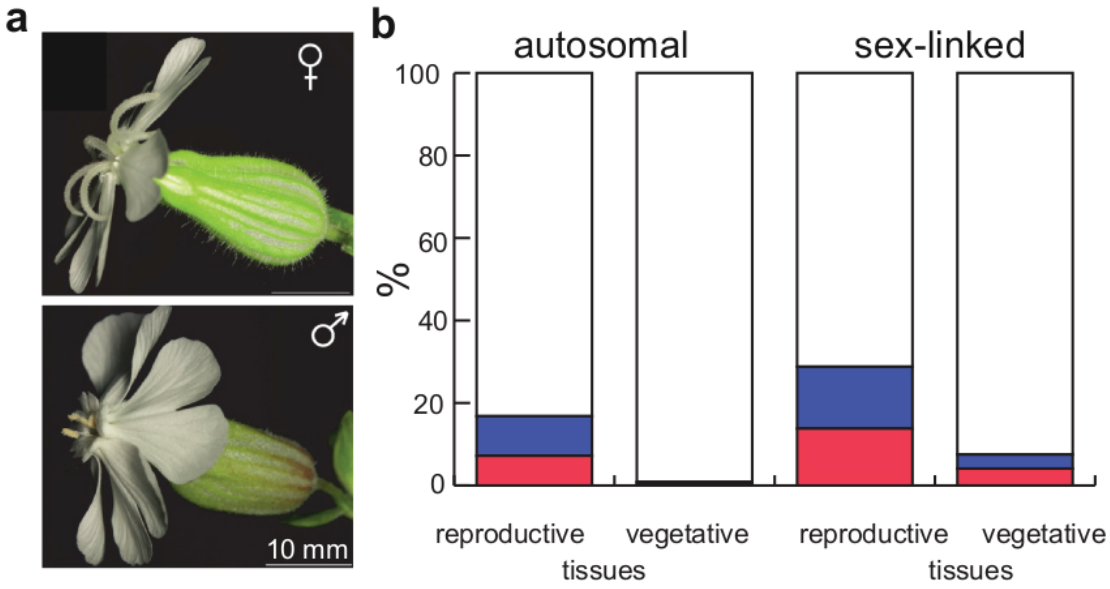


Figure 6.2: Sexual dimorphism and sex-biased gene expression in *S. latifolia*. **a**, Sexual dimorphism in female and male flowers of *S. latifolia*. **b**, Proportions of contigs with sex-biased (colored areas) and unbiased (white areas) expression for autosomal (12,708 contigs) and fully sex-linked contigs (936 contigs) in reproductive (flower buds) and vegetative (rosette leaves) tissues. The proportion of contigs with sex-biased expression was significantly higher among sex-linked contigs than among autosomal contigs (Fisher's exact-test, P-value ≤ 0.0001) in both reproductive and vegetative tissues. Proportions of contigs with female-biased (red areas) and male-biased expression are not larger on the sex chromosomes (blue areas; Fisher's exact test, P-value = 0.1) in reproductive tissues and vegetative tissues (see also Supplementary Table 6.S1).

and females of *S. latifolia* and used the data to test for differential representation of sex-biased genes on the sex chromosomes, and to investigate evolutionary changes in gene expression in *S. latifolia* males and females from the likely ancestral state represented by cosexual flowers of *S. vulgaris* hermaphrodites. Our expression analyses used mRNA iso-

lated from rosette leaves of *S. latifolia* and developing flower buds of both species. Overall, we obtained 145 Gb of RNA-seq data from Illumina 100 bp paired-end reads from *S. latifolia* flower buds (from seven male and seven female individuals), 41 Gb from *S. latifolia* rosette leaves (four individuals of each sex), and 33 Gb from flower buds of five *S. vulgaris* hermaphrodites. 58% and 57% of the *S. latifolia* reads from flower buds and rosette leaves, respectively, and 44% of the *S. vulgaris* reads mapped to the *S. latifolia* flower bud reference transcriptome (for details see Supplementary Table 6.S1). The lower percentage of mapped reads from *S. vulgaris* probably reflects sequence divergence between the two species (Marais et al., 2011).

6.3 The extent of sex-biased gene expression

We included flower buds in our expression analyses because sexual dimorphism in *S. latifolia* is strongest for flower and inflorescence traits (Delph et al., 2002). The strongest sex bias in expression is thus expected in flower buds (Ellegren and Parsch, 2007). However, many gene expression differences in buds may be associated with the presence or absence of the sex organs (we call these "primary differences" in Figure 6.1b). Before quantifying sex-biased gene expression, we therefore first excluded 903 contigs that showed sex-limited expression in *S. latifolia* flower buds, i.e. contigs that were expressed in buds of only one sex. These contigs were also excluded from counts of evolved expression differences between the sexes (Section 6.9.1). Among the remaining 11,366 *S. latifolia* contigs showing at least some expression in buds of both sexes, we identified contigs with significant sex differences in expression (Supplementary Figure 6.S1). The results are robust to different normalization procedures and estimators of gene expression differences (Supplementary Figure 6.S2), and agree well with qRT-PCR results (Supplementary Figure 6.S3; Spearman correlation; ρ=0.92; p-value: p<0.0001). GO analysis revealed that female-biased genes are enriched for transcription factors involved in cell-cycle and development but depleted in genes involved in catabolism (Supplementary Table 3). Male-biased genes are enriched in genes involved in metabolism (carbohydrates, lipids, secondary metabolites), transport and responses to various stimuli, and are depleted in genes involved in nucleic acid metabolism and protein synthesis and modification. Multiple biological processes that are over-represented among female-biased genes are under-represented among male-biased genes, or vice versa, including cell cycle, regulation of gene expression, and cellular protein modification process, suggesting that sex-biased expression has evolved to support contrasting biological functions in flower buds of males and females.

We divided the contigs with sex-biased expression in buds into autosomal, sex-linked (defined as contigs having alleles on both the X and Y chromosome) and X hemizygous contigs (with an expressed copy on the X only). These categories were inferred using a probabilistic model, from SNPs in parental plants that are segregating in their offspring

(Chapter 4). Inference of the frequencies of genes with sex-biased expression (Supplementary Table 6.S2) yielded an estimated 2,142 autosomal *S. latifolia* contigs with significant sex bias (16.8% of bud-expressed contigs, 7.2% with female biased expression, and 9.6% with male bias; Figure 6.2b). Sex-biased expression is much commoner (28.8%) among the 936 contigs inferred to be fully sex-linked (13.8% with female biased expression, and 15.0% with male bias Figure 6.2b and Supplementary Table 6.S2). The difference from autosomal contigs is highly significant (Fisher's exact test, $p < 0.0001$), even after excluding 18.8% of the sex-linked contigs whose female-biased expression may be caused by incomplete dosage compensation (Section 6.9.2).

In the data from rosette leaves (Fisher's exact test, p-value: $p < 0.0001$), many fewer genes showed sex-biased expression, (Delph et al., 2002; Ellegren and Parsch, 2007) and the expression biases for autosomal and sex-linked contigs were much lower than in buds (18.7-fold and 3.84-fold lower, respectively, see Figure 6.2b and Supplementary Table 6.S2). Our small estimated expression difference between the sexes in leaves of *S. latifolia* is compatible with findings for *Rumex hastatulus* (Hough et al., 2014) and suggests that sex bias in vegetative tissues is generally low (Ellegren and Parsch, 2007).

As in buds, genes with sex-biased expression in rosette leaves were over-represented on the sex chromosomes (Figure 6.2b). Contigs with female-biased expression represented 0.6% of autosomal and 4.1% of sex-linked contigs, significantly more than ones with male-biased expression (0.3% of autosomal and 3.4 % of sex-linked contigs, Supplementary Table 6.S2). The excess of female-biased contigs in leaves contrasts with the finding described above of more contigs with male-biased expression in buds (Fisher's exact test, P-value: $p < 0.0001$, Figure 6.2b,c; Supplementary Table 6.S2). In *Asparagus officinalis* flower buds, genes with higher male than female expression also predominated (Harkess et al., 2015). This predominance of genes with male-biased expression in reproductive tissues may be a consequence of sexual selection (Moore and Pannell, 2011). Because few genes show sex-biased expression in vegetative tissues, we focus below on reproductive tissues.

6.4 Evolution of sex-specific gene expression

To investigate the evolutionary processes that have led to the observed sex-biased gene expression in *S. latifolia*, we estimated gene expression in hermaphrodite flowers of the gynodioecious *S. vulgaris*. The expression levels of genes with no sex bias in expression (white bars in Figure 6.3) are largely unchanged in *S. latifolia* males and females, relative to *S. vulgaris* hermaphrodite flowers, indicating that much of the gene expression changes between the two species relates to the evolution of separate sexes. For both autosomal and sex-linked contigs in *S. latifolia* (Figure 6.3), the evolution of sex-biased expression mainly involves changes in females: female-biased expression (red bars in Figure 6.3) is due primarily to higher expression in *S. latifolia* females, and, notably, the

changes in the many genes with male-biased expression (blue bars in Figure 6.3) primarily involve lower expression in females, relative to expression in hermaphrodite *S. vulgaris*. Gene expression changes in males are much smaller than those in females, for both autosomal and sex-linked contigs, altthough the variances are high for the more limited number of sex-linked contigs (Figure 6.3). Similar patterns were found for both X-hemizygous contigs and contigs whose genomic location remained unknown (undefined contigs in Supplementary Figure 6.S8).

Why did these genes change expression? Given that we excluded genes likely to have undergone primary changes, due simply to loss of sex organs and functions, the sex differences in gene expression studied here should represent mainly secondary changes, including up- and down-regulation of genes following establishment of a unisexual type in a population (Figure 6.1b, Step ii, Vicoso et al., 2013b). Expression changes affecting fitness can be divided into three scenarios (Figure 6.1a). Increased or decreased expression of a gene in one sex, without change from the ancestral state in the other sex, is compatible with the notion that increased (or decreased) expression is advantageous for the former, but carries no major fitness cost for the latter sex. In Figure 1a, scenario I, the ancestral expression state is suboptimal for females, and an increase benefits females but is male-neutral. Similarly in scenario II the ancestral expression state is suboptimal for males, so that the change is female-neutral but benefits males. In contrast, scenario III illustrates the possibility that the expression levels of some genes in the ancestral hermaphrodite (before dioecy evolved) were suboptimal for both sexes, due to trade-offs, and that this was adjusted by evolutionary changes in both sexes after dioecy evolved. Large expression changes in opposite directions in the two sexes would suggest the evolution of beneficial changes in both sexes in response to this sexual antagonism, increasing expression of some genes in the sex where high expression is advantageous, and reducing it in the other sex (Figure 6.1a; scenario III). All scenarios are compatible with the hypothesis that sex-biased gene expression evolves to resolve sexually antagonistic effects (Bonduriansky and Chenoweth, 2009; Doorn and Kirkpatrick, 2007; Stewart et al., 2010).

We assessed directions of expression changes for genes with male- and female-biased expression in dioecious *S. latifolia*, relative to expression levels in hermaphroditic flowers of *S. vulgaris* (Bonduriansky and Chenoweth, 2009; Charlesworth and Charlesworth, 1978; Doorn, 2009; Moghadam et al., 2012; Ometto et al., 2011; Stewart et al., 2010; Vicoso et al., 2013a). Our analyses of putatively secondary expression changes (Figure 6.4) reveal that only a small proportion (14.9%) of contigs with male-biased expression have evolved through increased expression in males (Figure 6.4a, I, blue bar, see also Figure 6.1a), whereas 39.4% of them have undergone reduced expression in females (Figure 6.4a, II, blue bar). In marked contrast, a large percentage (42.1%) of female-biased genes are more strongly expressed in females, relative to cosexual flowers (Figure 6.4a, I, red bar), while only a few have undergone reduced expression in males (11.2%, Figure 6.4a, II, red

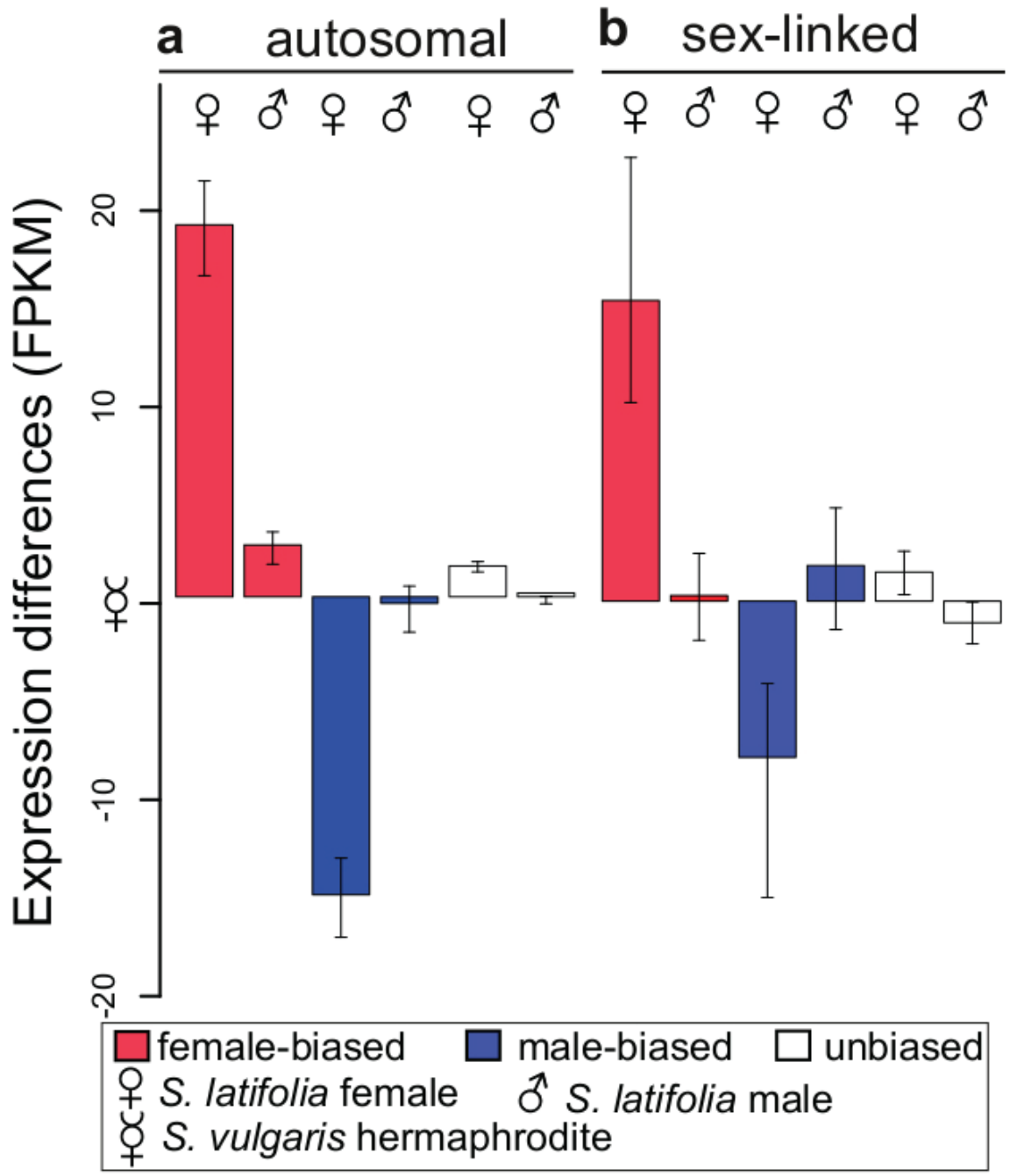


Figure 6.3: Expression changes in genes with sex-biased expression in *S. latifolia*. **a-b**, Expression differences (median with 95% confidence intervals) in (**a**) autosomal and (**b**) sex-linked contigs between *S. latifolia* females and males relative to *S. vulgaris* hermaphrodites for contigs with female-biased (red), male-biased (blue), and unbiased (white) expression in flower buds. Positive values, for example, indicate that genes have higher expression in *S. latifolia* females than in *S. vulgaris* hermaphrodites (Figure 6.3a; first bar from the left).

bar). The results are similar for sex-linked contigs (Figure 6.4a). Overall, sex-biased gene expression in *S. latifolia* therefore evolved primarily through expression changes in females: reduced expression in females led to male-biased expression, whereas increased expression in females led to genes with female-biased expression (Figure 6.4).

We inferred whether selection was involved in the expression changes using the ΔX ap-

proach (Moghadam et al., 2012; Ometto et al., 2011). The great majority of these expression changes in females appear to have been driven by selection (Figure 6.4a-b). Fewer expression changes occurred in males, where twice as many contigs showed increased than reduced expression, and the Δ X analysis finds that fewer than half of these changes were driven by selection.

Our finding that sex-biased gene expression in this dioecious plant has most often evolved as a consequence of transcriptional down-regulation in females, suggests that this is a response to sexual antagonism in which high expression levels are detrimental in females, but presumably benefited male functions in the hermaphrodite ancestor (Vicoso et al., 2013b). However, the observation that a substantial proportion of genes also underwent expression changes in males suggests that, for some genes, males benefit from changed expression of sexually antagonistic alleles. Together, these results suggest sexual conflict over gene expression in the hermaphrodite flowers of *S. vulgaris* has led to an outcome that is closer to the optimum for males than for females, and that this situation becomes resolved through the evolution of sex-biased gene expression following the evolution of separate sexes.

Our GO analysis indicates that resource re-allocation may have been involved in these gene expression changes. We found opposite patterns of GO term enrichment/depletion in female- and male-biased contigs (Supplementary Table 3), suggesting that genes important for each sex tend to be down-regulated in the other sex, consistent with differences in resource allocation. Our analysis of expression changes thus suggest that re-allocation during the evolution of dioecy was more important for females than for males, compatible with observations that female plants are often resource limited (Asikainen and Mutikainen, 2005; Barrett and Hough, 2013).

Few genes displayed significant changes in expression in both sexes (Figure 6.4a, scenario III) and the great majority that did so show male-biased expression (Figure 6.4c). While our results support the hypothesis that sex-biased expression has evolved to reduce intralocus sexual conflict, it remains unknown whether conflict resolution underlies all sex-biased expression, and sex-biased expression alone does not definitively imply the past existence of sexual antagonism (Innocenti and Morrow, 2010). All three of our scenarios are compatible with sexual conflict, and approximately 50% of contigs with female expression biases (53.7% for autosomal contigs and 51.0% for sex-linked ones) show one or other of these patterns, as do approximately 60% of contigs with male-biased expression (60.4% for autosomal and 64.2% for sex-linked contigs). Other genes may have evolved sex-biased expression as indirect consequences of the evolution of separate sexes, for example to compensate for negative fitness effects of sterility mutations and primary expression changes, or because upstream regulatory elements have evolved sex-biased expression.

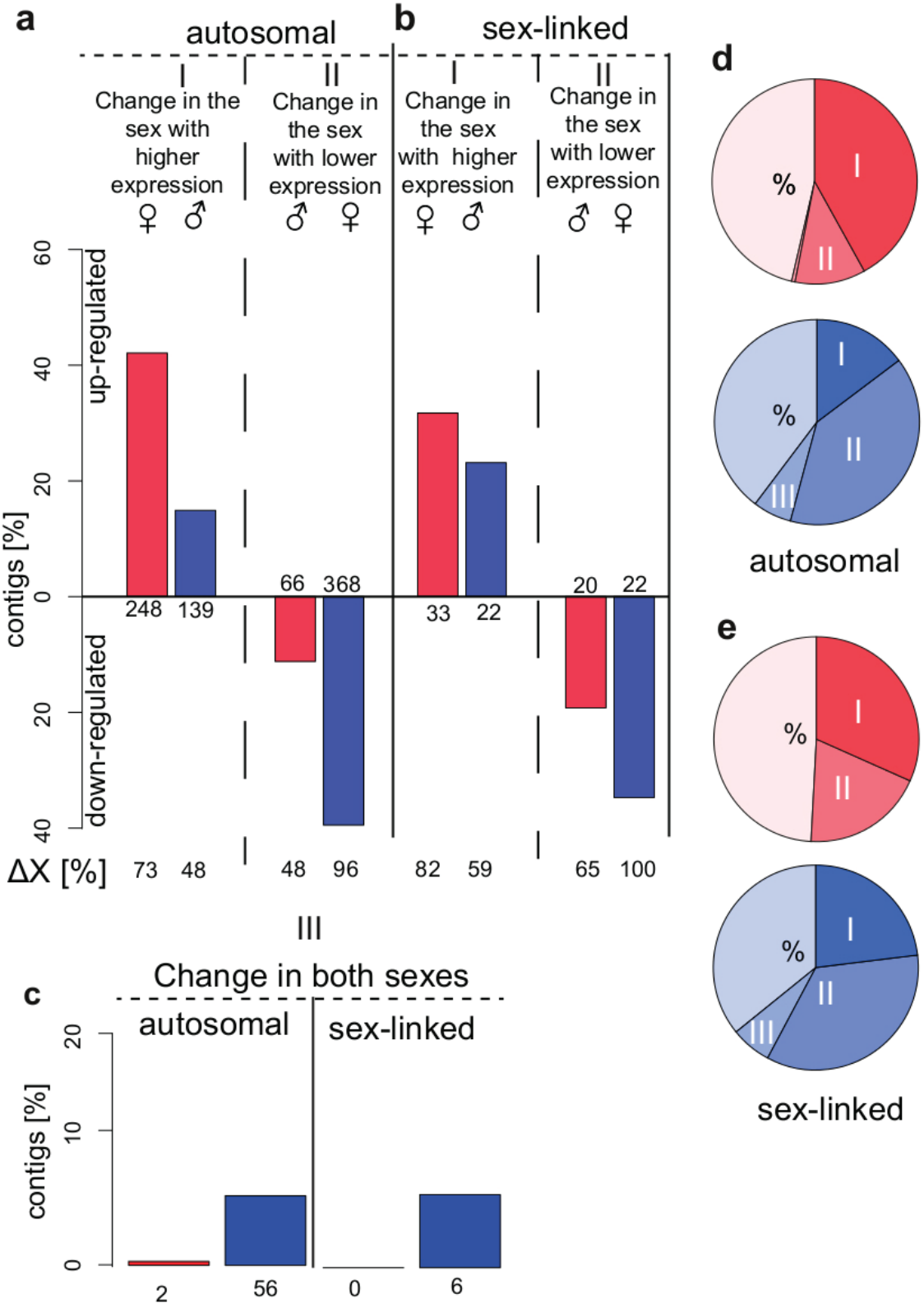


Figure 6.4: Direction of evolutionary change leading to sex-biased gene expression in *S. latifolia*. **a-b,** Percentages of autosomal (**a**) and sex-linked (**b**) contigs in flower buds that are I) significantly and exclusively expressed at higher levels in the same sex (scenario I in Figure 6.1a), II) expressed at significantly lower levels in the opposite sex (scenario II of Figure 6.1a). Red indicates contigs with female-biased, blue contigs with male-biased expression. Numbers indicate contigs per category. Values below bars are percentages of contigs in each category that are outliers for expression divergence ($\Delta X \geq$ 75th percentile across all contigs), indicative of directional changes in gene expression. **c**, Percentage of autosomal and sex-linked contigs whose expression has significantly changed in both sexes, relative to *S. vulgaris* hermaphrodites (scenario III in Figure 6.1a).

6.5 Sex-biased expression on sex chromosomes

In dioecious species, tertiary changes in gene expression may follow the evolution of sex chromosomes with non-recombining regions (Figure 6.1b, Step iii) and include expression changes that are specific for the X and Y chromosomes (Figure 6.4b). Overall, the expression changes inferred for sex-linked contigs are consistent with those for autosomal contigs, but the relative proportion of genes with changes in males was slightly higher than for autosomal contigs. Evidence for selective advantages of expression changes on the sex chromosomes was again strongest for changes in females (82% and 100% of contigs with higher and lower expression in females, respectively, are in the top 25% for Δ X; Figure 6.4b), and higher proportions of genes were inferred to be down-, rather than up-regulated as a consequence of selection (Figure 6.4b), implying strong selection to reduce fitness costs due to sexually antagonistic genes.

The evolutionarily much older sex chromosomes in *Drosophila* (Meisel et al., 2012; Reinius et al., 2012) have undergone degeneration and lost most Y-linked genes, rendering the majority of X-linked genes hemizygous in males. The X chromosomes of these species have evolved an overrepresentation of genes with female-biased expression (feminization), as predicted for hemizygous loci (reviewed in Ellegren and Parsch, 2007) and animal Y chromosomes are enriched for genes with male-biased expression among their few remaining genes (Bellott et al., 2014; Cortez et al., 2014; Skaletsky et al., 2003; Zhou and Bachtrog, 2012). Have the much younger *S. latifolia* sex chromosomes undergone similar changes? We found no significant overrepresentation of contigs with female-biased expression on the sex chromosomes (Figure 6.2b). However, allelic expression patterns did suggest possible feminization of the X chromosome and masculinization of the Y. We estimated the relative expression contribution of alleles of sex-linked genes in *S. latifolia* males (XY) and females (XX) using sex-linked SNPs (see Section 6.6 for details). Comparing female-biased contigs with unbiased contigs, to obtain relative expression levels from the X in females (estimated as half the expression in XX females) we found significantly higher values than in males, whereas, for male-biased sex-linked contigs, relative expression levels were significantly lower for males (Wilcoxon-test, P-values < 0.0001 for both comparisons; Figure 6.5a). Thus, expression from the *S. latifolia* X chromosome in females contributes strongly to female-biased expression, suggesting ongoing feminization and demasculinization of this X chromosome. This is consistent with the proposal that gene expression may evolve faster than structural changes (King and Wilson, 1975).

We also estimated allelic expression ratios for *S. latifolia* Y-linked alleles by counting SNP in males, and compared them to half the expression estimates from *S. vulgaris* (Y/0.5*AA, see Section 6.6 and Section 6.9.3). Contigs with male-biased expression showed higher counts from the Y (Wilcoxon-test, P-values < 0.01) and female-biased contigs lower counts (Wilcoxon-test, P-values < 0.001), compared to contigs without sex bias in expression (Figure 6.5b). Higher relative expression of Y-linked alleles thus contributes to male-

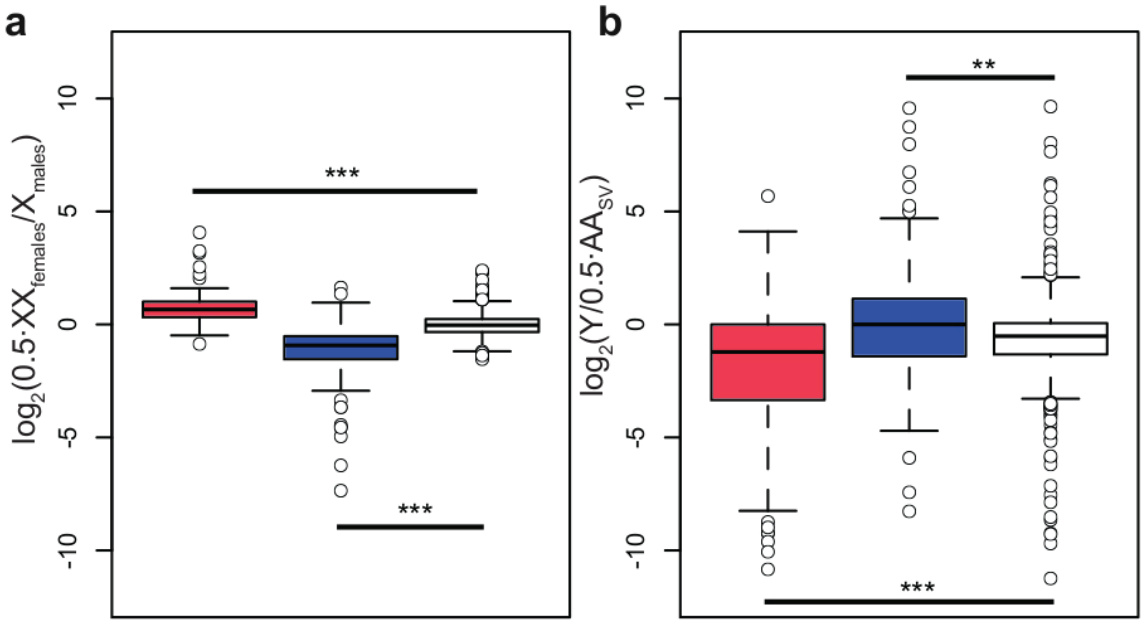


Figure 6.5: Feminization of the X and masculinization of the *S. latifolia* Y chromsomes. **a,** Boxplots of X expression ratios (on a $\log_2$ scale) for sex-linked contigs with female-biased, male-biased, and unbiased expression in flower buds. Ratios for contigs with female-biased expression are significantly larger than for unbiased contigs, whereas those of male-biased contigs were significantly smaller (Wilcoxon-test, P-value ≤ 0.0001), respectively, indicating feminization and demasculinization of the *S. latifolia* X chromosome. **b,** Boxplots of Y expression ratios (in $\log_2$ scale) for contigs with female-biased, male-biased, and unbiased expression. Ratios are significantly larger for male-biased contigs (Wilcoxon-test, P-value ≤ 0.01) and significantly smaller for female-biased contigs in comparison to contigs with unbiased expression (Wilcoxon-test, p ≤ 0.0001), indicating weak masculinization and stronger defeminization of the *S. latifolia* Y chromosome.

biased gene expression of fully sex-linked contigs (but not to female-biased expression); this is compatible with evidence that selection acting in males contributes to maintaining functional genes on the Y (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011). However, the X chromosome still contributes substantially to male-biased expression. Our results thus suggest opposite forces acting on the *S. latifolia* X and Y chromosomes, favouring tertiary gene expression divergence between the sexes which may eventually lead to the accumulation of genes with female-biased expression on the X chromosomes and male-biased genes on the Y chromosome.

Our findings illustrate the value of studying closely related species, one of which has evolved separate sexes and sex chromosomes. We show that sex-biased gene expression in dioecious *S. latifolia* has evolved primarily as a consequence of secondary expression changes in females. Our results support the long-standing hypothesis that sex-biased expression evolves to reduce the detrimental effects of sexually antagonistic alleles, both in autosomal and sex-linked genes, either by increasing expression in the favoured sex or by reducing expression in the disadvantaged sex.

6.6 Material and Methods

6.6.1 Plant material, RNA extraction and RNA-seq

We used multiple male and female individuals from two *S. latifolia* populations for this study. The first set included plants from an inbred line propagated by brother-sister mating for 10 (U10) generations. RNAseq data from three females (U10_34, U10_37 and U10_39) and three males (U10_11, U10_49 and U10_9) were already used in a previous study (Muyle et al., 2012, SRP010792). We added data from eight full-sib offspring from a cross between one of the female plants from the first cross (U10_37) and a male plant from a natural population (Leuk_144.1 from Leuk, Valais, Switzerland). These plants included four females (C1_26, C1_27, C1_29, C1_34) and four males (C1_01, C1_03, C1_04, C1_05). All plants were grown in a greenhouse at ETH research station Eschikon (Switzerland) under long-day conditions (16h light), and temperatures of 22°C during the day and 18°C at night. We further used two hermaphrodite individuals (See_1, from Seebach in Zurich, Switzerland, and Guard_1, from Guarda, Grisons, Switzerland) of the gynodioecious species *S. vulgaris* and three hermaphroditic offspring from the intraspecific cross between See_1 and Guard_1 (V1_1_herm, V1_2_herm, V1_4_herm).

6.6.2 RNA-seq experiments

High quality RNA was extracted from small flower buds without calyces at developmental stages B1-B2 (Matsunaga et al., 1996). Additionally, from the C1 individuals we extracted RNA from young fully developed rosette leaves as described in Muyle et al. (2012). cDNA libraries were prepared separately for each individual using random primers as Muyle et al. (2012). Tagged libraries from the eight *C1* individuals were sequenced in three channels, and Leuk_144.1 and the two *S. vulgaris* individuals (See_1, Guard_1) in two channels, on an Illumina Hiseq 2000 by GATC Biotech AG (Konstanz, Germany), using 100bp paired-end reads. Each lane included individuals from both sexes (Table 6.S1). For two of the libraries of each sex, sequencing was replicated in different lanes (Table 6.S1). The *S. vulgaris* V1 individuals and C1 leaf libraries were prepared with the Illumina TruSeq kit v2 and sequenced at the Quantitative Genomics Facility (QGF, ETH Zürich, Switzerland). RNA-seq data from sequencing runs were deposited at NCBI under accession number XXXXXX.

6.6.3 Normalized cDNA library production and sequencing

RNA from very small flower buds (~3-4 mm, including calyces), stage B1 (Matsunaga et al., 1996), small buds (as above), and large (~10mm, without calyces, late B1-B2, Matsunaga et al., 1996) flower buds, as well as flowers before anthesis (without calyces) were extracted from a female and a male individual of the inbred line. Full-length cDNA synthesis was performed for each developmental stage and sex (2μg total RNA) using the

Mint kit (Evrogen, Moscow). The used primers were slightly modified and instead of using 1μl of 3' primer (10uM), we used 1 μl of PolTdeg (10uM) (AAG CAGTGGTATCAACGCAGAGTACTTTTGTTTTTTTTTCTTTTTTTTTTVN) and 2μl of PCR Primer M1 (10μM) was substituted by 1μl of M1ACGG (10μM) (AAGCAGTGGTATCAACGCAGAGTACGG) and 1μl of polTM1 (10μM) (AAGCAGTGGTATCAACGCAGAGTACTTTTGTCTTTTGTTCTGTTTCTTTTVN). Products were purified with Invisorb Fragment CleanUp (Invitek, Germany). After double stranded cDNA synthesis, the four different developmental stages were equally pooled (4 x 200 ng) and normalized separately for each sex using the Trimmer cDNA normalization kit (Evrogen, Moscow). The normalization efficiency was tested by comparing quantitative real time (RT)-PCR results before and after the normalization with strongly and weakly expressed reference genes (data not shown). After a second purification, quality was checked on a Bioanalyzer (Agilent, USA). The two normalized cDNA libraries (0.5μg per library) were each sequenced in one half of a PTP plate using a shotgun approach on a Roche 454 Genome Sequencer FLX system at the FGCZ (Functional Genomics Centre Zurich, Zurich, Switzerland). The mean read length was 415 bp. 454 data were deposited under accession number XXXXXXX.

6.6.4 *De novo* assembly and identification of sex-linked contigs

We combined all flower bud RNAseq data from U10 and C1 individuals, including the normalized 454 reads, which were transformed to Illumina reads using simulate_illuminaPE_from_454ds.pl script in Trinity (Grabherr et al., 2011), for the *de novo* assembly. Reads were trimmed and duplicates removed using cutadapt (Martin, 2011) and filterPCRdupl (Smeds and Künstner, 2011). The *S. latifolia* flower bud reference transcriptome was assembled using Trinity r2012-10-05 (Grabherr et al., 2011). Reads were mapped back to the reference transcriptome, and only contigs with an expression threshold of 0.1 fragments per kilobase of exon per million fragments mapped (FPKM) were retained. Isoforms were collapsed using the rsem function, to yield a reference transcriptome. Poly-A tails, ribosomal and contigs smaller than 200 bp were removed, leaving 151,089 contigs with n50=1,796. Next, we used Trinity's orf-predictor and retained only the contigs with intact open reading frames (transcripts_to_best_scoring_ORFs.pl). In order to increase the probability of assembling gene sequences derived from the X and Y chromosomes in the same contigs, the ORFs were further assembled using custom perl scripts to run CAP3 (Huang and Madan, 1999, cap3 -p 70, Version Date: 10/15/07) on the sequences from the Trinity output. This reduced the number of contigs to 46,178. The llumina reads were then re-mapped onto the bud reference transcriptome using BWA (Li and Durbin, 2009 version 0.6.2, bwa aln -n 5 and bwa sampe, see Table 6.S1) and the libraries were merged using SAMTOOLS Version 0.1.18 (Li et al., 2009). The alignments obtained were locally realigned using GATK IndelRealigner (McKenna et al., 2010) and were analysed using reads2snps (Tsagkogeorga et al., 2012, Version 3.0, -fis 0 -model M2 -output_genotype best -multi_alleles acc -min_coverage 3 -par false) in order to genotype individuals at

each locus while allowing for biases in allele expression; in doing this, we chose not to remove putatively paralogous SNPs from the sequences by paraclean, because this tends to filter out X/Y SNPs. SEX-DETector (Chapter 4) was then used on the parents (U10_37 and LEUK_144.1) and their offspring (C1-individuals) to infer the genomic locations of the contigs from their segregation patterns; this was done after estimation of parameters using an SEM algorithm (Chapter 4). Contigs with posterior segregation type probabilities less than 0.8 or no informative SNPs were excluded from further analyses, and 31,014 contigs were considered as undefined by this criterion.
After this analysis, 13,807 contigs were inferred to be autosomal, 1,025 to be sex-linked (defined here as contigs with expressed X and Y gametologs), and 332 contigs as X-hemizygous. We removed contigs potentially containing transposable elements (TEs) by running RepeatMasker (Smit et al., 2013) with all plant specific and known *S. latifolia* TEs (Macas et al., 2011). This reduced the number of contigs to 40,822, of which 12,708 were autosomal, 936 fully sex-linked, 310 X-hemizygous and 26,872 undefined. Autosomal contigs had a mean contig length of 1151 with standard error ± 7 base pairs (bp) the values for sex-linked contigs were 1272 ± 40 bp, and for the X-hemizygous contigs 925 ± 37 and the undefined contigs 556 ± 3 bp.

6.6.5 Identifying genes with sex-biased expression in *S. latifolia*

We then counted the numbers of mapped reads per contig and library. To detect and quantify expression differences in flower buds, we used edgeR taking into account multiple biological replicates(Robinson et al., 2010) using the default settings and the TMM normalization method and a common dispersion. We used a false discovery rate of 5% to correct for multiple testing. To avoid difficulties with weakly expressed or chimeric contigs, this analysis included only contigs with an expression cutoff of 1 count per million (cpm) reads in at least half of the female and/or male libraries, leading to 33,551 contigs, of which 10,120 were classified as autosomal, 936 as sex-linked 310 as X-hemizygous. Statistically significant differences in gene expression between males and females were identified separately for the sex-linked, autosomal, X-hemizygous and undefined contigs. The numbers of contigs with sex-biased expression in vegetative rosette leaf tissues were determined similarly. To assess the robustness of our results, we implemented other normalization methods and estimators for the proportions of genes with sex-biased expression in flower buds, using non-default settings in edgeR, as well as the software package DESeq2 (Love et al., 2014).

6.6.6 qRT-PCR validation

We validated our RNAseq-based gene expression estimates with qRT-PCR assays for 18 contigs, including at least five that were categorized as each of female-biased, male-biased or unbiased, based on our RNA-seq results. We used the same *S. latifolia* RNA

extractions as for the Illumina RNAseq library preparation. cDNA was prepared using the GoScript Reverse Transcription System (Promega, USA). Primers for qRT-PCR assays were designed as descripted in (Zemp et al., 2014). To normalize expression levels, we used SL_ACTIN, SL_UBCE and SL_REF9 (M value < 0.5 and pairwise variation < 0.15 suggesting stable expression (Udvardi et al., 2008). qRT-PCR reactions were performed in triplicate on a 7500 Fast Real Time PCR system (ABI, USA) using TaqMan Gene Expression Mastermix (ABI, CA) after adding 0.075 μl 3:40 EvaGreen 20 X (Biotium, CA) per 1 μl Mastermix. Primer efficiencies were determined using LinReg and expression intensities were normalized separately for each sample and contig using QBase Plus (Hellemans et al., 2007) relative to the means of the reference genes. Means per contig and sex were calculated from the normalized qRT-PCR data and the normalized Illumina data to calculate the male/female expression ratios for each contig. The correlation between the qRT-PCR and Illumina data was calculated using R (R Development Core Team, 2012).

6.6.7 Analyses of sex bias

Among the undefined contigs, 839 were expressed exclusively in males, and 64 exclusively in females. These contigs were excluded from further analyses because they represent primary expression changes (Section 6.9.1). This step was not necessary for the other categories, because they were inferred from segregation analyses that required expression of alleles in both sexes.
To assess differences in the prevalence of sex-biased and unbiased, female-biased and male-biased contigs between the autosomal and sex-linked contigs, we used Fisher's exact tests implemented in R.

6.6.8 Gene Ontology (GO) analysis

The 46,178 RNA-seq contigs (see above) were blasted against the Uniprot database using blastx with an expectation cutoff value of 0.001, and an upper limit of hit results per sequence of 20. Significant blastx hits were obtained for 35,763 contigs. Gene Ontology terms were associated with the contigs using the Blast2GO PRO version V 2.7.2 (Conesa et al., 2005) and the Data Base version b2g_sep14, with 41,136 GO terms and 4,097 Enzymes available. Blast hits were related to the functional GO annotation by using the "Mapping" function of Blast2GO. The complete annotation of the sequences was performed using Annotation, ANNEX (increasing annotation step), GO-SLIM (using the goslim_plant.obo option to reduce the GO vocabulary) and InterPro (by choosing the whole set of InterPro annotation applications available: BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, HMMPanther, ProfileScan, HAMAP, PatternScan, SuperFamily, Gene3D, Phobius, Coils, SignalIPHMM and TMHMM). Annotation using the default parameters (E-Value-Hit-Filter 1.0E-6; Annotation CutOff 55; GO Weight 5; Hsp-Hit Coverage CutOff 0) yielded a total of 27,152 annotated contigs (after merging the InterPro

results with the GO annotation).
Enrichment tests for GO terms: Excluding the genes with female or male-limited expression (903 contigs - see above – of which 17 female-limited and 214 male-limited contigs were annotated) resulted in a total set of 26,921 annotated sequences used in the enrichment tests. Tests for enrichment were performed on all GO terms associated with the sequences, using a two-tailed Fisher's exact test with correction for multiple testing by using a false discovery rate of 5% (Benjamini and Hochberg, 1995). Double Ids (both on the test and on the reference set) were removed.

6.6.9 Allelic expression estimates for the sex-linked contigs in *S. latifolia* and the corresponding contigs in *S. vulgaris* hermaphrodites

For each *S. latifolia* sex-linked contig, we estimated expression levels using read counts from both X/Y and X-hemizygous informative SNPs. SNPs were attributed to an X/Y or X-hemizygous segregation type if the posterior probability in the test mentioned above (Chapter 4) was higher than 0.5, and X/Y SNPs were informative if the father (LEUK_144.1) was heterozygous and had a genotype different from the mother (U10_27) (otherwise it was not possible to distinguish the X and Y alleles and estimate separate X and Y expression values). All X-hemizygous SNPs are informative. X, Y, X+X and X+Y normalised expression levels (in RPKM, Oshlack et al., 2010) were computed for each contig by summing the read numbers for each X-linked or Y-linked allele for all SNPs for each individual plant separately, and the raw counts were then normalised as follows using the total number of mapped reads per individual (library size) and the number of sex-linked SNPs in the contig.

$$E = \frac{r}{n \times l} \tag{6.1}$$

where E is the normalised expression level, r is the sum of total read counts, n is the number of SNPs, and l is the normalised library size.
To quantify expression in *S. vulgaris* we mapped the *S. vulgaris* reads onto the *S. latifolia* flower bud reference transcriptome (see above) using BWA, using a maximum mismatch of 7 in the 100 bp reads to allow higher sequence divergence between the two species than within *S. latifolia*. Contigs were retained only if they had one count per million (cpm) reads in at least half of the libraries from *S. vulgaris* hermaphrodites. This yielded 16, 572 contigs (11, 564 autosomal in *S. latifolia*, 969 sex-linked and 267 X-hemizygous contigs).
In order to make *S. latifolia* and *S. vulgaris* allelic expression levels comparable for the sex-linked contigs, *S. vulgaris* allele-specific expression levels were estimated using only the *S. latifoli*a SNPs *that were* informative for X/Y inheritance or X-hemizygous. Read counts at each site in each contig were obtained for all *S. vulgaris* individuals using GATK DepthOfCoverage. Only sites corresponding to informative X/Y or X-hemizygous SNPs

in *S. latifolia* were used to compute contig-specific expression as explained in equation 6.1.

For each contig, we divided the expression estimates for the XX females by the estimated expression of the X-linked allele in *S. latifolia* males (see below); we refer to this expression value as the X expression ratio ($0.5*XX_{females}/X_{males}$). Similarly, we calculated the Y/0.5*AA expression ratios by dividing the expression of the Y alleles by half of the expression estimated for the corresponding genes in *S. vulgaris*. These ratios were computed for contigs with male- or female-biased expression and compared with contigs with unbiased expression, and the differences were tested using Wilcoxon- tests.

6.6.10 Expression divergence between *S. vulgaris* and *S. latifolia*

To quantify the evolutionary changes leading to sex-biased expression in *S. latifolia*, relative to the expression in *S. vulgaris* hermaphrodites, we calculated median expression levels and 95% confidence intervals for each contig; this was done separately for the *S. latifolia* males and females, after subtracting the means of the values from those estimated from the *S. vulgaris* hermaphrodites. We then classified the genes into different categories with respect to expression divergence (see Figure 6.1a). For female-biased contigs, we computed (i) the percentages of contigs with significantly increased expression in females but that have unchanged expression in males, (ii) contigs unchanged in females and with decreased expression in males (Figure 6.1a, scenario II) and (iii) contigs with significantly increased expression in females and decreased expression in males (Figure 6.1a, scenario III) using appropriate contrasts in edgeR with a false discovery rate of 5% separately for the autosomal and sex linked datasets. We also computed similar proportions for the contigs with male-biased expression.

To test whether directional selection has affected expression levels, we used the ΔX approach (Moghadam et al., 2012; Ometto et al., 2011), an analog of the McDonald-Kreitman test for sequence divergence (McDonald and Kreitman, 1991). Expression differences between *S. latifolia* and *S. vulgaris* were divided by the standard deviation for all contigs, estimated for the two sexes separately in *S. latifolia*. For categories I and II of Figure 6.1a, we computed the percentages of contigs displaying outlier expression divergence values (defined as $\Delta X \geq 75$ percentile across all contigs) between the two species.

6.7 Author contributions

6.8 References

Asikainen, E. and Mutikainen, P. (2005). Pollen and resource limitation in a gynodioecious species. eng. *American Journal of Botany* 92(3), 487–494. ISSN: 0002-9122. DOI: 10.3732/ajb.92.3.487.

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. eng. *Nature Reviews. Genetics* 14(2), 113–124. ISSN: 1471-0064. DOI: 10.1038/nrg3366.

Barrett, S. C. H. and Hough, J. (2013). Sexual dimorphism in flowering plants. eng. *Journal of Experimental Botany* 64(1), 67–82. ISSN: 1460-2431. DOI: 10.1093/jxb/ers308.

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T.-J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., Kremitzki, C., Fulton, R. S., Dugan, S., Ding, Y., Morton, D., Khan, Z., Lewis, L., Buhay, C., Wang, Q., Watt, J., Holder, M., Lee, S., Nazareth, L., Alföldi, J., Rozen, S., Muzny, D. M., Warren, W. C., Gibbs, R. A., Wilson, R. K., and Page, D. C. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. eng. *Nature* 508(7497), 494–499. ISSN: 1476-4687. DOI: 10.1038/nature13206.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1), 289–300. ISSN: 0035-9246.

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Qiu, S., and Charlesworth, D. (2015). Gene Loss from a Plant Sex Chromosome System. ENG. *Current biology: CB*. ISSN: 1879-0445. DOI: 10.1016/j.cub.2015.03.015.

Bernasconi, G., Antonovics, J., Biere, A., Charlesworth, D., Delph, L. F., Filatov, D., Giraud, T., Hood, M. E., Marais, G. a. B., McCauley, D., Pannell, J. R., Shykoff, J. A., Vyskot, B., Wolfe, L. M., and Widmer, A. (2009). Silene as a model system in ecology and evolution. eng. *Heredity* 103(1), 5–14. ISSN: 1365-2540. DOI: 10.1038/hdy.2009.34.

Bonduriansky, R. and Chenoweth, S. F. (2009). Intralocus sexual conflict. eng. *Trends in Ecology & Evolution* 24(5), 280–288. ISSN: 0169-5347. DOI: 10.1016/j.tree.2008.12.005.

Charlesworth, B. and Charlesworth, D. (1978). Model for Evolution of Dioecy and Gynodioecy. English. *American Naturalist* 112(988). WOS:A1978FX45000001, 975–997. ISSN: 0003-0147. DOI: 10.1086/283342.

Charlesworth, B. and Charlesworth, D. (2000). The degeneration of Y chromosomes. eng. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 355(1403), 1563–1572. ISSN: 0962-8436. DOI: 10.1098/rstb.2000.0717.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional

genomics research. eng. *Bioinformatics (Oxford, England)* 21(18), 3674–3676. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti610.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. eng. *Nature* 508(7497), 488–493. ISSN: 1476-4687. DOI: 10.1038/nature13151.

Delph, L. F., Knapczyk, F. N., and Taylor, D. R. (2002). Among-population variation and correlations in sexually dimorphic traits of Silene latifolia. en. *Journal of Evolutionary Biology* 15(6), 1011–1020. ISSN: 1420-9101. DOI: 10.1046/j.1420-9101.2002.00467.x.

Delph, L. F., Arntz, A. M., Scotti-Saintagne, C., and Scotti, I. (2010). The genomic architecture of sexual dimorphism in the dioecious plant Silene latifolia. eng. *Evolution; International Journal of Organic Evolution* 64(10), 2873–2886. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2010.01048.x.

Desfeux, C., Maurice, S., Henry, J. P., Lejeune, B., and Gouyon, P. H. (1996). Evolution of reproductive systems in the genus Silene. eng. *Proceedings. Biological Sciences / The Royal Society* 263(1369), 409–414. ISSN: 0962-8452.

Doorn, G. S. van and Kirkpatrick, M. (2007). Turnover of sex chromosomes induced by sexual conflict. eng. *Nature* 449(7164), 909–912. ISSN: 1476-4687. DOI: 10.1038/nature06178.

Doorn, G. S. van (2009). Intralocus sexual conflict. eng. *Annals of the New York Academy of Sciences* 1168, 52–71. ISSN: 1749-6632. DOI: 10.1111/j.1749-6632.2009.04573.x.

Ellegren, H. (2011). Emergence of male-biased genes on the chicken Z-chromosome: sex-chromosome contrasts between male and female heterogametic systems. eng. *Genome Research* 21(12), 2082–2086. ISSN: 1549-5469. DOI: 10.1101/gr.119065.110.

Ellegren, H. and Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. eng. *Nature Reviews. Genetics* 8(9), 689–698. ISSN: 1471-0056. DOI: 10.1038/nrg2167.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Palma, F. di, Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. eng. *Nature Biotechnology* 29(7), 644–652. ISSN: 1546-1696. DOI: 10.1038/nbt.1883.

Harkess, A., Mercati, F., Shan, H.-Y., Sunseri, F., Falavigna, A., and Leebens-Mack, J. (2015). Sex-biased gene expression in dioecious garden asparagus (Asparagus officinalis). ENG. *The New Phytologist*. ISSN: 1469-8137. DOI: 10.1111/nph.13389.

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and auto-

mated analysis of real-time quantitative PCR data. eng. *Genome Biology* 8(2), R19. ISSN: 1465-6914. DOI: 10.1186/gb-2007-8-2-r19.

Hough, J., Hollister, J. D., Wang, W., Barrett, S. C. H., and Wright, S. I. (2014). Genetic degeneration of old and young Y chromosomes in the flowering plant Rumex hastatulus. eng. *Proceedings of the National Academy of Sciences of the United States of America* 111(21), 7713–7718. ISSN: 1091-6490. DOI: 10.1073/pnas.1319227111.

Huang, X. and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. en. *Genome Research* 9(9), 868–877. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.9.9.868.

Innocenti, P. and Morrow, E. H. (2010). The sexually antagonistic genes of Drosophila melanogaster. eng. *PLoS biology* 8(3), e1000335. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000335.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Z., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., Weinberger, D. R., Mane, S., Hyde, T. M., Huttner, A., Reimers, M., Kleinman, J. E., and Sestan, N. (2011). Spatio-temporal transcriptome of the human brain. eng. *Nature* 478(7370), 483–489. ISSN: 1476-4687. DOI: 10.1038/nature10523.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. eng. *Science (New York, N.Y.)* 188(4184), 107–116. ISSN: 0036-8075.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. eng. *Bioinformatics (Oxford, England)* 25(14), 1754–1760. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp324.

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. eng. *Bioinformatics (Oxford, England)* 25(16), 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

Lipinska, A., Cormier, A., Luthringer, R., Peters, A. F., Corre, E., Gachon, C. M. M., Cock, J. M., and Coelho, S. M. (2015). Sexual Dimorphism and the Evolution of Sex-Biased Gene Expression in the Brown Alga Ectocarpus. ENG. *Molecular Biology and Evolution*. ISSN: 1537-1719. DOI: 10.1093/molbev/msv049.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. eng. *Genome Biology* 15(12), 550. ISSN: 1465-6914. DOI: 10.1186/s13059-014-0550-8.

Macas, J., Kejnovský, E., Neumann, P., Novák, P., Koblížková, A., and Vyskot, B. (2011). Next generation sequencing-based analysis of repetitive DNA in the model dioecious [corrected] plant Silene latifolia. eng. *PloS One* 6(11), e27335. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027335.

Mank, J. E. (2009). Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. eng. *The American Naturalist* 173(2), 141–150. ISSN: 1537-5323. DOI: 10.1086/595754.

Mank, J. E. (2013). Sex chromosome dosage compensation: definitely not for everyone. eng. *Trends in genetics: TIG* 29(12), 677–683. ISSN: 0168-9525. DOI: 10.1016/j.tig.2013.07.005.

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. eng. *PloS One* 6(8), e21915. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021915.

Marais, G. A. B., Nicolas, M., Bergero, R., Chambrier, P., Kejnovsky, E., Monéger, F., Hobza, R., Widmer, A., and Charlesworth, D. (2008). Evidence for degeneration of the Y chromosome in the dioecious plant Silene latifolia. eng. *Current biology: CB* 18(7), 545–549. ISSN: 0960-9822. DOI: 10.1016/j.cub.2008.03.023.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. en. *EMBnet.journal* 17(1), pages. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200.

Matsunaga, S., Kawano, S., Takano, H., Uchida, H., Sakai, A., and Kuroiwa, T. (1996). Isolation and developmental expression of male reproductive organ-specific genes in a dioecious campion, Melandrium album (Silene latifolia). eng. *The Plant Journal: For Cell and Molecular Biology* 10(4), 679–689. ISSN: 0960-7412.

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. eng. *Nature* 351(6328), 652–654. ISSN: 0028-0836. DOI: 10.1038/351652a0.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. eng. *Genome Research* 20(9), 1297–1303. ISSN: 1549-5469. DOI: 10.1101/gr.107524.110.

Meisel, R. P., Malone, J. H., and Clark, A. G. (2012). Disentangling the relationship between sex-biased gene expression and X-linkage. eng. *Genome Research* 22(7), 1255–1265. ISSN: 1549-5469. DOI: 10.1101/gr.132100.111.

Moghadam, H. K., Pointer, M. A., Wright, A. E., Berlin, S., and Mank, J. E. (2012). W chromosome expression responds to female-specific selection. eng. *Proceedings of the National Academy of Sciences of the United States of America* 109(21), 8207–8211. ISSN: 1091-6490. DOI: 10.1073/pnas.1202721109.

Moore, J. C. and Pannell, J. R. (2011). Sexual selection in plants. eng. *Current biology: CB* 21(5), R176–182. ISSN: 1879-0445. DOI: 10.1016/j.cub.2010.12.035.

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., and Marais, G. A. B. (2012). Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. eng. *PLoS biology* 10(4), e1001308. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001308.

Ometto, L., Shoemaker, D., Ross, K. G., and Keller, L. (2011). Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. eng. *Molecular Biology and Evolution* 28(4), 1381–1392. ISSN: 1537-1719. DOI: 10.1093/molbev/msq322.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. eng. *Genome Biology* 11(12), 220. ISSN: 1465-6914. DOI: 10.1186/gb-2010-11-12-220.

Rautenberg, A., Hathaway, L., Oxelman, B., and Prentice, H. C. (2010). Geographic and phylogenetic patterns in Silene section Melandrium (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. eng. *Molecular Phylogenetics and Evolution* 57(3), 978–991. ISSN: 1095-9513. DOI: 10.1016/j.ympev.2010.08.003.

Reinius, B., Johansson, M. M., Radomska, K. J., Morrow, E. H., Pandey, G. K., Kanduri, C., Sandberg, R., Williams, R. W., and Jazin, E. (2012). Abundance of female-biased and paucity of male-biased somatically expressed genes on the mouse X-chromosome. eng. *BMC genomics* 13, 607. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-607.

Robinson, K. M., Delhomme, N., Mähler, N., Schiffthaler, B., Onskog, J., Albrectsen, B. R., Ingvarsson, P. K., Hvidsten, T. R., Jansson, S., and Street, N. R. (2014). Populus tremula (European aspen) shows no evidence of sexual dimorphism. eng. *BMC plant biology* 14, 276. ISSN: 1471-2229. DOI: 10.1186/s12870-014-0276-5.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. eng. *Bioinformatics (Oxford, England)* 26(1), 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

Skaletsky, H. et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. eng. *Nature* 423(6942), 825–837. ISSN: 0028-0836. DOI: 10.1038/nature01722.

Smeds, L. and Künstner, A. (2011). ConDeTri - A Content Dependent Read Trimmer for Illumina Data. *PLoS ONE* 6(10), e26314. DOI: 10.1371/journal.pone.0026314.

Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0.

Steven, J. C., Delph, L. F., and Brodie, E. D. (2007). Sexual dimorphism in the quantitative-genetic architecture of floral, leaf, and allocation traits in Silene latifolia. eng. *Evolution; International Journal of Organic Evolution* 61(1), 42–57. ISSN: 0014-3820. DOI: 10.1111/j.1558-5646.2007.00004.x.

Stewart, A. D., Pischedda, A., and Rice, W. R. (2010). Resolving intralocus sexual conflict: genetic mechanisms and time frame. eng. *The Journal of Heredity* 101 Suppl 1, S94–99. ISSN: 1465-7333. DOI: 10.1093/jhered/esq011.

Team, R. D. C. (2012). R: A language and environment for statistical computing, reference index version 2.15.0. *R Foundation for Statistical Computing, Vienna, Austria*.

Tsagkogeorga, G., Cahais, V., and Galtier, N. (2012). The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in

the tunicate Ciona intestinalis. eng. *Genome Biology and Evolution* 4(8), 740–749. ISSN: 1759-6653. DOI: 10.1093/gbe/evs054.

Udvardi, M. K., Czechowski, T., and Scheible, W.-R. (2008). Eleven golden rules of quantitative RT-PCR. eng. *The Plant Cell* 20(7), 1736–1737. ISSN: 1040-4651. DOI: 10.1105/tpc.108.061143.

Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S., and Bachtrog, D. (2013a). Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. eng. *PLoS biology* 11(8), e1001643. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001643.

Vicoso, B., Kaiser, V. B., and Bachtrog, D. (2013b). Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. eng. *Proceedings of the National Academy of Sciences of the United States of America* 110(16), 6453–6458. ISSN: 1091-6490. DOI: 10.1073/pnas.1217027110.

Zemp, N., Minder, A., and Widmer, A. (2014). Identification of internal reference genes for gene expression normalization between the two sexes in dioecious white Campion. eng. *PloS One* 9(3), e92893. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0092893.

Zhou, Q. and Bachtrog, D. (2012). Sex-specific adaptation drives early sex chromosome evolution in Drosophila. eng. *Science (New York, N.Y.)* 337(6092), 341–345. ISSN: 1095-9203. DOI: 10.1126/science.1225385.

6.9 Supplementary information

6.9.1 Supplementary Note 1: Primary expression changes

Apparent sex-biased gene expression in flower buds of male and female plants may arise trivially, because genes with sex-limited expression are not expressed in the sex that does not form the corresponding tissue (for example, apparent male-biased gene expression may occur in *S. latifolia* for anther-specific genes, simply because no anthers are formed in female flowers and the corresponding genes are not expressed), or when genes are expressed at similar levels in both male and female organs and thus have reduced expression when the organs are not developed in one sex. We therefore excluded all contigs with strictly sex-limited expression patterns in *S. latifolia* from our analyses of sex-biased gene expression; we refer to these as primary expression changes (Figure 6.1b).

6.9.2 Supplementary Note 2: Incomplete dosage compensation of sex-linked contigs with female-biased expression

The higher proportion of genes with female-biased expression on the sex chromosomes may either be due a true overrepresentation of genes with female-biased expression on the sex chromosomes, or may alternatively be caused by incomplete dosage compensation (Mank, 2013). Evidence for dosage compensation in *S. latifolia* has been reported

(Muyle et al., 2012), and while it seems clear that not all genes are fully compensated (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011), partial dosage compensation is possible (Muyle et al., 2012). To assess the possibility of dosage compensation, we compared the female bias in expression between flower buds and vegetative tissues, since vegetative tissues will probably show less sex-biased expression. Of the 86 sex-linked contigs with female-biased expression in flower buds that were sufficiently expressed in rosette leaves, only 16 (18.6%) had female-biased expression in both tissues (Supplementary Figure 6.S4), and might be incompletely dosage compensated. To further test the robustness of our results in reproductive tissues, we excluded the 16 sex-linked contigs that showed female-biased expression in vegetative tissues. This reduced the proportion of genes with female-biased expression to 11.23 % (p-value<0.0001). These 16 genes may truly have female biased expression in rosette leaves (similar to the identified 18 contigs that displayed female-limited expression in rosette leaves), given that sexually dimorphic leaf traits have been observed in *S. latifolia* (Steven et al., 2007).

6.9.3 Supplementary Note 3: Are Y/X ratios good estimators for Y-chromosome degeneration?

Y/X ratios are frequently used to quantify the extent of degeneration of Y-linked alleles (Bachtrog, 2013; Charlesworth and Charlesworth, 2000). However, X-linked alleles may be up-regulated in males (dosage compensation), so this ratio is difficult to interpret. In *S. latifolia*, not all sex-linked genes appear to be dosage compensated (Bergero et al., 2015; Chibalina and Filatov, 2011), but some may show partial dosage compensation, with reduced expression of the Y alleles compensated by increased expression of the X-linked alleles in males (Muyle et al., 2012).

For assessing the Y/X expression ratios in *S. latifolia* we compare the expression ratios with the expression levels in *S. vulgaris*. We therefore calculated the Y to half of the corresponding expression level is *S. vulgaris* hermaphrodites (Y/ 0.5*A) (Figure 6.5b) and the Y over X ratio in males (Supplementary Figure 6.S5a). The two values were well correlated (Spearman correlation, ρ=0.598, P-value > 0.001; Supplementary Figure 6.S6) suggesting that the Y/X ratio is a good proxy for Y degeneration in *S. latifolia.* However, as just explained the expression intensity of the *S. vulgaris* hermaphrodite may be preferable for detecting Y degeneration for some of the contigs, because of up-regulation of the X in males.

Large Table available online:

Supplementary Table S3: Results of GO enrichment analysis for all genes showing sex biases in expression. `https://drive.google.com/file/d/0B7KhiXOz6_OLcEJsdkR5R3VtWVk/view?usp=sharing`

Table 6.S1: Mapping summary. Overview of the total numbers of reads obtained and mapped per library, species and tissue type.

Species	Library_ID	sex	tissue	chan-nel	Li-brary size (Gb)	Mapped reads (Gb)	% Mapped reads
S. latifolia	C1_01_male	male	bud	A	7.17	4.18	58.31
S. latifolia	C1_03_male	male	bud	A	8.22	4.93	59.99
S. latifolia	C1_04_male_1	male	bud	B	2.36	1.41	59.78
S. latifolia	C1_04_male_2	male	bud	C	17.3	9.78	56.59
S. latifolia	C1_05_male_1	male	bud	B	5.3	3.14	59.2
S. latifolia	C1_05_male_2	male	bud	C	12.83	7.51	58.5
S. latifolia	C1_26_female	female	bud	A	6.4	3.59	56.36
S. latifolia	C1_27_female	female	bud	A	14.53	8.83	60.74
S. latifolia	C1_29_female_1	female	bud	B	5.67	3.37	59.41
S. latifolia	C1_29_female_2	female	bud	C	10.48	6.13	58.47
S. latifolia	C1_34_female_1	female	bud	B	6.81	4.15	61.04
S. latifolia	C1_34_female_2	female	bud	C	12.29	7.47	60.78
S. latifolia	U10_11_male	male	bud	D	4.87	3.2	65.67
S. latifolia	U10_34_female	female	bud	E	4.57	3.01	65.95
S. latifolia	U10_37_female	female	bud	E	7.58	2.33	30.69
S. latifolia	U10_39_female	female	bud	E	4.47	2.84	63.6
S. latifolia	U10_09_male	male	bud	E	4.67	3.09	65.66
S. latifolia	U10_49_male	male	bud	D	8.99	3.44	38.26
S. latifolia	total		bud		144.5		57.72
S.vulgaris	See_1	hermaphrodite	bud	F	11.55	4.99	43.2
S.vulgaris	Guard_1	hermaphrodite	bud	G	6.42	3.09	48.13
S.vulgaris	V1_1	hermaphrodite	bud	H	3.28	1.23	37.5
S.vulgaris	V1_2	hermaphrodite	bud	H	5.99	2.78	46.41
S.vulgaris	V1_4	hermaphrodite	bud	H	5.69	2.56	44.99
S.vulgaris	total		bud		32.93		44
S. latifolia	C1_01_L_male	male	leaf	I	6.79	3.54	52.18
S. latifolia	C1_03_L_male	male	leaf	I	4.87	2.86	58.75
S. latifolia	C1_04_L_male	male	leaf	I	5.35	3.04	56.84
S. latifolia	C1_05_L_male	male	leaf	I	8.94	5.29	59.17
S. latifolia	C1_26_L_female	female	leaf	I	4.35	2.56	58.89
S. latifolia	C1_27_L_female	female	leaf	I	7.21	4.11	56.93
S. latifolia	C1_29_L_female	female	leaf	I	13.30	7.66	57.67
S. latifolia	C1_34_L_female	female	leaf	I	2.48	1.46	59.07
S. latifolia	total		leaf		41.32		57.44
S. latifolia	Leuk_144.1	male	bud	F	11.5	6.44	55.86
S. latifolia	U11_29_norm_female	female	floral	K_454	0.60	-	-
S. latifolia	U11_06_norm_male	male	floral	K_454	0.62	-	-
S. latifolia	total				1.22		

Table 6.S2: Comparison of the sex bias in bud and leaf tissues in *S. latifolia*. Proportion and absolute numbers of contigs with sex-biased (female-biased, male-biased and unbiased) expression for the contigs categorized as autosomal, sex-linked, X-hemizygous and undefined for flower buds and rosette leaf. In both cases, contigs with sex-biased expression are over-represented on the sex chromosomes (Wilcoxon-test, P-value ≤ 0.0001), but sex biases are much less abundant in vegetative leaf tissues than in flower buds.

Flower buds								
	Autosomal		Sex-linked		X-hemizygous		Undefined	
	%	#	%	#	%	#	%	#
Female-biased	7.2	919	13.8	129	5.8	18	2.9	591
Male-biased	9.6	1223	15.0	140	13.5	42	13.5	2738
Unbiased	83.2	10566	71.2	667	80.7	250	83.6	16932
Rosette leaves								
	Autosomal		Sex-linked		X-hemizygous		Undefined	
	%	#	%	#	%	#	%	#
Female-biased	0.6	70	4.1	34	0.8	2	1.3	91
Male-biased	0.3	34	3.4	28	0.8	2	2	147
Unbiased	99.1	10857	92.5	771	98.4	247	96.7	7028

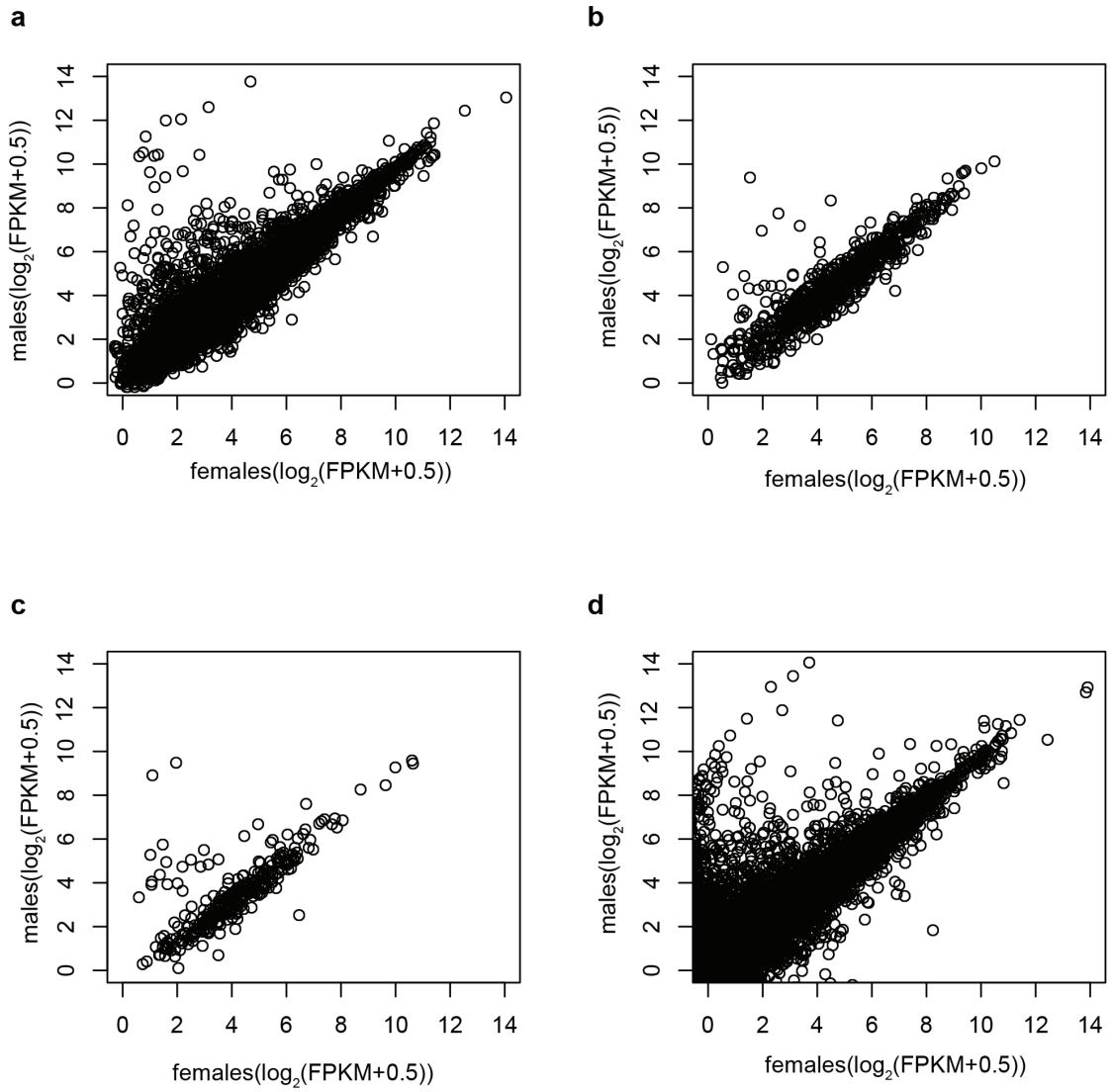


Figure 6.S1: Expression intensities in *S. latifolia* male and female flower buds. Mean expression intensity values (FPKM) in males are plotted against those in females for (**a**) autosomal contigs, (**b**) sex-linked, (**c**) X-hemizgous and (**d**) undefined contigs.

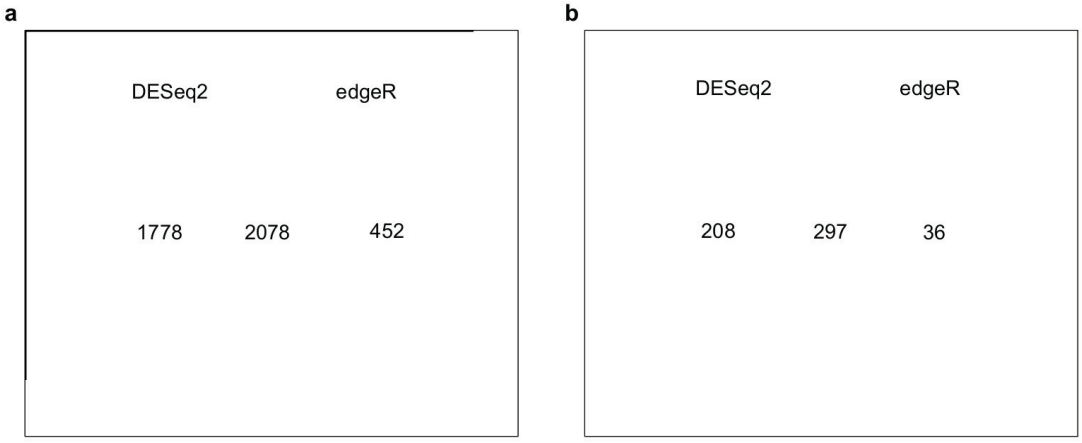


Figure 6.S2: Comparison between different normalization methods. Venn diagrams of all significantly sex-biased contigs in flower buds identified using both edgeR and DEseq2 for **a,** autosomal and, **b,** sex-linked contigs. We used the results obtained with edgeR for the analyses descrtibed in the main text, but an analysis of contigs that were identified as being sex-biased in expression by both methods also revealed a significant excess of sex-biased contigs on the sex chromosomes. Sex bias was significant for 1,765 (13.9%) of the autosomal contigs, and 936 (25.5%) of the sex-linked contigs. This difference is significant by a Fisher's exact test (p-value ≤ 0.0001).

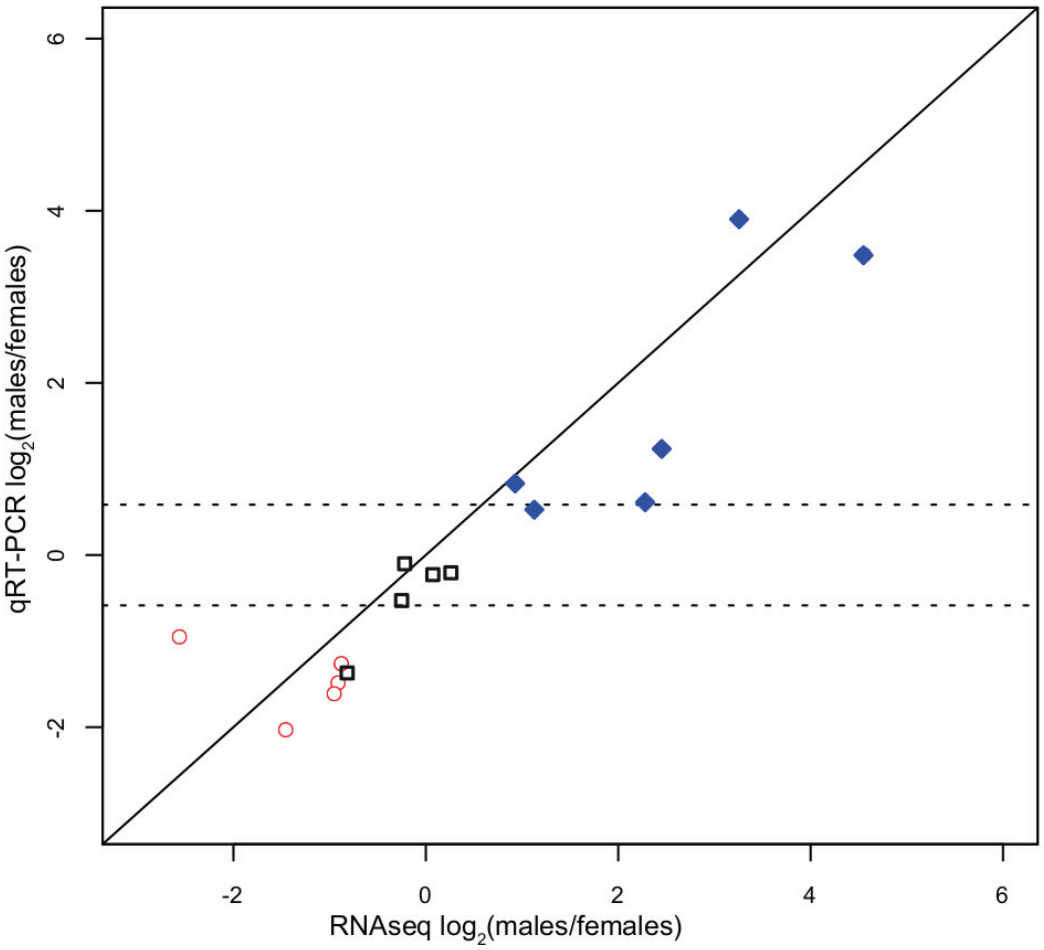


Figure 6.S3: qRT-PCR validation of contigs for which sex-biased expression was inferred using RNA-seq. Estimates of expression bias between the sexes based on qRT-PCR results are highly correlated with those from RNA-seq (Spearman correlation; ρ=0.92; p-value $\leq$ 0001). Contigs classified as either male-biased, female-biased, or unbiased in expression, based on RNA-seq, are indicated in blue (diamonds), red (circles) and white (squares), respectively. One contig classified by RNA-seq as unbiased showed female-biased expression in the qRT-PCR results, suggesting that our assessment of sex-biased expression based on RNA-seq data is conservative.

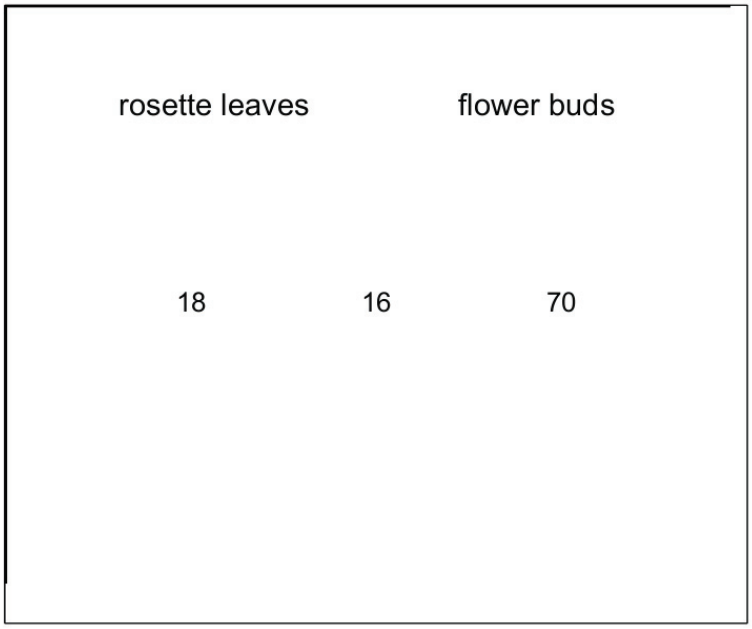


Figure 6.S4: Incomplete dosage compensation of female-biased, sex-linked contigs. We used comparisons of the expression levels of sex-linked contigs with female-biased expression in bud and leaf tissues to distinguish incomplete dosage compensation from female-biased expression. Vegetative tissues should not show strongly sex-biased expression because sexual dimorphism is expected to be less than in flowers (for more details see Section 6.9.2).

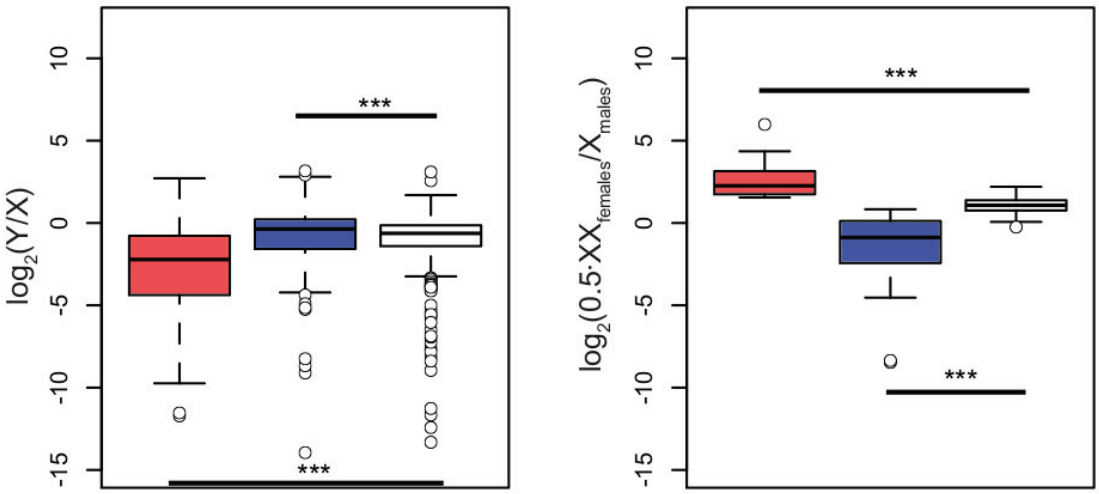


Figure 6.S5: Apparent masculinization of the Y chromosome, and X chromosome feminization of X-hemizygous contigs a, Boxplots showing Y/X expression ratios of sex-linked contigs with female-biased (red) , male-biased and unbiased (white) expression. The figure is organised similarly to Figure 6.5b, and also supports the suggested defeminization of the Y chromosome (Wilcoxon-test, p =0.001), but the degree of masculinization is slightly stronger than in Figure 6.5b (Wilcoxon-test, p =0.014) using the expression level from the single X in males to classify the gene as having partially degenerated, rather than the expression level in *S. vulgaris* hermaphrodites (Supplementary Figure 6.S8). **b,** Boxplot showing the ratios of 0.5*XX values in females divided by the X-allele expression in males for the contigs inferred to be X-hemizygous; female-biased contigs are shown in red, male-biased in blue and contigs with unbiased expression in white. In agreement with Figure 6.5b, the expression changes are consistent with both feminization (Wilcoxon-test, $p \leq 0.001$), and demasculinization (Wilcoxon-test, $p \leq 0.001$), of the *S. latifolia* X chromosome.

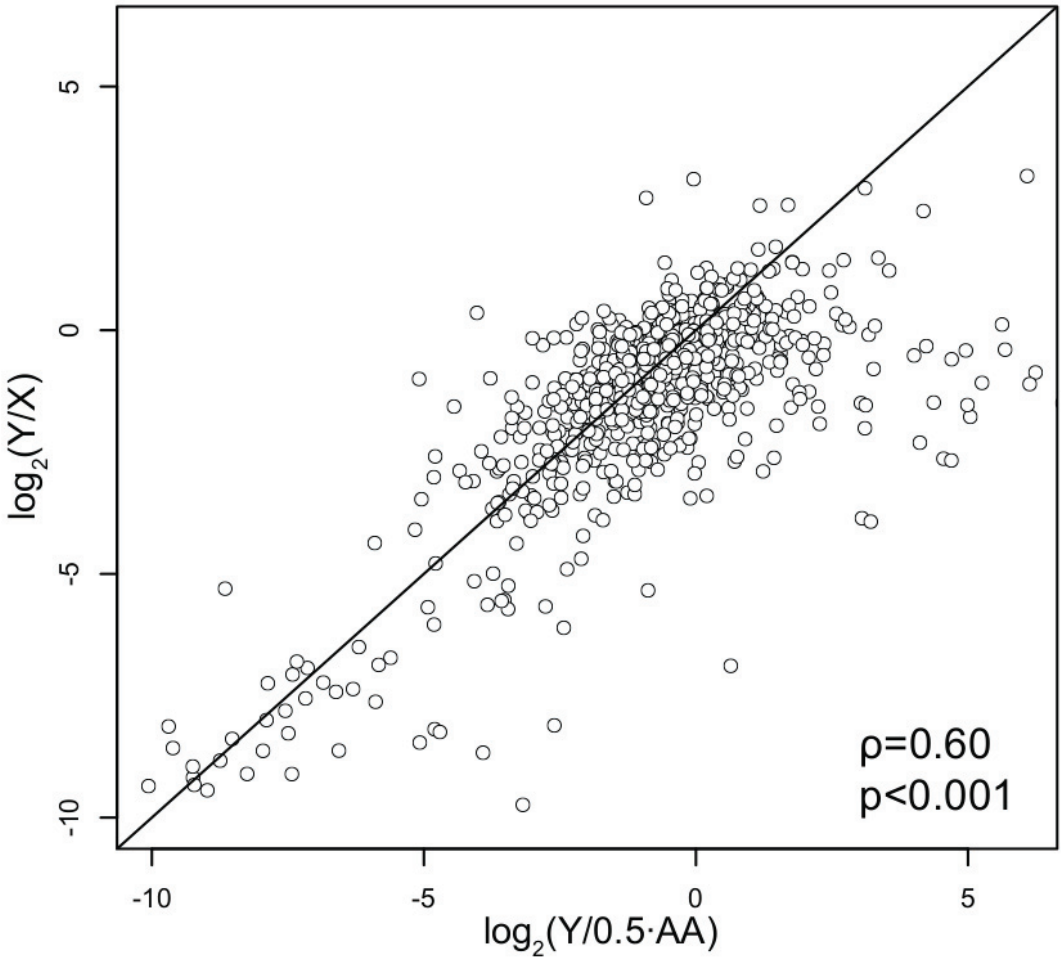


Figure 6.S6: Reference point for Y-degeneration. Correlation of the Y/X ratios in *S. latifolia* males and the Y values divided by *S. vulgaris* hermaphrodite 0.5*AA values for all *S. latifolia* sex-linked contigs (Spearman correlation, ρ=0.60 and P-value<0.0001). Most, but not all, contigs show good agreement between the two ratios.

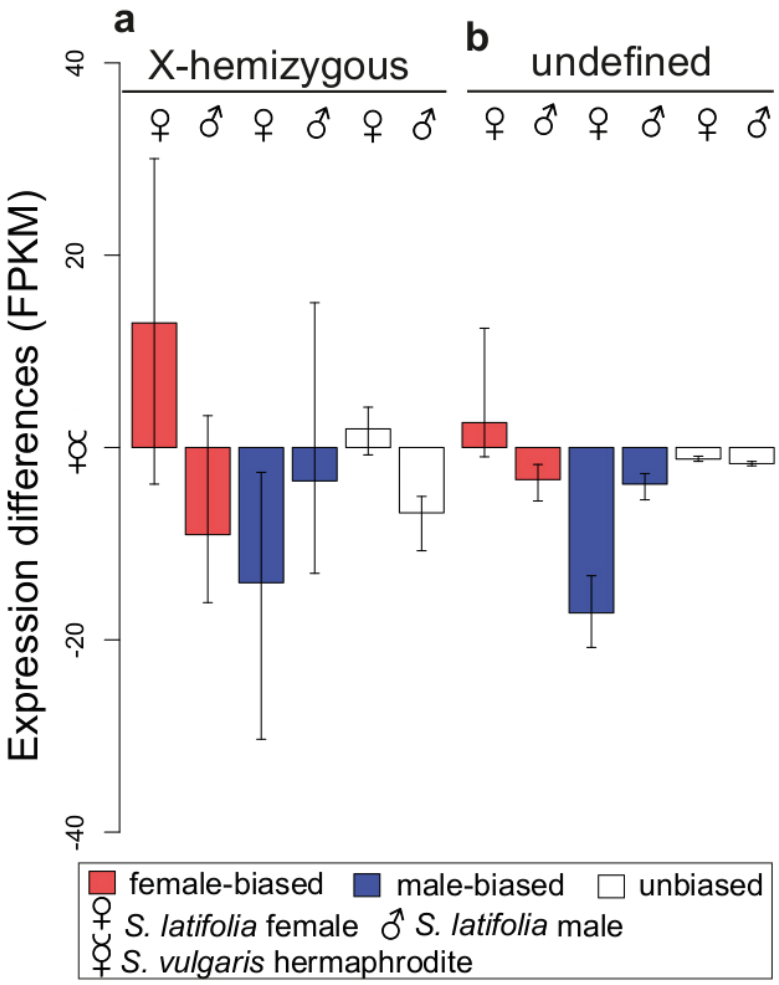


Figure 6.S7: Expression divergence for X-hemizygous and undefined contigs. Same graph as Figure 6.3 but expression divergence is shown for (**a**) X-hemizygous and (**b**) undefined contigs. The contigs whose genomic locations are undefined often behave like autosomal contigs, whereas the X-hemizygous contigs often show different patterns, presumably because of their low number.

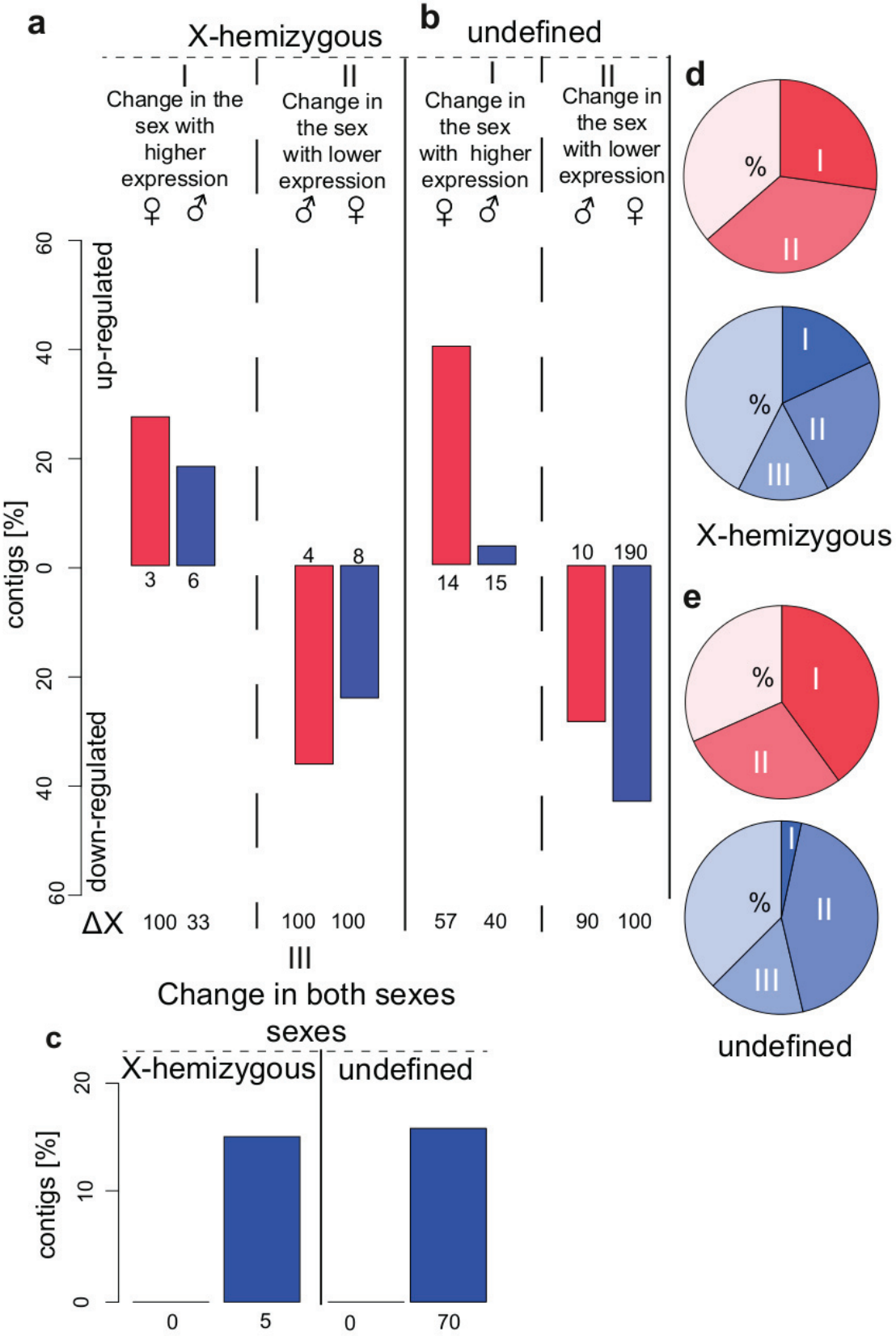


Figure 6.S8: Evolutionary expression patterns for X-hemizygous and undefined contigs. Same as Figure 6.4, but showing expression divergence values for (**a**) X-hemizygous and (**b**) undefined contigs. As in Supplementary Figure 6.S7, the contigs whose genomic locations are undefined often behave like autosomal contigs, whereas the X-hemizygous contigs often show different patterns.

Part III

General Conclusions and Perspectives

Obtaining sex chromosome sequences in non model organisms

My PhD work generated a highly reliable model-based method to infer sex-linked genes from RNA-seq data on a cross (parents and progeny of each sex), which is promising for the study of sex chromosomes in non model organisms as it is cheap (one Illumina lane for U/V systems and two for X/Y and Z/W systems) and can be applied without any former genetic knowledge (Chapter 4). I plan to do a post-doc with Gabriel Marais after my PhD and one aspect of my post-doc will be to apply SEX-DETector to various *Silene* species to compare the evolution of their sex chromosomes.

The fact that SEX-DETector requires a cross makes it non applicable to some non-model organisms for which crosses have not yet been generated and for which it would take a long time due to the delay before maturity (which can be several years for trees for example), also for some species it is not possible to generate a cross because they cannot reproduce and grow outside of their natural habitat. These limitations apply for example to the dioecious plant *Silene acaulis*, which is an alpine slow-growing cushion plant. Phylogeny suggests a very recent origin of dioecy in that plant since closely related species and subspecies are not dioecious. Jos Käfer and I made a field trip in the French Alps in order to sample males and females from a same population and to sequence them by RNA-seq. The data was analysed by Paul Jay, a Master intern, and I participated to the supervision of the internship. We searched the data for sex-linked genes by looking for genes where all males were heterozygous (XY) and all females homozygous (XX), and conversely in the case of a Z/W system. Both X/Y and Z/W analyses identified sex-linked genes, however when permuting the sexes of the individuals, only the observed number of X/Y genes was significantly higher than expected by chance, suggesting that *S. acaulis* has an X/Y system. These kind of analyses would be made easier if a model-based method was available to infer sex-linked genes on RNA-seq data for population data. Such a method already exists for DNA-seq data (Gautier, 2014) but is only adapted to old and diverged sex chromosomes and would need some adjustments to work on young chromosomes and RNA-seq data. A project is ongoing in Gabriel Marais' Lab to develop such a method.

The development of SEX-DETector, along with another project that generated BAC clone sequences (Chapter 3) made it possible for me to study the evolution of sex chromosomes in the dioecious plant *Silene latifolia* during my PhD.

Sex chromosome evolution in *Silene latifolia*

One of the aims of my PhD was to try and clarify the situation regarding the existence of dosage compensation in *Silene latifolia*, as contradicting results were published (Bergero et al., 2015; Chibalina and Filatov, 2011; Muyle et al., 2012). I could show that the apparent contradiction came from the use of different gene sets: either X/Y (Muyle et al., 2012) or X-hemizygous (without Y copy, Bergero et al., 2015; Chibalina and Filatov, 2011) genes. In the analyses I did, I included both type of genes and could confirm dosage compen-

sation for X/Y genes, as well as confirm that most X-hemizygous genes are not dosage compensated (Chapter 5). This suggests X-hemizygous genes are probably not dosage sensitive, which is consistent with the fact that they lost their Y copy. It should be kept in mind that, even in the canonical systems of dosage compensation such as placentals and *C. elegans*, dosage compensation is never complete for all genes, but rather happens only for dosage sensitive genes (Ercan, 2015). The fact that dosage compensation does not happen for X-hemizygous genes in *S. latifolia* should not lead to the conclusion that there is no dosage compensation in this plant, all genes should be analysed in order to see if a common pattern arises, such as the one we found in Chapter 5 which is a general hyperexpression of the maternal X in males and, to a lesser extent, in females, suggesting an imprinting mechanism for dosage compensation in *S. latifolia*. Future work could focus on epigenetic marks of *S. latifolia*'s sex chromosomes in order to elucidate the mechanism at the molecular level.

The genes identified through the sequencing of BAC clones allowed us to detect a bias in RNA-seq-based estimations of the percentage of Y gene loss. Indeed, RNA-seq was used to estimate the percentage of Y gene loss as the percentage of X-hemizygous genes among sex-linked genes (Bergero and Charlesworth, 2011; Bergero et al., 2015), however our results suggest that this percentage was underestimated because X-hemizygous genes tend to be less expressed than X/Y genes, which makes their detection harder with RNA-seq, and also because X-hemizygous genes can only be detected with SNPs on the X chromosomes, whereas X/Y genes can be identified both with X/Y and X/X SNPs. Unfortunately we did not have the proper statistical support to infer precisely a percentage of Y gene loss in *S. latifolia*, and more DNA-seq data will be required in order to test further whether plant Y chromosomes lose less genes than animal Y chromosomes. During my post-doc I also plan to use the availability of a new DNA-seq data to further address this issue.

The sex-linked genes I inferred with SEX-DETector and the allelic X and Y expression levels I generated were used in a project led by Niklaus Zemp in Alex Widmer's group (ETH Zurich) to study the evolution of sex-biased genes in *S. latifolia*. Results showed that sex chromosomes are a favoured region for the evolution of sex-biased expression, as expected due to linkage with the sex determining region, and also that sexual conflicts are mostly solved through changes in female expression levels.

Test of the dead end hypothesis of dioecy in *Silene*

During my thesis I grew plants in a greenhouse (using seeds collected by Gabriel Marais' group on the field or sent by European collaborators) and prepared samples to be sequenced by RNA-seq for three *Silene* species with contrasting mating systems: *S. latifolia* which is dioecious, *S. vulgaris* which is gynodioecious and *S. viscosa* which is hermaphroditic (all samples of *S. latifolia* and some samples of *S. viscosa* were generated by Niklaus Zemp in Zurich). I now have at my disposal RNA-seq data for multiple individuals from

populations spanning the geographic distribution of each species, which I will use to study the divergence as well as the polymorphism data in these species. I have already assembled the transcriptomes of each species and plan to continue the analyses during my post-doc in order to test whether dioecy is associated with higher rates of deleterious mutations accumulation, as predicted by the dead-end hypothesis (see Chapter 1).
This dataset will also allow me to test for faster-X evolution in *S. latifolia*, a phenomenon that was previously described in animals and that is due to the partial hemizygosity of the X chromosomes that allows for more efficient selection of recessive advantageous mutations.

References

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Qiu, S., and Charlesworth, D. (2015). Gene Loss from a Plant Sex Chromosome System. ENG. *Current biology: CB*. ISSN: 1879-0445. DOI: 10.1016/j.cub.2015.03.015.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Ercan, S. (2015). Mechanisms of x chromosome dosage compensation. eng. *Journal of Genomics* 3, 1–19. ISSN: 1839-9940. DOI: 10.7150/jgen.10404.

Gautier, M. (2014). Using genotyping data to assign markers to their chromosome type and to infer the sex of individuals: a Bayesian model-based classifier. eng. *Molecular Ecology Resources* 14(6), 1141–1159. ISSN: 1755-0998. DOI: 10.1111/1755-0998.12264.

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., and Marais, G. A. B. (2012). Rapid de novo evolution of X chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. eng. *PLoS biology* 10(4), e1001308. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001308.

Part IV

Annexes

Appendix A

Mating systems and selection efficacy: a test using chloroplastic sequence data in Angiosperms

This chapter is the result of a Master internship I did from September to December 2010 under the supervision of Sylvain Glémin (Monptellier, France). It aimed at testing whether the efficacy of selection is reduced in selfing angiosperm species. I worked again on this project during my PhD in order to publish it.

JOURNAL OF Evolutionary Biology

doi: 10.1111/jeb.12356

Mating systems and selection efficacy: a test using chloroplastic sequence data in Angiosperms

S. GLÉMIN* & A. MUYLE†

*Institut des Sciences de l'Evolution de Montpellier, UMR CNRS 5554, Montpellier, France
†Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Lyon, France

Keywords:

chloroplast;
deleterious mutations;
dN/dS ratio;
mating systems;
selection;
selfing.

Abstract

Selfing is assumed to reduce selection efficacy, especially purifying selection. This can be tested using molecular data, for example by comparing the Dn/Ds ratio between selfing and outcrossing lineages. So far, little evidence of relaxed selection against weakly deleterious mutations (as inferred by a higher Dn/Ds ratio) in selfers as compared to outcrossers has been found, contrary to the pattern often observed between asexual and sexual lineages. However, few groups have been studied to date. To further test this hypothesis, we compiled and analysed chloroplastic sequence data sets in several plant groups. We found a general trend towards relaxed selection in selfers in our data sets but with weak statistical support. Simulations suggested that the results were compatible with weak-to-moderate Dn/Ds ratio differences in selfing lineages. Simple theoretical predictions also showed that the ability to detect relaxed selection in selfers could strongly depend on the distribution of the effects of deleterious mutations on fitness. Our results are compatible with a recent origin of selfing lineages whereby deleterious mutations potentially have a strong impact on population extinction or with a more ancient origin but without a marked effect of deleterious mutations on the extinction dynamics.

Introduction

The evolution of selfing from outcrossing is one of the most frequent life-history trait transitions in angiosperms (Stebbins, 1957) and may also have recurrently occurred in many other groups of animals (Jarne & Auld, 2006) and fungi (Billiard *et al.*, 2011; Gioti *et al.*, 2012). Transitions from obligate outcrossing (self-incompatible species) to self-compatibility and eventually to selfing are mainly irreversible (Igic *et al.*, 2006, 2008; Igic & Busch, 2013) and come with higher extinction rates (Goldberg *et al.*, 2010). Selfing is thus supposed to be an evolutionary dead-end strategy (Stebbins, 1957; Takebayashi & Morrell, 2001; Igic & Busch, 2013). Although there is increasing evidence in favour of the dead-end hypothesis, the underlying causes of higher extinction rates in selfers are still being debated (Glémin & Galtier, 2012; Wright *et al.*, 2013).

Selfing automatically reduces the effective population size, N_e, by two-fold because of nonindependent gamete sampling during reproduction (Pollak, 1987; Nordborg, 1997), while also reducing effective recombination because it mainly occurs between homozygote sites (Nordborg, 2000). Genetic hitchhiking effects due to higher genetic linkage (Maynard-Smith & Haigh, 1974; Charlesworth *et al.*, 1993) are thus expected to reduce N_e below the automatic two-fold level. Finally, recurrent bottlenecks that are considered to be more frequent in selfers since a single seed can find a new population (Schoen & Brown, 1991; Ingvarsson, 2002) can also further reduce N_e. The overall reduction in N_e could be formulated as follows (Glémin, 2007):

$$N_e(F) = \alpha(F)\frac{N}{1+F} \quad (1)$$

where F is Wright's fixation index (F_{IS}), and $0 < \alpha(F) \leq 1$ summarizes the hitchhiking and bottleneck effects. The main hypotheses put forward to explain the dead-end theory are thus based on the premise that,

Correspondence: Sylvain Glémin, Institut des Sciences de l'Evolution. CC64. Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France.
Tel.: +33 (0)4 67 14 35 87; fax: +33 (0)4 67 14 36 10;
e-mail: sylvain.glemin@univ-montp2.fr

because of the low N_e and weak recombination efficiency, selfing should reduce the selection efficacy, potentially leading to the accumulation of mildly deleterious mutations (Lynch *et al.*, 1995; Schultz & Lynch, 1997; Glémin, 2007) and limiting adaptation to new biotic and abiotic environments (Stebbins, 1957; Agrawal & Lively, 2001; Morran *et al.*, 2011; Glémin & Ronfort, 2013). However, the respective roles of reduced adaptive ability versus genetic deterioration are still unclear. Importantly, selfing also exposes alleles in homozygotes, thus increasing their apparent dominance levels and facilitating both positive and negative selection. For intermediate dominance (i.e. additive selection, whereby the homozygotic effect of a mutation is twice the heterozygotic effect), the apparent dominance is increased by a $1 + F$ factor, which exactly compensates for the automatic N_e reduction (Caballero & Hill, 1992; Charlesworth, 1992; Pollak & Sabran, 1992). If mutations are recessive or partially recessive, as presumed for deleterious mutations, N_e must be reduced beyond the two-fold level – that is, $\alpha(F) < 1$ – for selection to be reduced in selfers (Glémin, 2007). This is also true for the fixation of new beneficial mutations, but not for adaptation from standing variation, which is less dependent on dominance levels, and which is more efficient in outcrossers than in selfers (Glémin & Ronfort, 2013).

The accumulation of deleterious mutations can be tested through molecular approaches by comparing the rates of nonsynonymous versus synonymous change in divergence (Dn/Ds, ω hereafter) or polymorphism (Pn/Ps). If most changes in amino acids are neutral or deleterious, ω and Pn/Ps should be higher in selfing lineages due to the relaxation of selection against these deleterious mutations. Such approaches have been conducted in several species (Table 1) and gave mixed results. Most studies based on divergence failed to detect relaxed selection in selfers, except in *Neurospora*, where tiny differences in ω ($\omega_{out} \sim 0.14$ vs. $\omega_{self} \sim 0.17$) have been detected using a very large data set of more than 2700 genes (Gioti *et al.*, 2012). On the contrary, most studies based on polymorphism detected relaxed selection in selfers compared to outcrossers. This suggests that selfing, hence relaxed selection, is of recent origin and can only be properly detected on a short time scale. This has been clearly illustrated in a study of the recently derived selfing species *Capsella rubella*, where a new method to identify founding haplotypes of ancestral populations allowed separation of ancestral and novel variations (Brandvain *et al.*, 2013). The authors detected a strong signature of relaxed selection in newly emerged variation (high Pn/Ps ratio within founding haplotypes) compared to ancestral variation (low Pn/Ps ratio between founding haplotypes).

In contrast, studies comparing asexual and sexual species more often detected relaxed selection in asexual lineages using divergence measures, even with relatively small data sets (Table 1). Note, however, that the assumption that higher ω values in asexual lineages can be directly attributed to clonality has been recently challenged in a new analysis of *Daphnia* genomes (Tucker *et al.*, 2013). This suggests that deleterious mutations should accumulate at a slower rate in selfers than in asexuals and could not be the main cause of the extinction of the former (see review and discussion in Glémin & Galtier, 2012). However, data are still insufficient to draw a firm conclusion, especially because the increase in ω in selfing lineages has not been tested in many plant groups.

Here, we have therefore extended the testing of this hypothesis to new plant groups, while focusing on two chloroplastic genes (*matK* and *rbcL*) for several reasons. These genes have been widely sequenced for phylogenetic studies and are thus available for many angiosperm species, especially in groups whose mating systems have been mapped onto a phylogeny. As chloroplastic genomes are haploid, they are not subject to dominance effects, thus sidestepping possible confounding effects of homozygosity on selection efficacy (see above). The two-fold reduction in N_e due to homozygosity does not affect haploid genomes so that only bottleneck and hitchhiking effects do matter:

$$N_e^{chloro}(F) = \alpha(F) N_{female} \quad (2)$$

.

As stated above, this additional more than two-fold reduction in N_e ($\alpha(F) < 1$) is pivotal to the argument that selection is less efficient in selfers (Glémin, 2007). Bottlenecks should affect cytoplasmic and nuclear genes similarly, whereas nuclear genes should be more sensitive to hitchhiking effects. Nuclear genes are affected by selection due to sites physically linked on the same chromosome and also to sites on other chromosomes that are genetically – but not physically – linked through selfing. Conversely, selection on the nuclear genome affects chloroplastic genes only through linkage due to selfing. For species with few chromosomes and/or a short nuclear genetic map, $\alpha(F)$ should be higher for chloroplastic than for nuclear genes. However, for species with numerous chromosomes and/or a long nuclear genetic map, physical linkage is negligible for nuclear genes and $\alpha(F)$ should be similar for chloroplastic and for nuclear genes. This could be more formally illustrated by comparing $\alpha(F)$ under the background selection model for chloroplastic and nuclear genes (see Appendix 1), as shown in Fig. 1. The difference is substantial only for very short genetic maps (dotted lines equivalent to only one chromosome of 100 cM). For larger maps, differences between nuclear and chloroplastic genes can occur when the genomic mutation rate is high and average selection coefficient against deleterious mutations is rather low (Fig. 1c). Another potential problem of our approach is that the data sets are limited to only two genes (but also see some studies

Table 1 Summary of studies comparing molecular evolution patterns between selfing and outcrossing or asexual and sexual species (updated from Glémin & Galtier, 2012).

Taxonomic group	Groups compared	Data set	dN/dS	pN/pS	Positive selection	Codon usage	Reference
Selfing vs. outcrossing							
Angiosperms	29 selfers/42 outcrossers	Meta-analysis (polymorphism)		±	±		Glémin *et al.* (2006)
Arabidopsis	1 selfer/1 outcrosser	23 nuc genes + 1 chloro gene	–			–	Wright *et al.* (2002)
Arabidopsis	1 selfer/1 outcrosser	675/62 nuc genes		–			Foxe *et al.* (2008)
Arabidopsis/ Brassica	1 selfer/2 outcrossers	185 nuc genes				–	Wright *et al.* (2007)
Arabidopsis/ Capsella	1 selfer/1 outcrosser	257 nuc genes		**+**	**+**		Slotte *et al.* (2010)
Arabidopsis/ Capsella	2 selfers/2 outcrossers	780, 89/120, 257 nuc genes		**+**			Qiu *et al.* (2011)
Capsella	1 selfer/1 outcrosser	Complete genome/transcriptome		**+**			Slotte *et al.* (2013)
Caenorhabditis	2 selfers/4 outcrossers	> 1000 nuc genes	–			±	Cutter *et al.* (2008)
Collinsia	1 selfer/1 outcrosser	17 nuc genes + transcriptomes		**+**		–	Hazzouri *et al.* (2013)
Eichornia	3 selfers/1 outcrosser	~ 8000 nuc genes		**+**		**+**	Ness *et al.* (2012)
Neurospora	32 homothallic/ 17 heterothallic	7 nuc genes	**+***				Nygren *et al.* (2011)
Neurospora	1 pseudohomothallic/ 1 heterothallic	> 2000 nuc genes				–	Whittle *et al.* (2011)
Neurospora	4 homothallic/ 31 heterothallic	> 2700 nuc genes	**+**			±	Gioti *et al.* (2013)
Triticeae	2 selfers/2 outcrossers	52 nuc genes + 1 chloro gene	–		±		Haudry *et al.* (2008)
Triticeae	9 selfers/ 10 outcrossers	27 nuc genes	–				Escobar *et al.* (2010)
Asexuals vs. sexuals							
Aphids	4 sexuals/4 asexuals	255 nuc genes + 10 mito genes	–				Ollivier *et al.* (2012)
Campeloma	6 asexuals/12 sexuals	1 mito gene (*Cytb*)	**+**				Johnson & Howard (2007)
Daphnia	14 asexuals/14 sexuals	Complete mito genome	**+**				Paland & Lynch (2006)
Daphnia	11 asexuals/11 sexuals	Complete mito genome	–				Tucker *et al.* (2013)
Oenothera	16 asexuals/16 sexuals	1 nuc gene (*chiB*)			**+**		Hersch-Green *et al.* (2012)
Potamopyrgus	14 asexuals/14 sexuals	Complete mito genome	**+**				Neiman *et al.* (2010)
Rotifers	3 asexuals/2 sexuals	1 nuc gene (*Hsp 82*)	**–**				Mark Welch & Meselson (2001)
Rotifers	3 asexuals/4 sexuals	1 mito gene (*Cox I*)	±	**+**			Barraclough *et al.* (2007)
Timema	6 asexuals/7 sexuals	2 nuc genes + 1 mito gene	**+**				Henry *et al.* (2012)

*Terminal vs. internal branches not controlled. Positive results are in bold.

in asexuals, Table 1), which limits the statistical power of individual analyses. However, any general trends should be detected by the combination of data sets. We thus also conducted simulations to test the power of our analyses to detect reductions in the selection efficacy in selfers.

Materials and methods

Data sets

We built the data sets by combining two sources of information. First, we searched in the literature for publications in which the mapping of contrasted mating systems on a phylogeny was documented. We used information from studies comparing selfers and outcrossers, self-compatible and self-incompatible species, and homostylous and heterostylous species. Hereafter, SELF denotes any one of selfing, self-compatible or homostylous mating system, and OUT refers to any outcrossing, self-incompatible or heterostylous mating system. Although direct tests of the hypothesis of the accumulation of deleterious mutations should concern selfing vs. outcrossing species, the self-compatibility status is usually associated with rather high selfing rates, and it could be a short intermediate step towards selfing (Igic *et al.*, 2008; Igic & Busch, 2013). For instance, this rationale has been used to test the dead-end hypothesis in Solanaceae (Goldberg *et al.*, 2010). We then conducted a GenBank search for *matK* and *rbcL* sequences corresponding to the species present in these phylogenies. Sequences were aligned with MUSCLE (Edgar, 2004) and manually checked and cleaned. When sequence lengths were too heterogeneous between species, we kept the species with the longest sequences to avoid numerous alignment gaps.

Previously published trees served as reference trees for each group of species. When sequences were lacking, we removed the corresponding species from

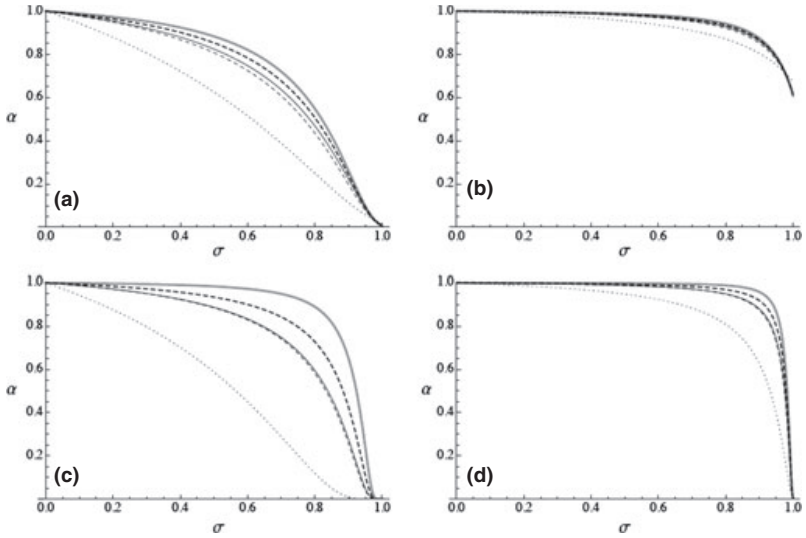


Fig. 1 Reduction in the effective population size due to background selection as compared to an outcrossing population as a function of the selfing rate for nuclear and chloroplastic genes. The equations are given in Appendix 1. Total genomic mutation rate is $U = 1$ (a,c) or $U = 0.1$ (b,d), dominance coefficient is $h = 0.2$ and selection coefficient $s = 0.1$ (a,b) or $s = 0.01$ (c,d). Thick grey line: chloroplastic genes. Dotted line: one chromosome of genetic map Cr = Gr = 1. Dashed line: one chromosome of genetic map Cr = Gr = 5. Thin line: five chromosomes of equal size and equal map Cr = 1, so that Gr = 5. Thick dashed line: 10 chromosomes of equal size and equal map Cr = 1, so that Gr = 10. Note that we did not plot larger genetic maps as they quickly converge to the chloroplastic case.

the corresponding tree. Methods implemented to map mating system evolution on phylogenies may differ between publications. To homogenize data sets and determine the status of all branches, as required in dN/dS analyses, we opted to map mating systems using parsimony and assuming unidirectional shifts from outcrossing (resp. self-incompatibility or heterostyly) to selfing (resp. self-compatibility or homostyly). The resulting maps were usually very close to those proposed in the original publications.

The list of species with the corresponding GenBank accessions is given in Table S1. Phylip files of alignments and trees with mapped mating systems are provided in Dryad repository.

Sequence analyses

We used the *codeml* program of the PAML package (Yang, 2007) to perform various tests of codon evolution along phylogenies. Before testing the effect of mating systems on ω, we assessed the possible occurrence of positive selection in the sequences, as it has been shown that positive selection can affect *rbcL* (Kapralov & Filatov, 2007). We used the 'site models' implemented in *codeml* to detect potential sites under positive selection. We compared the M7 model (with a beta distribution of ω values between 0 and 1) and the M8 model (with a beta distribution plus an additional category with $\omega > 1$) by a likelihood ratio test (LRT) with two degrees of freedom. When sites under positive selection were detected, we built a second data set by removing from the alignment the sites belonging to the $\omega > 1$ class with high posterior probability (Bayes empirical Bayes prob. > 0.95).

For each data set, including those without sites under positive selection, we ran several nested 'branch models' (Fig. 2). We first ran a null model with a single ω for all branches (M_0) and a model with two ω, that is, one for SELF branches and one for OUT branches (including internal branches; $M_{\text{self-out}}$ Fig. 2a). However, changes in terminal branches may correspond to polymorphic mutations, not substitutions, yielding higher ω because weakly deleterious mutations can be transiently polymorphic before eventually being lost. This can be misleading because selfing, SC or homostylous is mainly found on terminal branches. Moreover, assignation of mating systems on internal branches can be less accurate than on terminal branches. We thus ran two alternative models: a model with two ω, one for internal and the other for external branches ($M_{\text{int-ext}}$ Fig. 2b), and also a model with three ω, one for internal branches, one for SELF external branches and one for OUT external branches (M_3 Fig. 2c). Finally, some data sets presented internal SELF branches. We thus ran a model with four ratios for SELF internal, OUT internal, SELF external and OUT external branches (M_4 Fig. 2d). Nested models were tested by LRTs with the appropriate number of degrees of freedom.

To combine information from the different data sets, we performed binomial sign tests to compare ω

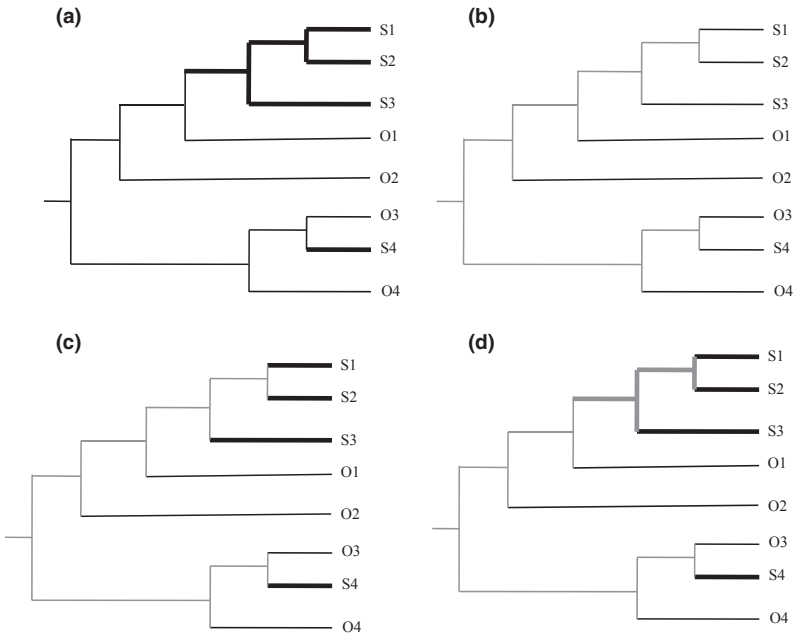


Fig. 2 Branch models used in the analyses. (a) $M_{\text{self-out}}$: SELF (thick) and OUT (thin) branches (two ω). (b) $M_{\text{int-ext}}$: internal (grey) and external (black) branches (two ω). (c) M_3: internal branches (grey), SELF (thick black) and OUT (thin black) external branches (three ω). (d) M_4: SELF (thick grey) and OUT (thin grey) internal, and SELF (thick black) and OUT (thin black) external branches (four ω). Outcrossing, resp. selfing, species are indicated by O, resp. S.

between internal and external branches and between SELF and OUT branches. We also combined likelihood between data sets for each model by summing log-likelihoods. As we wanted to perform unilateral tests, that is, $\omega_{\text{ext}} > \omega_{\text{int}}$ and $\omega_{\text{SELF}} > \omega_{\text{OUT}}$, we added the log-likelihood of the alternative model of a given data set only if the order of ω values corresponded to the alternative hypothesis. Otherwise, if the order of ω was contrary to that expected, we added the log-likelihood of the corresponding null model. In such cases, the data set did not contribute to increasing the likelihood of the alternative model but cost one degree of freedom (Escobar *et al.*, 2010).

Simulations

To assess the power of our analyses to detect relaxed selection in selfers, we performed simulations to determine which differences in ω were detectable in our analyses. For simplicity, we chose two reference trees with only 12 species: one with six SELF species on the leaves and one with a clade of three SELF species (Fig. 3). In both trees, the total length of SELF branches corresponded to 25% of the total tree. We only considered two ω and did not distinguish between internal and external branches. We used the *evolver* program of the PAML package (Yang, 2007) to simulate sequences along these two phylogenies with two different ω. We simulated sequences of 999 bp (333 codons) with a codon composition and a Ts/Tv parameter (=2.37) corresponding to the *matK* data set in *Veronica*. To vary the simulated data set sizes, we set the total tree length at 0.5, 1 and 5. We used three ω values for the OUT branches: $\omega_{\text{OUT}} = 0.1$, 0.25 and 0.5. For the SELF branches, we increased the ω_{OUT} values by the following quantities $\Delta\omega = \omega_{\text{SELF}} - \omega_{\text{OUT}} = 0.05$, 0.1, 0.15, 0.2 and 0.25. We simulated 100 data sets for every combination of parameters (two trees, three total tree lengths, three ω_{OUT} and five ω_{SELF}). Then, we ran the M_0 and $M_{\text{self-out}}$ models defined above with *codeml*.

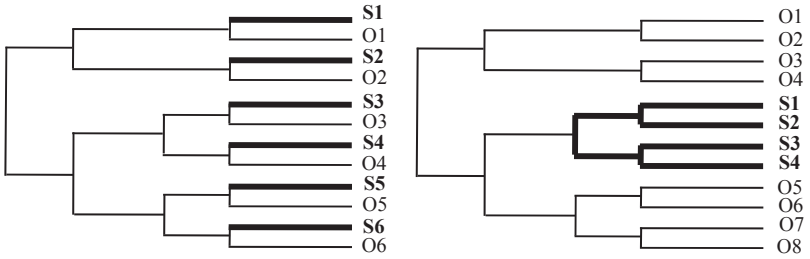


Fig. 3 Reference trees used in the simulations. For both trees, the length of SELF branches (thick lines), is 25% of the total tree length. Outcrossing, resp. selfing, species are indicated by O, resp. S.

We recorded the ω_{SELF} and ω_{OUT} values estimated in the $M_{self\text{-}out}$ model and the *P*-value of the LRT of the two models.

Results

We obtained 19 data sets corresponding to 13 species groups representative of several angiosperm families (Table 2). The smallest data set contained 10 species with a 466 bp alignment and the largest contained 175 species with a 1515 bp alignment. Three data sets (for *Exochaenium* and *Schiedea* genera) had fewer than 50 substitutions (NdN + SdS in Table 2), so some groups of branches had very few or no synonymous and/or nonsynonymous substitutions, and some models had an aberrant or undefined ω ratio. The results are presented in Table 3, but these three data sets were not taken into further account in the analyses.

The two genes mostly evolved under purifying selection. They both had $\omega < 1$, and, on average, *rbcL* was more constrained than *matK* (mean $\omega = 0.20$ vs. $\omega = 0.45$). External branches also showed higher ω than internal branches in 15 data sets of 16 (unilateral sign test: $P = 0.0003$; combined LRT: $\chi^2 = 60.34$, d.f. = 16, $P < 10^{-6}$, Table 3), which is the expected pattern under purifying selection. However, significant positive selection was found in most data sets (12 of 16), but very few codons (< 2.5%) were found to evolve under positive selection (BEB > 0.95; Table 2 and see details in Table S2). These sites were removed to build the second data sets. As detailed below, the results were similar when comparing the two data sets (compare Tables 3 and 4).

Effect of mating systems on ω

When comparing the $M_{self\text{-}out}$ and the M_0 model, 13 of 16 data sets showed higher ω_{SELF} than ω_{OUT} (unilateral sign test: $P = 0.010$; combined LRT: $\chi^2 = 34.06$, d.f. = 16, $P = 0.005$). Using the combination of data sets, we thus found evidence of higher ω in SELF than in OUT branches. However, when taken individually, only two data sets of 16 (Asteraceae *matK* and Solanaceae *rbcL*) showed significantly higher ω_{SELF} than ω_{OUT}. Similar results were obtained when codons potentially evolving under positive selection were removed (second data sets), but the power to detect differences by LRT was lower (Table 4): ω_{SELF} was higher than ω_{OUT} for 13 of 16 data sets (unilateral sign test: $P = 0.010$; combined LRT $\chi^2 = 18.48$, d.f. = 16: $P = 0.297$), and the difference was significant for only two different data sets (Pontedariaceae *rbcL* and Primula *rbcL*). However, as the SELF branches were mostly terminal, the previous results could be explained by the differences between internal and external branches that we detected in most data sets (see above). We thus tested the difference between mating systems on external branches only (models M_3 vs. $M_{int\text{-}ext}$). A majority of data sets (11 of 16) still showed higher external ω_{SELF} than ω_{OUT}, but the difference was not significant (unilateral sign test:

Table 2 Summary of the data sets used in the analyses: taxonomic group, mating system type.

Taxonomic group	MS type	No. of otal	No. of S, SC, or homo	No. of O, SI or het	Gene	No. of sites	NdN	SdS	No. of codons $\omega > 1$	Reference
Asteraceae	SC/SI	55	18	37	*matK*	1536	442.2	266.2	2	Ferrer & Good-Avila (2007)
	SC/SI	19	11	8	*rbcL*	1455	144.4	226.3	9	
Exochaenium	homo/het	10	3	7	*matK*	466	31.5	18.5	0	Kissling & Barrett (2013)
Geraniaceae	S/O	48	14	34	*rbcL*	1425	132.9	498.2	8	Fiz *et al.* (2008)
Linum	homo/het	11	6	5	*rbcL*	1425	16.2	108.6	0	Armbruster *et al.* (2006)
Medicago	S/O	31	25	6	*matK*	1518	225.9	108.5	6	Bena *et al.* (1998)
Polemoniaceae	S/O	50	19	31	*matK*	1074	400.1	221.2	8	Barrett *et al.* (1996)
Pontederiaceae	homo/het	23	8	15	*rbcL*	1344	35.9	145.2	3	Kohn *et al.* (1996)
Primula	homo/het	175	50	125	*matK*	1515	1127.8	702.2	10	Mast *et al.* (2006)
	homo/het	45	9	36	*rbcL*	1416	196.7	323.2	8	
Psychotria	homo/het	15	4	11	*matK*	864	98.5	64.4	0	Sakai & Wright (2008)
	homo/het	15	4	11	*rbcL*	553	40.8	30.8	4	
Schiedea	S/O	27	6	21	*matK*	1077	24.9	7.1	3	Sakai *et al.* (2006)
	S/O	22	4	18	*rbcL*	1362	22.1	12	6	
Solanaceae	SC/SI	83	61	22	*matK*	1527	266	151.8	6	
	SC/SI	31	21	10	*rbcL*	579	61.4	58.2	2	
Triticeae	S/O	19	9	10	*matK*	1539	76.6	80.7	4	Escobar *et al.* (2010)
Veronica	S/O	17	11	6	*matK*	951	180.7	175.6	0	Muller & Albach (2010)
	S/O	20	7	13	*rbcL*	1305	31.1	81.5	1	

SI, self-incompatible; SC, self-compatible; homo, homostyly; het, heterostyly; S, selfing; O, outcrossing; total, total number of species; NdN, total number of nonsynonymous substitutions; SdS, total number of synonymous substitutions.

Table 3 Results of ω = Dn/Ds analyses for branch models.

		M_0		$M_{int\text{-}ext}$			$M_{self\text{-}out}$			*P*-values	
Taxonomic group	Gene	ω	lnL	ω_{int}	ω_{ext}	lnL	ω_{out}	ω_{self}	lnL	$M_{int\text{-}ext}$ vs. M_0	$M_{self\text{-}out}$ vs. M_0
Asteraceae	*matK*	0.409	−5544.04	0.314	0.482	−5536.53	0.384	0.464	−5538.63	**< 0.0001**	**0.001**
Asteraceae	*rbcL*	0.188	−3988.59	0.188	0.188	−3988.59	0.174	0.206	−3988.31	0.999	0.459
Geraniaceae	*rbcL*	0.089	−5464.84	0.064	0.123	−5459.96	0.074	0.161	−5459.50	**0.002**	0.001
Linum	*rbcL*	0.046	−2574.38	0.030	0.057	−2573.82	0.039	0.062	−2574.06	0.289	0.422
Medicago	*matK*	0.418	−4206.90	0.285	0.461	−4205.71	0.313	0.429	−4206.65	0.124	0.482
Polemoniaceae	*matK*	0.464	−5311.78	0.446	0.480	−5311.69	0.436	0.503	−5311.45	0.677	0.420
Pontederiaceae	*rbcL*	0.077	−2835.61	0.106	0.061	−2834.57	0.068	0.105	−2835.06	0.149	0.294
Primula	*matK*	0.441	−14 147.27	0.430	0.451	−14 147.16	0.455	0.405	−14 146.76	0.637	0.309
Primula	*rbcL*	0.172	−4829.76	0.112	0.211	−4825.85	0.166	0.193	−4829.56	**0.005**	0.521
Psychotria	*matK*	0.436	−2013.58	0.398	0.496	−2013.39	0.444	0.371	−2013.53	0.539	0.755
Psychotria	*rbcL*	0.395	−1195.39	0.211	0.667	−1192.78	0.331	0.714	−1194.55	**0.022**	0.196
Solanaceae	*matK*	0.442	−5140.10	0.393	0.477	−5138.24	0.413	0.467	−5139.93	0.054	0.558
Solanaceae	*rbcL*	0.316	−1437.52	0.238	0.338	−1437.27	0.215	0.531	−1435.01	0.482	**0.025**
Triticeae	*matK*	0.256	−3078.97	0.132	0.384	−3074.20	0.238	0.283	−3078.84	**0.002**	0.605
Veronica	*matK*	0.284	−2997.67	0.265	0.300	−2997.56	0.246	0.326	−2996.96	0.638	0.235
Veronica	*rbcL*	0.114	−2437.01	0.062	0.185	−2434.89	0.131	0.094	−2436.74	**0.039**	0.462
Combined			−67 203.42			−67 173.25			−67 186.39	**< 0.0001**	**0.005**
Exochaenium	*matK*	0.435	−859.62	0.370	0.464	−859.56	0.359	*	−857.22	0.741	**0.029**
Schiedea	*matK*	0.666	−1624.83	*	0.348	−1621.56	0.952	0.094	−1623.22	**0.011**	0.074
Schiedea	*rbcL*	0.622	−2080.18	1.660	0.233	−2072.63	0.684	*	−2074.67	**0.0001**	**0.001**

		M_3				M_4					M_3 vs.	M_4 vs.
Taxonomic group	Gene	ω_{int}	$\omega_{ext\text{-}out}$	$\omega_{ext\text{-}self}$	lnL	$\omega_{int\text{-}out}$	$\omega_{int\text{-}self}$	$\omega_{ext\text{-}out}$	$\omega_{ext\text{-}self}$	lnL	$M_{int\text{-}ext}$	M_3
Asteraceae	*matK*	0.313	0.455	0.524	−5536.36	0.320	0.265	0.455	0.528	−5536.28	0.563	0.691
Asteraceae	*rbcL*	0.194	0.135	0.202	−3987.91						0.245	
Geraniaceae	*rbcL*	0.064	0.096	0.161	−5458.26						0.065	
Linum	*rbcL*	0.030	0.052	0.062	−2573.78						0.781	
Medicago	*matK*	0.285	0.293	0.485	−4205.11	*	0.277	0.293	0.485	−4204.60	0.273	0.312
Polemoniaceae	*matK*	0.446	0.422	0.515	−5311.39	0.443	0.460	0.422	0.515	−5311.38	0.436	0.908
Pontederiaceae	*rbcL*	0.107	0.029	0.106	−2832.00						**0.023**	
Primula	*matK*	0.431	0.466	0.417	−14 146.90	0.444	0.390	0.466	0.417	−14 146.61	0.465	0.452
Primula	*rbcL*	0.112	0.222	0.190	−4825.66						0.540	
Psychotria	*matK*	0.398	0.547	0.372	−2013.20						0.542	
Psychotria	*rbcL*	0.211	0.635	0.712	−1192.77						0.873	
Solanaceae	*matK*	0.393	0.337	0.544	−5136.96	0.466	0.255	0.336	0.544	−5135.59	0.109	0.098
Solanaceae	*rbcL*	0.238	0.225	0.565	−1435.17	0.182	0.386	0.225	0.565	−1434.86	**0.040**	0.429
Triticeae	*matK*	0.132	0.484	0.318	−3073.74	0.111	0.184	0.486	0.316	−3073.48	0.339	0.472
Veronica	*matK*	0.276	0.268	0.312	−2997.55	0.174	0.345	0.295	0.312	−2996.05	0.912	0.083
Veronica	*rbcL*	0.066	0.328	0.127	−2433.90	0.068	0.065	0.328	0.127	−2433.90	0.160	
Combined					−67 162.76						0.179	
					−40 071.97					−40 069.89		0.842
Exochaenium	*matK*	0.430	0.316	*	−857.11						**0.027**	
Schiedea	*matK*	*	0.475	0.093	−1620.79						0.214	
Schiedea	*rbcL*	1.660	0.266	*	−2076.44						1.000	

M_0, one ω; $M_{int\text{-}ext}$, internal and external branches, two ω; $M_{self\text{-}out}$, SELF and OUT branches, two ω; M_3, internal branches; SELF and OUT external branches, three ω; M_4, SELF and OUT internal, SELF and OUT external branches, four ω; lnL, log-likelihood. The combined likelihood and *P*-value computations are explained in the main text. *P*-value lower than 0.05 are in bold.
*Undefined ω ratio.

$P = 0.105$; combined LRT: $\chi^2 = 20.98$, d.f. = 16, $P = 0.179$), and only two data sets were individually significant (Pontederiaceae *rbcL* and Solanaceae *rbcL*). Similar results are obtained for the second data sets: external ω_{SELF} was higher than external ω_{OUT} for 11 of 16 data sets (unilateral sign test: $P = 0.105$; combined LRT $\chi^2 = 11.5$, d.f. = 16, $P = 0.777$), and only one data set was individually significant (Pontederiaceae *rbcL*).

Table 4 As in Table 3 for data sets without sites under significant positive selection.

Taxonomic group	Gene	M_0		$M_{int\text{-}ext}$			$M_{self\text{-}out}$			*P*-values	
		ω	lnL	ω_{int}	ω_{ext}	lnL	ω_{out}	ω_{self}	lnL	$M_{int\text{-}ext}$ vs. M_0	$M_{self\text{-}out}$ vs. M_0
Asteraceae	*matK*	0.403	−5466.47	0.314	0.472	−5464.15	0.376	0.466	−5465.87	**0.031**	0.273
Asteraceae	*rbcL*	0.133	−3620.15	0.098	0.161	−3618.50	0.108	0.167	−3618.72	0.127	0.216
Geraniaceae	*rbcL*	0.054	−4954.51	0.037	0.076	−4950.61	0.049	0.076	−4953.56	**0.005**	0.167
Medicago	*matK*	0.362	−3965.64	0.233	0.403	−3964.19	0.289	0.370	−3965.49	0.089	0.582
Polemoniaceae	*matK*	0.442	−4808.33	0.442	0.443	−4808.33	0.416	0.480	−4808.04	0.993	0.444
Pontederiaceae	*rbcL*	0.050	−2657.25	0.053	0.048	−2657.23	0.032	0.099	−2654.62	0.840	**0.022**
Primula	*matK*	0.402	−13 132.22	0.386	0.417	−13 131.95	0.41	0.381	−13 132.02	0.464	0.529
Primula	*rbcL*	0.118	−4320.65	0.051	0.160	−4312.04	0.103	0.170	−4318.72	**< 0.0001**	**0.050**
Psychotria	*rbcL*	0.219	−1018.48	0.120	0.368	−1017.00	0.176	0.462	−1017.60	0.085	0.185
Solanaceae	*matK*	0.393	−4737.78	0.360	0.417	−4737.23	0.392	0.395	−4740.07	0.296	1.000
Solanaceae	*rbcL*	0.181	−1284.46	0.000	0.231	−1281.18	0.133	0.292	−1283.08	**0.010**	0.097
Triticeae	*matK*	0.204	−2950.12	0.096	0.311	−2945.44	0.183	0.236	−2949.86	**0.002**	0.469
Veronica	*rbcL*	0.099	−2393.39	0.031	0.165	−2390.57	0.103	0.096	−2393.38	**0.017**	0.886
Combined			−62 895.08			−62 863.21			−62 885.84	**< 0.0001**	0.297

Taxonomic group	Gene	M_3				M_4					M_3 vs. $M_{int\text{-}ext}$	M_4 vs. M_3
		ω_{int}	$\omega_{ext\text{-}out}$	$\omega_{ext\text{-}self}$	lnL	$\omega_{int\text{-}out}$	$\omega_{int\text{-}self}$	$\omega_{ext\text{-}out}$	$\omega_{ext\text{-}self}$	lnL		
Asteraceae	*matK*	0.313	0.439	0.525	−5463.88	0.439	0.529	0.319	0.268	−5463.81	0.461	0.710
Asteraceae	*rbcL*	0.098	0.139	0.167	−3618.39						0.647	
Geraniaceae	*rbcL*	0.037	0.077	0.075	−4950.60						0.934	
Medicago	*matK*	0.233	0.268	0.423	−3963.72	0.268	0.423	*	0.224	−3963.12	0.328	0.276
Polemoniaceae	*matK*	0.442	0.390	0.474	−4808.06	0.390	0.474	0.429	0.506	−4807.96	0.460	0.649
Pontederiaceae	*rbcL*	0.053	0.013	0.099	−2652.73						**0.003**	
Primula	*matK*	0.386	0.422	0.403	−13 131.91	0.398	0.352	0.422	0.403	−13 131.67	0.089	0.491
Primula	*rbcL*	0.051	0.154	0.171	−4311.96						0.703	
Psychotria	*rbcL*	0.120	0.305	0.462	−1016.88						0.633	
Solanaceae	*matK*	0.348	0.326	0.480	−4738.85	0.326	0.480	0.439	0.198	−4736.57	1.000	0.033
Solanaceae	*rbcL*	0.000	0.180	0.331	−1280.42	0.180	0.309	0.000	0.144	−1280.33	0.218	0.665
Triticeae	*matK*	0.096	0.395	0.254	−2944.98	0.398	0.253	0.049	0.186	−2943.65	0.338	0.102
Veronica	*rbcL*	0.050	0.248	0.129	−2390.55	0.248	0.130	0.033	0.066	−2390.27	0.828	0.459
Combined†					−62 857.46						0.777	
					−34 758.60					−34 756.10		0.758

†For species without codons under positive selection, values of Table 3 were used to compute the combined likelihoods. *Undefined ω ratio.

Finally, we found neither individual nor general effects of mating systems on internal branches (Tables 3 and 4). Overall, we noted a general trend towards a higher ω in SELF branches, but the statistical support was weak.

Simulations

The simulation results were very similar between the topologies tested (Fig. 3), so we only present the results for the first topology with only SELF terminal branches (Fig. 4). The alternative topology was slightly more favourable for the detection of relaxed selection in SELF lineages (not shown). As expected, the power to detect differences between SELF and OUT branches increased with the total tree length. Throughout the different simulations, the average number of substitutions was 59 for the total tree length = 0.5, 118 for the total tree length = 1 and 595 for the total tree length = 5. Most of the real data sets were thus encompassed between the grey and black dots in Fig. 4 (see Table 2 for substitution values). Note, however, that this power was reduced for the comparison of external branches (M_3 vs. $M_{int\text{-}ext}$) due to the reduction in the total number of substitutions. As expected, the power also increased with the difference between ω_{SELF} and ω_{OUT}. However, the power seemed to scale with the $r_\omega = \omega_{SELF}/\omega_{OUT}$ ratio (as plotted in Fig. 4) rather than with the $\Delta\omega = \omega_{SELF} - \omega_{OUT}$ difference. For the same $\Delta\omega$, differences were thus more easily detected for highly constrained genes with low ω_{OUT} (see details in Table S3). These simulations showed that finding the expected pattern in 13 data sets of 16 (Fig. 4a) and a significant effect in two of 16 (Fig. 4b) was thus compatible with a moderate effect of selfing, with an r_ω ratio roughly between 1.2 and 1.5.

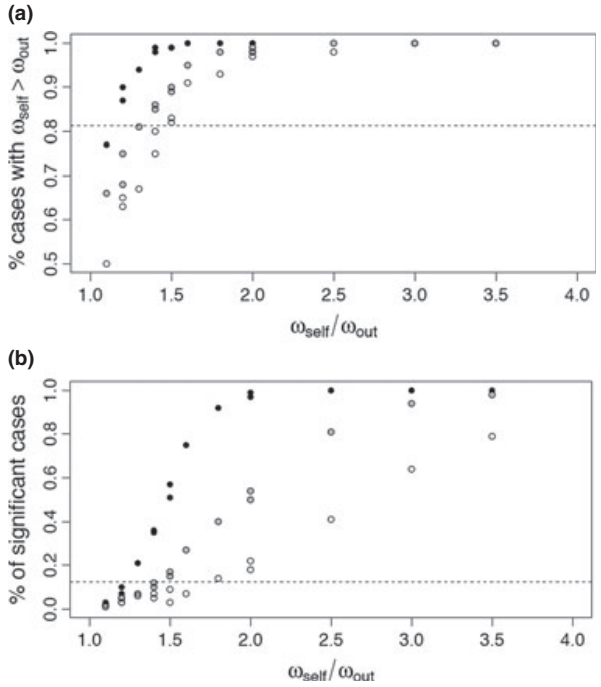


Fig. 4 Summary of the simulation results. Proportion of simulations with $\omega_{SELF} > \omega_{OUT}$ (a) and proportion of simulations with significant differences at the 5% level (b) as a function of the $r_\omega = \omega_{SELF}/\omega_{OUT}$ ratio. Dashed lines correspond to the observed proportions in the real data sets: 13/16 (a) and 2/16 (b). Total branch length = 0.5 (white symbols), 1 (grey symbols) and 5 (black symbols). As the results mainly depend on r_ω, but not directly on $\omega_{OUT} = 0.1$, the results for the different ω_{OUT} were pooled to enhance the legibility of the figure.

To what variation in N_e – that is, to what α in eqn (1) – does a given r_ω ratio correspond to? Assuming the distribution of fitness effects of deleterious mutations (DFEM) follows a gamma distribution with mean $\gamma = Ns$ and shape β, we can show that (see Appendix 2):

$$r_\omega = \frac{\omega_{SELF}}{\omega_{OUT}} \approx \alpha^{-\beta} \quad (3)$$

For instance, for $\beta = 0.25$, $r_\omega = 1.5$ corresponds to $\alpha = 0.2$ if selfing occurred at the same time as speciation. However, if we assume that selfing only appeared on the last f fraction of the branches, we simply have:

$$r_\omega \approx 1 - f + f\alpha^{-\beta} \quad (4)$$

For instance, for $\beta = 0.25$ and $\alpha = 0.2$ as above, r_ω drops to 1.1 if selfing actually corresponded to only 20% of the end of SELF branches, as could be the case in *Arabidopsis thaliana* (Tang *et al.*, 2007; but see Bechsgaard *et al.*, 2006; for more recent estimated origin). More generally, Fig. 5 shows the α or f values needed to get various r_ω ratios as a function of β. As β decreases, that is, the DFEM becomes more leptokurtic, a more marked reduction in N_e (lower α) or more ancient transitions (higher f) are necessary to detect differences in ω between SELF and OUT branches. This is because most mutations have very small effects and behave almost neutrally in both selfing and outcrossing species.

Discussion

Weak signature of relaxed selection in selfers

We compiled 19 chloroplastic data sets (16 of which provided enough information to infer realistic ω values) in 13 angiosperm groups. Only a few (one to three, depending on the analysis) data sets showed significant evidence of elevated ω in SELF compared to OUT branches. By combining data sets, a general trend emerged, but it was partly due to the fact that SELF

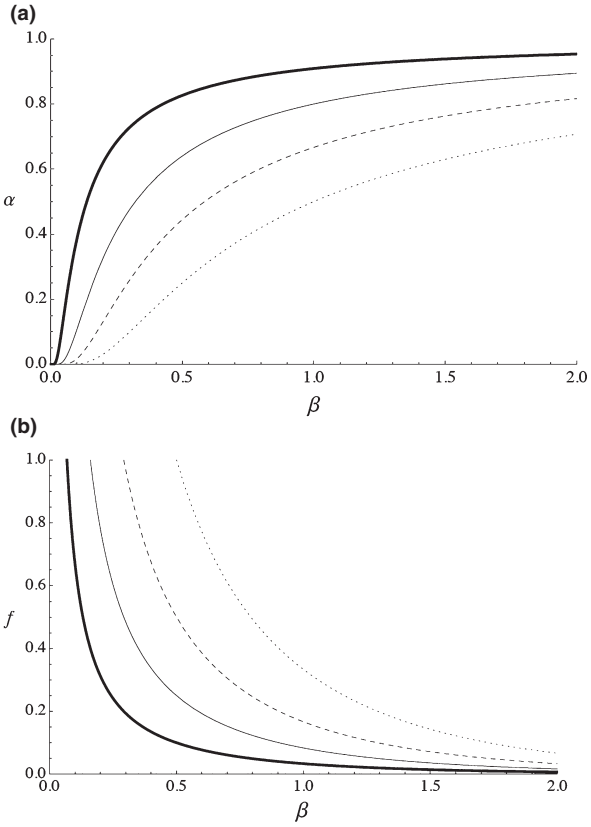


Fig. 5 Reduction in N_e (α) (a) and percentage of the branch affected by selfing (f) (b) needed to obtain an $r_\omega = \omega_{SELF}/\omega_{OUT}$ ratio of 1.1 (bold lines), 1.25 (thin lines), 1.5 (dashed lines) and 2 (dotted lines) as a function of β. In (a), f was set at 1, and in (b), α was set at 0.25.

branches were mostly terminal. Overall, we found weak statistical support for increased ω in SELF branches. However, the simulation results suggested that these results were compatible with true weak-to-moderate differences in ω (Fig. 4).

Previous studies using similar divergence analyses also found a nil or weak signature of relaxed selection in selfers, whereas relaxed selection seemed to be more easily detected using polymorphism data (see references in Table 1). This could simply be due to the fact that many other factors can mask the relationship between mating systems and ω in large-scale phylogenetic analyses. Here, we also used proxies of selfing rates, such as self-compatibility and homostyly. Self-compatible species are likely on the way to high selfing but could transiently have intermediate selfing rates (Igic *et al.*, 2008; Igic & Busch, 2013). Under mixed mating, reduced selection efficacy is not expected to be very strong, unless bottleneck effects predominate. However, the signature of relaxed selection was not lower for these data sets than for data sets with 'true' selfing species. As we analysed chloroplastic genes, the expected reduction in N_e could be a bit lower than for nuclear genes in species with few chromosomes and/or a short nuclear genetic map, as discussed in the introduction. This could also contribute to the weak signature of relaxed selection. For instance, the selfer *Arabidopsis thaliana* has a small genome with a rather short genetic map (five chromosomes of average length 100 cM, which roughly corresponds to the thin black lines in Fig. 1). In this species, the reduction in effective population size as compared to the outcrosser *A. lyrata* was found to be less severe in chloroplastic than in nuclear genes, in agreement with lower background selection affecting chloroplast (Wright *et al.*, 2008). Chromosome numbers are higher in most angiosperm species (including those studied here), and the differences between chloroplast and nuclear genes could be rather low. However, the difference also depends on the rate and effect of deleterious mutations (Fig. 1). A better characterization of both genetic maps and the distribution of mutational effects would be needed to evaluate the respective impact of background selection on chloroplastic and nuclear genes. It would also be interesting to confirm (or not) our results using nuclear data.

The lack of any sign of relaxed selection has also been interpreted as evidence of a recent origin of selfing lineages. Our results seem to support this, but we also found that this interpretation closely depends on the underlying distribution of fitness effects of mutations (DFEM) affecting the genes used in the Dn/Ds analyses (Eqns 3 and 4). Figure 5 shows that the interpretation holds only if this distribution is not too leptokurtic (β not too small). We do not know the β parameters for the *matK* and *rbcL* genes in the different species groups we used. However, several genome-wide estimates based on sequence polymorphism data suggest that β is rather small (e.g. 0.23 in humans, see Eyre-Walker *et al.*, 2006; between 0.08 and 0.21 in several plant species T. Gossmann pers. comm. and Gossmann *et al.*, 2010). With such low β values, a marked N_e reduction in selfers (small α) over a long time (f close to 1 in eqn 4) will only lead to slight differences in ω, as observed in the data sets. A more ancient selfing origin, that is, from the beginning of the branches leading to current selfing species, would be in line with the recent finding that a change in mating systems more frequently occurs in association with speciation events ('cladogenetic' mode of change) than within lineages ('anagenetic' mode) (Goldberg & Igic, 2012).

Several biological factors have been proposed to explain the weaker signature of selfing than asexuality in ω analyses (Glémin & Galtier, 2012). Differences in the shape of the DFEM of genes used in the different analyses (mostly mitochondrial in asexual animals vs. mostly nuclear or chloroplastic in selfing plants) could also play a role. If possible, choosing genes with high β should facilitate the detection of relaxed selection caused by a reduction in population size.

Selfing evolution implications

Based on the assumptions underlying eqns (3) and (4), our results are thus compatible with either small β and possibly a strong reduction in N_e over a long time or a higher β and a recent origin of selfing lineages. But does β matter? Considering the risk of population extinction due to the accumulation of deleterious mutations, and assuming that mutation effects have a gamma distribution, Lande (1994) showed that the mean time to extinction is approximately proportional to $N_e^{\beta+1}$ for $\gamma > 1$ (his equation 10). A reduction in N_e by an α factor thus reduces the mean time to extinction by an $\alpha^{\beta+1}$ factor. If β is weak, a reduction in N_e reduces the mean time to extinction almost linearly. For higher β, a reduction in N_e has a more drastic effect on the mean time to extinction. For example, for $\beta = 1$ (i.e. an exponential distribution of deleterious effects), reducing the effective population size by two-fold reduces the mean time to extinction by four-fold. If β is small, our results are compatible with a marked reduction in N_e in selfers, but the impact on extinction through mutational meltdown may be rather weak. If β is higher, deleterious mutations may have a stronger impact on species extinction, but they likely accumulate for a much shorter time. Overall, although the reasons are maybe more complex than previously thought (Glémin & Galtier, 2012), this suggests that deleterious mutations may not be the main cause of extinction in selfing species.

However, understanding the causes of higher extinction rates in selfing species is still a difficult task, and

the possible role of deleterious mutations should be tested more quantitatively. Our results suggest that better characterization of DFEM and mutation rates in selfing and outcrossing species could help in addressing this issue. Moreover, theoretical work is still needed to test whether the accumulation of deleterious mutations could be sufficient to cause species extinction but without leaving a strong molecular signature. Finally, it would also be crucial to clarify the demographic consequences of the mutation load in an ecological context (Agrawal & Whitlock, 2012) to resolve this question.

Acknowledgements

We thank Stephen Wright and an anonymous reviewer for their helpful comments and suggestions on a previous version of the manuscript. We declare no conflict of interest. This publication is the contribution ISEM 2014-013 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554 – CNRS). This work was supported by the French Centre National de la Recherche Scientifique and Agence Nationale de la Recherche (ANR-11-BSV7-013-03).

Authors' contribution

SG designed the project. AM and SG built the data sets and analysed data. SG developed theoretical predictions. SG and AM wrote the manuscript.

References

Abramowitz, M. & Stegun, I.A. 1970. *Handbook of Mathematical Functions*, 9th edn. Dover, New York.

Agrawal, A.F. & Lively, C.M. 2001. Parasites and the evolution of self-fertilization. *Evolution* **55**: 869–879.

Agrawal, A.F. & Whitlock, M.C. 2012. Mutation load: the fitness of individuals in populations where deleterious alleles are abundant. *Annu. Rev. Ecol. Evol. Syst.* **43**: 115–135.

Armbruster, W.S., Perez-Barrales, R., Arroyo, J., Edwards, M.E. & Vargas, P. 2006. Three-dimensional reciprocity of floral morphs in wild flax (*Linum suffruticosum*): a new twist on heterostyly. *New Phytol.* **171**: 581–590.

Barraclough, T.G., Fontaneto, D., Ricci, C. & Herniou, E.A. 2007. Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. *Mol. Biol. Evol.* **24**: 1952–1962.

Barrett, S.C., Harder, L.D. & Worley, A.C. 1996. The comparative biology of pollination and mating in flowering plants. *Philos. Trans. Roy. Soc. London B* **351**: 1271–1280.

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

Bena, G., Lejeune, B., Prosperi, J.M. & Olivieri, I. 1998. Molecular phylogenetic approach for studying life-history evolution: the ambiguous example of the genus *Medicago* L. *Proc. Biol. Sci.* **265**: 1141–1151.

Billiard, S., Lopez-Villavicencio, M., Devier, B., Hood, M.E., Fairhead, C. & Giraud, T. 2011. Having sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating types. *Biol. Rev. Camb. Philos. Soc.* **86**: 421–442.

Brandvain, Y., Slotte, T., Hazzouri, K.M., Wright, S.I. & Coop, G. 2013. Genomic identification of founding haplotypes reveal the history of the selfing species *Capsella rubella*. *PLoS Genet.* **9**: e1003754.

Caballero, A. & Hill, W.G. 1992. Effects of partial inbreeding on fixation rates and variation of mutant genes. *Genetics* **131**: 493–507.

Charlesworth, B. 1992. Evolutionary rates in partially self-fertilizing species. *Am. Nat.* **140**: 126–148.

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

Cutter, A.D. & Payseur, B.A. 2003. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol. Biol. Evol.* **20**: 665–673.

Cutter, A.D., Wasmuth, J.D. & Washington, N.L. 2008. Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. *Genetics* **178**: 2093–2104.

Edgar, R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Escobar, J.S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J. *et al.* 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (poaceae). *Evolution* **64**: 2855–2872.

Eyre-Walker, A., Woolfit, M. & Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**: 891–900.

Ferrer, M.M. & Good-Avila, S.V. 2007. Macrophylogenetic analyses of the gain and loss of self-incompatibility in the Asteraceae. *New Phytol.* **173**: 401–414.

Fiz, O., Vargas, P., Alarcon, M., Aedo, C., Garcia, J.L. & Aldasoro, J.J. 2008. Phylogeny and historical biogeography of Geraniaceae in relation to climate changes and pollination ecology. *Syst. Bot.* **33**: 326–342.

Foxe, J.P., Dar, V.U., Zheng, H., Nordborg, M., Gaut, B.S. & Wright, S.I. 2008. Selection on amino acid substitutions in Arabidopsis. *Mol. Biol. Evol.* **25**: 1375–1383.

Gioti, A., Mushegian, A.A., Strandberg, R., Stajich, J.E. & Johannesson, H. 2012. Unidirectional evolutionary transitions in fungal mating systems and the role of transposable elements. *Mol. Biol. Evol.* **29**: 3215–3226.

Gioti, A., Stajich, J. & Johannesson, H. 2013. *Neurospora* and the dead-end hypothesis: genomic consequences of selfing in the model genus. *Evolution* **67**: 3600–3616.

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

Glémin, S. & Galtier, N. 2012. Genome evolution in outcrossing versus selfing versus asexual species. *Methods Mol. Biol.* **855**: 311–335.

Glémin, S. & Ronfort, J. 2013. Adaptation and maladaptation in selfing and outcrossing species: new mutations versus standing variation. *Evolution* **67**: 225–240.

Glémin, S., Bazin, E. & Charlesworth, D. 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc. Biol. Sci.* **273**: 3011–3019.

Goldberg, E.E. & Igic, B. 2012. Tempo and mode in plant breeding system evolution. *Evolution* **66**: 3701–3709.

Goldberg, E.E., Kohn, J.R., Lande, R., Robertson, K.A., Smith, S.A. & Igic, B. 2010. Species selection maintains self-incompatibility. *Science* **330**: 493–495.

Gossmann, T.I., Song, B.H., Windsor, A.J., Mitchell-Olds, T., Dixon, C.J., Kapralov, M.V. *et al.* 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.* **27**: 1822–1832.

Haudry, A., Cenci, A., Guilhaumon, C., Paux, E., Poirier, S., Santoni, S. *et al.* 2008. Mating system and recombination affect molecular evolution in four Triticeae species. *Genet. Res.* **90**: 97–109.

Hazzouri, K.M., Escobar, J.S., Ness, R.W., Killian Newman, L., Randle, A.M., Kalisz, S. *et al.* 2013. Comparative population genomics in *Collinsia* sister species reveals evidence for reduced effective population size, relaxed selection, and evolution of biased gene conversion with an ongoing mating system shift. *Evolution* **67**: 1263–1278.

Henry, L., Schwander, T. & Crespi, B.J. 2012. Deleterious mutation accumulation in asexual *Timema* stick insects. *Mol. Biol. Evol.* **29**: 401–408.

Hersch-Green, E.I., Myburg, H. & Johnson, M.T. 2012. Adaptive molecular evolution of a defence gene in sexual but not functionally asexual evening primroses. *J. Evol. Biol.* **25**: 1576–1586.

Hudson, R.R. & Kaplan, N.L. 1995. Deleterious background selection with recombination. *Genetics* **141**: 1605–1617.

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

Igic, B., Bohs, L. & Kohn, J.R. 2006. Ancient polymorphism reveals unidirectional breeding system shifts. *Proc. Natl. Acad. Sci. USA* **103**: 1359–1363.

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

Ingvarsson, P.K. 2002. A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. *Evolution: Int. J. Org. Evolution* **56**: 2368–2373.

Jarne, P. & Auld, J.R. 2006. Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. *Evolution: Int. J. Org. Evolution* **60**: 1816–1824.

Johnson, S.G. & Howard, R.S. 2007. Contrasting patterns of synonymous and nonsynonymous sequence evolution in asexual and sexual freshwater snail lineages. *Evolution: Int. J. Org. Evolution* **61**: 2728–2735.

Kapralov, M.V. & Filatov, D.A. 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol. Biol.* **7**: 73.

Kissling, J. & Barrett, S.C. 2013. Variation and evolution of herkogamy in Exochaenium (Gentianaceae): implications for the evolution of distyly. *Ann. Bot.* **112**: 95–102.

Kohn, J.R., Graham, S.W., Morton, B., Doyle, J.J. & Barrett, S.C.H. 1996. Reconstruction of the evolution of reproductive characters in Pontederiaceae using phylogenetic evidence from chloroplast DNA restriction-site variation. *Evolution* **50**: 1454–1469.

Lande, R. 1994. Risk of population extinction from fixation of new deleterious mutations. *Evolution* **48**: 1460–1469.

Lynch, M., Conery, J. & Bürger, R. 1995. Mutational meltdowns in sexual populations. *Evolution* **49**: 1067–1080.

Mark Welch, D.B. & Meselson, M.S. 2001. Rates of nucleotide substitution in sexual and anciently asexual rotifers. *Proc. Natl. Acad. Sci. USA* **98**: 6720–6724.

Mast, A.R., Kelso, S. & Conti, E. 2006. Are any primroses (Primula) primitively monomorphic? *New Phytol.* **171**: 605–616.

Maynard-Smith, J. & Haigh, D. 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**: 23–35.

Morran, L.T., Schmidt, O.G., Gelarden, I.A., Parrish, R.C. & Lively, C.M. 2011. Running with the Red Queen: host-parasite coevolution selects for biparental sex. *Science* **333**: 216–218.

Muller, K. & Albach, D.C. 2010. Evolutionary rates in *Veronica* L. (Plantaginaceae): disentangling the influence of life history and breeding system. *J. Mol. Evol.* **70**: 44–56.

Neiman, M., Hehman, G., Miller, J.T., Logsdon, J.M. Jr & Taylor, D.R. 2010. Accelerated mutation accumulation in asexual lineages of a freshwater snail. *Mol. Biol. Evol.* **27**: 954–963.

Ness, R.W., Siol, M. & Barrett, S.C. 2012. Genomic consequences of transitions from cross- to self-fertilization on the efficacy of selection in three independently derived selfing plants. *BMC Genomics* **13**: 611.

Nordborg, M.D.P. 1997. The coalescent process with selfing. *Genetics* **146**: 1185–1195.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**: 923–929.

Nygren, K., Strandberg, R., Wallberg, A., Nabholz, B., Gustafsson, T., Garcia, D. *et al.* 2011. A comprehensive phylogeny of Neurospora reveals a link between reproductive mode and molecular evolution in fungi. *Mol. Phylogenet. Evol.* **59**: 649–663.

Ollivier, M., Gabaldon, T., Poulain, J., Gavory, F., Leterme, N., Gauthier, J.P. *et al.* 2012. Comparison of gene repertoires and patterns of evolutionary rates in eight aphid species that differ by reproductive mode. *Genome Biol. Evol.* **4**: 155–167.

Paland, S. & Lynch, M. 2006. Transitions to asexuality result in excess amino acid substitutions. *Science* **311**: 990–992.

Pollak, E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics* **117**: 353–360.

Pollak, E. & Sabran, M. 1992. On the theory of partially inbreeding finite populations. III. Fixation probabilities under partial selfing when heterozygotes are intermediate in viability. *Genetics* **131**: 979–985.

Qiu, S., Zeng, K., Slotte, T., Wright, S. & Charlesworth, D. 2011. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. *Genome Biol. Evol.* **3**: 868–880.

Sakai, S. & Wright, S.J. 2008. Reproductive ecology of 21 coexisting Psychotria species (Rubiaceae): when is heterostyly lost? *Biol. J. Linn. Soc.* **93**: 125–134.

Sakai, A.K., Weller, S.G., Wagner, W.L., Nepokroeff, M. & Culley, T.M. 2006. Adaptive radiation and evolution of breeding systems in Schiedea (Caryophyllaceae), an endemic *Hawaiian* genus. *Ann. Mo. Bot. Gard.* **93**: 49–63.

Schoen, D.J. & Brown, A.H.D. 1991. Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proc. Natl. Acad. Sci. USA* **88**: 4494–4497.

Schultz, S.T. & Lynch, M. 1997. Mutation and extinction: the role of variable mutational effects, synergistic epistasis, beneficial mutations, and degree of outcrossing. *Evolution* **51**: 1363–1371.

Slotte, T., Foxe, J.P., Hazzouri, K.M. & Wright, S.I. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Mol. Biol. Evol.* **27**: 1813–1821.

Slotte, T., Hazzouri, K.M., Agren, J.A., Koenig, D., Maumus, F., Guo, Y.L. *et al.* 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**: 831–835.

Stebbins, G.L. 1957. Self fertilization and population variability in higher plants. *Am. Nat.* **91**: 337–354.

Takebayashi, N. & Morrell, P.L. 2001. Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. *Am. J. Bot.* **88**: 1143–1150.

Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y.L., Hu, T.T. *et al.* 2007. The evolution of selfing in *Arabidopsis thaliana*. *Science* **317**: 1070–1072.

Tucker, A.E., Ackerman, M.S., Eads, B.D., Xu, S. & Lynch, M. 2013. Population-genomic insights into the evolutionary origin and fate of obligately asexual *Daphnia pulex*. *Proc. Natl. Acad. Sci. USA*, **39**: 15740–15745.

Welch, J.J., Eyre-Walker, A. & Waxman, D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *J. Mol. Evol.* **67**: 418–426.

Whittle, C.A., Sun, Y. & Johannesson, H. 2011. Evolution of synonymous codon usage in *Neurospora tetrasperma* and *Neurospora discreta*. *Genome Biol. Evol.* **3**: 332–343.

Wright, S.I., Lauga, B. & Charlesworth, D. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Mol. Biol. Evol.* **19**: 1407–1420.

Wright, S.I., Iorgovan, G., Misra, S. & Mokhtari, M. 2007. Neutral evolution of synonymous base composition in the Brassicaceae. *J. Mol. Evol.* **64**: 136–141.

Wright, S.I., Nano, N., Foxe, J.P. & Dar, V.U. 2008. Effective population size and tests of neutrality at cytoplasmic genes in *Arabidopsis*. *Genet. Res. (Camb)* **90**: 119–128.

Wright, S.I., Kalisz, S. & Slotte, T. 2013. Evolutionary consequences of self-fertilization in plants. *Proc. Biol. Sci.* **280**: 20130133.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**: 1586–1591.

Appendix 1
Effect of background selection on nuclear and chloroplastic genes

In this JEB issue, Kamran-Disfani and Agrawal obtained an expression for the reduction in N_e caused by background selection in a partially fertilizing population, which is more accurate than the approximated one obtained previously (Cutter & Payseur, 2003; Glémin, 2007; Glémin & Ronfort, 2013). Consider a chromosome of size C and a focal gene located at position z and assuming the same deleterious mutation rates, u, dominance, h, and selection effect, s, for all selected genes on the chromosome, they found that:

$$\rho_{\text{linked}}(F,C,z)=\text{Exp}\left[-\int_0^{Cz}\frac{u(h+F-hF)s}{2((h+F-hF)s+(1-F)R(x))^2}\text{d}x-\int_0^{C(1-z)}\frac{u(h+F-hF)s}{2((h+F-hF)s+R(1-F)(x))^2}\text{d}x\right] \quad \text{(A1.1)}$$

where F is Wright's fixation index and $R(x) = \frac{1}{2}(1 - \text{Exp}[-2rx])$ is Haldane's mapping function, with r being the recombination rate between two adjacent sites. Kamran-Disfani and Agrawal gave the exact analytical expression for (A1.1) in supplementary material, but the equation is unwieldy and not reproduced here. This result, which is based on the findings of Hudson & Kaplan (1995), can be easily extended to the effect of deleterious alleles segregating at loci not physically linked to the focal gene, simply using $R(x) = 1/2$ and integrating between 0 and $G - C$, where G is the total genome size. $G - C$ thus corresponds to the other chromosome(s) of the genome.

$$\rho_{\text{unlinked}}(F,G-C)=\text{Exp}\left[-\int_0^{G-C}\frac{u(h+F-hF)s}{2((h+F-hF)s+(1-F)/2)^2}\text{d}x\right]=\text{Exp}\left[-\frac{u(G-C)(h+F-hF)s}{2((h+F-hF)s+(1-F)/2)^2}\right] \quad \text{(A1.2)}$$

The total effect of background selection is thus as follows:

$$\rho_{\text{nuc}}(F,G,C,z)=\rho_{\text{unlinked}}(F,G-C)\rho_{\text{linked}}(F,C,z) \quad \text{(A1.3)}$$

and the α term is given by as follows:

$$\alpha_{\text{nuc}}(F,G,C,z)=\frac{\rho_{\text{nuc}}(F,G,C,z)}{\rho_{\text{nuc}}(0,G,C,z)} \quad \text{(A1.4)}$$

For a chloroplastic gene, we can neglect the effect of other linked chloroplastic genes, so we simply have as follows:

$$\rho_{\text{chloro}}(F,G)=\rho_{\text{unlinked}}(F,G) \quad \text{(A1.5)}$$

and

$$\alpha_{\text{chloro}}(F,G)=\frac{\rho_{\text{chloro}}(F,G)}{\rho_{\text{chloro}}(0,G)} \quad \text{(A1.6)}$$

We can show that $\lim_{\text{Cr}\to\infty}\rho_{\text{linked}}(F,C,z)=\rho_{\text{unlinked}}(F,C)$, so when the chromosome genetic map is large, Cr >> 1, $\rho_{\text{nuc}}(F, G, C, z)$ tends towards $\rho_{\text{unlinked}}(F, G) = \rho_{\text{chloro}}(F, G)$. Similarly, when $G >> C$, $\rho_{\text{nuc}}(F, G, C, z) \simeq \rho_{\text{unlinked}}(F, G) = \rho_{\text{chloro}}(F, G)$. The additional reduction in N_e due to background selection in selfing species, $\alpha(F)$, is thus similar for chloroplastic and nuclear genes except for genomes with few chromosomes and a short genetic map. This is illustrated in Fig. 1.

Appendix 2
Derivation of eqns (3) and (4)

Assuming a gamma distribution of deleterious effects of mutations with mean $\gamma = Ns$ and shape β, Welch *et al.* (2008) showed that ω can be well approximated by:

$$\omega(\gamma, \beta) \approx \beta^{1+\beta}\gamma^{-\beta}\zeta(1+\beta) \quad \text{(A2.1)}$$

where ζ is the Riemann zeta function (Abramowitz & Stegun, 1970). Here, $\gamma = Ns$ and not $\gamma = 4Ns$ as in Welch *et al.* (2008) because the chloroplastic genome is haploid and uniparentally transmitted. If we assume that the shape β is the same between outcrossing and selfing species, we simply have as follows:

$$r_\omega = \frac{\omega_{\text{SELF}}}{\omega_{\text{OUT}}} \approx \frac{\omega(\alpha\gamma, \beta)}{\omega(\gamma, \beta)} = \alpha^{-\beta} \quad \text{(A2.2)}$$

hence eqns (3) and (4). For a fixed r_ω ratio, we thus have the following relationship between α and β as plotted in Fig. 5a:

$$\alpha = r_\omega^{-1/\beta} \quad \text{(A2.3)}$$

and between f and β, as plotted in Fig. 5b:

$$f = \frac{\alpha^\beta(1 - r_\omega)}{\alpha^\beta - 1} \quad \text{(A2.4)}$$

Supporting information

Additional Supporting Information may be found in the online version of this article:
Table S1 List of species and Genbank accessions.
Table S2 Details results of the comparison between M7 and M8 models in codeml to detect codons under positive selection.
Table S3 Details of simulation results.

Data deposited at Dryad: doi:10.5061/dryad.618fv

Received 1 October 2013; revised 10 February 2014; accepted 12 February 2014

A.1 Supporting information

Table S1 Available online: https://drive.google.com/file/d/0B7KhiXOz6_OLMHRsM0x1TnFpcXc/view?usp=sharing

Table A.S2: Details results of the comparison between M7 and M8 models in codeml to detect codons under positive selection.

Taxonomic group	Gene	M7 lnL	M8 lnL	P-value
Asteraceae	matK	-5487.62588	-5480.141886	5.62E-04
Asteraceae	rbcL	-1418.416906	-1413.571591	7.87E-03
Exochaenium	matK	-857.349831	-855.898292	2.34E-01
Geraniaceae	rbcL	-5299.8346	-5280.601703	4.44E-09
Linum	rbcL	-2565.742418	-2562.919713	5.94E-02
Medicago	matK	-4143.361266	-4126.658996	5.58E-08
Polemoniaceae	matK	-5249.181551	-5220.193146	2.57E-13
Pontederiaceae	rbcL	-2787.524902	-2772.028887	1.86E-07
Primula	matK	-13897.36161	-13849.02473	1.00E-21
Primula	rbcL	-4669.227351	-4647.171095	2.64E-10
Psychotria	matK	-2012.21321	-2012.212798	1.00E+00
Psychotria	rbcL	-1168.514347	-1150.313839	1.25E-08
Schiedea	matK	-1618.217983	-1608.692223	7.29E-05
Schiedea	rbcL	-2065.735222	-2038.368321	1.30E-12
Solanaceae	matK	-5083.315706	-5045.596761	4.16E-17
Solanaceae	rbcL	-1407.062171	-1381.899579	1.18E-11
Triticeae	matK	-3051.029828	-3046.732596	1.36E-02
Veronica	matK	-2975.763824	-2972.991865	6.25E-02
Veronica	rbcL	-2420.71157	-2418.361929	9.54E-02

Table A.S3: Details of simulation results.

Tree_length	textbfW_out	textbfDelta_W	textbf#Simul_Wself>Wout	textbf#Significant_Simul	textbf#Total_Simul	textbfMean_dS	textbfMean_dN
0,1	0,1	0,025	46	2	99	0,1085777778	0,0120363636
0,1	0,1	0,05	59	3	99	0,1070141414	0,0125414141
0,1	0,1	0,1	69	5	99	0,1031484848	0,0134151515
0,1	0,1	0,15	82	7	99	0,1005808081	0,0140858586
0,1	0,1	0,2	85	13	99	0,0987818182	0,0145424242
0,1	0,1	0,25	86	16	97	0,0964979381	0,0150453608
0,1	0,25	0,025	45	1	99	0,0750858586	0,0208959596
0,1	0,25	0,05	49	1	99	0,0742121212	0,0211464646
0,1	0,25	0,1	58	2	97	0,0722443299	0,021643299
0,1	0,25	0,15	70	3	96	0,0710604167	0,0220833333
0,1	0,25	0,2	71	3	94	0,0703308511	0,0222531915
0,1	0,25	0,25	70	2	90	0,06866	0,0225811111
0,1	0,5	0,025	41	0	91	0,0506175824	0,0277703297
0,1	0,5	0,05	40	0	90	0,0504344444	0,0278722222
0,1	0,5	0,1	41	0	89	0,0495325843	0,0280179775
0,1	0,5	0,15	40	0	86	0,0495023256	0,0282104651
0,1	0,5	0,2	44	0	84	0,0489571429	0,0283357143
0,1	0,5	0,25	47	0	84	0,0483130952	0,0285821429
0,5	0,1	0,025	65	1	100	0,563444	0,058806
0,5	0,1	0,05	83	3	100	0,554692	0,061487
0,5	0,1	0,1	97	18	100	0,539708	0,065711
0,5	0,1	0,15	98	41	100	0,527106	0,068887
0,5	0,1	0,2	100	64	100	0,516168	0,0717
0,5	0,1	0,25	100	79	100	0,50799	0,07392
0,5	0,25	0,025	51	3	100	0,402372	0,103869
0,5	0,25	0,05	65	3	100	0,396752	0,105295
0,5	0,25	0,1	80	5	100	0,388929	0,107479
0,5	0,25	0,15	91	7	100	0,382498	0,109374
0,5	0,25	0,2	93	14	100	0,376198	0,110867
0,5	0,25	0,25	98	22	100	0,37117	0,112381
0,5	0,5	0,025	47	2	100	0,274367	0,137956
0,5	0,5	0,05	50	2	100	0,27185	0,138613
0,5	0,5	0,1	63	3	100	0,267317	0,139697
0,5	0,5	0,15	67	6	100	0,263684	0,140617
0,5	0,5	0,2	75	7	100	0,259951	0,141345
0,5	0,5	0,25	82	9	100	0,257552	0,141948
1	0,1	0,025	74	5	100	1,116689	0,119712
1	0,1	0,05	90	15	100	1,100463	0,124268
1	0,1	0,1	99	50	100	1,070118	0,132197
1	0,1	0,15	100	81	100	1,047522	0,138809
1	0,1	0,2	100	94	100	1,031117	0,144077
1	0,1	0,25	100	98	100	1,01483	0,148743
1	0,25	0,025	58	5	100	0,804263	0,207002
1	0,25	0,05	68	5	100	0,796467	0,209373
1	0,25	0,1	86	12	100	0,78051	0,213896
1	0,25	0,15	95	27	100	0,766947	0,217983
1	0,25	0,2	98	40	100	0,753489	0,221143
1	0,25	0,25	98	54	100	0,742312	0,223774
1	0,5	0,025	63	1	100	0,547412	0,276608
1	0,5	0,05	66	1	100	0,543254	0,277787
1	0,5	0,1	75	5	100	0,534282	0,279693
1	0,5	0,15	81	7	100	0,528247	0,282208
1	0,5	0,2	85	10	100	0,522141	0,283743
1	0,5	0,25	89	17	100	0,514859	0,285279
5	0,1	0,025	90	10	100	5,645174	0,601123
5	0,1	0,05	99	51	100	5,527279	0,625741
5	0,1	0,1	100	97	100	5,346986	0,666041
5	0,1	0,15	100	100	100	5,213256	0,699108
5	0,1	0,2	100	100	100	5,086669	0,726306
5	0,1	0,25	100	100	100	4,979777	0,750186
5	0,25	0,025	67	3	100	4,060843	1,044296
5	0,25	0,05	90	7	100	4,000682	1,057838
5	0,25	0,1	99	36	100	3,902733	1,082319
5	0,25	0,15	100	75	100	3,825448	1,101048
5	0,25	0,2	100	92	100	3,753236	1,119086
5	0,25	0,25	100	99	100	3,695753	1,134388
5	0,5	0,025	65	0	100	2,770487	1,404112
5	0,5	0,05	77	3	100	2,742296	1,410669
5	0,5	0,1	87	10	100	2,698703	1,421683
5	0,5	0,15	94	21	100	2,661617	1,431316
5	0,5	0,2	98	35	100	2,624176	1,441338
5	0,5	0,25	99	57	100	2,592071	1,449067

Appendix B

Rapid de novo evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes

This chapter is the result of a Master internship I did from February to June 2011 under the supervision of Gabriel Marais (Lyon, France). It aimed at testing whether dosage compensation exists in *Silene latifolia*.

OPEN ACCESS Freely available online

PLoS BIOLOGY

Rapid De Novo Evolution of X Chromosome Dosage Compensation in *Silene latifolia*, a Plant with Young Sex Chromosomes

Aline Muyle[1☯], Niklaus Zemp[2☯], Clothilde Deschamps[3], Sylvain Mousset[1], Alex Widmer[2¶*], Gabriel A. B. Marais[1¶*]

1 Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France, **2** Institute of Integrative Biology (IBZ), ETH Zurich, Zürich, Switzerland, **3** Pôle Rhône-Alpes de Bioinformatique (PRABI), Villeurbanne, France

Abstract

Silene latifolia is a dioecious plant with heteromorphic sex chromosomes that have originated only ~10 million years ago and is a promising model organism to study sex chromosome evolution in plants. Previous work suggests that *S. latifolia* XY chromosomes have gradually stopped recombining and the Y chromosome is undergoing degeneration as in animal sex chromosomes. However, this work has been limited by the paucity of sex-linked genes available. Here, we used 35 Gb of RNA-seq data from multiple males (XY) and females (XX) of an *S. latifolia* inbred line to detect sex-linked SNPs and identified more than 1,700 sex-linked contigs (with X-linked and Y-linked alleles). Analyses using known sex-linked and autosomal genes, together with simulations indicate that these newly identified sex-linked contigs are reliable. Using read numbers, we then estimated expression levels of X-linked and Y-linked alleles in males and found an overall trend of reduced expression of Y-linked alleles, consistent with a widespread ongoing degeneration of the *S. latifolia* Y chromosome. By comparing expression intensities of X-linked alleles in males and females, we found that X-linked allele expression increases as Y-linked allele expression decreases in males, which makes expression of sex-linked contigs similar in both sexes. This phenomenon is known as dosage compensation and has so far only been observed in evolutionary old animal sex chromosome systems. Our results suggest that dosage compensation has evolved in plants and that it can quickly evolve de novo after the origin of sex chromosomes.

Citation: Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, et al. (2012) Rapid De Novo Evolution of X Chromosome Dosage Compensation in *Silene latifolia*, a Plant with Young Sex Chromosomes. PLoS Biol 10(4): e1001308. doi:10.1371/journal.pbio.1001308

Academic Editor: Detlef Weigel, Max Planck Institute for Developmental Biology, Germany

Received September 22, 2011; **Accepted** March 1, 2012; **Published** April 17, 2012

Funding: This study was supported by ETH Zurich and SNF grants (TH-07 06-3 and 31003A-116455) to A.W., and the work done in Lyon by Agence Nationale de la Recherche (ANR) to G.A.B.M. (grant number ANR-08-JCJC-0109). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gabriel.marais@univ-lyon1.fr (GABM); alex.widmer@env.ethz.ch (AW)

☯ These authors contributed equally to this work as first authors

¶ These authors contributed equally to this work as senior authors.

Introduction

In humans, where the evolution of sex chromosomes is probably best known, the XY chromosome pair was originally a recombining pair of autosomes that progressively stopped recombining, most likely because of a series of inversions on the Y chromosome [1–4]. This started ~150 million years ago [5,6] and the non-recombining human Y chromosome subsequently suffered from degenerating processes known as Hill-Robertson effects (inefficient selection and reduced polymorphism, see [7–9]), which explain the massive loss of Y genes (~97%) and the concomitant accumulation of DNA repeats on the non-recombining Y compared to the X chromosome and the still recombining pseudoautosomal regions (PARs) [2,3]. Even the few genes that persisted on the Y show signs of degeneration [10,11]. The classical view is that the massive loss of Y-linked genes has been balanced by the evolution of dosage compensation (equal dosage of X and autosomal transcripts in both males and females [12–14]), which is achieved by the inactivation of one X chromosome in females [15]. The question whether this three-step scenario (X–Y recombination suppression, Y degeneration, X dosage compensation) is similar for all species with sex chromosomes, in particular those with much younger sex chromosomes, has received much attention from evolutionary biologists, and several alternative model organisms to study the evolution of sex chromosomes have emerged, some of them very recently [9,16–18].

S. latifolia (white campion) is one such model organism. It is a dioecious plant from the Caryophyllaceae family with heteromorphic sex chromosomes that have originated only ~10 million years ago [19–22] and is a promising model organism to study sex chromosome evolution in plants [23,24]. Previous work suggests that *S. latifolia* XY chromosomes have stopped recombining gradually [21,22,25] and that the Y is undergoing degeneration (gene loss, reduced polymorphism, accumulation of repeats, maladapted proteins, reduced gene expression) as in animal sex chromosomes [26–34]. Despite these highly interesting results, work on sex chromosome evolution in *S. latifolia* has been limited by the slow pace of sex-linked gene identification (one to two new

Author Summary

The mammalian sex chromosomes originated from an ancestral pair of autosomes about 150 million years ago and the Y chromosome subsequently degenerated, losing most of its genes. During this process, a phenomenon called dosage compensation evolved to compensate for the gene loss on the Y chromosome and to equalize expression of X-linked genes in the two sexes. In humans, this is achieved by inactivating one of the two X chromosomes in females. Dosage compensation has also been reported in other animal XY systems such as fruit flies and worms, each 100 million years old or more. Here we studied dosage compensation in plants. We used high-throughput RNA sequencing in male and female *Silene latifolia* (white campion)—a dioecious plant whose XY chromosomes originated only about 10 million years ago—to identify hundreds of sex-linked genes. Analysis of their expression patterns in males and females revealed equal doses of sex-linked transcripts in both sexes, regardless of the degree of reduction of Y expression due to degeneration. Our results thus show that dosage compensation occurs in plants and is thus not an animal-specific phenomenon. They also reveal that proportionate dosage compensation can evolve rapidly de novo after the origin of sex chromosomes.

genes/year) [21,25,35–40]. This situation is now changing rapidly, thanks to next-generation sequencing (NGS) approaches, which are helping reveal the strong potential of the *S. latifolia* model [23,24,41–43].

Here we report a study using such an NGS approach, RNA-seq, applied to several males and females of an *S. latifolia* inbred line. Using a de novo assembly strategy followed by SNP analysis, we identified >1,700 sex-linked contigs, increasing by almost 100-fold the number of sex-linked sequences available until recently in *S. latifolia*. Studying these 1,700 sex-linked contigs, we found that expression of alleles on the Y is significantly reduced compared to those on the X chromosome, providing evidence for large-scale ongoing degeneration of the *S. latifolia* Y chromosome. By comparing the expression of X-linked alleles in males and females, which differ in the number of X chromosomes, we further found evidence of equal dosage of X transcripts among sexes for sex-linked genes showing Y degeneration, a phenomenon known as dosage compensation. To our knowledge, this is the first evidence for dosage compensation in plants and reveals that dosage compensation is not an animal-specific phenomenon. Moreover, the finding of dosage compensation in evolutionary young sex chromosomes has novel implications for the evolution of sex chromosomes because it shows that 10 million years are sufficient to evolve dosage compensation de novo. By contrast, dosage compensation in animals has to date been documented only in >100-million-year-old sex chromosome systems.

Results

Identification and Validation of New Sex-Linked Genes

We used RNA-seq—a next-generation transcriptome-sequencing approach—to identify new sex-linked genes and to study gene expression (find more details in Text S1). We obtained ~35 Gb of sequence data from three males and three females from a ten-generation inbred population of *S. latifolia* using Illumina technology (Table S1). Male and female reads were pooled and assembled de novo (see Material and Methods) (Figure S1), and we obtained 141,855 contigs (Table S2). From these, we identified sex-linked contigs using a segregation analysis similarly to [42,43] and found 1,736 contigs with at least one sex-linked SNP (Table S2). We tested the reliability of our inference of sex-linkage by first using known autosomal genes [44] to see whether sex-linked SNPs have been wrongly inferred for these, but could not find any for the ten autosomal genes tested (Table S3). This very low rate of false positives was confirmed when running our scripts to detect sex-linked SNPs on a set of simulated autosomal SNPs (Text S2). We thus concluded that our inferences of sex-linkage are highly reliable. To estimate how many sex-linked contigs we missed with our method, we checked how many of the previously identified sex-linked genes were among our sex-linked contigs (Table S3). 42% of these were not found, which means that our rate of false negatives is quite high, and we identified a subset (probably about half; see Figure S2; Text S1) of the sex-linked genes in *S. latifolia*. Many of our sex-linked contigs should be full-length transcripts as suggested by the size distribution plot (Figure S3).

Expression Analysis of X-Linked and Y-Linked Alleles

We used read numbers to estimate expression levels of the sex-linked contigs (see Material and Methods). We first compared expression levels of X-linked and Y-linked alleles in males. The read numbers were normalized to be able to combine data from different male individuals. As shown in Figure 1, we found that the Y/X expression ratio is significantly less than 1 (median 0.77, mean 0.89, significant Wilcoxon paired test $p<10^{-16}$). This is in agreement with previous work on six experimentally identified sex-linked genes [33] and also with recent work using RNA-seq data [42,43]. Why Y expression is reduced over evolutionary time is not fully understood. It could be because of the accumulation of

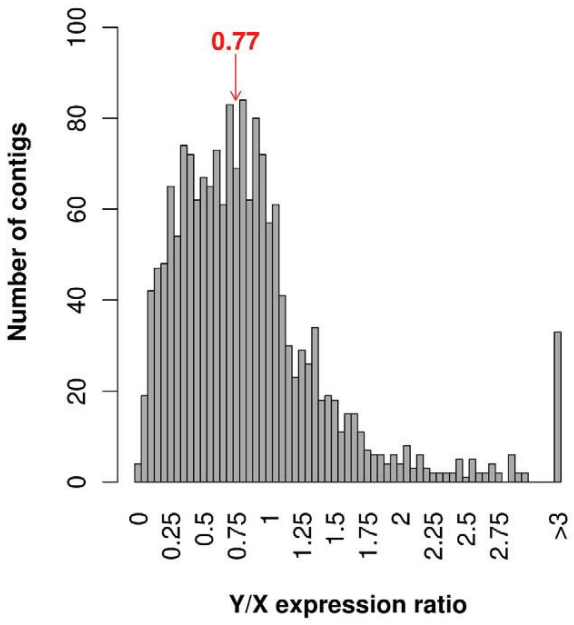


Figure 1. Distribution of Y/X expression ratios in *S. latifolia* males for the 1,736 sex-linked contigs. Total Y and X read numbers were summed at sex-linked SNP locations for each contig and normalized for each male separately, then averaged across males to obtain the Y/X ratio. The median is shown in red.
doi:10.1371/journal.pbio.1001308.g001

slightly deleterious mutations in promoters and cis-regulatory elements, and/or the insertion of transposable elements when the methylation of these elements spreads to nearby genes. However, this trend is considered a hallmark of Y chromosome degeneration and has been observed in several animal systems [45,46]. Y degeneration is thus clearly visible in *S. latifolia* but may not be as pronounced as expected because of haploid selection on pollen preventing the degeneration of many pollen-expressed Y genes [42] (but see [43,47]).

The observation that many X/Y pairs show reduced Y expression (Figure 1) raises the question whether dosage compensation has evolved in *S. latifolia*. To test this, we compared expression levels of sex-linked genes between males and females following a normalization procedure that allows comparing different individuals (see Material and Methods). First, we computed the ratio of the expression intensities of X-linked contigs in males and females and called this the Xmale/2Xfemale expression ratio (to stress the difference in gene copy number between male and female). In the absence of dosage compensation, the Xmale/2Xfemale expression ratio is expected to be 0.5, simply because males (XY) have one X-linked copy and females (XX) have two. This is what we observe for contigs that do not show reduced expression of the Y-linked allele relative to the X-linked allele, i.e., that have a Y/X expression ratio close to 1 (median of Xmale/2Xfemale ratio is 0.51 for contigs with $1 \leq Y/X < 1.5$; see Figure 2). However, for contigs with reduced Y expression and therefore low Y/X ratios, we observe an Xmale/2Xfemale expression ratio very close to 1 (median of contigs with $Y/X < 0.5$ is 0.93; see Figure 2). This suggests that for contigs with reduced Y expression, for which expression of sex-linked genes would thus be unbalanced between males and females, a mechanism has evolved that compensates for the reduced Y expression by increasing X expression in males.

To study this phenomenon further, we compared expression of X-linked and Y-linked alleles in males and females for different Y/X expression ratio categories (Figure 3). We excluded sex-linked

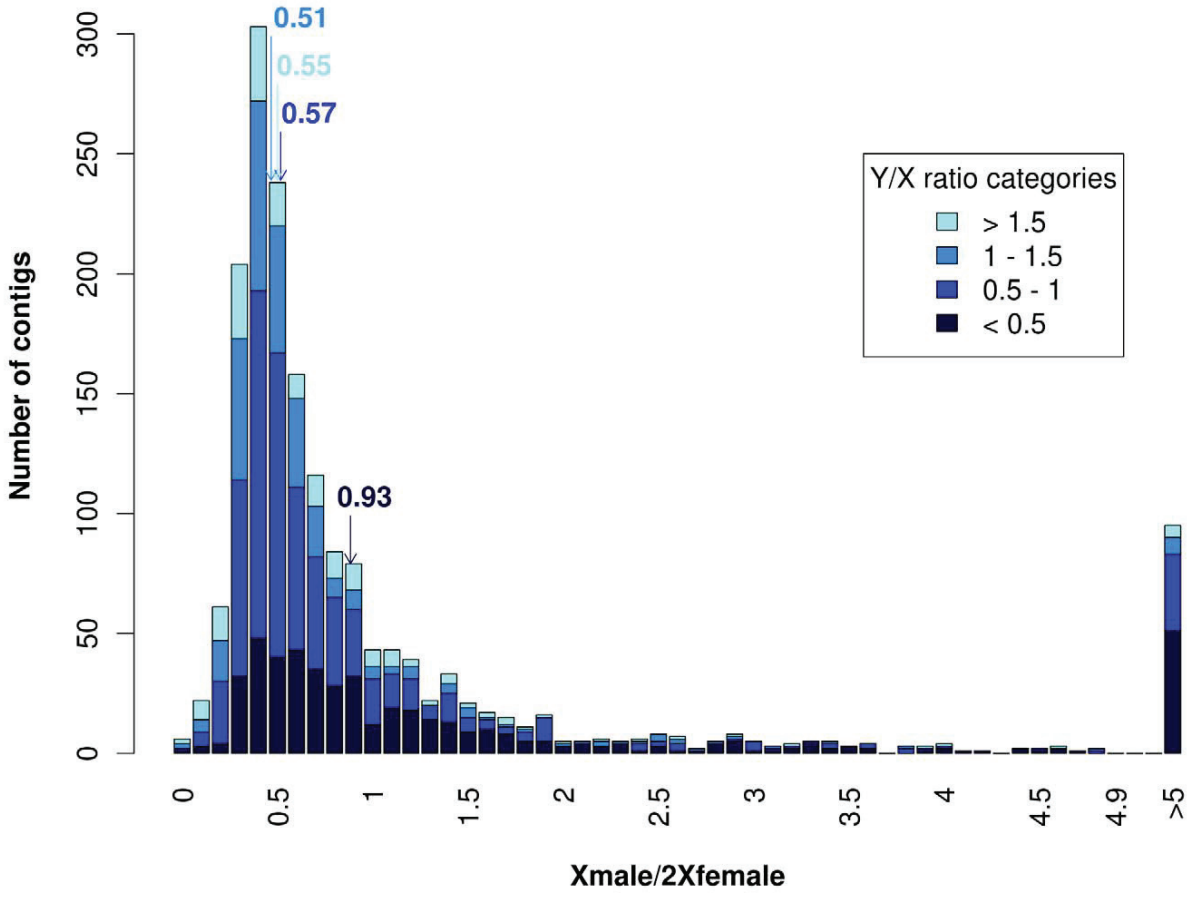


Figure 2. Distribution of the ratio between the expression of the single X in males and the two X copies in females (Xmale/2Xfemale) for all sex-linked contigs. Different categories of sex-linked contigs are shown: Y/X ratio below 0.5 (379 contigs), Y/X ratio between 0.5 and 1 (656 contigs), Y/X ratio between 1 and 1.5 (315 contigs), Y/X ratio above 1.5 (195 contigs). Medians are indicated in the colour corresponding to each Y/X ratio category. When the contigs with high Xmale/2Xfemale ratios are removed as in Figure 3 (see text for explanations) the medians remain unaltered except for the category $Y/X < 0.5$ where it changes to 0.76 but is still significantly different from 0.5 (Wilcoxon test, $p < 10^{-16}$). Total X read numbers were summed at sex-linked SNP locations in each contig and normalized for each individual separately, then averaged among males and females to get the Xmale/2Xfemale ratio.
doi:10.1371/journal.pbio.1001308.g002

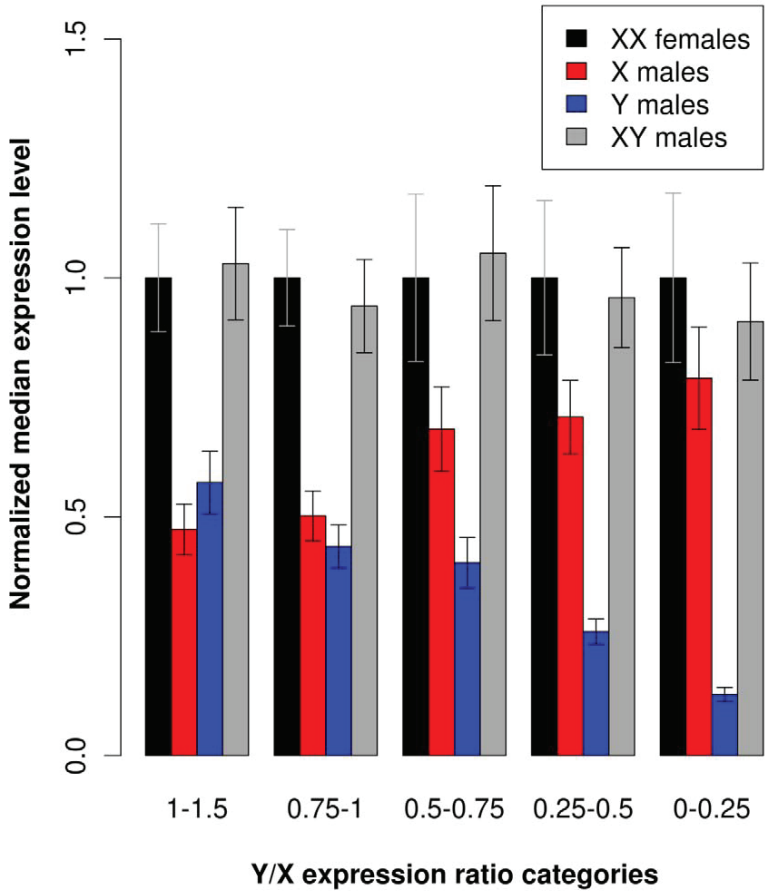


Figure 3. Expression levels of sex-linked contigs in both sexes for different Y/X expression ratio categories. Total read numbers were summed at sex-linked SNP locations and normalized for each individual and contig separately; medians for all contigs and individuals of the same sex were then obtained. Contigs with Y/X expression ratios above 1.5 were excluded, as well as contigs with Xmale/2Xfemale ratios above 2 (see text for explanations), which reduces the dataset to 1,346 sex-linked contigs. XX females, median expression level of both X-linked alleles in females; X males, median expression level of the single X-linked allele in males; Y males, median expression level of the Y-linked allele in males; XY males, median expression level of the X-linked plus Y-linked alleles in males. To compare different Y/X expression ratio categories, medians were normalized using the XX expression levels in females. Sample sizes are: 0–0.25, 110; 0.25–0.5, 269; 0.5–0.75, 315; 0.75–1, 341; 1–1.5, 315. Note that we do not have any contig with Y/X = 0 as our method did not allow us to detect such contigs (see Material and Methods). Error bars indicate 95% confidence intervals.
doi:10.1371/journal.pbio.1001308.g003

contigs that showed either an elevated Y expression (high Y/X ratios) or male-biased X expression (high Xmale/2Xfemale ratios). Such male-biased expression patterns suggest that these genes may be sexually antagonistic genes. The evolutionary dynamics of such genes is known to be distinct from other sex-linked genes and no dosage compensation is expected [48,49]. Figure 3 shows the results for the remaining 75% of sex-linked genes. We found that X expression in males increases with decreasing Y expression, which results in similar expression levels of sex-linked contigs in both sexes and provides further evidence of dosage compensation in *S. latifolia*. Importantly, this result is consistent even when we include only sex-linked contigs with at least two sex-linked SNPs, for which we estimated the rate and number of erroneous sex-linked contigs to be extremely low (0.001 and 1.38, respectively; see Figure S4). We also looked at expression patterns of the contigs corresponding to known sex-linked genes. Although this analysis can only be qualitative due to the small number of such genes, we found that Y/X ratios for most genes are consistent with previous work [33] and some known sex-linked genes show evidence for dosage compensation (Table S4).

Discussion

Evidence for X Chromosome Dosage Compensation in *S. latifolia*

There was a recent claim of absence of dosage compensation in *S. latifolia* [42], which seems to contradict our findings. However, the test for dosage compensation performed in this recent work is very different from ours. As Chibalina and Filatov (2011) analyzed crosses (parents and progeny), they were able to identify X-linked genes without detectable homologous Y-linked copies (called hemizygous genes). They compared the expression levels of these hemizygous genes between sexes, found a significantly reduced expression in males compared to females, and concluded that this was evidence for the absence of dosage compensation in *S. latifolia* [42]. Their test however may be overly conservative, as it requires

a strict Xmale/2Xfemale ratio of 1 to infer for dosage compensation. Their figure 4 suggests the Xmale/2Xfemale ratio is not 0.5, as expected under a complete absence of dosage compensation, but instead is close to 0.7, which is consistent with many hemizygous genes being dosage compensated. Importantly, the hemizygous genes were interpreted as sex-linked genes with fully degenerated Y copies, which may not always be the case as genes that have recently moved from the autosomes to the X chromosome will also be detected as hemizygous genes but dosage compensation is clearly not expected for those genes [43]. Such gene movement has been documented in *S. latifolia* [39] and may account for the intermediate Xmale/2Xfemale value (between 0.5 and 1) found in [42]. By contrast, we looked for departure from a Xmale/2Xfemale of 0.5 and did not restrict the test to sex-linked genes with no Y expression but included the many sex-linked genes with reduced but still detectable Y expression. We thus performed a more permissive test for dosage compensation, which may be more suitable in the case of young sex chromosomes with incipient X chromosome dosage compensation.

Sex Bias in Gene Expression and Dosage Compensation

Dosage compensation is not the only sex-specific gene expression regulation that is expected on the X chromosome. Indeed, X-linked genes involved in sexual conflicts—for instance those underlying sexual dimorphism and having sexually antagonistic effects—can show sex-biased expression and this can substantially affect the global X expression pattern in both sexes if these genes are numerous [50]. A way to distinguish dosage compensation from such sex-specific expression regulation is to look at the X over autosome (X/A) expression ratio as only dosage compensation predicts a X/A expression of 1 [50]. However, this test is difficult to perform here for several reasons. First, our set of sex-linked genes is expected to exclude those with very low expression levels because the detection of sex-linked SNPs requires reasonably high read coverage. This should bias upward the average expression level of sex-linked genes compared to the "autosonal" set, which is what we actually found (the mean number of reads per base is 466.7 for sex-linked contigs and 101.4 for non–sex-linked contigs). Second, we do not have a reliable "autosomal" set as this includes a mixture of autosomal contigs and sex-linked contigs not detected by our method (~40% of all sex-linked genes, see above). Although we excluded possible candidates for sexually antagonistic genes (some of the contigs with high Xmale/2Xfemale may be "male-beneficial and female-detrimental" genes), we cannot completely rule out the possibility that others remained in the set of contigs used to assess dosage compensation (especially some contigs with low Xmale/2Xfemale may be "female-beneficial and male-detrimental" genes). However, Figure 3 shows that the increase of X expression in males follows the level of degeneration of Y expression, which is not expected in case of sexually antagonistic selection. Moreover, increased expression of the X-linked allele in males always compensates for the reduced Y expression, such that the total expression of these sex-linked genes is similar in both sexes (i.e., X+Y expression in males = X+X expression in females), which is not in agreement with sexually antagonistic selection. On the contrary, sexually antagonistic selection predicts between-sex differences in expression of sex-linked genes. The results presented in Figure 3 are thus better explained by dosage compensation than by sexually antagonistic selection.

Dosage Compensation in XY and ZW Systems

Global dosage compensation has previously been documented in male heterogametic systems (XY) such as *Drosophila*, *Caenorhabditis elegans*, and mammals [14,51], whereas only partial (or no) dosage compensation has been found in female heterogametic systems (ZW) [52]. Indeed, in zebra finch, chicken, and crow, no global mechanism to balance avian Z chromosome gene dosage (such as X chromosome inactivation) has been found [53–56] and in chicken, dosage compensation seems to be local, with only few Z-linked genes being dosage compensated [57]. Similar observations have been made in silkworm [58,59], indicating that the lepidopteran Z is not fully dosage compensated, and also in the parasite *Schistosoma mansoni* [60]. Moreover, studies on the platypus [61,62] and on sticklebacks [63] suggest that partial dosage compensation can also exist in male heterogametic systems (XY). Overall, these new data suggest that full dosage compensation is not a necessary outcome of sex chromosome evolution [50]. An important point of whether dosage compensation will evolve or not is the presence of dosage-sensitive genes on the proto-sex chromosomes, as these genes are the only ones for which dosage compensation is vital [50,64]. Although we do not have any data about the fraction of dosage-sensitive genes in the different sex chromosome systems, it has been suggested that resistance to aneuploidy and polyploidization may indicate whether the genome as a whole includes many such genes or not [50]. Polyploidization is known to be common in plants [65]. However, plant polyploids do have dosage problems that cause endosperm development failure and reduced fertility [64,66]. Following polyploidization events, the retention of plant duplicate genes seems to be driven by dosage constraints as in animals [64]. All this suggests that the success of polyploids in plants may not be related to lack of dosage constraints but to other reasons (e.g., vegetative propagation). It is also known that aneuploidy has more severe phenotypic consequences than polyploidy in plants, which further supports the idea of strong dosage constraints in plant genomes [64]. As far as we know, there is no documented case of fertile polyploids in dioecious *Silene* species and it is possible that the *S. latifolia* genome includes enough dosage-sensitive genes for dosage compensation to evolve.

Mechanisms of Dosage Compensation in Plants

Our results reveal that dosage compensation is not restricted to animals but also occurs in plants and raise questions about the mechanisms underlying dosage compensation. In animals, three different dosage compensation mechanisms have been uncovered (reviewed in [67]): hyper-expression of X-linked alleles in male *Drosophila*, down-regulation of the two X-linked alleles in hermaphrodites of *C. elegans*, and inactivation of one of the two female X chromosomes in mammals. We tested whether such a chromosome-wide inactivation exists in *S. latifolia* by checking whether both X-linked alleles are expressed in females. Although heterozygosity is low in our X-linked alleles because our individuals are inbred, we found that the level of heterozygosity of the X-linked alleles is similar for sex-linked contigs with dosage compensation and those without dosage compensation (Table S5). This suggests that both X-linked alleles are expressed, whatever the level of dosage compensation is, and does not support an X-inactivation-like mechanism in *S. latifolia*. Further work will be needed to identify the molecular mechanism underlying dosage compensation in *S. latifolia*.

De Novo Evolution of Dosage Compensation in a Young XY System

Previous work in animals has reported dosage compensation in old X chromosomes (see above) and also in young neoX chromosomes such as the *D. miranda* neoX. The fusion between X and the autosome that formed the *D. miranda* neoX is very recent (1.5 million years old), but dosage compensation is achieved by a protein complex (the MSL complex) that pre-dates neoX formation and has been shown to be very old [68]. Evidence for de novo

evolution of dosage compensation in evolutionary young animal sex chromosomes is therefore lacking [50]. In the *Silene* genus, most species are hermaphroditic or gynodioecious and do not have sex chromosomes. Sex chromosomes have evolved recently in two independent lineages, one including *S. latifolia* and one containing *S. colpophylla* [20,44,69]. Our results therefore reveal that dosage compensation has evolved de novo in evolutionarily young sex chromosomes in probably less than 10 million years. Furthermore, Figure 2 shows that many dosage-compensated contigs have an Xmale/2Xfemale ratio that is not exactly 1 (although the median is close to 1, there is no peak at 1 for $Y/X<0.5$ contigs). This is consistent with the mechanism being evolutionarily young and not optimized yet. Our results also reveal that dosage compensation can evolve as soon as Y expression starts declining. This way, dosage compensation already exists when the Y copy is ultimately lost (and can even facilitate such loss, see [70]). Instead of being a later step of sex chromosome evolution following Y degeneration, our results suggest that the evolution of dosage compensation and Y degeneration probably occur at the same time.

Material and Methods

Plant Material, RNA Extraction, Sequencing, and Assembly of Illumina Data

Plants used in this study belong to a population of *S. latifolia* that has been inbred for ten generations with brother-sister mating: three males (U10_11, U10_49, and U10_09) and three females (U10_34, U10_37, and U10_39) that were grown in a temperature-controlled greenhouse. The QiagenRNeasy Mini Plant extraction kit was used to extract total RNA two times separately from four flower buds at developmental stages B1–B2 after removing the calyx. Samples were treated additionally with QiagenDNase. RNA quality was assessed with an Aligent Bioanalyzer ($RIN>9$) and quantity with an Invitrogen Qubit. An intron-spanning PCR product was checked on an agarose gel to exclude the possibility of genomic DNA contamination. Then, the two extractions of the same individual were pooled. Samples were sequenced by FASTERIS SA on an Illumina HiSeq2000 following an Illumina paired-end protocol (fragment lengths 150–250 bp, 100 bp sequenced from each end). Individuals were tagged and pooled for sequencing in two different runs (U10_49 male and U10_37 female in the first run and the others in the second). See Table S1 for sizes of the different libraries. Our Illumina reads are available in the GEO database (through the GEO Series GSE35563).

De novo assembly was conducted on a computer cluster (Figure S1). Illumina reads from all individuals were pooled together for assembly with AbySS 1.2.5 ($E=10$, $n=5$) [71] with the paired-end option and with all k-mers ranging from 51 to 96 in order to address variable transcript expression [72]. A k-mer length equal to 51 was the minimum possible to avoid contigs shorter than the reads, and 96 is the maximum allowed by AbySS. Only contigs were kept at this stage, singlets were discarded. Contigs that exactly matched another longer contig were then removed by pairwise comparison of AbySS outputs using Trans-ABySS 1.2.0 [72]. A non-redundant set of contigs was thus obtained and further assembled through two runs of CAP3 version 12/21/07 [73]. Singlets and contigs were conserved after each CAP3 run. CAP3 runs increased the chance for X and Y copies to be assembled into the same contig, which is crucial for further sex-linked SNP detection. Contigs shorter than 200 bp were not included in the final set of contigs.

Mapping, SNPs Analysis, and Sex-Linkage Detection

Illumina reads were mapped onto reference sequences (final set of contigs and also CDS from known sex-linked genes retrieved from GenBank for adjusting SNP detection, see below) for each individual separately using BWA 0.5.9 [74] (using default parameters for paired-end reads, and gap and mismatch maximum number of 5 as suggested for 100 bp reads in [74]), which was shown to be efficient and to use much less RAM than other programs for Illumina read mapping [75]. Alignments of all individuals were then merged together using Samtoolsmerge version 0.1.12 [76]. The percentage of mapped reads was assessed using Samtoolsflagstat version 0.1.12 [76] and the average coverage was determined using the Genome Analysis Toolkit (GATK 1.0.5315) Depth of Coverage [77].

SNPs were detected with the GATK Unified Genotyper (using the following parameters: -stand_call_conf 4 -stand_emit_conf 0 -mbq 17 -mmq 0 -mm40 40 -bad_mates -dcov 2000) [77], which is considered the best currently available tool for SNP detection [78]. Thresholds for the different SNP detection parameters were set to be very low (except for the base quality parameter) in order not to disfavour Y SNPs that are expected to be found in low numbers and low mapping quality if a contig contains mainly X reads, which can happen when X-linked alleles are more strongly expressed than Y-linked alleles [33].

The detected SNPs were then filtered using Perl scripts to retrieve SNPs for which all males are heterozygous (XY) and all females homozygous (XX). All contigs with at least one SNP showing this pattern were considered sex-linked. For females, the genotypes inferred by GATK were directly used for analysis. For males, this information is not reliable since the Y-linked allele is expected to be less expressed than the X-linked allele [33] while GATK genotyper makes the assumption that both alleles are expressed at a similar level. The read numbers of each SNP were thus used to infer male genotypes (see Text S3 for details).

Polymorphism on the X chromosome (at least one male or female heterozygous or all individuals homozygous but not for the same polymorphism) was detected on sex-linked contigs with a similar filter as the one described above.

Estimates of Expression Levels of the Sex-Linked Contigs

Expression levels of the X-linked and Y-linked alleles in males and both X copies in females were computed by counting reads at sex-linked SNP locations only, and not for the entire contigs, in order to clearly distinguish between X and Y reads. Total read numbers of all X or Y SNPs provided by the GATK Unified Genotyper [77] were summed for each X-linked or Y-linked alleles and each individual separately and then normalized using the total number of mapped reads per individuals (library size) and the number of sex-linked SNPs in the contigs:

$$E = \frac{r}{n \times l}$$

With E = normalized expression level, r = sum of total read counts, $n = n$ sex-linked SNPs, l = normalized library size.

The library size of the six individuals was normalized to take into account the difference in mitochondrial, chloroplast, and transposable element (TE) transcript quantity between sexes and the difference in rRNA quantity between the first and the second Illumina run. The *Arabidopsis thaliana* rRNA genes, complete *S. latifolia* mtDNA genome [79], *S. latifolia* chloroplast genes *rpoB*, *rpoC1*, *rpoC2*, *rps2*, *atpI*, *atpH*, *atpF*, *atpA*, *psbI*, *psbK*, *rps16*, *matK*, *psbA*, *rpl2*, *ycf2*, *ndhB*, *rps7*, and the TEs known in *Silene* [80] were retrieved from GenBank. The read numbers of rRNA, TEs and mtRNA, and cpRNA were determined by mapping the Illumina reads onto the known CDS sequences of these elements using the default parameters in BWA (results presented in Table S1).

The expression levels were normalized for each contig and for each individual in number of reads per kilobase per million mapped reads (RPKM) [81], and then the mean for each sex was computed.

Supporting Information

Figure S1 Assembly, mapping, and SNP analysis. Steps of the de novo assembly. From left to right: during first assembly with ABySS, k-mers ranging from 51 from 96, only contigs were kept. Pairwise comparisons of contigs were then done by TransABySS in order to remove small contigs that exactly matched longer contigs. Contigs were then further assembled by two runs of CAP3 (mismatches and partial overlaps allowed); singlets and contigs were kept after each run. Illumina reads were mapped onto the contigs with BWA and SNPs were detected with GATK. SNPs were then analyzed in order to detect sex-linked SNPs (all males heterozygous XY, and all females homozygous XX).
(TIFF)

Figure S2 Number of sex-linked SNPs detected and coverage for known sex-linked genes. cDNA sequences of previously identified sex-linked genes were retrieved from GenBank. Illumina reads were mapped on the cDNA sequences using BWA and SNP detection was done as in Material and Methods. We then computed the number of sex-linked SNPs detected over the number of known sex-linked SNPs for these genes and compared this with the number of reads (= coverage) for each X/Y gene pairs. Sex-linked genes were grouped by strata as in [82].
(TIFF)

Figure S3 Size (bp) distribution of sex-linked contigs.
(TIFF)

Figure S4 Expression levels of sex-linked contigs in both genders for different Y/X expression ratio categories for contigs with ≥2 sex-linked SNPs (1,009 contigs). The legend is the same as for Figure 3 except for contig numbers: 0–0.25, 66; 0.25–0.5, 165; 0.5–0.75, 248; 0.75–1, 279; 1–1.5, 251.
(TIFF)

Table S1 Raw Illumina data and results of the assembly.
(DOC)

Table S2 Contig statistics.
(DOC)

Table S3 Results of SNP analysis for known autosomal and sex-linked genes.
(RTF)

Table S4 Analysis of expression patterns in known sex-linked genes.
(RTF)

Table S5 Levels of heterozygosity of the X-linked alleles with and without dosage compensation.
(DOC)

Text S1 Identification and validation of new sex-linked genes.
(RTF)

Text S2 Simulations to estimate the rate of false positive sex-linked genes.
(DOC)

Text S3 SNP detection and filtering.
(DOC)

Acknowledgments

We thank the Genetic Diversity Centre (GDC) for support.

Author Contributions

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: AW GABM SM. Performed the experiments: NZ. Analyzed the data: AM NZ CD GABM. Contributed reagents/materials/analysis tools: AW. Wrote the paper: GABM AW.

References

1. Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. Science 286: 964–967.
2. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.
3. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al. (2005) The DNA sequence of the human X chromosome. Nature 434: 325–337.
4. Lemaitre C, Braga MD, Gautier C, Sagot MF, Tannier E, et al. (2009) Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. Genome Biol Evol 1: 56–66.
5. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res 18: 965–973.
6. Potrzebowski L, Vinckenbosch N, Jégou B, Marques AC, Chalmel F, et al. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biology 6: e80. doi:10.1371/journal.pbio.0060080.
7. Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. Philos Trans R Soc Lond B Biol Sci 355: 1563–1572.
8. Bachtrog D (2008) The temporal dynamics of processes underlying Y chromosome degeneration. Genetics 179: 1513–1525.
9. Wilson MA, Makova KD (2009) Genomic analyses of sex chromosome evolution. Annu Rev Genomics Hum Genet 10: 333–354.
10. Wyckoff GJ, Li J, Wu C-I (2002) Molecular evolution of functional genes on the mammalian y chromosome. Mol Biol Evol 19: 1633–1636.
11. Wilson MA, Makova KD (2009) Evolution and survival on eutherian sex chromosomes. PLoS Genet 5: e1000568. doi:10.1371/journal.pgen.1000568.
12. Charlesworth B (1978) A model for the evolution of Y chromosomes and dosage compensation. Proc Natl Acad Sci U S A 75: 5618–5622.
13. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434: 400–404.
14. Vicoso B, Bachtrog D (2009) Progress and prospects toward our understanding of the evolution of dosage compensation. Chromosome Res 17: 585–602.
15. Chow J, Heard E (2009) X inactivation and the complexities of silencing a sex chromosome. Curr Opin Cell Biol 21: 359–366.
16. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity 95: 118–128.
17. Marshall Graves JA (2008) Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. Annu Rev Genet 42: 565–586.
18. Ellegren H (2011) Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. Nat Rev Genet 12: 157–166.
19. Westergaard M (1958) The mechanism of sex determination in dioecious flowering plants. Adv Genet 9: 217–281.
20. Desfeux C, Maurice S, Henry JP, Lejeune B, Gouyon PH (1996) Evolution of reproductive systems in the genus *Silene*. Proc Biol Sci 263: 409–414.
21. Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, et al. (2005) A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. PLoS Biol 3: e4. doi:10.1371/journal.-pbio.0030004.
22. Filatov DA (2005) Evolutionary history of *Silene latifolia* sex chromosomes revealed by genetic mapping of four genes. Genetics 170: 975–979.
23. Bernasconi G, Antonovics J, Biere A, Charlesworth D, Delph LF, et al. (2009) *Silene* as a model system in ecology and evolution. Heredity 103: 5–14.
24. Ming R, Bendahmane A, Renner S (2011) Sex chromosomes in land plants. Annu Rev Plant Biol 62: 485–514.
25. Bergero R, Forrest A, Kamau E, Charlesworth D (2007) Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: evidence from new sex-linked genes. Genetics 175: 1945–1954.
26. Guttman DS, Charlesworth D (1998) An X-linked gene has a degenerate Y-linked homologue in the dioecious plant *Silene latifolia*. Nature 393: 263–266.
27. Filatov DA, Moneger F, Negrutiu I, Charlesworth D (2000) Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. Nature 404: 388–390.

28. Filatov DA, Charlesworth D (2002) Substitution rates in the X- and Y-linked genes of the plants, *Silene latifolia* and *S. dioica*. Mol Biol Evol 19: 898–907.
29. Pritham EJ, Zhang YH, Feschotte C, Kesseli RV (2003) An Ac-like transposable element family with transcriptionally active y-linked copies in the white campion, *Silene latifolia*. Genetics 165: 799–807.
30. Laporte V, Filatov DA, Kamau E, Charlesworth D (2005) Indirect evidence from DNA sequence diversity for genetic degeneration of the Y-chromosome in dioecious species of the plant *Silene*: the SlY4/SlX4 and DD44-X/DD44-Y gene pairs. J Evol Biol 18: 337–347.
31. Kejnovsky E, Hobza R, Kubat Z, Widmer A, Marais GAB, et al. (2007) High intrachromosomal similarity of retrotransposon long terminal repeats: Evidence for homogenization by gene conversion on plant sex chromosomes? Gene 390: 92–97.
32. Bergero R, Forrest A, Charlesworth D (2008) Active miniature transposons from a plant genome and its nonrecombining Y chromosome. Genetics 178: 1085–1092.
33. Marais GA, Nicolas M, Bergero R, Chambrier P, Kejnovsky E, et al. (2008) Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. Curr Biol 18: 545–549.
34. Qiu S, Bergero R, Forrest A, Kaiser VB, Charlesworth D (2010) Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes. Proc Biol Sci 277: 3283–3290.
35. Delichère C, Veuskens J, Hernould M, Barbacar N, Mouras A, et al. (1999) SlY1, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein. EMBO Journal 18: 4169–4179.
36. Atanassov I, Delichere C, Filatov DA, Charlesworth D, Negrutiu I, et al. (2001) Analysis and evolution of two functional Y-linked loci in a plant sex chromosome system. Mol Biol Evol 18: 2162–2168.
37. Moore RC, Kozyreva O, Lebel-Hardenack S, Siroky J, Hobza R, et al. (2003) Genetic and functional analysis of DD44, a sex-linked gene from the dioecious plant *Silene latifolia*, provides clues to early events in sex chromosome evolution. Genetics 163: 321–334.
38. Filatov DA (2005) Substitution rates in a new *Silene latifolia* sex-linked gene, SlssX/Y. Mol Biol Evol 22: 402–408.
39. Kaiser VB, Bergero R, Charlesworth D (2009) Slcyt, a newly identified sex-linked gene, has recently moved onto the X chromosome in *Silene latifolia* (Caryophyllaceae). Mol Biol Evol 26: 2343–2351.
40. Kaiser VB, Bergero R, Charlesworth D (2011) A new plant sex-linked gene with high sequence diversity and possible introgression of the X copy. Heredity 106: 339–347.
41. Blavet N, Charif D, Oger-Desfeux C, Marais GA, Widmer A (2011) Comparative high-throughput transcriptome sequencing and development of SiESTa, the *Silene* EST annotation database. BMC Genomics 12: 376.
42. Chibalina MV, Filatov DA (2011) Plant Y chromosome degeneration is retarded by haploid purifying selection. Curr Biol 21: 1475–1479.
43. 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.
44. Marais GA, Forrest A, Kamau E, Kafer J, Daubin V, et al. (2011) Multiple nuclear gene phylogenetic analysis of the evolution of dioecy and sex chromosomes in the genus Silene. PLos One 6: e21915. doi:10.1371/journal.-pone.0021915.
45. Bachtrog D (2006) Expression profile of a degenerating neo-y chromosome in *Drosophila*. Curr Biol 16: 1694–1699.
46. Zhou Q, Wang J, Huang L, Nie W, Liu Y, et al. (2008) Neo-sex chromosomes in the black muntjac recapitulate incipient evolution of mammalian sex chromosomes. Genome Biol 9: R98.
47. Bachtrog D (2011) Plant sex chromosomes: a non-degenerated Y? Curr Biol 21: R685–688.
48. Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8: 689–698.
49. Bachtrog D, Toda NR, Lockton S (2010) Dosage compensation and demasculinization of X chromosomes in *Drosophila*. Curr Biol 20: 1476–1481.
50. Mank JE, Hosken DJ, Wedell N (2011) Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. Evolution 65: 2133–2144.
51. Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, et al. (2011) Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila* melanogaster. Nat Genet 43: 1179–1185.
52. Mank J (2009) The W, X, Y and Z of sex-chromosome dosage compensation. Trends Genet 25: 226–233.
53. Itoh Y, Melamed E, Yang X, Kampf K, Wang S, et al. (2007) Dosage compensation is less effective in birds than in mammals. J Biol 6: 2.
54. Ellegren H, Hultin-Rosenberg L, Brunstrom B, Dencker L, Kultima K, et al. (2007) Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. BMC Biol 5: 40.
55. Itoh Y, Replogle K, YH Kim, Wade J, Clayton D, et al. (2010) Sex bias and dosage compensation in the zebra finch versus chicken genomes: general and specialized patterns among birds. Genome Res 20: 512–518.
56. Wolf JB, Bryk J (2011) General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. BMC Genomics 12: 91.
57. Mank JE, Ellegren H (2009) All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. Heredity 102: 312–320.
58. Zha X, Xia Q, Duan J, Wang C, He N, et al. (2009) Dosage analysis of Z chromosome genes using microarray in silkworm, Bombyx mori. Insect Biochem Mol Biol 39: 315–321.
59. Walters JR, Hardcastle TJ (2011) Getting a full dose? Reconsidering sex chromosome dosage compensation in the silkworm, *Bombyx mori*. Genome Biol Evol 3: 491–504.
60. Vicoso B, Bachtrog D (2011) Lack of global dosage compensation in *Schistosoma mansoni*, a female-heterogametic parasite. Genome Biol Evol 3: 230–235.
61. Deakin JE, Hore TA, Koina E, Marshall Graves JA (2008) The status of dosage compensation in the multiple X chromosomes of the platypus. PLoS Genet 4: e1000140. doi:10.1371/journal.pgen.1000140.
62. Deakin J, J C, Hore T, Graves J (2009) Unravelling the evolutionary origins of X chromosome inactivation in mammals: insights from marsupials and monotremes. Chromosome Res 15: 671–685.
63. Leder EH, Cano JM, Leinonen T, O'Hara RB, Nikinmaa M, et al. (2010) Female-biased expression on the X chromosome as a key step in sex chromosome evolution in threespine sticklebacks. Mol Biol Evol 27: 1495–1503.
64. Birchler JA, Veitia R (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol 186: 54–62.
65. Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet 34: 401–437.
66. Kohler C, Mittelsten Scheid O, Erilova A (2010) The impact of the triploid block on the origin and evolution of polyploid plants. Trends Genet 26: 142–148.
67. Straub T, Becker PB (2007) Dosage compensation: the beginning and end of generalization. Nat Rev Genet 8: 47–57.
68. Marin I, Franke A, Bashaw GJ, Baker BS (1996) The dosage compensation system of *Drosophila* is co-opted by newly evolved X chromosomes. Nature 383: 160–163.
69. Mrackova M, Nicolas M, Hobza R, Negrutiu I, Moneger F, et al. (2008) Independent origin of sex chromosomes in two species of the genus *Silene*. Genetics 179: 1129–1133.
70. Engelstädter J (2008) Muller's ratchet and the degeneration of Y chromosomes: a simulation study. Genetics 180: 957–967.
71. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117–1123.
72. Robertson G, Schein J, Chiu R, Corbett R, Field M, et al. (2010) De novo assembly and analysis of RNA-seq data. Nat Methods 7: 909–912.
73. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868–877.
74. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
75. Bao S, Jiang R, Kwan W, Wang B, Ma X, et al. (2011) Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56: 406–414.
76. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.
77. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.
78. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443–451.
79. Sloan DB, Alverson AJ, Storchova H, Palmer JD, Taylor DR (2010) Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. BMC Evol Biol 10: 274.
80. Cermak T, Kubat Z, Hobza R, Koblizkova A, Widmer A, et al. (2008) Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. Chromosome Res 16: 961–976.
81. Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. Genome Biol 11: 220.
82. Bergero R, Charlesworth D, Filatov DA, Moore RC (2008) Defining regions and rearrangements of the *Silene latifolia* Y chromosome. Genetics 178: 2045–2053.

B.1 Supplementary information

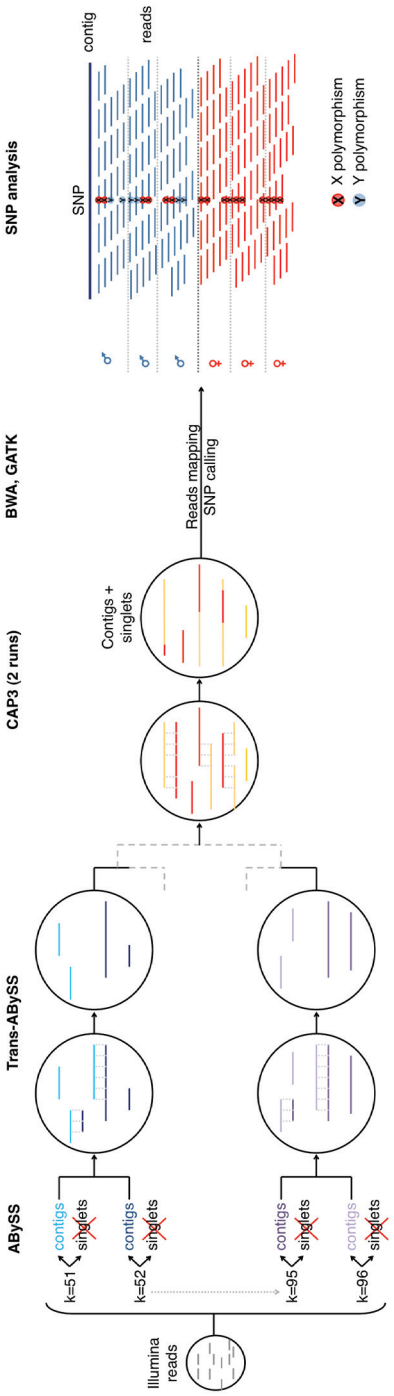


Figure B.S1: Assembly, mapping, and SNP analysis. Steps of the de novo assembly. From left to right: during first assembly with ABySS, k-mers ranging from 51 from 96, only contigs were kept. Pairwise comparisons of contigs were then done by Trans-ABySS in order to remove small contigs that exactly matched longer contigs. Contigs were then further assembled by two runs of CAP3 (mismatches and partial overlaps allowed); singlets and contigs were kept after each run. Illumina reads were mapped onto the contigs with BWA and SNPs were detected with GATK. SNPs were then analyzed in order to detect sex-linked SNPs (all males heterozygous XY, and all females homozygous XX).

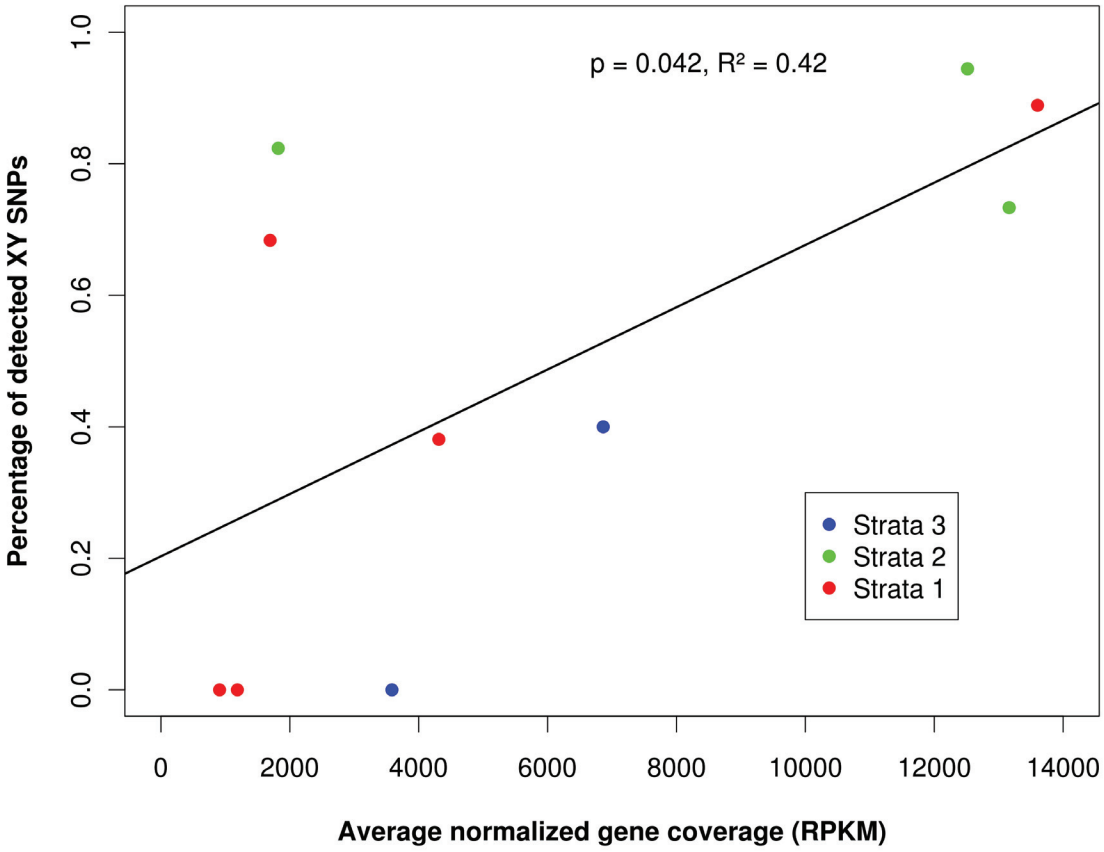


Figure B.S2: Number of sex-linked SNPs detected and coverage for known sex-linked genes. cDNA sequences of previously identified sex-linked genes were retrieved from GenBank. Illumina reads were mapped on the cDNA sequences using BWA and SNP detection was done as in Material and Methods. We then computed the number of sex-linked SNPs detected over the number of known sex-linked SNPs for these genes and compared this with the number of reads (coverage) for each X/Y gene pairs. Sex-linked genes were grouped by strata as in (Bergero et al., 2008)

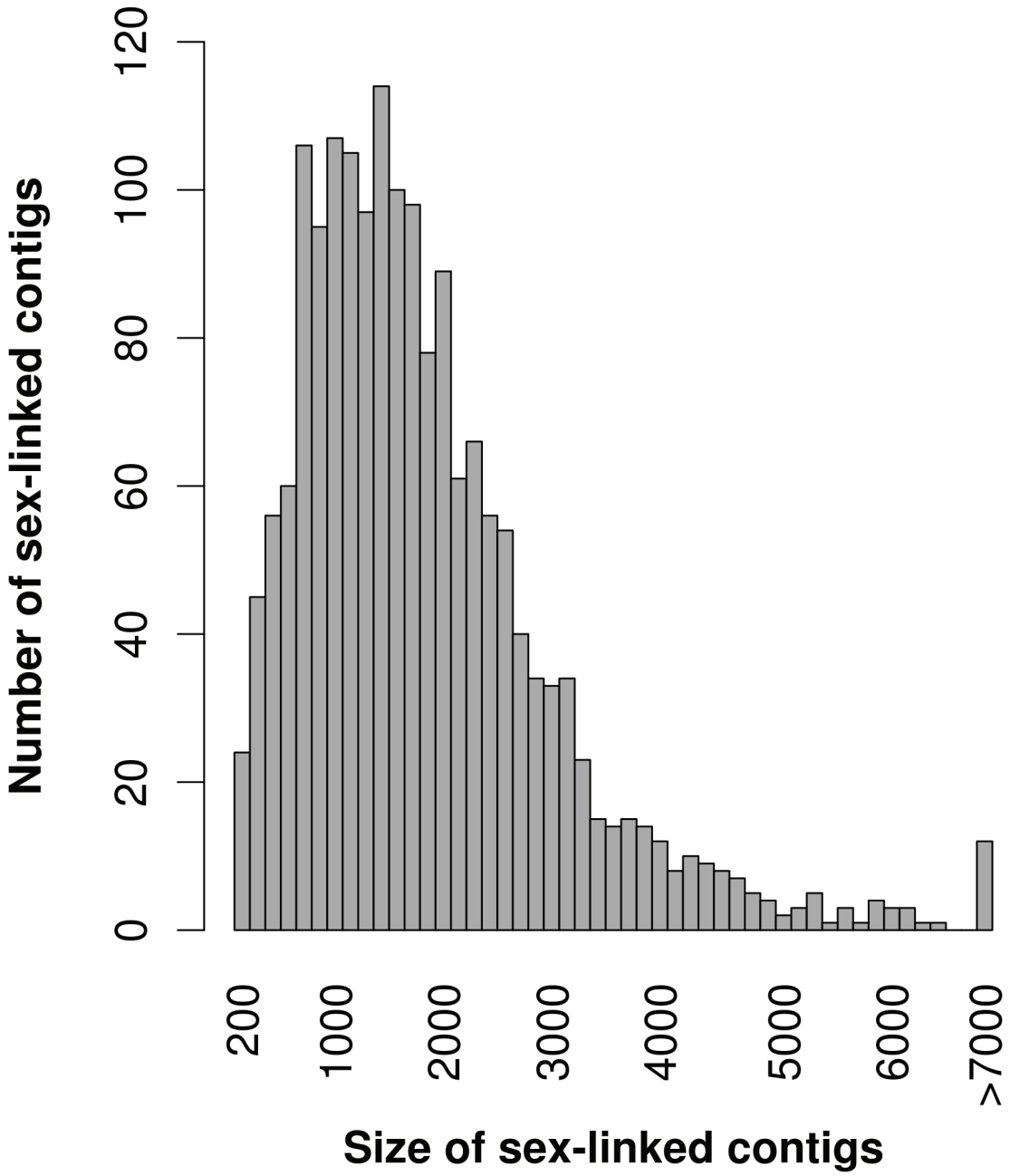


Figure B.S3: Size (bp) distribution of sex-linked contigs.

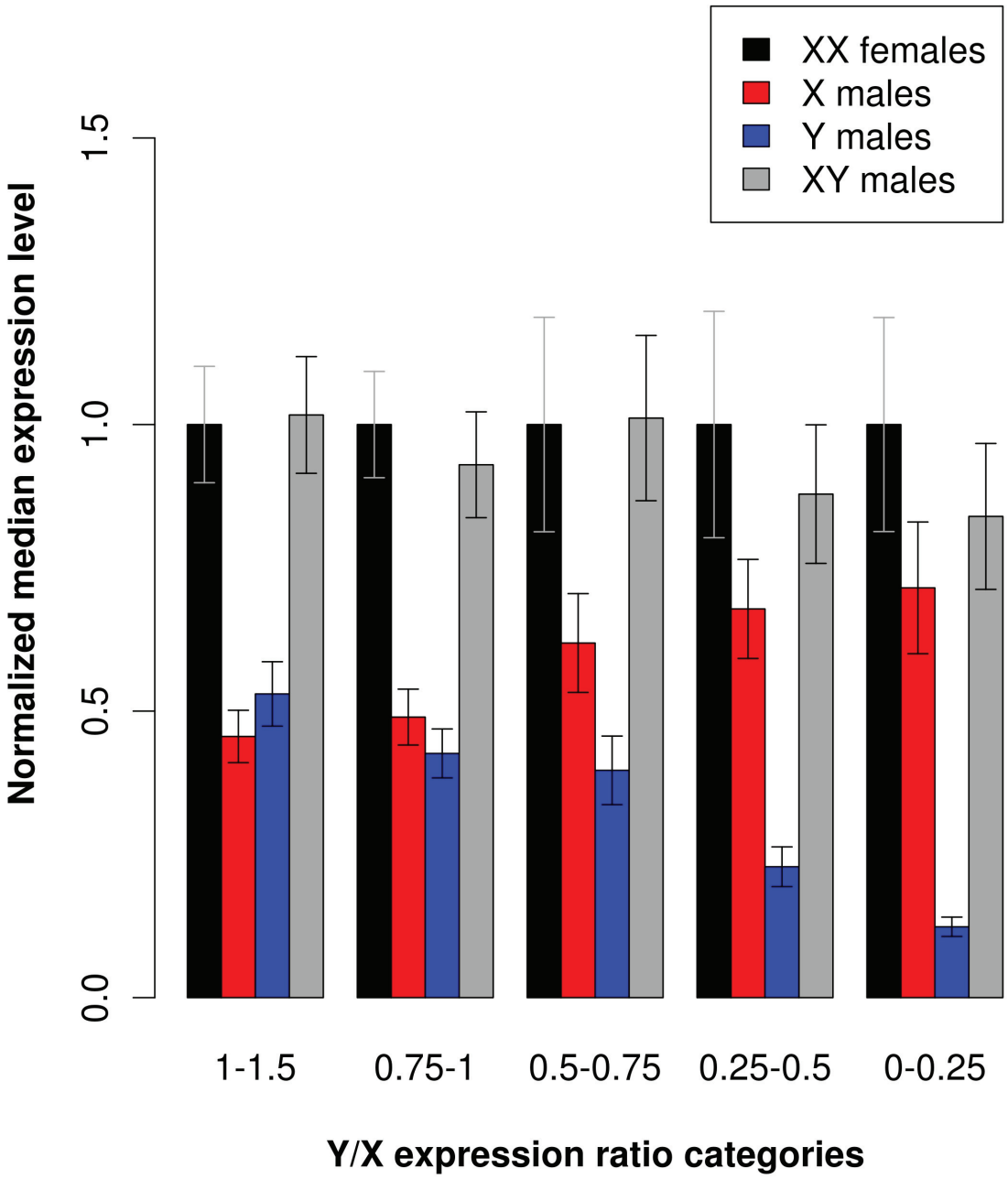


Figure B.S4: Expression levels of sex-linked contigs in both genders for different Y/X expression ratio categories for contigs with at least 2 sex-linked SNPs (1,009 contigs). The legend is the same as for Figure 3 except for contig numbers: 0-0.25, 66; 0.25-0.5, 165; 0.5-0.75, 248; 0.75-1, 279; 1-1.5, 251.

Table B.S1: Raw Illumina data and results of the assembly. Size of the dataset (numbers of reads and numbers of bases), numbers of mapped reads, coverage and normalized library size for each individual. The coverage is the average number of reads per base and the normalized library size corresponds to the number of mapped reads minus the number of reads that belong to rRNA, TEs, cpRNA, mtRNA (see Material and Methods). The normalized library sizes were used to compute normalized expression levels.

Individual	U10_37 (female)	U10_49 (male)	U10_11 (male)	U10_9 (male)	U10_39 (female)	U10_34 (female)	Total
Number of reads	75800468	89940748	48720356	46671858	44704710	45662378	351500518
Number of bases (Gb)	7.58	8.99	4.87	4.67	4.47	4.57	35.15
Number of mapped reads	66368044	80010603	45966660	44133217	41605007	43128090	321211621
Coverage	29.96	39.22	33.98	33.09	32	33.03	201.3
Normalized library size	58048299	73475899	45617786	43766170	41190095	42773518	304871767

Table B.S2: Contig statistics.

	Number of contigs	Mean size (bp)	Median size (bp)	Max size (bp)	Min size (bp)	Mean coverage (X)	Mean SNP # per contig
All contigs	141,855	748.5	410	20,988	200	201.3	10.26
Sex-linked contigs	1736	1931	1673	16,746	207	427.2	22.97

Table B.S3: Results of SNP analysis for known autosomal and sex-linked genes. Numbers of sex-linked SNPs identified for known autosomal and sex-linked genes in S. latifolia. The cDNA sequences retrieved from GenBank were blasted onto the final contig set of the assembly to identify the corresponding contig of each gene and SNPs were studied to retrieve those with a sex-linked pattern. Autosomal cDNA sequences are from Marais et al. (2011). Sex-linked cDNA sequences are from GenBank.

Genes	Number of sex-linked SNPs	Type	References
SlXY7	38	Sex-linked gene	Bergero et al., 2007
SlXY1	5	Sex-linked gene	Delichère et al., 1999
SlXY9	6	Sex-linked gene	Kaiser et al., 2011
SlMF1[1]	0	Sex-linked gene	Matsunaga et al., 2005
SlssXY	4	Sex-linked gene	Filatov, 2005
SlCypXY	38	Sex-linked gene	Bergero et al., 2007
DD44XY	26	Sex-linked gene	Moore et al., 2003
SlXY3	2	Sex-linked gene	Nicolas et al., 2005
SlAP3XY[2]	0	Sex-linked gene	Cegan et al., 2010; Matsunaga et al., 2003
SlXY4[2]	0	Sex-linked gene	Atanassov et al., 2001
SlX6a[3]	0	Sex-linked gene	Bergero et al., 2007
SlX6b[3]	0	Sex-linked gene	Bergero et al., 2007
2A10	0	Autosomal gene	Marais et al., 2011
ABCtr	0	Autosomal gene	Marais et al., 2011
ADPGph	0	Autosomal gene	Marais et al., 2011
ATUB-A	0	Autosomal gene	Marais et al., 2011
ClpP3	0	Autosomal gene	Marais et al., 2011
ELF	0	Autosomal gene	Marais et al., 2011
LIP21	0	Autosomal gene	Marais et al., 2011
OxRZn	0	Autosomal gene	Marais et al., 2011
PGK	0	Autosomal gene	Marais et al., 2011
PSIcentII	0	Autosomal gene	Marais et al., 2011

[1] *SlMF1* did not have sufficient coverage (it is a very weakly expressed gene) to study polymorphisms and thus could not be identified as sex-linked.
[2] *SlAP3XY* and *SlXY4* could not be detected as sex-linked because their X and Y copies were assembled into different contigs, which is not surprising, given that these genes have highly divergent X/Y copies (Bergero et al., 2007; Nicolas et al., 2005).
[3] *SlX6a* and *SlX6b* were not identified as sex-linked because they are duplicate genes and were assembled into the same contig, which prevented detection of XY SNPs.

Table B.S4: Analysis of expression patterns in known sex-linked genes. These sex-linked genes have been identified in previous work (see Table S3 (Table B.S3) for references). Data are not available for *SlMF1*, *SlX6a* and *SlX6b* (see Table S3 (Table B.S3) for more details) and these genes are not included in the table. Y/X ratios from this study and from experimental data (Marais et al., 2008) are well correlated (except for *SlCypXY*). We computed the ratio of male expression (X + Y) over female expression (both X-linked copies). In absence of dosage compensation (DC), this ratio should be 1 for genes without Y degeneration (Y/X ratio around 1) and 0.5 for genes with Y degeneration (low Y/X ratios). In presence of DC, this ratio should be 1 for both genes without Y degeneration (Y/X ratio around 1) and with Y degeneration (low Y/X ratios) as in Figure 3. Here we considered that genes with high Y/X ratios do not need DC, genes with low Y/X ratios and male expression (X + Y) over female expression (both X-linked copies) ratios over 0.8 and under 1.2 are consistent with DC. Other genes are considered equivocal.

Genes	Y/X ratio (Illumina data)	Y/X ratio (experimental data)	Male exp (X+Y) / Female exp (2X)	Observations
SlXY3	0.06	0.56	1.70	Equivocal
SlXY7	0.13	0.32	0.87	Consistent with DC
SlXY4[1]	0.45	0.33	0.89	Consistent with DC
SlXY9	0.49	na	0.66	Equivocal
DD44XY	0.61	0.59	0.95	Consistent with DC
SlAP3XY[1]	0.62	na	1.78	Equivocal
SlssXY	0.70	na	1.01	DC not needed
SlXY1	0.83	1.13	0.59	DC not needed
SlCypXY	1.07	0.14	0.68	DC not needed

Table B.S5: Levels of heterozygosity of the X-linked alleles with and without dosage compensation. The level of heterozygosity (number of heterozygous sites/bp) was computed for each Y/X expression ratio category. The fraction of polymorphic sites can vary from one category to another and we also computed the fraction of heterozygous SNPs among the sex-linked SNPs for each Y/X expression ratio category. Only sex-linked contigs in low Y/X ratio categories show dosage compensation. Results are shown for each female individually and averaged for all females.

Y/X ratio	< 0.25	0.25 - 0.5	0.5 - 0.75	0.75 - 1	1 - 1.5	> 1.5
Level of heterozygosity for female U10_34	0.0059	0.0054	0.0034	0.0038	0.0035	0.0029
Level of heterozygosity for female U10_37	0.0054	0.0050	0.0033	0.0036	0.0034	0.0030
Level of heterozygosity for female U10_39	0.0057	0.0055	0.0035	0.0038	0.0036	0.0031
Average level of heterozygosity	0.0057	0.0053	0.0034	0.0037	0.0035	0.0030
% of heterozygous SNPs for female U10_34	44,00%	42,00%	30,00%	32,00%	29,00%	32,00%
% of heterozygous SNPs for female U10_37	41,00%	39,00%	28,00%	30,00%	28,00%	33,00%
% of heterozygous SNPs for female U10_39	43,00%	43,00%	30,00%	32,00%	30,00%	34,00%
Average % of heterozygous SNPs	43,00%	42,00%	29,00%	31,00%	29,00%	33,00%

[1] *SlAP3XY* and *SlXY4* are not among the sex-linked contigs that we identified because X and Y copies are too divergent and were assembled independently (see Table S3 (Table B.S3)). We used *SlAP3X* and *SlX4* from Genbank to map X and Y Illumina reads and estimate expression.

Text S1: Identification and validation of new sex-linked genes

Identification of new sex-linked genes

The *S. latifolia* haploid genome has been estimated to be around 3 Gb and to have a high DNA repeat content (Bernasconi et al., 2009). Because the genome is complex and has not been sequenced, we used RNA-seq -a next-generation transcriptome-sequencing approach- that is increasingly being used in non-laboratory organisms to identify new genes (Bräutigam and Gowik, 2010; Gibbons et al., 2009; Vera et al., 2008) and is ideally suited to study gene expression including that of sex-linked genes (Deng et al., 2011; Lott et al., 2011; Xiong et al., 2010). We used 3 males and 3 females from a *S. latifolia* population that has been inbred for 10 generations. For each individual, RNA extracted from flower buds was tagged and sequenced using Illumina paired-end technology (2 x 100 bp) for ~35 Gb of sequences in total (Table B.S1). Male and female reads were pooled and assembled *de novo* together using a pipeline (described in Material and Methods and Figure B.S1) and we obtained 141,855 contigs with a mean size of 748.5 bp (Table B.S2). From these, we identified sex-linked contigs using segregation analysis.

The *S. latifolia* X and Y chromosomes are expected to include several gene categories: many (if not most) sex-linked genes should derive from the original autosomal pair that gave rise to *S. latifolia* X and Y chromosomes (Bergero et al., 2007; Filatov, 2005; Nicolas et al., 2005), the other genes resulting from translocation from autosomes to the sex chromosomes (one such case is known in *S. latifolia*, see Kaiser et al., 2009) or from chromosome-specific duplications. How many of the *S. latifolia* sex-linked genes (of autosomal origin) have lost their Y copy following X-Y recombination suppression is currently unknown. However, given that *S. latifolia* sex chromosomes have originated recently, we expect many of these to have a functional Y-linked copy or at least a recognizable pseudogenized Y-linked copy, as has been found for the *MROS3* gene (Guttman and Charlesworth, 1998). We thus expect many of the sex-linked genes to have both an X-linked and a Y-linked copy. These genes are sometimes referred to as gametologs (= homologs arising through lack of recombination and subsequent differentiation of sex chromosomes, see García-Moreno and Mindell, 2000). Here we will use the terms X-linked and Y-linked alleles for convenience as our method relies on Single Nucleotide Polymorphism (SNP) detection (see below), although the synonymous divergence between these "alleles" ranges from 5-20% (Bergero et al., 2007; Filatov, 2005; Nicolas et al., 2005) and thus can be significantly higher than the level of polymorphism among autosomal alleles. The overall X-Y divergence (including both synonymous and non-synonymous sites) is, however, low in all known sex-linked genes, the X-linked and Y-linked alleles should thus assemble into a single contig for most sex-linked genes, with perhaps a decreased probability when the synonymous X-Y divergence is close to the upper limit detected so far (20%) and when X-Y divergence is notable at the non-synonymous level. We thus mapped all Illumina reads back onto the contigs (91.4% were successfully mapped, see Table B.S1 and searched for SNPs in the read alignments (see Material and Methods). A

subset of SNPs showed patterns typical of sex-linkage (heterozygous in males and homozygous in females). This way, we found 16,308 sex-linked SNPs in 1736 contigs with at least one sex-linked SNP (Table B.S2). Using a similar approach, two recent studies found ~400 (Chibalina and Filatov, 2011) and ~1800 (Bergero and Charlesworth, 2011) sex-linked genes. Differences in RNA-seq assembly and SNP filtering procedures may explain those different numbers. However, our estimate falls within the range of values from this recent work, and all studies do not differ greatly.

Validation of the new sex-linked genes

Establishing sex-linkage usually requires more than 6 individuals. However, we here used individuals from a highly inbred line whose level of polymorphism is expected to be very low. Indeed, the estimate for the level of polymorphism in our inbred line (= total number of detected SNPs divided by the total number of bp, taking all contigs into account) is 0.0074, and is thus about one order of magnitude lower than what has been reported for autosomal or sex-linked loci in natural populations of *S. latifolia* (Laporte et al., 2005; Qiu et al., 2010). We thus expect SNPs in our contigs to arise mainly because of assembling together X-linked and Y-linked alleles, or recently duplicated genes, but only the former should show sex-linkage patterns. We nevertheless tested the reliability of our inference of sex-linkage in several ways. First, we used known autosomal genes (Marais et al., 2011) to see whether sex-linked SNPs have been wrongly inferred for these, but could not find any for the 10 autosomal genes tested (Table B.S3). Second, we simulated genotypes for 3 males and 3 females for ~40,000 autosomal SNPs using information on polymorphism level and sequencing errors from our data to check whether a pattern of sex-linkage could be obtained by chance for autosomal SNPs (Text S2, Appendix B.1). We found the rate of false sex-linked contigs to be very low (0.02) and as few as 37 sex-linked contigs may be erroneous. We therefore concluded that our inferences of sex-linkage are highly reliable. This is consistent with the very low rates of false positives found by testing experimentally 10 to 18 new sex-linked genes in the recent work of (Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011).

To estimate how many sex-linked contigs we missed with our method, we checked how many of the previously identified sex-linked genes (see list and references in Table B.S3 were among our sex-linked contigs. 42% of these were not found in our set of sex-linked contigs. Further analyses showed, as expected, that our method is less efficient when the X-linked and/or the Y-linked allele have low expression and thus low coverage, because this makes SNP detection unlikely (Figure B.S2) or when X-linked and Y-linked alleles are too divergent and are not combined into one single contig but instead into two distinct contigs (see Table B.S3 footnotes), which prevents inference of sex linkage. The actual number of sex-linked genes in *S. latifolia* could be ~4000 (Bergero and Charlesworth, 2011). Part of the X is pseudoautosomal and still recombines with the Y during male meiosis. Genetic data suggest that the pseudoautosomal region (PAR) may be as large as 30 cM (Charlesworth D., personal communication) and makes up as

much as 1/3 of the X chromosome if we make the very simplistic assumption of similar recombination rates along the X chromosome. Only ~2600 putative genes on the X should then be located in the non-recombining region and could then be identified as sex-linked with our segregation analysis. Given our rate of missed sex-linked genes (see above), we expect to miss ~40% of the true sex-linked genes (i.e ~1040). This is consistent with the fact that about half of the X-specific region harbours genes with maximum (20%) X-Y synonymous divergence, which our method is likely to miss. This also means that we expect to find ~1560 sex-linked genes with our method. This figure relies on a very rough idea of how many sex-liked genes there could be in *S. latifolia* (Bergero and Charlesworth, 2011), but is nevertheless consistent with the actual number of sex-linked contigs that we found. Of course, some of our contigs may be fragments of the same transcripts. However, the size distribution of contigs (Figure B.S3) suggests that many of them are full-length transcripts as expected given the average coverage that we have here (~400X, see Table B.S2).

Text S2: Simulations to estimate the rate of false-positive sex-linked contigs

We ran simulations in order to estimate the rate of wrongly inferred sex-linked contigs. The idea is to simulate autosomal SNPs, then to run our procedure for detecting sex-linked SNPs (see Material and Methods and Text S3, Appendix B.1) on these simulated SNPs, and to check how many SNPs are inferred as being sex-linked when none are expected, which will give the rate of false positives sex-linked SNPs (and false positive sex-linked contigs). We did not use non-sex-linked contigs since this includes a mixture of autosomal contigs and sex-linked contigs undetected by our method and is not a true autosomal dataset. To be as close as the real data as possible, our strategy was to simulate autosomal SNPs with characteristics similar to the SNPs on the sex-linked contigs that we detected. To do this, we simulated autosomal SNPs using all 39,569 polymorphisms in our 1736 sex-linked contigs (sex-linked plus other SNPs) and sequencing errors data provided by FASTERIS.

In order to simulate autosomal SNPs, we used female genotypes as X and autosomes have similar levels of polymorphism in *S. latifolia* (Qiu et al., 2010). We computed the percentage of each genotypes in females using all SNPs from sex-linked contigs (% of females homozygous for the reference allele, % of females homozygous the alternative allele, % of heterozygous females). Six genotypes were randomly sampled from these genotype frequencies and assigned to three males and three females. For each individual, read numbers were obtained from the observed numbers at real SNPs from sex-linked contigs (all 39,569 SNPs were used one after the other). This means that our set of simulated autosomal SNPs takes into account the differences in expression levels that we observe between males and females and also between contigs (some being lowly expressed and

other highly expressed). For heterozygous genotypes, read numbers were drawn from a binomial distribution with each polymorphism having an equal probability to be drawn, assuming they are equally expressed because autosomal. We then added sequencing errors by randomly exchanging alleles among each other using the PhiX error rates provided by FASTERIS. We thus obtained 39,569 simulated autosomal SNPs with, for each individual, the genotype and the read numbers of each allele.

Then, the scripts used to detect XY polymorphisms were run on the simulated autosomal SNP data and the proportion of false sex-linked contigs was computed. The proportion α of false XY SNPs observed in the simulated SNP data was used to compute the probability P_i for each sex-linked contig i to be a false positive, given n_i the number of XY SNPs of the contig and m_i the other SNPs:

$$P_i = \binom{n_i + m_i}{n_i} \alpha^{n_i} (1-\alpha)^{m_i} \tag{B.1}$$

This probability was computed for each identified sex-linked contig and will depend on how polymorphic a contig is (accounting for differences in polymorphism levels between contigs). The expected number of false sex-linked contigs F is the sum of probabilities for each contig to be false:

$$F = \sum_{i=1}^{1736} P_i \tag{B.2}$$

Text S3: SNP detection and filtering

The empirical filter that follows was established after running GATK on known XY genes for training. At least 3 reads of good quality of each polymorphism were required for a male to be considered heterozygous. In case of a highly expressed gene (more than 1000 reads), the threshold was at least 10 reads of good quality for each polymorphism. If one male had 3 or less than 3 reads of good quality we could not infer its genotype reliably but if the other male and female individuals showed a clear sex-linked pattern, the SNP was considered XY all the same. In some cases, homozygous females had a few reads with the Y polymorphism (this observation was also made on known XY genes and can be attributed to sequencing or tag assignment errors). In such cases, the maximum number of reads observed with the Y polymorphism in females was used as a threshold to infer whether males were heterozygous. This wasset to 5X/5Y for 1 Y read in females (20.1% of sex-linked SNPs) and respectively 20X/20Y for <10 Y reads in females (9.2% of sex-linked SNPs), 100X/100Y for <20 Y reads in females (0.3% of sex-linked SNPs) and 200X/200Y for <30 Y read in females (0.008% of sex-linked SNPs) using empirical information from known XY genes analysis. Cases where homozygous females had more than 30 reads with the Y allele were excluded.

B.2 References

Atanassov, I., Delichère, C., Filatov, D. A., Charlesworth, D., Negrutiu, I., and Monéger, F. (2001). Analysis and evolution of two functional Y-linked loci in a plant sex chromosome system. eng. *Molecular Biology and Evolution* 18(12), 2162–2168. ISSN: 0737-4038.

Bergero, R., Charlesworth, D., Filatov, D. A., and Moore, R. C. (2008). Defining regions and rearrangements of the Silene latifolia Y chromosome. eng. *Genetics* 178(4), 2045–2053. ISSN: 0016-6731. DOI: 10.1534/genetics.107.084566.

Bergero, R. and Charlesworth, D. (2011). Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. eng. *Current biology: CB* 21(17), 1470–1474. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.032.

Bergero, R., Forrest, A., Kamau, E., and Charlesworth, D. (2007). Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: evidence from new sex-linked genes. eng. *Genetics* 175(4), 1945–1954. ISSN: 0016-6731. DOI: 10.1534/genetics.106.070110.

Bernasconi, G., Antonovics, J., Biere, A., Charlesworth, D., Delph, L. F., Filatov, D., Giraud, T., Hood, M. E., Marais, G. a. B., McCauley, D., Pannell, J. R., Shykoff, J. A., Vyskot, B., Wolfe, L. M., and Widmer, A. (2009). Silene as a model system in ecology and evolution. eng. *Heredity* 103(1), 5–14. ISSN: 1365-2540. DOI: 10.1038/hdy.2009.34.

Bräutigam, A. and Gowik, U. (2010). What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. eng. *Plant Biology (Stuttgart, Germany)* 12(6), 831–841. ISSN: 1438-8677. DOI: 10.1111/j.1438-8677.2010.00373.x.

Cegan, R., Marais, G. A., Kubekova, H., Blavet, N., Widmer, A., Vyskot, B., Dolezel, J., Safár, J., and Hobza, R. (2010). Structure and evolution of Apetala3, a sex-linked gene in Silene latifolia. eng. *BMC plant biology* 10, 180. ISSN: 1471-2229. DOI: 10.1186/1471-2229-10-180.

Chibalina, M. V. and Filatov, D. A. (2011). Plant Y chromosome degeneration is retarded by haploid purifying selection. eng. *Current biology: CB* 21(17), 1475–1479. ISSN: 1879-0445. DOI: 10.1016/j.cub.2011.07.045.

Delichère, C., Veuskens, J., Hernould, M., Barbacar, N., Mouras, A., Negrutiu, I., and Monéger, F. (1999). SlY1, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein. eng. *The EMBO journal* 18(15), 4169–4179. ISSN: 0261-4189. DOI: 10.1093/emboj/18.15.4169.

Deng, X., Hiatt, J. B., Nguyen, D. K., Ercan, S., Sturgill, D., Hillier, L. W., Schlesinger, F., Davis, C. A., Reinke, V. J., Gingeras, T. R., Shendure, J., Waterston, R. H., Oliver, B., Lieb, J. D., and Disteche, C. M. (2011). Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila

melanogaster. eng. *Nature Genetics* 43(12), 1179–1185. ISSN: 1546-1718. DOI: 10.1038/ng.948.

Filatov, D. A. (2005). Substitution rates in a new Silene latifolia sex-linked gene, SlssX/Y. eng. *Molecular Biology and Evolution* 22(3), 402–408. ISSN: 0737-4038. DOI: 10.1093/molbev/msi003.

García-Moreno, J. and Mindell, D. P. (2000). Rooting a phylogeny with homologous genes on opposite sex chromosomes (gametologs): a case study using avian CHD. eng. *Molecular Biology and Evolution* 17(12), 1826–1832. ISSN: 0737-4038.

Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P., and Rokas, A. (2009). Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. eng. *Molecular Biology and Evolution* 26(12), 2731–2744. ISSN: 1537-1719. DOI: 10.1093/molbev/msp188.

Guttman, D. S. and Charlesworth, D. (1998). An X-linked gene with a degenerate Y-linked homologue in a dioecious plant. eng. *Nature* 393(6682), 263–266. ISSN: 0028-0836. DOI: 10.1038/30492.

Kaiser, V. B., Bergero, R., and Charlesworth, D. (2011). A new plant sex-linked gene with high sequence diversity and possible introgression of the X copy. eng. *Heredity* 106(2), 339–347. ISSN: 1365-2540. DOI: 10.1038/hdy.2010.76.

Kaiser, V. B., Bergero, R., and Charlesworth, D. (2009). Slcyt, a newly identified sex-linked gene, has recently moved onto the X chromosome in Silene latifolia (Caryophyllaceae). eng. *Molecular Biology and Evolution* 26(10), 2343–2351. ISSN: 1537-1719. DOI: 10.1093/molbev/msp141.

Laporte, V., Filatov, D. A., Kamau, E., and Charlesworth, D. (2005). Indirect evidence from DNA sequence diversity for genetic degeneration of the Y-chromosome in dioecious species of the plant Silene: the SlY4/SlX4 and DD44-X/DD44-Y gene pairs. eng. *Journal of Evolutionary Biology* 18(2), 337–347. ISSN: 1010-061X. DOI: 10.1111/j.1420-9101.2004.00833.x.

Lott, S. E., Villalta, J. E., Schroth, G. P., Luo, S., Tonkin, L. A., and Eisen, M. B. (2011). Noncanonical Compensation of Zygotic X Transcription in Early Drosophila melanogaster Development Revealed through Single-Embryo RNA-Seq. *PLoS Biol* 9(2), e1000590. DOI: 10.1371/journal.pbio.1000590.

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. eng. *PloS One* 6(8), e21915. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021915.

Marais, G. A. B., Nicolas, M., Bergero, R., Chambrier, P., Kejnovsky, E., Monéger, F., Hobza, R., Widmer, A., and Charlesworth, D. (2008). Evidence for degeneration of the Y chromosome in the dioecious plant Silene latifolia. eng. *Current biology: CB* 18(7), 545–549. ISSN: 0960-9822. DOI: 10.1016/j.cub.2008.03.023.

Matsunaga, S., Isono, E., Kejnovsky, E., Vyskot, B., Dolezel, J., Kawano, S., and Charlesworth, D. (2003). Duplicative transfer of a MADS box gene to a plant Y chromosome. eng. *Molecular Biology and Evolution* 20(7), 1062–1069. ISSN: 0737-4038. DOI: 10.1093/molbev/msg114.

Matsunaga, S., Lebel-Hardenack, S., Kejnovsky, E., Vyskot, B., Grant, S. R., and Kawano, S. (2005). An anther- and petal-specific gene SlMF1 is a multicopy gene with homologous sequences on sex chromosomes. eng. *Genes & Genetic Systems* 80(6), 395–401. ISSN: 1341-7568.

Moore, R. C., Kozyreva, O., Lebel-Hardenack, S., Siroky, J., Hobza, R., Vyskot, B., and Grant, S. R. (2003). Genetic and functional analysis of DD44, a sex-linked gene from the dioecious plant Silene latifolia, provides clues to early events in sex chromosome evolution. eng. *Genetics* 163(1), 321–334. ISSN: 0016-6731.

Nicolas, M., Marais, G., Hykelova, V., Janousek, B., Laporte, V., Vyskot, B., Mouchiroud, D., Negrutiu, I., Charlesworth, D., and Monéger, F. (2005). A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. eng. *PLoS biology* 3(1), e4. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0030004.

Qiu, S., Bergero, R., Forrest, A., Kaiser, V. B., and Charlesworth, D. (2010). Nucleotide diversity in Silene latifolia autosomal and sex-linked genes. eng. *Proceedings. Biological Sciences / The Royal Society* 277(1698), 3283–3290. ISSN: 1471-2954. DOI: 10.1098/rspb.2010.0606.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., and Marden, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. eng. *Molecular Ecology* 17(7), 1636–1647. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2008.03666.x.

Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Wang, X., Zhang, J., and He, X. (2010). RNA sequencing shows no dosage compensation of the active X-chromosome. eng. *Nature Genetics* 42(12), 1043–1047. ISSN: 1546-1718. DOI: 10.1038/ng.711.

Appendix C

GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice)

This chapter is the result of a Master internship I did from February to June 2010 under the supervision of Sylvain Glémin (Monptellier, France). It aimed at studying the evolutionary forces shaping base composition in the *Oryza* genus (rice).

GC-Biased Gene Conversion and Selection Affect GC Content in the *Oryza* Genus (rice)

Aline Muyle,[1] Laurana Serres-Giardi,[1] Adrienne Ressayre,[2] Juan Escobar,[1] and Sylvain Glémin*,[1]

[1]Institut des Sciences de l'Evolution, UMR 5554 CNRS, Université Montpellier II, France

[2]INRA, UMR de Génétique Végétale, INRA/CNRS/Univ Paris-Sud/AgroParistech, Ferme du Moulon, Gif sur Yvette, France

***Corresponding author:** E-mail: glemin@univ-montp2.fr.

Associate editor: Naoki Takebayashi

Abstract

Base composition varies among and within eukaryote genomes. Although mutational bias and selection have initially been invoked, more recently GC-biased gene conversion (gBGC) has been proposed to play a central role in shaping nucleotide landscapes, especially in yeast, mammals, and birds. gBGC is a kind of meiotic drive in favor of G and C alleles, associated with recombination. Previous studies have also suggested that gBGC could be at work in grass genomes. However, these studies were carried on third codon positions that can undergo selection on codon usage. As most preferred codons end in G or C in grasses, gBGC and selection can be confounded. Here we investigated further the forces that might drive GC content evolution in the rice genus using both coding and noncoding sequences. We found that recombination rates correlate positively with equilibrium GC content and that selfing species (*Oryza sativa* and *O. glaberrima*) have significantly lower equilibrium GC content compared with more outcrossing species. As recombination is less efficient in selfing species, these results suggest that recombination drives GC content. We also detected a positive relationship between expression levels and GC content in third codon positions, suggesting that selection favors codons ending with G or C bases. However, the correlation between GC content and recombination cannot be explained by selection on codon usage alone as it was also observed in noncoding positions. Finally, analyses of polymorphism data ruled out the hypothesis that genomic variation in GC content is due to mutational processes. Our results suggest that both gBGC and selection on codon usage affect GC content in the *Oryza* genus and likely in other grass species.

Key words: codon usage, GC-biased gene conversion, GC content, mating systems, *Oryza*, recombination..

Introduction

Genomic nucleotide landscapes strongly vary among organisms. In eukaryotes, mean GC content ranges from about 20% to 60% and from 30% to 50% in animals and land plants (Lynch 2007). Moreover, several groups show strong heterogeneity of base composition along chromosomes, the so-called isochore structure, initially discovered in vertebrates (for review, see Eyre-Walker and Hurst 2001). For example, the GC content of 100-kb regions in the human genome ranks from 35% to 65%. This structure affects both coding and noncoding sequences and the GC content of a gene is highly correlated to the GC content of its flanking regions. GC content is also associated with several genome features: GC-rich regions tend to have more genes, with shorter introns; GC-rich regions replicate earlier and some classes of transposable elements are only found in regions of particular base composition (Eyre-Walker and Hurst 2001). Resolving the origin and causes of isochores is thus of major importance to understand genomes organization.

Three major kinds of hypotheses have been proposed to explain isochores evolution: selection, mutational biases, and GC-biased gene conversion (hereafter gBGC) (reviewed in Eyre-Walker and Hurst 2001). Selectionist theories argue that high GC content can be selected for by their thermal stability (Bernardi 2000). However, this theory assumes significant differences in thermal stability, hence fitness, between two individuals differing only by a few point mutations even in noncoding sequences. Given the low effective population sizes N_e and hence the low efficacy of natural selection of some groups exhibiting isochores like mammals, such assumption seems unrealistic (reviewed in Duret and Galtier 2009). In addition, this theory fails to explain why selection for thermal stability should only affect some regions of the genome and the occurrence of isochores in cold-blooded vertebrates (Kuraku et al. 2006; Chojnowski et al. 2007). There are, however, other selective forces that can influence base composition. For instance, alternative codon usage in protein-coding genes can be driven by selection in order to increase the speed or accuracy of translation (Akashi 2001). If the preferred codons have a particular base composition, then selection on codon usage can drive GC content of coding sequences (CDS), especially for highly expressed genes.

Mutational theories suggest that heterogeneous mutational biases along genomes could drive GC content. Several mutational biases have been proposed to influence base composition. High GC content was suggested to favor GC mutations, thus driving isochores evolution thanks to a positive feedback loop on base composition (Fryxell and Zuckerkandl 2000). G and C nucleotides were also proposed to be more often misincorporated in early-replicating regions, when the pool of free nucleotides is still GC rich (Eyre-Walker and Hurst 2001). Finally, as recombination

Table 1. Predictions of the Three Hypotheses to Explain GC Content Evolution. GC*: equilibrium GC content (Sueoka 1962)

	gBGC	Selection on Codon Usage*	Mutational Bias
Recombination	**GC* increased with recombination**	**GC3* increased with recombination**	**?**
Mating systems	**GC* higher in outcrossers**	**GC3* higher in outcrossers**	**—**
Expression level	**—**	**GC3* increased with expression**	**—**
DAF spectra	**Increased GC DAF**	**Increased GC3 DAF**	**Similar spectra for AT and GC derived alleles**

*Predictions valid if most preferred codons end in G or C.

was proposed to be mutagenic (Lercher and Hurst 2002; Hellmann et al. 2003), recombination could also affect base composition, provided that recombination affects both mutation rates and bias.

The third hypothesis involves gBGC, a process taking place during recombination (see review in Marais 2003; Duret and Galtier 2009). During meiosis, after double-strand break, single-stranded DNA invades the homologous sequence to form a heteroduplex. When parental alleles are different, the mismatch created in the heteroduplex is repaired by changing one of the nucleotides. Experimental results in yeast and human culture cell showed that this repair is biased toward GC alleles (Birdsell 2002). This was recently confirmed by a broad survey of segregation patterns in individual meioses in yeast (Mancera et al. 2008). Biased gene conversion can also arise from a bias in initiation of double-strand breaks at meiosis, but this process is less documented and not necessarily GC biased (Marais 2003). From the point of view of population genetics, gBGC is a kind of meiotic drive that mimics selection in favor of G and C alleles and increases their probability of fixation (Nagylaki 1983). However, gBGC can be viewed as a neutral process because it does not depend on the fitness effect of alleles on the individuals that carry them. As the intensity of gBGC directly depends on recombination in heterozygotes, it will be more efficient in large outcrossing populations and in highly recombining regions along genomes.

Selection and gBGC can be distinguished from mutational bias using polymorphism data (table 1). Both gBGC and selection increase the probability of fixation of W → S mutations (weak to strong: from A or T to G or C) and decrease it for S → W mutations (strong to weak: from G or C to A or T), whereas mutational bias does not affect probabilities of fixation. Under selection or gBGC, W → S mutations are thus expected to segregate at higher frequencies than S → W mutations. gBGC can be distinguished from selection because it affects both coding and noncoding regions, whereas selection on codon usage is mainly restricted to third codon positions (table 1).

There is evidence for genomic effect of gBGC, especially in yeast (Birdsell 2002), mammals (Duret and Arndt 2008; Duret and Galtier 2009), and birds (Webster et al. 2006), but also in turtle, nematode, Drosophila, paramecium, and green algae, supporting the gBGC hypothesis (reviewed in Duret and Galtier 2009). In these organisms, correlations between recombination rate (or proxies of recombination) and GC content were observed for both coding and noncoding regions, ruling out selection on codon usage as the sole cause. More precisely, in human, the better correlation with the equilibrium GC content than the actual GC content supports the fact that recombination drives GC content, not the reverse (Meunier and Duret 2004; Duret and Arndt 2008). Polymorphism analyses in humans also showed an excess of derived G and C alleles in high frequency compared with A and T alleles, ruling out the mutation bias hypothesis. Moreover, this excess is greater in highly recombining single-nucleotide polymorphisms (SNPs) compared with low recombining ones (Spencer et al. 2006).

Among plants, grasses present a peculiar genomic structure: they exhibit a remarkable GC content heterogeneity (Carels and Bernardi 2000; Wang et al. 2004) that parallels the mammalian genome organization (e.g., regions with high GC content having numerous genes with short introns). More strikingly, GC3 distribution is strongly bimodal, as observed in rice (Shi et al. 2006). However, the causes of this high and heterogeneous GC content are still unclear. Several studies suggested that selection on codon usage occurs in rice as codon bias is related to gene expression levels (Guo et al. 2007; Mukhopadhyay et al. 2007), and most preferred codons end in G or C (Wang and Roossinck 2006). However, codon usage alone can hardly explain the overall high GC content and the strong heterogeneity in base composition observed in grasses. This is because first and second codon positions, as well as introns, are also affected (Carels and Bernardi 2000; Shi et al. 2006) and lowly expressed genes also show a strongly bimodal GC3 distribution (Mukhopadhyay et al. 2007). Other authors have thus invoked a general mutational bias toward GC to explain the isochore structure in rice and more generally in grasses (Wang et al. 2004; Wang and Hickey 2007). However, these studies are mainly based on the comparison of *Arabidopsis* and rice genes, making difficult to infer the underlying processes. Alternatively, comparative and phylogenetic approaches within grasses showed that GC content and the equilibrium GC content correlated with recombination and mating system, suggesting that gBGC could be involved (Glémin et al. 2006; Haudry et al. 2008; Escobar et al. 2010). gBGC is also a strong alternative candidate to mutational bias to explain the global shift in GC content in grasses.

The aim of this study is to assess which evolutionary forces have a significant impact on GC content in grasses (Poaceae). We chose to focus on the rice genus (*Oryza*) because it is currently one of the few grasses for which genomic data, genetic and physical maps, and polymorphism data are available. Furthermore, the existence of selfers and outcrossers among the *Oryza* genus allows for additional tests: selfing species show high levels of homozygosity such that gBGC should be inefficient. Selection is also predicted to be less efficient in selfing species because the effective size and effective recombination are reduced (Charlesworth 1992; Glémin 2007). To test the three

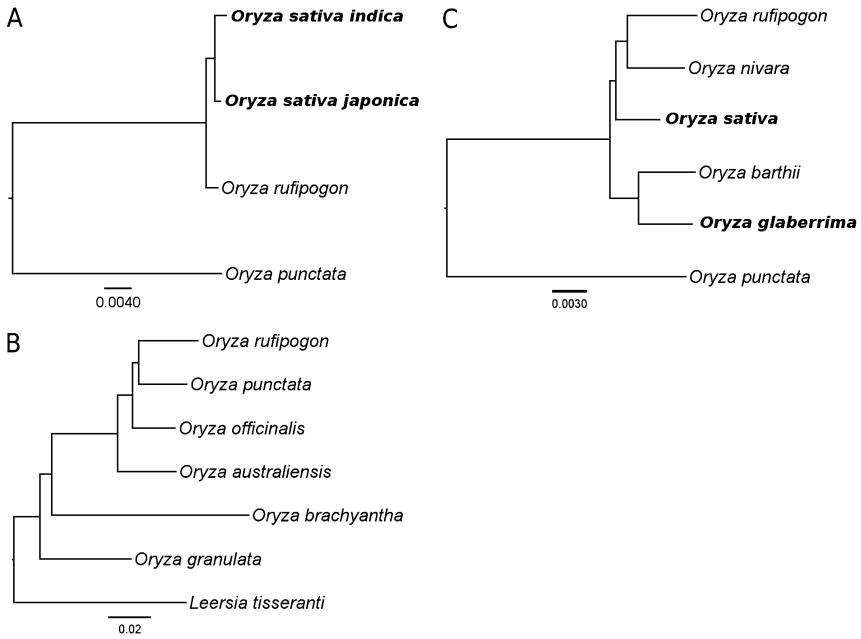


Fig. 1. Trees used in phylogenetic analyses. Bolded names indicate self-fertilizing species. (*A*) Data set 1 (132 genes, 166,648 bp). (*B*) Data set 2 (102 genes, 114,868 bp). (*C*) Data set 3 (307 genes, 136,686 bp). Branch lengths are given in units of per site substitution rate.

hypotheses mentioned above, we studied correlations between GC content, the equilibrium GC content, and recombination rates along the genomes of several *Oryza* species and compared the equilibrium GC content between selfing and outcrossing species. To contrast gBGC and selection hypotheses, we used intronic sequences as a neutral control. We also studied correlations between GC equilibrium at third codon position and expression level to assess whether selection was at work on codon usage. To discriminate between gBGC/selection and mutational biases, we studied fixation probabilities with polymorphism data.

Materials and Methods

Test of the Effect of Recombination and Expression on GC Content

Data Sets

We used two data sets published by Zou et al. (2008) and kindly provided by Fu-Min Zhang. They included exon and intron sequences. The first one contains four taxa, the two cultivated rices, *Oryza sativa* ssp. *japonica* and *O. sativa* ssp. *indica*, their wild ancestor, *O. rufipogon*, and an outgroup *O. punctata* (fig. 1*A*). It is composed of 132 genes scattered along the genome for a total length of 166,648 bp. This data set was used to study correlations between recombination rates, expression levels, and equilibrium GC content at small phylogenetic scale. The second data set gives an overview of the rice genus and consists of seven species covering major clades of the genus, including the outgroup *Leersia tisseranti* (fig. 1*B*). This data set contains 102 genes (most of which are in common with the first data set) and has a total length of 114,868 bp. This data set was used to study correlations between recombination rates, expression levels, and equilibrium GC content for each gene at a larger phylogenetic scale. For the two data sets, sequences were aligned with Muscle v3.8.31 (Edgar 2004) and the resulting alignments were manually checked. These data sets were already annotated by the authors (Zou et al. 2008), and we directly used these annotations (coding vs. noncoding, positions 1, 2, and 3).

Expression Level Data

We retrieved expression level data from the Rice Genome Annotation Project Web site (http://rice.plantbiology.msu.edu/expression_anatomy.shtml). To get the number of EST counts, we used locus names directly provided in the data sets. EST counts were used as the proxy for expression level.

Recombination Rate Estimates

To estimate recombination rates, we built a genetic versus physical distances map (Marey's map) using the 3,267 markers of the latest high density rice genetic map available on the Rice Genome Program Web site (http://rgp.dna.affrc.go.jp/E/publicdata/geneticmap2000/index.html). We then retrieved the physical position of these markers from the Rice Genome Annotation Project Web site (http://rice.plantbiology.msu.edu/annotation_insilico_genmarkers.shtml) and excluded those with multiple locations. We got 2114 markers. On the RGAP Web site, some markers are redundant because both the 5′ and the 3′ ends of the markers have been physically mapped. In such a case, we only kept one marker, which reduces the data set to 1219 markers. For each marker, we attributed the middle position of the marker (or the average of the middle position for redundant markers) as its physical location. After visual inspection of the Marey's maps, we excluded 17 markers with anomalous positions. We thus finally obtained 1202 markers along the rice genome

(see supplementary table S1, Supplementary Material online). Recombination rates were computed with the MareyMap program (Rezvoy et al. 2007). To fit the data, we first used the loess function that locally adjusts a polynomial second-degree curve, using a weight attributed to each marker depending on how far it is from the center of the window (Rezvoy et al. 2007). The size of the window is defined as a percentage of the total number of markers. We used 20% to get rather smooth recombination rate curves. We also used the cubic splines method with the cross-validation option (for details, see Rezvoy et al. 2007), which captures variation in recombination rates at a more local scale (but it is also more sensible to imprecision and errors in the genetic map). We then estimated the local recombination rate of each marker from the fitted curves using the physical location of each gene according to the release 6.1 of MSU Rice Genome Annotation. Marey's maps and recombination maps are available on supplementary figure S1 (Supplementary Material online).

GC Estimates and Analyses

GC content was computed for all genes. To investigate the forces affecting the evolution of GC content, we also computed the equilibrium GC content (hereafter GC*) reached by a sequence if substitution rates remain constant over time (Sueoka 1962):

$$GC^* = \frac{U_{W\to S}}{U_{W\to S} + U_{S\to W}}, \quad (1)$$

where $U_{W\to S}$ is the W to S substitution rate and $U_{S\to W}$ the reverse one. We used the BPPML program (Dutheil and Boussau 2008) which uses a maximum-likelihood approach to estimate GC* along a given phylogeny. It allows nonstationarity (ancestral and current GC content can differ) and nonhomogeneity (branches can have distinct GC*), which allows to test whether branches underwent similar evolution. We used the tree topologies given by Zou et al. (2008) (see fig. 1). To get a reasonable starting tree for ML search, we re-estimated branch lengths using the HKY model using PhyML v3.0 (Guindon and Gascuel 2003). In BPPML, we used the T92 model of sequence evolution (Tamura 1992) with a uniform + invariant distribution for rate evolution. This distribution was preferred to the more classical gamma distribution because data sets contain few variable sites and the gamma shape parameter was wrongly estimated for some genes. In all cases, the transition/transversion ratio was shared by all branches, whereas GC* was allowed to vary between branches or groups of branches. In nonstationary models, the root position is usually badly estimated and we preferred to fix it as the one obtained with PhyML. For each gene separately, we computed GC* for all sites (coding and noncoding), GC* at third codon position (GC3*), and at noncoding positions alone (GC* noncoding).

For the first data set, GC* were computed using *O. punctata* as outgroup and by grouping *O. rufipogon*, *O. sativa indica*, and *O. sativa japonica* together to get sufficient power as individual branches are short (fig. 1*A*). With the second data set, we investigated whether GC* evolved at a larger phylogenetic scale. We used *L. tisseranti* as outgroup and estimated either one GC* for the whole *Oryza* clade or one GC* for each species (terminal branch) separately: *O. rufipogon*, *O. punctata*, *O. officinalis*, *O. australiensis*, *O. brachyantha*, and *O. granulata* (fig. 1*B*). For these analyses, we used the species phylogeny and, as a control, the gene phylogenies provided in the original publication.

GC and GC* were arcsin squared root transformed and regressed against recombination rates and expression levels. Genes for which GC* = 0 or 1 were excluded of these analyses. Kendall nonparametric correlations were also performed. Note that recombination rates and expression levels were only available for *O. sativa* and were used for all other species.

Test of the Effect of Mating System on GC Content

Data Set

To test for the effect of mating system on the GC content dynamic, we used the data set published by Cranston et al. (2009). We directly used the alignment provided in their supplementary material after having annotated it (see below, Supplementary Material online). It consists of CDS and flanking noncoding sequences (introns, untranslated regions) corresponding to 307 genes and has a total length of 136,686 bp. It includes the two cultivated and mostly self-fertilizing species *O. sativa* (Asian rice) and *O. glaberrima* (African rice) and four wild species (*O. rufipogon*, *O. nivara*, *O. barthii* and *O. punctata*) (fig. 1*C*). *Oryza barthii*, *O. nivara*, and *O. rufipogon* have mixed-mating system with outcrossing rates ranging from 10 to 50% (Sweeney and McCouch 2007), and *O. punctata* could also exhibit significant level of outcrossing, at least the perennial forms (Sano 1980). For concision, all these four species were considered as outcrossers thereafter. As evolution is mostly unidirectional from outcrossing to selfing (Takebayashi and Morrell 2001; Igic et al. 2006), all internal branches were assumed to be outcrossing.

Gene Annotation

We annotated this data set by blasting sequences on the complete rice genome (version 6.1) provided by the Rice Genome Annotation Project Web site. For this, we used BlastALL (Altschul et al. 1990) (options: −p BlastN − e 1×10^{-5}) and PERL and BASH homemade scripts. We only kept hits with 100% identity. Sequence fragments that had no match on the whole CDS data were considered as noncoding. Reading frames of the CDS fragments were determined using the distance that separated their own start from the real start (ATG) of the whole-gene CDS. This distance was obtained with a blast of the CDS fragments on the whole-gene CDS. We confirmed the accuracy of this method by checking for stop codons along the fragments of CDS once they were framed.

Statistical Analyses

We used the ML framework described above to estimate GC* and compared various nested hierarchical models of sequence evolution using likelihood ratio tests (LRTs). Note that we were not interested in comparing GC* among genes but among branches. We thus concatenated all sequences (or all noncoding and all third codon positions) to get sufficient power. We could thus infer whether selfing versus outcrossing branches, and internal versus external

branches, underwent significantly different evolution (table 1). Under gBGC or selection, GC* is expected to be higher on internal than on external branches because polymorphic mutations can be counted as substitutions. As $W \rightarrow S$ mutations have a higher probability of fixation than $S \rightarrow W$ ones, $U_{W \rightarrow S}$is underestimated and $U_{S \rightarrow W}$ overestimated. In addition to the null model, M0, with homogeneous GC*, we run three nonhomogenous models. In model M1a, selfing and outcrossing branches have a different GC*. In model M1b, internal and external branches have a different GC*. In model M2, external selfing, external outcrossing, and internal outcrossing branches have different GC* (note that there are no internal selfing branches). To test the effect of internal *versus* external branches, model M2 was compared with model M1a (LRT with 1 degree of freedom). To test the effect of mating systems, model M2 was compared with model M1b (LRT with 1 degree of freedom).

Analyses of Polymorphism Data

Data Set

We used the data set published by Caicedo et al. (2007) who analyzed polymorphism in 111 gene fragments in the wild rice species *O. rufipogon* (21 individuals) and the two cultivated rice subspecies, *O. sativa* ssp. *japonica* (46 individuals) and *O. sativa* ssp. *indica* (26 individuals). *Oryza barthii* and *O. meridionalis*, their sister species, were used as outgroups to orientate polymorphism as in Caicedo et al. (2007). We retrieved the sequences corresponding to a total length of 55,562 bp from GeneBank and aligned them with Muscle v3.8.31 (Edgar 2004). All the resulting alignments were manually checked.

Determination of the Derived Allele Frequency Spectrum

We used the derived allele frequency (DAF) spectrum to quantify the strength of gBGC or selection: $B = 4N_e b$, where N_e is the effective population size and b the gBGC or selection coefficient. We analyzed separately the *O. rufipogon*, *O. s. indica*, and *O. s. japonica* data sets. To build DAF spectra, we inferred the ancestral states of SNPs using the sequences of *O. barthii* or *O. meridionalis* as outgroups, as in Caicedo et al. (2007). Because CpG hypermutability can lead to orientation problems and to spurious signatures of selection and/or gBGC (Hernandez et al. 2007), we used the method developed by Duret and Arndt (2008) that takes CpG into account to infer ancestral states, thanks to a program kindly provided by Peter Arndt. As this method requires rather large alignments, we concatenated all genes. Then we built a triple sequence alignment including an outgroup and two sequences corresponding to the two polymorphic alleles randomly assigned to one sequence or to the other. We then obtained a probability distribution for the identity of the ancestral state for each position in the alignment. SNP counts were thus weighted by these probabilities. For instance, consider a site with 12 A and four G with the ancestral state being A with a probability 0.9 and G with a probability 0.1. This site counts as 0.9 W $\rightarrow$ S SNP with DAF = 4/16 and as 0.1 S $\rightarrow$ W SNP with DAF = 12/16. We ignored missing data when it occurred in polymorphic sites, so that the number of sampled sequences, n, varied from site to site. We thus transformed our data by resampling polymorphic sites in order to have the same sample size, n', for all sites. As in Caicedo et al. (2007), we used $n' = 16$. The few SNPs with $n < 16$ were removed. The frequency of the derived allele in the reduced sample follows a hypergeometric distribution. Given k the frequency of the derived allele in the original sample of size n, the probability that i copies were observed in the reduced sample of size n' is (Hernandez et al. 2007):

$$P(i/n'|k/n) = \frac{C_k^i C_{n-k}^{n'-i}}{C_n^{n'}}. \quad (2)$$

Fit of a gBGC/Selection Population Genetics Model

We fitted a population genetic model to the DAF spectrum to estimate B. To take demography into account, we adapted the model developed by Eyre-Walker et al. (2006) to estimate selection against deleterious mutations. Although Eyre-Walker et al. (2006) distinguished the nonsynonymous and the synonymous spectra, we distinguished the W $\rightarrow$ S, the S $\rightarrow$ W, and the W $\leftrightarrow$ W + S $\leftrightarrow$ S spectra, the latter two being used as a neutral reference. Polymorphic sites were then divided into these three groups. The probability of observing k_i SNPs having i derived alleles out of n follows a Poisson distribution, $P(\mu,k_i)$, with mean:

$$\mu_{\text{neutral}} = \frac{4N_e \nu L r_i}{i}, \quad (3a)$$

for neutral W $\leftrightarrow$ W and S $\leftrightarrow$ S mutations,

$$\mu_{W \rightarrow S}(i) = 2N_e u L (1 - p_{GC}) r_i \int_0^1 C_n^i x^i (1-x)^{n-i} H(b,x) dx, \quad (3b)$$

for W $\rightarrow$ S mutations, and

$$\mu_{S \rightarrow W}(i) = 2N_e \lambda u L p_{GC} r_i \int_0^1 C_n^i x^i (1-x)^{n-i} H(-b,x) dx, \quad (3c)$$

for S $\rightarrow$ W mutations, where

$$H(b,x) = 2\frac{1 - e^{-4N_e b(1-x)}}{(1 - e^{-4N_e b})(1-x)} \quad (3c)$$

is the time that a mutation with gBGC coefficient b spends between x and $x + dx$. The first term within the integral corresponds to the binomial sampling of i alleles over n given their frequency x. N_e is the effective population size, ν the mutation rate from W to W and from S to S mutations, u the mutation rate from W to S, $u\lambda$ the mutation rate from S to W, λ being the mutational bias toward AT, L the sequence length, and p_{GC} the GC content of the sequence. We assumed that p_{GC} is constant, that is the ongoing substitution process does not significantly affect base composition at the polymorphism time scale. The r_i have been introduced by Eyre-Walker et al. (2006) to take demography and/or population structure (and sampling) into account. There is one r_i for each SNP class,

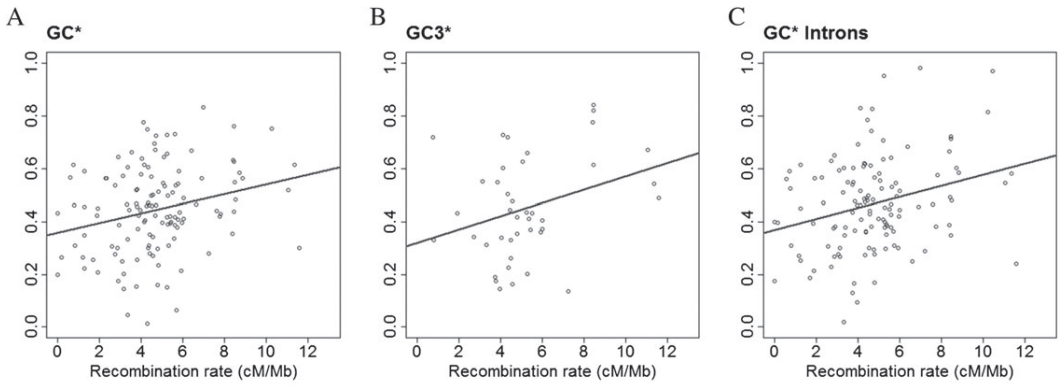


FIG. 2. Correlations between recombination rates, estimated by the loess method, and GC*. Each point corresponds to one gene in data set 1. Genes with GC* = 0 or 1 have been excluded. Regression lines are plotted when significant. (*A*) Total GC*. $n = 127$, $R^2 = 0.062$, *P* value = 0.005, $\tau_{Kendall} = 0.155$, *P* value = 0.010 for the loess method, and $n = 130$, $R^2 = 0.062$, *P* value = 0.004, $\tau_{Kendall} = 0.156$, *P* value = 0.009 for the cubic splines method. (*B*) GC3*. $n = 41$, $R^2 = 0.099$, *P* value = 0.044, $\tau_{Kendall} = 0.177$, *P* value = 0.103 for the loess method, and $n = 127$, $R^2 = 0.077$, *P* value = 0.079, $\tau_{Kendall} = 0.170$, *P* value = 0.118 for the cubic splines method. C: GC* in noncoding positions. $n = 124$, $R^2 = 0.080$, *P* value = 0.001, $\tau_{Kendall} = 0.146$, *P* value = 0.016 for the loess method, and $n = 41$, $R^2 = 0.054$, *P* value = 0.008, $\tau_{Kendall} = 0.142$, *P* value = 0.018 for the cubic splines method.

corresponding to the deviation from the standard equilibrium model relative to the singleton class for which r_1 is set to one. This parameterization makes the assumption that deviation is the same for neutral and selected sites. Although this is not necessarily true, Eyre-Walker et al. (2006) showed this approximation is robust. This is especially true in our case where low gBGC/selection coefficients are expected. To simplify the analysis and to get sufficient power, we assumed that the parameters are shared by all SNPs. Assuming independence between SNPs, the likelihood of the model can thus be written down as:

$$\Gamma = \prod_{i=1}^{n-1} P(\mu_{\text{neutral}}, k_i^{W\leftrightarrow W, S\leftrightarrow S}) P(\mu_{W\to S}, k_i^{W\to S}) P(\mu_{S\to W}, k_i^{S\to W}). \quad (4)$$

The log-likelihood was maximized by using the FindMaximum function of the Mathematica software v6 (Wolfram 1996) with the MaxIteration option set to 8. $4N_e u$, $4N_e v$, λ, and the r_i were constrained to be positive, whereas $B = 4N_e b$ was let free. *B* was estimated for all SNPs, for SNPs in third codon position, and for SNPs in noncoding positions. Then we used an LRT with one degree of freedom to compare the gBGC model for which *B* is estimated to the neutral model for which *B* is fixed to zero. Because they were sampled in several gene fragments in which linkage disequilibrium can occur (especially in the *O. sativa* datasets because of selfing), not all SNPs are fully independent. r_i coefficients potentially capture the effects of the sampling scheme; however, *P* values given by LRT must be viewed with caution (especially for *O. sativa*). We then determined the expected GC* given the estimates of *B* and λ, using the so-called Li-Bulmer equation, initially given for codon usage (Bulmer 1991):

$$GC^* = \frac{1}{1 + \lambda e^{-B}}. \quad (5)$$

Results

Analyses on a Small Phylogenetic Scale

We first carried out analyses at a small phylogenetic scale, that is, for the species closely related to *O. sativa*, the cultivated Asian rice species. We computed GC* for each of the 132 genes using the species tree provided in figure 1*A* (data available on supplementary table S2, Supplementary Material online). We first investigated the relationship between GC*, current GC content, and recombination rates by regression and its strength using the Kendall nonparametric correlation coefficient. We found a weak but significant correlation between current GC content and recombination rates estimated both with the loess (R^2 of the regression = 0.064, *P* value = 0.003, $\tau_{Kendall} = 0.145$, *P* value = 0.0213) and the cubic splines methods ($R^2 = 0.057$, *P* value = 0.005, $\tau_{Kendall} = 0.131$, *P* value = 0.022). We also found a correlation between GC* and recombination rates ($R^2 = 0.062$, *P* value = 0.005, $\tau_{Kendall} = 0.155$, *P* value = 0.010 and see figure 2*A* for the loess method, and $R^2 = 0.062$, *P* value = 0.004, $\tau_{Kendall} = 0.156$, *P* value = 0.009 for cubic splines method). The correlation between GC* and recombination was still significant when looking only at noncoding positions ($R^2 = 0.080$, *P* value = 0.001, $\tau_{Kendall} = 0.146$, *P* value = 0.016 and see figure 2*C* for the loess method, and $R^2 = 0.054$, *P* value = 0.008, $\tau_{Kendall} = 0.142$, *P* value = 0.018 for the cubic splines method) and only marginally significant at third codon positions ($R^2 = 0.099$, *P* value = 0.044, $\tau_{Kendall} = 0.177$, *P* value = 0.103 and see figure 2*B* for the loess method, and $R^2 = 0.077$, *P* value = 0.079, $\tau_{Kendall} = 0.170$, *P* value = 0.118 for the cubic splines method), probably because of lower statistical power ($n = 41$ for third positions vs. $n = 124$ for noncoding positions).

We also tested for the effect of recombination on GC* through the comparison of outcrossing and selfing lineages using the third data set (fig. 1*C*). As there are only terminal selfing branches while there are both internal and terminal

Table 2. Estimates of GC* on the Branches of the Tree Corresponding to Figure 2C and *P* values of the Nested Hierarchical Models of Sequence Evolution Likelihood-Ratio Tests.

		GCtotal*	GC3*	GCnoncoding*
	Number of sites	**136,685**	**41,730**	**12,772**
GC* on branches	**External selfing**	**0.35**	**0.34**	**0.27**
	External outcrossing	**0.39**	**0.37**	**0.32**
	Internal outcrossing	**0.49**	**0.47**	**0.45**
***P* values**	**Selfing vs. outcrossing**	***<10^{-6}***	***0.003***	***0.04***
	External selfing vs. external outcrossing	***<10^{-9}***	***<10^{-4}***	***0.03***
	External vs. internal	***<10^{-4}***	***0.04***	***0.06***
	External outcrossing vs. internal outcrossing	***<10^{-7}***	***<10^{-3}***	***0.04***

Significant values at the 5% level are italic.

outcrossing branches, we compared external selfing and external outcrossing branches to test for the effect of mating system. Similarly, we compared internal outcrossing and external outcrossing branches to test for the possible effect of branch positions. External selfing branches have a significantly lower GC* compared with external outcrossing branches, for all positions, third codon positions, and noncoding positions (table 2). Likewise, for all subsets, outcrossing external branches have a lower GC* compared with internal outcrossing branches, as expected if either gBGC or selection on codon usage is driving GC content (table 2).

To test for possible effects of selection on third codon position, we investigated the relationship between GC3* and noncoding GC* with the expression level using data set 1. A positive correlation between GC3* and the expression level was detected ($R^2 = 0.209$, *P* value = 0.0044, $\tau_{Kendall} = 0.233$, *P* value = 0.044, fig. 3*A*), whereas the expression level and noncoding GC* were not correlated (fig. 3*B*).

Analyses on a Larger Phylogenetic Scale

We then explored the relationship between recombination, expression, and GC content at the whole *Oryza* genus scale. We computed GC* for each gene of the second data set, using the species tree provided in fig. 1*B* and allowing one GC* for all the *Oryza* clade and one for the outgroup, *L. tisserandi* (data available on supplementary table S2, Supplementary Material online). We reasoned that if the same processes have been acting to modify the GC content at this phylogenetic scale, GC* would be better estimated because of the higher number of substitutions in this data set compared with the first data set. We observed no correlation between GC* and recombination rates nor between GC3* and expression levels (data not shown). We controlled that these results were not affected by the topology of the tree by using gene trees rather than the species tree.

In order to assess at which point of the phylogeny the correlation between GC* and recombination or expression is lost, we computed GC* and GC3* for each gene and for each species of the second data set and studied correlations with recombination and expression. GC* are correlated with recombination rates in *O. rufipogon* but not in other species, except in one case (*O. brachyanta*, table 3). We observed similar results for correlations between expression levels and GC3* (table 3).

Polymorphism Data Analyses

From the Caicedo et al. (2007) data set, we obtained W → S, S → W, and neutral DAF spectra for wild and cultivated rices. In *O. rufipogon*, we observed that W → S SNPs tend to segregate at higher frequencies than S

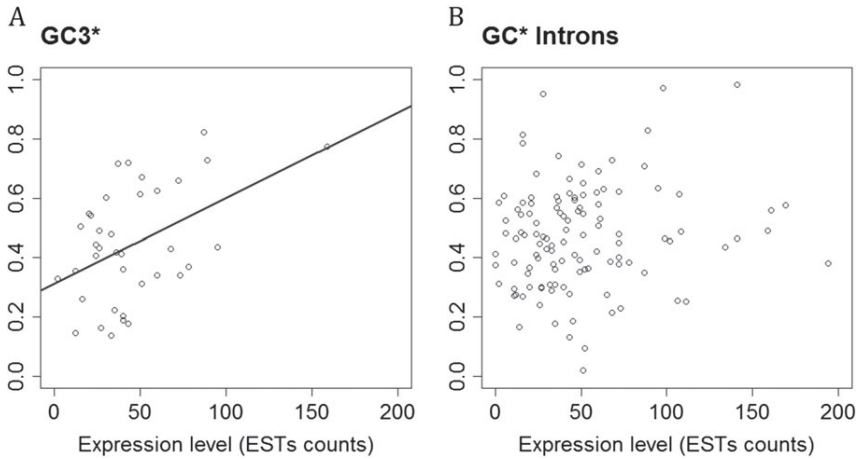


Fig. 3. Correlations between raw expression level (EST counts) and GC*. Each point corresponds to one gene in data set 1. Genes with GC* = 0 or 1 have been excluded. Regression lines are plotted when significant. (*A*) GC3*. $n = 41$, $R^2 = 0.209$, *P* value = 0.0044, $\tau_{Kendall} = 0.233$, *P* value = 0.044. (*B*) GC* in noncoding positions. $n = 127$, $R^2 = 0.0002$, *P* value = 0.8805, $\tau_{Kendall} = 0.070$, *P* value = 0.4443.

Table 3. Correlation between Recombination and GC* and Expression and GC3* for Each Branch of the Tree Corresponding to figure 2*B*.

	Correlation between GC* and Recombination				Correlation between GC3* and Expression			
	Linear Regression		Kendall Correlation		Linear Regression		Kendall Correlation	
	R^2	*P* Value	τ	*P* Value	R^2	*P* Value	τ	*P* Value
Oryza rufipogon	***0.049***	***0.028***	***0.138***	***0.043***	***0.101***	***0.009***	***0.227***	***0.007***
O. punctata	**0.017**	**0.201**	**−0.096**	**0.166**	**0.001**	**0.780**	**0.039**	**0.650**
O. officinalis	**0.006**	**0.460**	**−0.050**	**0.472**	**0.034**	**0.150**	**−0.074**	**0.390**
O. australiensis	**0.018**	**0.191**	**−0.119**	**0.082**	**0.002**	**0.750**	**−0.019**	**0.820**
O. brachyantha	**0.033**	**0.072**	***0.149***	***0.029***	**0.015**	**0.240**	**0.072**	**0.310**
O. granulata	**0.010**	**0.320**	**0.112**	**0.104**	**0.007**	**0.460**	**0.070**	**0.360**

Significant values at the 5% level are in italic type.

→ W SNPs, especially in third codon positions (fig. 4 and supplementary fig S2, Supplementary Material online). For all positions, the mean frequency of the derived allele is 0.28 for S → W SNPs versus 0.31 for W → S SNPs when SNPs were orientated using *O. meridionalis* as outgroup, and 0.29 and 0.34, respectively, when orientated using *O. barthii*. These differences are, however, not significant (sign-rank test *P* value = 0.474 and *P* value = 0.090, respectively). In *O. sativa*, patterns are less biased and less clear because of the strong impact of demography (see Caicedo et al. 2007) and because of the lower number of SNPs (see supplementary table S2, Supplementary Material online).

To better capture the potential differences between the two spectra (and not only differences between mean), we fitted a gBGC/selection model on the spectra and compared it with a neutral model. In *O. rufipogon*, the estimated *B* coefficient is positive in most cases. However, the gBGC/selection model was significantly better than the neutral model only for third codon positions (table 4). In contrast, in *O. sativa*, *B* is lower and never significant. In all cases, the mutational bias, λ, is higher than 1, showing a global mutational bias toward A and T bases. In *O. rufipogon*, λ is consistently similar for all sites, with values around 3. In *O. sativa*, estimates are lower and less stable. Using *B* and λ estimates, we computed the predicted GC* (table 5). Results suggest that GC is decreasing in introns in both *O. rufipogon* and *O. sativa*, whereas in third codon positions it is increasing in *O. rufipogon* and seems to be at equilibrium in *O. sativa*.

Discussion

Among plants, grasses show rich and highly heterogeneous GC content along their genomes (Carels and Bernardi 2000), which is reminiscent of the mammalian isochores. Although gBGC is thought to play a crucial role in shaping mammalian nucleotide landscape (reviewed in Duret and Galtier 2009), which forces act on grass genomes is still a controversial issue (Wang et al. 2004; Glémin et al. 2006; Guo et al. 2007; Mukhopadhyay et al. 2007; Wang and Hickey 2007; Haudry et al. 2008; Escobar et al. 2010). Here we tested in the *Oryza* genus the various

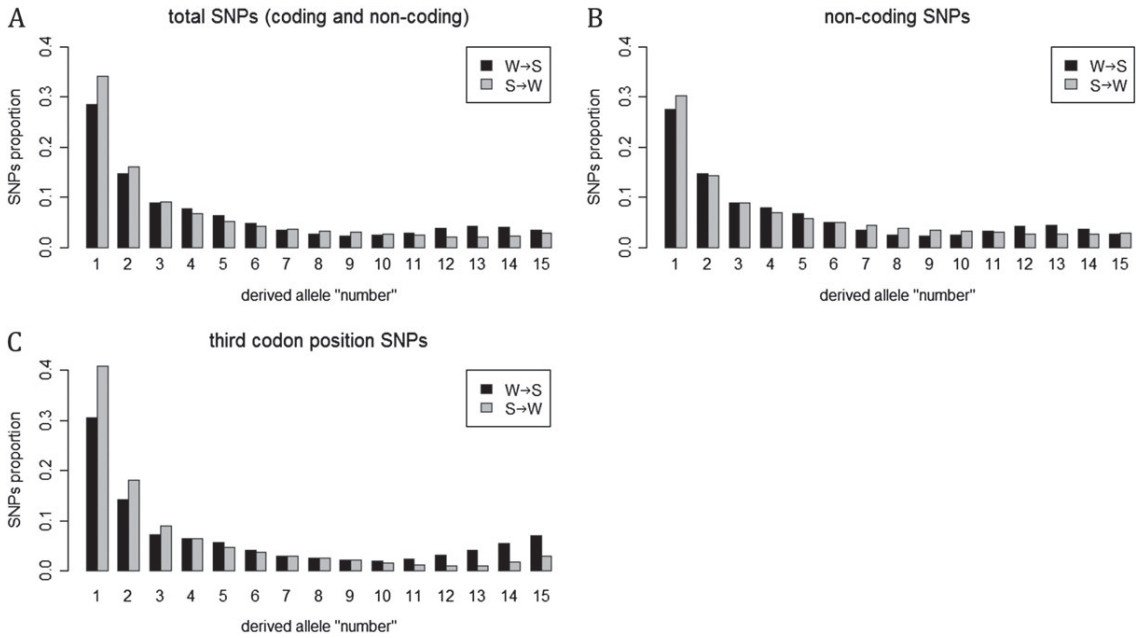


Fig. 4. DAF spectra of *Oryza rufipogon* SNPs orientated with *O. barthii*. (*A*) Total SNPs (coding and noncoding). (*B*) Noncoding SNPs. (*C*) Third codon position SNPs. To cope with SNPs with missing data, spectra have been computed for a subsample of size $n = 16$ as in Caicedo et al. (2007).

Table 4. Estimations of the Intensity of gBGC/selection ($B = 4N_e b$) and mutation bias, λ, from DAF spectra. *n*: number of W → S and S → W SNPs used, $n_{(WW + SS)}$: number of "neutral" SNPs used (see main text for details).

		GCtotal				GC3				GCnoncoding				
Species	Outgroup	*B*	λ	*P* value	*n*	*B*	λ	*P* value	*n*	*B*	λ	*P* value	*n*	$n_{(WW+SS)}$
Oryza rufipogon	*O. meridionalis*	0.463	2.934	0.225	517	*1.677*	2.932	*0.030*	112	−0.001	3.234	0.999	326	82
	O. barthii	0.672	3.014	0.062	536	*1.864*	2.774	*0.017*	114	0.273	3.512	0.535	345	94
O. sativa japonica	*O. meridionalis*	−0.018	2.263	0.967	224	0.341	1.206	0.721	47	0.219	3.388	0.682	143	30
	O. barthii	0.055	2.605	0.897	233	1.176	2.133	0.205	49	−0.107	3.290	0.833	148	36
O. sativa indica	*O. meridionalis*	−0.473	2.156	0.347	223	0.002	1.425	0.999	50	−0.777	2.509	0.216	141	41
	O. barthii	0.026	2.708	0.957	237	0.820	1.846	0.416	51	−0.420	3.158	0.494	144	54

Significant values at the 5% level are in italic type.

hypotheses previously proposed to explain GC content variations in other species: biased gene conversion, selection, and mutational biases (Eyre-Walker and Hurst 2001).

Recombination and Mating Systems Affect GC Content

We found a weak positive correlation between GC content and recombination rates in *Oryza*. However, such a correlation cannot distinguish between recombination driving GC content and the reverse. For this reason, we studied correlations between recombination and equilibrium GC content (GC* or "future" GC content). We also found a positive and significant correlation between GC* and recombination, suggesting that recombination drives GC content (Duret and Arndt 2008). In addition, we found that selfing species have lower GC* values compared with outcrossing species (table 2). Because of their high homozygosity, selfing species should be weakly affected by gBGC (Marais et al. 2004). This result suggests that recombination, through mating systems, drives GC content and not the other way around as GC content cannot drive mating systems.

These observations could be explained by a mutagenic effect of recombination. If recombination induced more W → S than S → W mutations, GC content would mechanically increase with recombination. However, several results are not consistent with this hypothesis. First, selfing species have lower GC* compared with outcrossing species (table 2). Recombination is still at work in selfing species, whereas it is not genetically efficient because of homozygosity. If recombination were mutagenic we would not expect an effect of mating system. Second, under the mutational bias hypothesis, we expect mutations to have similar fixation probabilities. The fact that external branches have lower GC* compared with internal branches (table 2) shows that polymorphism present in external branches is enriched in W alleles that are eventually lost more frequently than S alleles, that is, they are selected against. Polymorphism data also show that W → S mutations segregate at higher frequency than S → W mutations, at least for third codon positions (fig. 3 and table 4). As expected, this effect is higher in *O. rufipogon*, a species that exhibit higher outcrossing rate than *O. sativa*. Third, we also showed that mutations are biased towards A and T bases ($\lambda \approx 3$), which cannot explain the high GC content of rice genome and grass genomes in general. Only two hypotheses thus remain to explain the effect of recombination: gBGC and selection.

gBGC, Selection, or Both?

Previous studies have already shown that mating system and recombination affect GC content in grasses (Haudry et al. 2008; Escobar et al. 2010). However, as these studies only used CDS, they could not disentangle selective and gBGC effects. Indeed, recombination and outcrossing affect selection and gBGC in a similar way: gBGC is driven by recombination and selection is more efficient in outcrossing and recombining genomes (Charlesworth and Wright 2001; Gordo and Charlesworth 2001). We thus expect selection on codon usage and gBGC to be less efficient in genomic regions of low recombination and in selfing species. To distinguish gBGC from selection on codon usage, we used noncoding sequences, affected only by gBGC, and expression levels, which mainly affects third codon positions.

GC* is positively correlated with recombination both for third codon positions and noncoding positions (fig. 2), showing that recombination affects GC content at all positions, with no distinction whether they are coding or noncoding. Similarly, GC* values for both noncoding and third codon positions are significantly lower in selfing than in outcrossing branches (table 2). GC* is also significantly lower in external than internal branches for all site categories (table 2). These observations cannot be explained by selection on codon usage alone as only third codon positions would be affected. We thus have strong evidence for gBGC in the *Oryza* genus.

Besides, we observed a positive correlation between GC3* and the expression level but not with noncoding GC* (fig. 3). This clearly suggests selection on codon usage for translation efficiency, as almost all favored codons in grasses, and in rice in particular, end with a G or C base (Wang and Roossinck 2006). Another result suggesting selection on codon usage comes from the polymorphism analysis. The gBGC/selection model is significantly better than the neutral model only for third codon positions and the estimated coefficient is higher than for noncoding sites (table 4). This strongly suggests that selection on codon usage affects third codon positions in addition of any possible gBGC effect. These results show that selection is at work on codon usage and that it also drives the GC3-content in the *Oryza* genomes. The fact that these effects are weaker and not significant in *O. sativa* is also consistent

Table 5. Estimate of GC* from equation (5) and estimates given in Table 4.

Species		GCtotal	GC3	GCnoncoding
	Current GC	0.44	0.55	0.36
Oryza rufipogon	GC*/*O. meridionalis*	0.35	0.65	0.24
	GC*/*O. barthii*	0.39	0.70	0.27
Oryza sativa japonica	GC*/*O. meridionalis*	0.30	0.54	0.27
	GC*/*O. barthii*	0.29	0.60	0.21
O. sativa indica	GC*/*O. meridionalis*	0.22	0.41	0.15
	GC*/*O. barthii*	0.27	0.55	0.17

with a reduced efficacy of both selection and gBGC due to the high selfing rate.

Nonequilibrium Processes

We observed significant correlations between GC* and recombination only at a small phylogenetic scale, when we focused on *O. sativa* and *O. rufipogon* species. As recombination rates were obtained from *O. sativa* ssp. *japonica*, this suggests that recombination patterns vary quite quickly along the *Oryza* genus, corresponding to about 10 My (Zou et al. 2008; Kellog 2009). If the recombination pattern and the associated gBGC dynamics were mostly conserved, we would have expected a stronger relationship between recombination and GC* at the whole genus scale, simply because, if constant, GC* is better estimated in a larger phylogeny. In contrast, in Triticeae, recombination pattern is thought to be partly conserved. Accordingly, Escobar et al. (2010) found a significant positive correlation between recombination and GC* at the whole Triticeae scale, corresponding to about 12–15 Mya, similar to the divergence of the *Oryza* genus. We observed similar results for correlations between GC3* and expression levels (obtained from *O. sativa*), suggesting that expression levels may also vary quickly at this evolutionary scale.

The comparison between current GC content and GC* in the fourth data set (table 5) also shows that the *O. rufipogon* genome is not at equilibrium. Noncoding GC content seems to be decreasing. This is consistent with a mutational bias toward A and T bases. However, to allow this decrease, the GC content of noncoding positions must have previously increased. Only gBGC can cause an increase in the GC content in noncoding regions. gBGC strength could have gone down recently. A change in gBGC strength is quite plausible considering that recombination is known to vary quite quickly (Coop and Przeworski 2007). Moreover, if recombination occurs in short-lived hot spots as in primates (Spencer et al. 2006), only a small fraction of polymorphism should be driven by gBGC, making its detection difficult with as few as 800 SNPs (compared with the few millions of SNPs available in humans). In noncoding regions, selection may favor AT variants in some positions because of their involvement in regulatory functions, which would oppose to gBGC.

Comparison with Mammalian Isochores

Nucleotide landscapes of grass genomes show striking similarities with the isochore structure found in mammals and other vertebrates. Assuming our results in *Oryza* could be generalized to other grass species, we can draw parallels between the mechanisms involved in the two groups. As in mammals, recombination seems to drive GC content, likely through the effect of gBGC. Our results are less clear than in mammals, simply because the data sets we used are much smaller than those used in mammals (e.g. Duret and Arndt 2008). Although we found evidence of gBGC in noncoding regions, we were not able to estimate significant gBGC coefficients (table 4), which suggests gBGC is too weak to be detectable with the SNPs data set we used. In humans, the gBGC intensity is also low, $4N_eb \approx 1.3$ for the 20% of the genome with the highest recombination rates (Spencer et al. 2006). Grouping SNPs for all recombination categories as we did would lead to $4N_eb \approx 0.3$, which is indeed too low to be detectable with the data set we used. However, even such a low level is sufficient to explain GC content patterns in primates (Duret and Arndt 2008).

Contrary to mammals, we also found evidence of efficient selection on codon usage in the *Oryza* genus, whereas it is very low, if any, in mammals. As most preferred codons end in G or C in grasses (Wang and Roossinck 2006), selection on codon usage may also contribute to the high GC3 content found in grasses. Moreover, analyses of polymorphism data suggest that selection on codon usage is stronger than gBGC, as positive and significant selection/gBGC coefficients were only found at third codon positions (table 4).

Conclusions

Our results shed light on the likely causes of the peculiar nucleotide landscape in rice and very likely in other grass species. We showed that gBGC is an active process in this group and may have driven GC content evolution in addition to selection on codon usage. Because it is a selection-like process, gBGC can lead to spurious signatures of relaxed or even positive selection (Galtier and Duret 2007; Galtier et al. 2009; Ratnakumar et al. 2010). We thus strongly suggest taking gBGC into account in future genomic studies in grasses. Quantifying the genomic and the possible fitness impacts of gBGC will be now possible thanks to the increasing amount of genomic data available in several species. Finally, why gBGC seems to be more active in grasses than in other angiosperms studied so far remains an open and challenging question.

Supplementary Material

Supplementary tables S1and S2 and figures S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Fu-Min Zhang and Song Ge who kindly provided us with their data set and Peter Arndt who kindly provided his programs for CpG analyses. Julien Dutheil gave very helpful guidelines for the use of BPPML and Gabriel Marais

for the use of MareyMap. We also thank Nicolas Galtier and Laurent Duret for helpful discussions and two anonymous reviewers for their constructive comments. This publication is the contribution ISEM 2011-030 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554—CNRS). This work was supported by the French *Centre National de la Recherche Scientifique* and *Agence Nationale de la Recherche* (ANR-08-GENM-036-01).

References

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev*. 11:660–666.
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.
Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3–17.
Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol*. 19:1181–1197.
Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.
Caicedo AL, Williamson SH, Hernandez RD, et al (12 co-authors) 2007. Genome-wide patterns of nucleotide polymorphism in domesticated. rice. *PLoS Genetics* 3:1745–1756.
Carels N, Bernardi G. 2000. Two classes of genes in plants. *Genetics* 154:1819–1825.
Charlesworth B. 1992. Evolutionary rates in partially self-fertilizing species. *Am Nat*. 140:126–148.
Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. *Curr Opin Genet Dev*. 11:685–690.
Chojnowski JL, Franklin J, Katsu Y, Iguchi T, Guillette LJ Jr., Kimball RT, Braun EL. 2007. Patterns of vertebrate isochore evolution revealed by comparison of expressed mammalian, avian, and crocodilian genes. *J Mol Evol*. 65:259–266.
Coop G, Przeworski M. 2007. An evolutionary view of human recombination. *Nat Rev Genet*. 8:23–34.
Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA. 2009. Species trees from highly incongruent gene trees in rice. *Syst Biol*. 58:489–500.
Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4:e1000071.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*. 10:285–311.
Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol*. 8:255.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5:113.
Escobar JS, Cenci A, Bolognini JS, Haudry A, Laurent S, David J, Glemin S. 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (poaceae). *Evolution* 64:2855–2872.
Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet*. 2:549–555.
Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900.
Fryxell KJ, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol*. 17:1371–1383.
Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet*. 23:273–277.
Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*. 25:1–5.
Glémin S. 2007. Mating systems and the efficacy of selection at the molecular level. *Genetics* 177:905–916.
Glémin S, Bazin E, Charlesworth D. 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc Biol Sci*. 273:3011–3019.
Gordo I, Charlesworth B. 2001. Genetic linkage and molecular evolution. *Curr Biol*. 11:R684–R686.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.
Guo X, Bao J, L Fan . 2007. Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes. *FEBS Lett*. 581:1015–1021.
Haudry A, Cenci A, Guilhaumon C, Paux E, Poirier S, Santoni S, David J, Glémin S . 2008. Mating system and recombination affect molecular evolution in four Triticeae species. *Genet Res*. 90:97–109.
Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet*. 72:1527–1535.
Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. *Mol Biol Evol*. 24:2196–2202.
Igic B, Bohs L, Kohn JR. 2006. Ancient polymorphism reveals unidirectional breeding system shifts. *Proc Natl Acad Sci U S A*. 103:1359–1363.
Kellog EA. 2009. The evolutionary history of Ehrhartoideae, Oryzeae, and Oryza. *Rice* 2:1–14.
Kuraku S, Ishijima J, Nishida-Umehara C, Agata K, Kuratani S, Matsuda Y. 2006. cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a chromosomal size-dependent GC bias shared by sauropsids. *Chromosome Res*. 14:187–202.
Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*. 18:337–340.
Lynch M. 2007. The origin of genome architecture. Sunderland (MA): Sinauer.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454:479–485.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet*. 19:330–338.
Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. *Genome Biol*. 5:R45.
Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*. 21:984–990.
Mukhopadhyay P, Basak S, Ghosh TC. 2007. Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes. *Gene* 400:71–81.
Nagylaki T. 1983. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A*. 80:6278–6281.
Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci*. 365:2571–2580.
Rezvoy C, Charif D, Gueguen L, Marais GA. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*. 23:2188–2189.
Sano Y. 1980. Adaptive strategies compared between the diploid and tetraploid forms of Oryza punctata. *Bot Mag*. 93:171–180.

Shi X, Wang X, Li Z, Zhu Q, Tang W, Ge S, Luo J. 2006. Nucleotide substitution pattern in rice paralogues: implication for negative correlation between the synonymous substitution rate and codon usage bias. *Gene* 376:199–206.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet*. 2:e148.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A*. 48:582–592.

Sweeney M, McCouch S. 2007. The complex history of the domestication of rice. *Ann Bot*. 100:951–957.

Takebayashi N, Morrell PL. 2001. Is self-fertilization an evolutionary deed end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. *Am J Bot*. 88:1143–1150.

Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*. 9:678–687.

Wang HC, Hickey DA. 2007. Rapid divergence of codon usage patterns within the rice genome. *BMC Evol Biol*. 7 (Suppl 1):S6.

Wang HC, Singer GA, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. *Mol Biol Evol*. 21:90–96.

Wang LJ, Roossinck MJ. 2006. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. *Plant Mol Biol*. 61:699–710.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol*. 23:1203–1216.

Wolfram S. 1996. The Mathematica book. Cambridge: Cambridge University Press.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang LL, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol*. 9:R49.

C.1 Supplementary Material

Available online:

Figure S1: http://mbe.oxfordjournals.org.gate1.inist.fr/content/suppl/2011/04/18/msr104.DC1/FigureS1_MBE35.doc

Figure S2: http://mbe.oxfordjournals.org.gate1.inist.fr/content/suppl/2011/04/18/msr104.DC1/FigureS2_MBE35.doc

Table S1: http://mbe.oxfordjournals.org.gate1.inist.fr/content/suppl/2011/04/18/msr104.DC1/TableS1_MBE35.xls

Table S2: http://mbe.oxfordjournals.org.gate1.inist.fr/content/suppl/2011/04/18/msr104.DC1/TableS2_MBE35.xls