Diss. ETH N° 20391

Silene latifolia (Caryophyllaceae) sex chromosome evolution

A dissertation submitted to

ETH ZURICH

for the degree of

DOCTOR OF SCIENCES

presented by

Nicolas Blavet

Master of Science in

Mathematical and Informatical Analysis of Life

born March 6th, 1983

citizen of France

accepted on the recommendation of

Prof. Dr. Alex Widmer, examiner

Prof. Dr. Paul Schmid-Hempel, co-examiner

Dr. Gabriel Marais, co-examiner

2012

Table of contents

Summary

Sex chromosomes evolved more than a hundred million years ago in mammals and birds but much more recently in some flowering plant lineages. Studies on the evolution of sex chromosomes are important to understand sex determination mechanisms and are of general evolutionary interest because sex chromosomes have evolved independently numerous times in different lineages of animals, fungi and plants. Flowering plants are particularly suited for investigating sex chromosome evolution because sex chromosomes emerged relatively recently in different groups and because of the presence of closely related species lacking sex chromosomes. In this thesis, I contribute to the understanding of sex chromosome evolution by analyzing the recently emerged sex chromosomes of the white campion, *Silene latifolia*.

First, I investigated transcriptomic data coming from eight individuals representing five closely related Caryophyllaceae species, including four *Silene* and one *Dianthus* species (Chapter I). I found about 74000 genes in the studied species with about 1400 genes specific to the *Silene* genus. Moreover the detection of thousands of single nucleotide polymorphisms (SNPs) provides new molecular resources for linkage mapping and population genetic analyses. This study showed the value of comparative transcriptome analyses based on next generation sequencing data for the characterization of genetic variation in non-model species.

In a second step, I analyzed a large piece of genomic DNA located in a pseudoautosomal region (PAR) of plant sex chromosomes (Chapter II). Comparison of *S. latifolia* and *S. vulgaris* homologous bacterial artificial chromosome (BAC) sequences identified new pseudoautosomal genes in *S. latifolia*. These genes are conserved in size and linear arrangement in both species, which indicates small-scale gene collinearity between the PAR region of sex chromosomes in *S. latifolia* and the

corresponding autosomal region in *S. vulgaris*. Contrary to expectations, I found no increase in GC and GC3 content of the pseudoautosomal genes, in contrast to the situation in mammals, but I found evidence for a moderate size increase of the *S. latifolia* PAR compared to the *S. vulgaris* autosomes. This increase is similar to the size difference observed between *S. latifolia* and *S. vulgaris* autosomes and can not explain the specific size increase seen in the *S. latifolia* sex chromosomes.

In a third study, I focused on the non-recombining part of the *S. latifolia* sex chromosomes (Chapter III). Similarly to Chapter II, I compared BAC sequences from *S. latifolia* and *S. vulgaris*, identified new sex-linked genes and found conserved small-scale collinearity between both *S. latifolia* X and Y chromosomes and the corresponding *S. vulgaris* autosomal region for genes located in the oldest evolutionary stratum. The absence of pseudogenes on the Y chromosome BACs and the accumulation of transposable elements were notable as they indicate that *S. latifolia* is rarely losing genes but is accumulating transposable elements on the Y chromosome. An analysis of the distribution of genes with reduced expression of the allele on the Y chromosome further revealed that Y gene degeneration is a random process in *S. latifolia*.

This thesis provides new insights to the understanding of the early stages of sex chromosome evolution in plants and shows the value of next generation sequencing technologies for the study of genomic and transcriptomic variation in formerly largely intractable non-model organisms.

Résumé

L’évolution des chromosomes sexuels a débuté il y a des centaines de millions d’années chez les mammifères et les oiseaux mais est apparue plus récemment dans certains taxons de plantes à fleurs. L'analyse de l'évolution des chromosomes sexuels est importante pour la compréhension des mécanismes de la détermination sexuelle et est d’intérêt général pour l’évolution, car les chromosomes sexuels ont évolué indépendamment à plusieurs reprises dans des lignées différentes d’animaux, de champignons et de plantes. Les plantes à fleurs sont particulièrement intéressantes pour la recherche sur l’évolution de ces chromosomes du fait de la récente émergence de ceux-ci et de la présence d’espèces proches étant dépourvues de chromosome sexuel. J’ai contribué par cette thèse à la compréhension de l’évolution des chromosomes sexuels, en analysant ceux, récemment apparus, du compagnon blanc *Silene latifolia*.

Premièrement, j’ai étudié les données transcriptomique provenant de huit individus représentant cinq Caryophyllacées phylogénétiquement proches dont quatre espèces de *Silene* et une espèce de *Dianthus* (Chapitre I). Nous avons trouvé environ 74000 gènes dans les espèces étudiées et environ 1400 gènes spécifiques du genre *Silene*. De plus, la détection de milliers de site de polymorphisme nucléotidique va apporter de nouvelles ressources moléculaires pour la cartographie génétique et les analyses de génétique des populations. J’ai également mis en évidence l’intérêt des analyses comparatives de transcriptomes d'espèces non modèles à grande échelle en utilisant les techniques de séquençage de nouvelle génération.

En deuxième étape, j'ai analysé un large fragment d'ADN génomique provenant de la région pseudoautosomale de chromosome sexuel d’une plante (Chapitre II). La comparaison de séquences homologues de chromosomes artificiels de bactéries (BAC) venant de *S. latifolia* et *S. vulgaris* a révélé de nouveaux gènes

pseudoautosomaux dont la taille et l'arrangement linéaire est conservé entre les deux espèces, ce qui indique une colinéarité à petite échelle entre le chromosome X et la région autosomal correspondante. Contrairement aux prévisions, je n'ai trouvé aucune augmentation des taux de GC et GC3 sur les gènes pseudoautosomaux de *S. latifolia*, à la différence des mammifères, mais j'ai mis en évidence une augmentation modérée de la taille de la région pseudoautosomal de *S. latifolia* comparé à la séquence autosomal de *S. vulgaris*. Cette augmentation est similaire à la différence existant entre les autosomes des deux espèces et ne peut pas expliquer l'augmentation spécifique de la taille des chromosomes sexuels de *S. latifolia*.

Dans la troisième étude, je me suis concentré sur la partie non recombinante des chromosomes sexuels de *S. latifolia* (Chapitre III). De même qu'au Chapitre II, nous avons comparé des séquences de BAC provenant de *S. latifolia* et *S. vulgaris*. J'ai identifié de nouveaux gènes liés aux sexes et j'ai trouvé une colinéarité entre des gènes situés dans la plus vieille strate évolutive des chromosomes X et Y de *S. latifolia* et de l'autosome correspondant de *S. vulgaris*. L'absence de pseudogène et l'accumulation d'éléments transposables, m'ont permis d'indiquer que les gènes du chromosome Y de *S. latifolia* sont rarement perdus et que ce dernier accumule des éléments transposables. Une analyse portant sur l'expression réduite des allèles du chromosome Y ont révélé que la dégénération des gènes Y de *S. latifolia* est un processus aléatoire.

Cette thèse apporte de nouveaux éléments pour la compréhension des premiers stages de l'évolution des chromosomes sexuels de plantes et montre l'importance des techniques de séquençage de nouvelle génération pour l'étude des variations génomiques et transcriptomiques d'organismes non modèles.

General introduction

Sex chromosome evolution

Sex chromosomes have been intensively studied since 1891 when Hermann Henking first found an X chromosome in the bug species *Pyrrhocoris apterus* [Henking, 1891]. In 1905, Nettie M. Stevens and Edmund B. Wilson, independently identified Y chromosomes in *Tenebrio molitor*, and in *Lygaeus turcicus*, *Euschistus fissilis* and *Coenus delius*, respectively [Stevens, 1905; Wilson, 1905], while Theophilus S. Painter discovered them in mammals in 1921 [Painter, 1921]. Two years later, Kathleen B. Blackburn found the first X and Y chromosomes in flowering plants [Blackburn, 1923]. Sex chromosome analysis is important to understand sex determination mechanisms and is of general evolutionary interest because sex chromosomes have evolved independently numerous times in different lineages of animals, fungi and plants [Fraser and Heitman, 2004; Graves and Peichel, 2010; Ming, et al., 2011]. Despite the many independent origins of sex chromosomes, they have several characteristics in common and are an interesting example of evolutionary convergence [Bachtrog, et al., 2011; Ellegren, 2011].

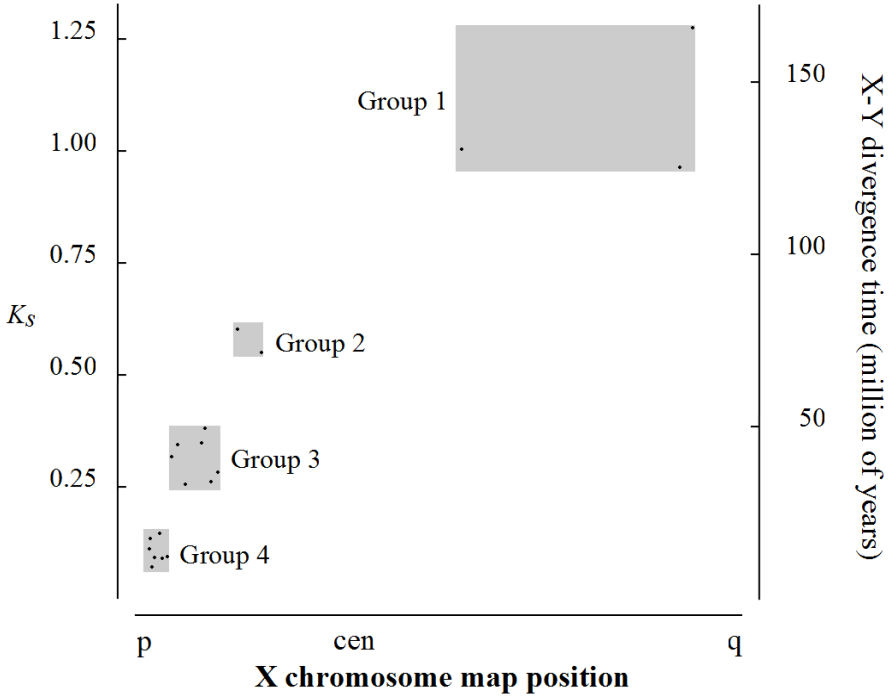
Recombination suppression

One of the main characteristics of fully developed sex chromosomes is the suppression of recombination between the sex chromosomes in the heterogametic sex (XY or ZW). Indeed, in males of most eutherians, recombination between the X and the Y chromosomes during meiosis is restricted to short chromosomal regions that are known as pseudoautosomal regions (PARs) [Burgoyne, 1982]. In marsupials, no PARs exist and the X and Y chromosomes are not recombining at all [Sharp, 1982; Page, et al., 2006]. In *Drosophila melanogaster*, no meiotic recombination occurs in males, again leading to an absence of recombination between the X and the Y chromosomes [Morgan, 1914; McKee and Handel, 1993].

Evolutionary strata

Along the human sex chromosomes, different degrees of divergence between X and Y chromosome loci have been reported. The number of synonymous substitutions per synonymous site (K_S) is a measure of nucleotide divergence between X and Y-linked alleles since they started to diverge. Estimates of K_S are positively correlated with the position of genes along the X chromosome [Lahn and Page, 1999] (Figure 1).

Figure 1: Plot of K_S versus X-chromosome map position for 19 human X-Y gene pairs. On the x-axis are indicated the p-arm (p), the centromere (cen) and the q-arm (q) of the X chromosome. *(source: adapted from [Lahn and Page, 1999])*



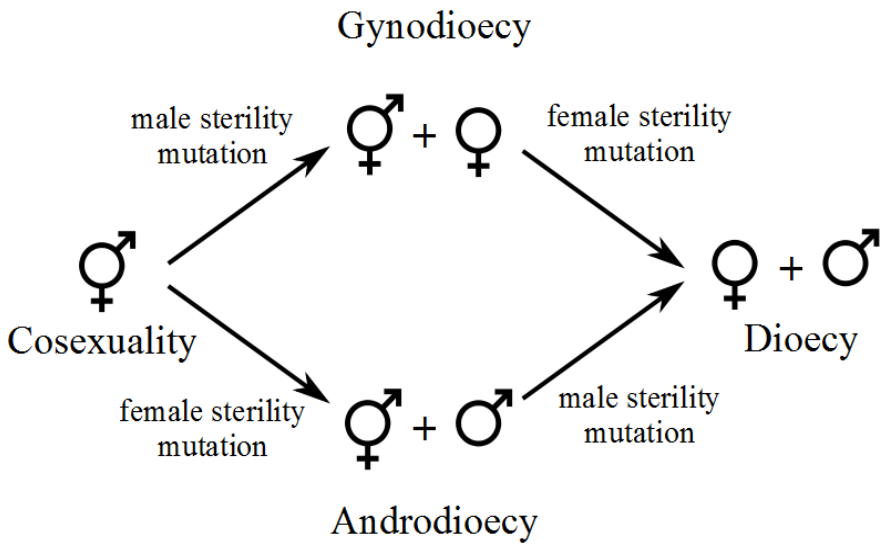
The further genes are located from the PAR on the p-arm of the X chromosome, the higher their divergence from the Y-allele. Divergence (as measured by Ks) is highest close to the primary female sex determining gene *SOX3* [Lahn and Page, 1999] which is located in the first stratum (Group 1 in Figure 1) and marks the area where

recombination initially stopped between the X and the Y chromosomes. The cessation of X and Y recombination dates to the emergence of eutherian sex chromosomes, about 150 million years ago (Mya) [Veyrunes, et al., 2008]. Chromosomal inversions, which are known to suppress recombination [Jaarola, et al., 1998], are expected to have played a major role in X and Y chromosome differentiation by stopping gradually the recombination and to have led to the formation of evolutionary strata [Lahn and Page, 1999]. Similarly, evolutionary strata were found in other mammals, such as in mice [Sandstedt and Tucker, 2004], but also in birds [Handley, et al., 2004].

Origins of sex chromosomes

Sex chromosomes have most likely evolved from a pair of autosomes as evidenced by homologies between the X and Y chromosomes [Graves, 1996]. Fixation of sex determination loci on autosomes is expected to follow two possible pathways of two steps when starting from a hermaphroditic system. The first pathway implies that a male sterility mutation arises and is subsequently fixed on the proto-X chromosome, which may lead to a gynodioecious population (mixture of hermaphrodite and female individuals). Then, a female sterility mutation may arise and become fixed on the proto-Y chromosome, which leads to the establishment of males. Similarly, the second pathway starts with a female sterility mutation fixation yielding to an androdioecious population (mixture of hermaphrodite and male individuals), in which a male sterility mutation fixation would lead to the establishment of females. Then, in both case, hermaphroditism may be lost through natural selection because of the costs to maintain cosexuality [Ohno, 1967; Bull, 1983; Charlesworth, 1991] (Figure 2). These steps are thought to be a general pattern in the evolution of separate sexes. However, while they cannot be observed in organisms with evolutionary old sex determination systems, they can be directly observed and tested in some lineages of flowering plants [Westergaard, 1958; Charlesworth, 1991].

Figure 2: Steps in the evolution of dioecy. *(source: adapted from [Charlesworth, 1991; Lynch, 2007])*



Degeneration of the Y chromosome

The cessation of recombination leads to the degeneration of Y (and W) chromosomes. In humans, the X chromosome is 155 Mbp long and carries 52 genes, while the Y chromosome is only 59 Mbp long and carries 16 genes (numbers take in account only genes presenting protein evidence, Human genome Build 37.3 *http://www.ncbi.nlm.nih.gov*). This degeneration process includes the loss of genes, the accumulation of repeated elements, and reduced gene expression [Bachtrog, et al., 2011]. The reduction or loss of gene expression on the human Y chromosome, which is expected to be caused by different processes (Muller's ratchet, background selection, the Hill-Robertson effect with weak selection, and the hitchhiking of deleterious alleles by favorable mutations [Felsenstein, 1974; Charlesworth and Charlesworth, 2000; Hedrick, 2005]), has led to the evolution of a system to balance the expression between females that carry two X chromosomes and males carrying a single X chromosome. This phenomenon is called dosage compensation (reviewed by [Straub and Becker, 2007; Casci, 2011]). The degeneration of the Y chromosome and presence of dosage compensation are also found in *Drosophila melanogaster* [Marín,

et al., 2000; Carvalho, 2002] as well as the ZW system in chicken [Ezaz, et al., 2006; Arnold, et al., 2008].

Plant sex chromosomes

Angiosperms (flowering plants) present a large diversity of mating systems and individual plants may carry either male or female, or both types of reproductive organs in their flowers. In addition, different combinations of sexes may occur together in the same populations (Table 1). Despite this high diversity of sexual systems, sex chromosomes have been found so far only in a limited number of dioecious plants from various genera (e.g. *Cannabis*, *Silene*, *Rumex*, *Carica* and *Asparagus* [Vyskot and Hobza, 2004]). These different species are all male heterogametic. The sex chromosomes in their respective genomes present different stages of differentiation between the X and the Y chromosome (Figure 3). The sex chromosomes found in *Asparagus officinalis* resemble those expected during the first step of emergence of sex chromosomes. Here, X and Y chromosomes are homomorphic, and YY genotypes are viable, indicating that the Y chromosome is not degenerated and can recombine with the X. These sex chromosomes have been identified because the Y chromosome carries both a dominant male activator (M) and a female repressor (F) gene, so males are heterozygous MF/mf whereas females are homozygous mf/mf [Jamilena, et al., 2008]. The second step could be represented by *Carica papaya* sex chromosomes. In this species, the sex chromosomes are homomorphic and the X and Y are able to recombine along most of their length (90%), but YY mutants are lethal due to a highly divergent sex specific part [Liu, et al., 2004; Ma, et al., 2004]. The third step may be represented by the heteromorphic sex chromosomes found in *Marchantia polymorpha* or *Silene latifolia*. The Y chromosome in *S. latifolia* is about 40% larger than the X chromosome [Ming and Moore, 2007]. Recombination between the X and Y during male meiosis is restricted to a pseudoautosomal region [Lengerova, et al., 2003] that has been estimated to be about 10% of the Y chromosome [Čermak, et al., 2008; Filatov, et al., 2009].

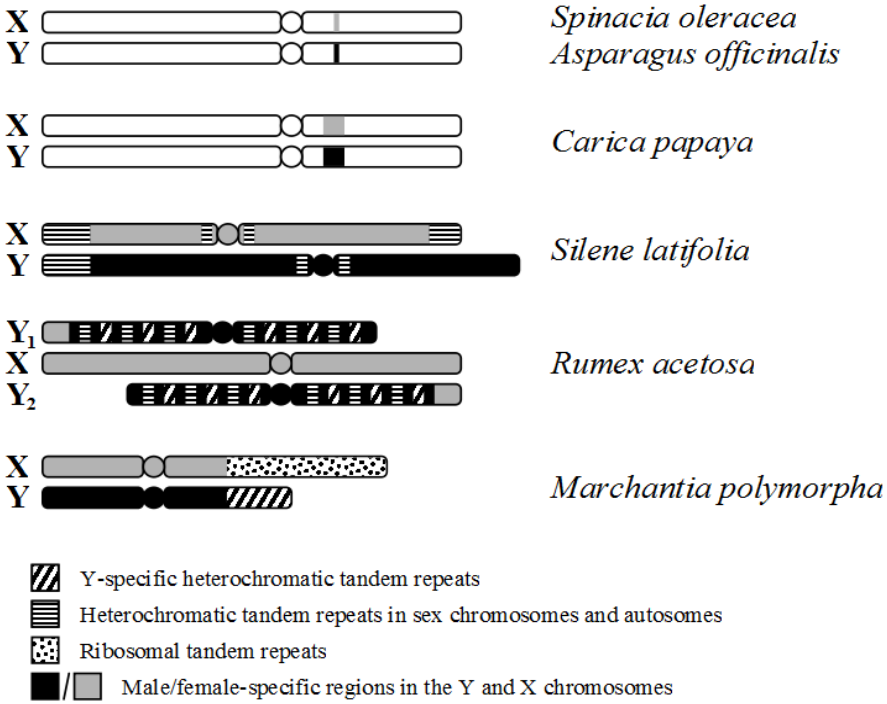
Table 1: Modes of Sexuality in Flowers, Plants and Populations. Phenotype: h = hermaphrodite, f = female, m = male. *(source: adapted from [Dellaporta and Calderon-Urrea, 1993])*

Sexuality	Phenotype	Description
Individual flowers		
Hermaphrodite (bisexual)	h	Bisexual flower with both stamens and pistil
Diclinous (unisexual)	f or m	Unisexual flowers
Pistillate (carpillate)	f	Unisexual flower with pistil only (female flowers)
Staminate	m	Unisexual flower with stamens only (male flowers)
Individual plants		
Hermaphrodite	h	Only hermaphrodite flowers
Monoecious	f, m	Both pistillate and staminate flowers on the same plant
Dioecious	f or m	Staminate and pistillate flowers borne on different plants
Gynoecious	f	Plant bears only pistillate flowers
Androecious	m	Plant bears only staminate flowers
Gynomonoecious	h, f	Plant bears both hermaphrodite and pistillate flowers
Andromonoecious	h, m	Plant bears both hermaphrodite and staminate flowers
Trimonoecious (polygamous)	h, f, m	Hermaphrodite, pistillate, and staminate flowers on the same plant
Plant populations		
Hermaphrodite	h	Only hermaphrodites
Monoecious	f, m	Only monoecious plants
Dioecious	f and m	Only dioecious plants
Gynodioecious	h and f	Both hermaphrodite and gynoecious individuals
Androdioecious	h and m	Both hermaphrodite and androecious individuals
Trioecious (subdioecious)	h and f and m	Hermaphrodite, pistillate, and staminate individuals

Evolutionary strata have been investigated in *S. latifolia* sex chromosomes, using the few genes identified until recently, and the results suggest that two degeneration events occurred during sex chromosome evolution [Marais, et al., 2008]. In *M. polymorpha* the Y chromosome is smaller than the X and carries Y-specific repeats on

about 25% of its length. Moreover, in this haploid species X and Y chromosomes are not recombining [Okada, et al., 2001].

Figure 3: Schematic diagram of flowering plant sex chromosomes. *Spinacia oleracea, A. officinalis* and *C. papaya* present short X- and Y-specific regions, while the sex-specific regions in *S. latifolia, R. acetosa* and *M. polymorpha* cover most of the sex chromosomes. *(source: adapted from [Jamilena, et al., 2008])*



Another evolutionary pathway for plant sex chromosomes could be represented by the sex chromosomes of *Rumex acetosa*. In this species, males carry two Y chromosomes and these are highly differentiated from the X chromosomes. Nevertheless, both Y chromosomes can pair with the X (each at one tip of the X). In *R. acetosa*, Y chromosomes are not involved in sex determination. In this species, sex is determined by the ratio of X chromosomes to autosomes. Plants with a ratio superior to 1 are females, plants with a ratio inferior to 0.5 are males and plants with a ratio in between

0.5 and 1 are hermaphrodites. However, Y chromosomes are required for the successful progress of meiosis in pollen mother cells [Parker and Clark, 1991].

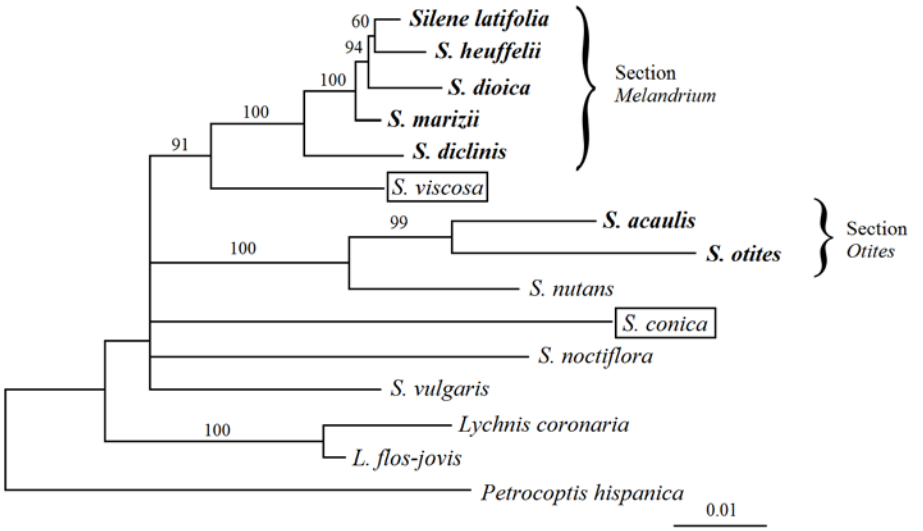
Study species

Silene latifolia and S. vulgaris

The white campion, *Silene latifolia*, is a dioecious plant (male and female individuals) growing in most of Europe, northern Africa and western Asia [Baker, 1947]. The bladder campion, *S. vulgaris*, is a gynodioecious species growing in Europe and North America. Both species belong to the family Caryophyllaceae. In the genus *Silene*, gynodioecy is the most common reproductive system and has been proposed to represent the ancestral mating system in the genus [Desfeux, et al., 1996], while dioecious species are rare and have evolved independently in two distinct sections: *Otites* and *Melandrium* (*S. latifolia*, *S. dioica* and *S. diclinis*) [Desfeux, et al., 1996; Kejnovský and Vyskot, 2010; Marais, et al., 2011] (Figure 4). It has been proposed than hermaphroditic species evolved into gynodioecious or androdioecious species after fixation of a male or female sterility mutation, respectively. Subsequently, fixation of a female sterility mutation in gynodioecious populations or fixation of a male sterility mutation in androdoecious populations, lead to dioecious species [Charlesworth, 1991].

Silene latifolia and *S. vulgaris* share the same chromosome number (n=12), but have different genome sizes. *S. latifolia* has a haploid genome size of about 2,646 Mbp [Costich, et al., 1991] while the genome size of *S. vulgaris* (1,103 Mbp) [Široký, et al., 2001] is less than half that of *S. latifolia*. *Silene vulgaris* lacks sex chromosomes, but *S. latifolia* is one of the best known dioecious plant species and has heteromorphic sex chromosomes [Blackburn, 1923; Winge, 1923]. Comparative analysis of *S. latifolia* and *S. vulgaris* may thus further our understanding of the emergence and evolution of plant sex chromosomes [Vyskot and Hobza, 2004]. First analyses of *S. latifolia* sex-linked gene divergence estimated the age of the sex chromosomes around 10 million years [Filatov, 2005].

Figure 4: Phylogenic tree of the genus *Silene*. The phylogenetic tree is based on 12 genes and resulted from a consensus of two methods: maximum likelihood and Super Distance Matrix (SDM). Only bootstrap values >50% are shown. Dioecious species are written in bold, hermaphrodites are written in a box and all other species are gynodioecious species. *(source: adapted from [Marais, et al., 2011])*



Features of *Silene* sex chromosomes can further be compared with other well-studied sex chromosomes, e.g. in mammals or *Drosophila*. Similarly to mammalian sex chromosomes, *Silene latifolia* is male heterogametic (females are XX and males are XY), and the X and Y chromosomes are not recombining except in pseudoautosomal regions (PARs) [Di Stilio, et al., 1998; Scotti and Delph, 2006].

Silene latifolia sex chromosomes are both larger than the autosomes and the Y chromosome is the largest chromosome of the genome. The *Silene latifolia* Y chromosome is about 570Mb, the X is about 400Mb [Ming and Moore, 2007] and the autosomes about 200Mb, while *S. vulgaris* autosomes are about 100Mb [Široký, et al., 2001]. The chromosome number and the size of both *S. latifolia* sex chromosomes suggest that translocations of whole chromosomes have not contributed to the enlarged X and Y [Charlesworth, 2002]. Possible reasons for *S. latifolia* sex chromosome enlargement have been studied repeatedly [Kejnovský, et al., 2006; Čermak, et al., 2008; Kubat, et al., 2008; Kejnovský, et al., 2009] and it has been

suggested that the size increase is due to the accumulation of repeated elements. This accumulation of elements is expected to occur primarily in the early stages of sex chromosome evolution and could have been a reason for the degeneration of the Y chromosome in mammals.

Silene latifolia sex chromosomes have most likely evolved from a single pair of autosomes. This is supported by the observation that one autosome of *S. vulgaris* carries homologues of *S. latifolia* sex-linked genes [Filatov, 2005]. Comparisons between these two *Silene* species have been used successfully many times to investigate sex chromosome evolution [Delichère, et al., 1999; Matsunaga, et al., 2003; Filatov, 2005; Matsunaga, et al., 2005; Nicolas, et al., 2005; Matsunaga, 2006; Bergero, et al., 2008; Marais, et al., 2008; Kaiser, et al., 2009; Čegan, et al., 2010; Qiu, et al., 2010; Bergero and Charlesworth, 2011; Chibalina and Filatov, 2011; Kaiser, et al., 2011] and we use the same comparison for our analysis of sex chromosome evolution in *S. latifolia*.

Main methods used

Sequencing using Roche 454 pyrosequencing (Chapters I, II, III):

The 454 pyrosequencing method appeared in 2005 and was the first next-generation sequencing method available [Margulies, et al., 2005]. This method is based on two steps: First, the DNA (genomic or complementary DNA) is cut and adaptors are attached in order to link the nucleotide fragments to a bead where it is amplified by PCR reaction (one fragment per bead). Then, in a second step, the beads are captured in picoliter-sized wells and pyrosequencing (pyrophosphate-based sequencing) is performed in parallel on each DNA fragment. Nucleotide incorporation is detected by the release of inorganic pyrophosphate (PPi), which is converted into ATP and then used by the enzyme luciferase to generate a light signal. The signal is proportional to the amount of consecutive nucleotides of the same kind (e.g. if there are three consecutive cytosine in the single-stranded fragment, the amount of light generated would be three times that of a single cytosine in the fragment), and the cycle is iteratively repeated for each of the four bases. The average read length was between 110 bp to approximately 250 bp, and nowadays the sequence reads are up to 1,000 bp long [Medini, et al., 2008; Metzker, 2010]. 454 sequencing is advantageous compared to conventional Sanger sequencing as it generates a large amount of data per run (about a million reads) and that cloning is not required as an initial step. These factors reduce the time, labor and cost of projects [Noonan, et al., 2006; Wicker, et al., 2006; Sato, et al., 2011].

Database and web site development (Chapter I):

The amount of genomic or transcriptomic data are increasing exponentially since the development of the next generation sequencing methods such as 454 sequencing. Creation of publicly available databases are essential for sharing data between scientists, and the main database, GenBank, is the most important repository for nucleotide sequences in the world. Nevertheless, the creation of specialized databases, focusing on one species or lineage (e.g. *Arabidopsis thaliana*

[*http://www.arabidopsis.org*] or butterflies [Papanicolaou, et al., 2008]) can combine valuable information and provide specific tools to investigate the data efficiently (e.g. gene annotations, chromosome maps).

Bacterial artificial chromosome (BAC) library development (Chapter II, III):

BAC library construction follows four steps: First, long molecular weighted DNA (several million nucleotides) is isolated. Then, the DNA is partially digested using a restriction enzyme (e.g. HindIII). In the next step, the fragmented DNA (100-200 kb) is inserted by ligation in a plasmid vector. The final step is the transformation of *Escherichia coli* with the plasmid. A BAC library contains up to one million large DNA fragments in a stable stock of *E. coli*. BAC libraries are a powerful tool for genomic research and gene identification, and have successfully been used to sequence several genomes (e.g. hemp [van Bakel, et al., 2011], woodland strawberry [Shulaev, et al., 2011]) or chromosomal regions (e.g. bovine PAR [Das, et al., 2009])

Main objectives and research questions

The main objective of this thesis was to study sex chromosome evolution in *Silene latifolia*, using *S. vulgaris* as a reference. For this purpose, we used a bioinformatic approach. We first analyzed the transcriptomes of closely related species from the genus *Silene* and developed an EST database (Chapter I). Using BAC sequences, we then focused on the analysis of the *S. latifolia* pseudoautosomal region and its evolution by comparison with *S. vulgaris* homologous autosome (Chapter II) and then we analyzed evolution and divergence of the sex-specific region of X and Y chromosomes (Chapter III).

The main research goals were the following:

1) Develop a database of transcriptomic resources for the genus *Silene* in order to better characterize the transcriptome and estimate genetic variation in five closely-related species. (Chapter I)

2) Characterize the *Silene latifolia* pseudoautosomal region and compare it with the homologous autosome in *S. vulgaris*. (Chapter II)

3) Identify new sex-linked genes, estimate sex chromosome divergence and study evolutionary processes acting on *S. latifolia* sex chromosomes by comparing *S. latifolia* X and Y chromosomes and the homologous autosome in *S. vulgaris*. (Chapter III)

References

Arnold, A. P., Itoh, Y. & Melamed, E. (2008) A Bird's-Eye View of Sex Chromosome Dosage Compensation. *Annual Review of Genomics and Human Genetics* **9**, 109-127.

Bachtrog, D. *et al.* (2011) Are all sex chromosomes created equal? *Trends in genetics : TIG* **27**, 350-357.

Baker, H. (1947) *Melandrium Album*. *The Journal of Ecology* **35**, 274-282.

Bergero, R. & Charlesworth, D. (2011) Preservation of the Y Transcriptome in a 10-Million-Year-Old Plant Sex Chromosome System. *Current biology : CB* **21**, 1470-1474.

Bergero, R., Charlesworth, D., Filatov, D. A. & Moore, R. C. (2008) Defining regions and rearrangements of the *Silene latifolia* Y chromosome. *Genetics* **178**, 2045-2053.

Blackburn, K. B. (1923) Sex chromosomes in plants. *Nature* **112**, 687-688.

Bull, J. J. (1983) *Evolution of sex determinig mechanisms*. (Benjamin Cummings).

Burgoyne, P. S. (1982) Genetic homology and crossing over in the X and Y chromosomes of mammals. *Human Genetics* **61**, 85-90.

Carvalho, A. B. (2002) Origin and evolution of the Drosophila Y chromosome. *Current Opinion in Genetics & Development* **12**, 664-668.

Casci, T. (2011) Dosage compensation: What dosage compensation? *Nature Reviews Genetics* **12**, 2-2.

Čegan, R. *et al.* (2010) Structure and evolution of Apetala3, a sex-linked gene in *Silene latifolia*. *Bmc Plant Biology* **10**, 180.

Čermak, T. *et al.* (2008) Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. *Chromosome Research* **16**, 961-976.

Charlesworth, B. (1991) The evolution of sex chromosomes. *Science* **251**, 1030-1033.

Charlesworth, B. & Charlesworth, D. (2000) The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **355**, 1563-1572.

Charlesworth, D. (2002) Plant sex determination and sex chromosomes. *Heredity* **88**, 94-101.

Chibalina, Margarita V. & Filatov, Dmitry A. (2011) Plant Y Chromosome Degeneration Is Retarded by Haploid Purifying Selection. *Current biology : CB* **21**, 1475-1479.

Costich, D., Meagher, T. & Yurkow, E. (1991) A rapid means of sex identification in *Silene latifolia* by use of flow cytometry. *Plant Molecular Biology Reporter* **9**, 359-370.

Das, P. J., Chowdhary, B. P. & Raudsepp, T. (2009) Characterization of the Bovine Pseudoautosomal Region and Comparison with Sheep, Goat, and Other Mammalian Pseudoautosomal Regions. *Cytogenetic and Genome Research* **126**, 139-147.

Delichère, C. *et al.* (1999) SIY1, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein. *Embo Journal* **18**, 4169-4179.

Dellaporta, S. L. & Calderon-Urrea, A. (1993) Sex determination in flowering plants. *Plant Cell* **5**, 1241-1251.

Desfeux, C., Maurice, S., Henry, J.-P., Lejeune, B. & Gouyon, P.-H. (1996) Evolution of Reproductive Systems in the Genus Silene. *Proceedings of the Royal Society of London Series B-Biological Sciences* **263**, 409-414.

Di Stilio, V. S., Kesseli, R. V. & Mulcahy, D. L. (1998) A pseudoautosomal random amplified polymorphic DNA marker for the sex chromosomes of *Silene dioica*. *Genetics* **149**, 2057-2062.

Ellegren, H. (2011) Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Nature Reviews Genetics* **12**, 157-166.

Ezaz, T., Stiglec, R., Veyrunes, F. & Marshall Graves, J. A. (2006) Relationships between vertebrate ZW and XY sex chromosome systems. *Current biology : CB* **16**, R736-743.

Felsenstein, J. (1974) The evolutionary advantage of recombination. *Genetics* **78**, 737-756.

Filatov, D. A. (2005) Evolutionary history of *Silene latifola* sex chromosomes revealed by genetic mapping of four genes. *Genetics* **170**, 975-979.

Filatov, D. A. (2005) Substitution rates in a new Silene latifolia sex-linked gene, *SlssX/Y*. *Molecular Biology and Evolution* **22**, 402-408.

Filatov, D. A., Howell, E. C., Groutides, C. & Armstrong, S. J. (2009) Recent Spread of a Retrotransposon in the *Silene latifolia* Genome, Apart From the Y Chromosome. *Genetics* **181**, 811-817.

Fraser, J. A. & Heitman, J. (2004) Evolution of fungal sex chromosomes. *Molecular Microbiology* **51**, 299-306.

Graves, J. A. M. (1996) Mammals that break the rules:Genetics of Marsupials and Monotremes. *Annual Review of Genetics* **30**, 233-260.

Graves, J. A. M. & Peichel, C. (2010) Are homologies in vertebrate sex determination due to shared ancestry or to limited options? *Genome Biology* **11**, 205.

Handley, L.-J. L., Ceplitis, H. & Ellegren, H. (2004) Evolutionary Strata on the Chicken Z Chromosome: Implications for Sex Chromosome Evolution. *Genetics* **167**, 367-376.

Hedrick, P. W. (2005) *Genetics of populations, third edition*. (Jones and Bartlett Publishers, Inc.).

Henking, H. (1891) Ueber Spermatogenese und deren Beziehung zur Eientwicklung bei *Pyrrhocoris apterus*. *Zeitschrift fur Wissenschaftliche Zoologie* **51**.

Jaarola, M., Martin, R. H. & Ashley, T. (1998) Direct Evidence for Suppression of Recombination within Two Pericentric Inversions in Humans: A New Sperm-FISH Technique. *American journal of human genetics* **63**, 218-224.

Jamilena, M., Mariotti, B. & Manzano, S. (2008) Plant sex chromosomes: molecular structure and function. *Cytogenetic and Genome Research* **120**, 255-264.

Kaiser, V. B., Bergero, R. & Charlesworth, D. (2009) Slcyt, a Newly Identified Sex-Linked Gene, Has Recently Moved onto the X Chromosome in *Silene latifolia* (Caryophyllaceae). *Molecular Biology and Evolution* **26**, 2343-2351.

Kaiser, V. B., 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.

Kejnovský, E., Hobza, R., Čermak, T., Kubat, Z. & Vyskot, B. (2009) The role of repetitive DNA in structure and evolution of sex chromosomes in plants. *Heredity* **102**, 533-541.

Kejnovský, E. *et al.* (2006) Accumulation of chloroplast DNA sequences on the Y chromosome of <i>Silene latifolia</i&gt. *Genetica* **128**, 167-175.

Kejnovský, E. & Vyskot, B. (2010) *Silene latifolia*: The Classical Model to Study Heteromorphic Sex Chromosomes. *Cytogenetics and Genome Research* **129**, 250-262.

Kubat, Z., Hobza, R., Vyskot, B. & Kejnovský, E. (2008) Microsatellite accumulation on the Y chromosome in *Silene latifolia*. *Genome* **51**, 350-356.

Lahn, B. T. & Page, D. C. (1999) Four Evolutionary Strata on the Human X Chromosome. *Science* **286**, 964-967.

Lengerova, M., Moore, R. C., Grant, S. R. & Vyskot, B. (2003) The sex chromosomes of *Silene latifolia* revisited and revised. *Genetics* **165**, 935-938.

Liu, Z. Y. *et al.* (2004) A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* **427**, 348-352.

Lynch, M. (2007) *The Origins of Genome Architecture*. (Sinauer Associates, Inc. Publishers).

Ma, H. *et al.* (2004) High-Density Linkage Mapping Revealed Suppression of Recombination at the Sex Determination Locus in Papaya. *Genetics* **166**, 419-436.

Marais, G. A. B. *et al.* (2011) Multiple Nuclear Gene Phylogenetic Analysis of the Evolution of Dioecy and Sex Chromosomes in the Genus *Silene*. *PLoS ONE* **6**, e21915.

Marais, G. A. B. *et al.* (2008) Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. *Current Biology* **18**, 545-549.

Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Marín, I., Siegal, M. L. & Baker, B. S. (2000) The evolution of dosage-compensation mechanisms. *BioEssays* **22**, 1106-1114.

Matsunaga, S. (2006) Sex chromosome-linked genes in plants. *Genes & Genetic Systems* **81**, 219-226.

Matsunaga, S. *et al.* (2003) Duplicative transfer of a MADS box gene to a plant Y chromosome. *Molecular Biology and Evolution* **20**, 1062-1069.

Matsunaga, S. *et al.* (2005) An anther- and petal-specific gene SlMF1 is a multicopy gene with homologous sequences on sex chromosomes. *Genes & Genetic Systems* **80**, 395-401.

McKee, B. D. & Handel, M. A. (1993) Sex chromosomes, recombination, and chromatin conformation. *Chromosoma* **102**, 71-80.

Medini, D. *et al.* (2008) Microbiology in the post-genomic era. *Nature Reviews Microbiology* **6**, 419-430.

Metzker, M. L. (2010) Sequencing technologies [mdash] the next generation. *Nature Reviews Genetics* **11**, 31-46.

Ming, R., Bendahmane, A. & Renner, S. S. (2011) Sex Chromosomes in Land Plants. *Annual Review of Plant Biology* **62**, 485-514.

Ming, R. & Moore, P. H. (2007) Genomics of sex chromosomes. *Current Opinion in Plant Biology* **10**, 123-130.

Morgan, T. H. (1914) No crossing over in the male of *Drosophila* of genes in the second and third pairs of chromosomes. *The Biological Bulletin* **26**, 195-204.

Nicolas, M. *et al.* (2005) A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *Plos Biology* **3**, 47-56.

Noonan, J. P. *et al.* (2006) Sequencing and Analysis of Neanderthal Genomic DNA. *Science* **314**, 1113-1118.

Ohno, S. (1967) *Sex chromosomes and Sex-Linked Genes*. (Springer-Verlag).

Okada, S. *et al.* (2001) The Y chromosome in the liverwort Marchantia polymorpha has accumulated unique repeat sequences harboring a male-specific gene. *Proceedings of the National Academy of Sciences* **98**, 9454-9459.

Page, J. *et al.* (2006) Involvement of Synaptonemal Complex Proteins in Sex Chromosome Segregation during Marsupial Male Meiosis. *Plos Genetics* **2**, e136.

Painter, T. S. (1921) The Y-Chromosome in Mammals. *Science* **53**, 503-504.

Papanicolaou, A., Gebauer-Jung, S., Blaxter, M. L., Owen McMillan, W. & Jiggins, C. D. (2008) ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Research* **36**, D582-D587.

Parker, J. S. & Clark, M. S. (1991) Dosage sex-chromosome systems in plants. *Plant Science* **80**, 79-92.

Qiu, S., Bergero, R., Forrest, A., Kaiser, V. B. & Charlesworth, D. (2010) Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes. *Proceedings of the Royal Society B-Biological Sciences* **277**, 3283-3290.

Sandstedt, S. A. & Tucker, P. K. (2004) Evolutionary Strata on the Mouse X Chromosome Correspond to Strata on the Human X Chromosome. *Genome Research* **14**, 267-272.

Sato, K., Motoi, Y., Yamaji, N. & Yoshida, H. (2011) 454 sequencing of pooled BAC clones on chromosome 3H of barley. *BMC Genomics* **12**, 246.

Scotti, I. & Delph, L. F. (2006) Selective Trade-offs and Sex-Chromosome Evolution in *Silene latifolia*. *Evolution* **60**, 1793-1800.

Sharp, P. (1982) Sex chromosome pairing during male meiosis in marsupials. *Chromosoma* **86**, 27-47.

Shulaev, V. *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics* **43**, 109-116.

Široký, J., Lysák, M. A., Doležel, J., Kejnovský, E. & Vyskot, B. (2001) Heterogeneity of rDNA distribution and genome size in *Silene* spp. *Chromosome Research* **9**, 387-393.

Stevens, N. M. (1905) *Studies in Spermatogenesis with Especial Reference to the "Accessory Chromosome"*. Vol. 36 (Carnegie Institution of Washington).

Straub, T. & Becker, P. B. (2007) Dosage compensation: the beginning and end of generalization. *Nature Reviews Genetics* **8**, 47-57.

van Bakel, H. *et al.* (2011) The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology* **12**, R102.

Veyrunes, F. *et al.* (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Research* **18**, 965-973.

Vyskot, B. & Hobza, R. (2004) Gender in plants: sex chromosomes are emerging from the fog. *Trends in Genetics* **20**, 432-438.

Westergaard, M. (1958) in *Advances in Genetics* Vol. 9 (ed M. Demerec) 217-281 (Academic Press).

Wicker, T. *et al.* (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**, 275.

Wilson, E. B. (1905) Studies on chromosomes. II. The paired microchromosomes, idiochromosomes and heterotropic chromosomes in hemiptera. *Journal of Experimental Zoology* **2**, 507-545.

Winge, O. (1923) On sex chromosomes, sex determination and preponderance of females in some dioecious plants. *Comptes Rendus des Travaux du Laboratoire Carlsberg* **15**, 1-26.

Chapter I

Published article:

Comparative high-throughput transcriptome sequencing and development of SiESTa, the *Silene* EST annotation database

Nicolas Blavet [1], Delphine Charif [2], Christine Oger-Desfeux[3], Gabriel AB Marais [2], Alex Widmer [1]

BMC Genomics (2011), 12: 376

[1]Institute of Integrative Biology (IBZ), ETH Zurich, Universitaetstrasse 16, Zürich, 8092, Switzerland
[2]Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR 5558, Université Lyon 1, Villeurbanne, F-69622 cedex, France
[3]DTAMB/PRABI, IFR41, Université Lyon 1, Bâtiment Gregor Mendel, Villeurbanne, F-69622 cedex, France

Abstract

Background:

The genus *Silene* is widely used as a model system for addressing ecological and evolutionary questions in plants, but advances in using the genus as a model system are impeded by the lack of available resources for studying its genome. Massively parallel sequencing cDNA has recently developed into an efficient method for characterizing the transcriptomes of non-model organisms, generating massive amounts of data that enable the study of multiple species in a comparative framework. The sequences generated provide an excellent resource for identifying expressed genes, characterizing functional variation and developing molecular markers, thereby laying the foundations for future studies on gene sequence and gene expression divergence. Here, we report the results of a comparative transcriptome sequencing study of eight individuals representing four *Silene* and one *Dianthus* species as outgroup. All sequences and annotations have been deposited in a newly developed and publicly available database called SiESTa, the *Silene* EST annotation database.

Results:

A total of 1,041,122 EST reads were generated in two runs on a Roche GS-FLX 454 pyrosequencing platform. EST reads were analyzed separately for all eight individuals sequenced and were assembled into contigs using TGICL. These were annotated with results from BLASTX searches and Gene Ontology (GO) terms, and thousands of single-nucleotide polymorphisms (SNPs) were characterized. Unassembled reads were kept as singletons and together with the contigs contributed to the unigenes characterized in each individual. The high quality of unigenes is evidenced by the proportion (49%) that have significant hits in similarity searches with the *A. thaliana* proteome. The SiESTa database is accessible at http://www.siesta.ethz.ch.

Conclusion:

The sequence collections established in the present study provide an important genomic resource for four *Silene* and one *Dianthus* species and will help to further develop *Silene* as a plant model system. The genes characterized will be useful for future research not only in the species included in the present study, but also in related species for which no genomic resources are yet available. Our results demonstrate the efficiency of massively parallel transcriptome sequencing in a comparative framework as an approach for developing genomic resources in diverse groups of non-model organisms.

Keywords: cDNA library, database, EST, SNP, *Silene*.

Background

The genus *Silene* (Caryophyllaceae) consists of several hundred species with a mainly holarctic distribution. Because species vary widely in their breeding system, sex determination and ecology, the genus has historically played an important role in genetic and ecological studies dating back to Mendel and Darwin. More recently the genus has emerged as a model system in plant ecology, evolution, genetics and developmental biology [1]. However, a major limitation of using *Silene* as model system is the near absence of genomic information pertaining to the genus. Recently the first EST library was published [2], based on normalized cDNA sequences derived from different reproductive tissues of the dioecious species *Silene latifolia*. In *S. latifolia*, sex is determined by heteromorphic sex chromosomes. As in mammals, *S. latifolia* males are heterogametic (XY) and females are homogametic (XX). In contrast to the evolution of mammalian sex chromosomes which evolved about 150 million years ago (my) [3], the age of *S. latifolia* sex chromosomes has been estimated to be about 10 my [4]. The overwhelming majority of *Silene* species are however not dioecious and lack sex chromosomes. These species are either hermaphroditic or gynodioecious, such as in the case of the widely distributed bladder campion *S. vulgaris*. The relatively recent evolution of sex chromosomes in *S. latifolia* and the availability of closely related species without sex chromosomes, make the genus an ideal target for studying the evolution of sex chromosomes.

The closest relatives of *S. latifolia* are a group of dioecious species, including *S. marizii* and *S. dioica*, with which *S. latifolia* often hybridizes upon secondary contact [5]. The two species occupy different habitats [6] and differ in flower color and odor [7]. As is frequently the case in pairs of closely related plant species where multiple barriers contribute to reproductive isolation [8, 9], reproductive isolation between *S. dioica* and *S. latifolia* is incomplete, and the occurrence of gene flow across species boundaries leads to porous genomes [10]. A recent population genomic analysis revealed that neutral processes, introgression and adaptive divergence shape species differences [11]. However, the extent to which genes underlying floral trait or habitat differences contribute to adaptive divergence has never been investigated. A major hindrance to investigate the genetic causes of adaptive divergence is that the *Silene* genome remains largely unexplored. The present study tackles this problem by comparative high-throughput transcriptome sequencing of *Silene latifolia*, *S. dioica*, *S. marizii*, *S. vulgaris* and *Dianthus superbus*. The sequences generated in this study are annotated and publicly available through SiESTa, the *Silene* EST annotation database (www.siesta.ethz.ch).

Silene and *Dianthus* species vary greatly in genome size and have different haploid chromosome numbers (n=12 and n=15 respectively). With a haploid genome size of about 2,646 Mbp [12], the *S. latifolia* genome is similar in size to corn (about 2,671 Mbp) [13]. In contrast, the genome size of *S. vulgaris* (1,103 Mbp) [14] is less than half that of *S. latifolia* and some *Dianthus* species have even smaller genomes (613 Mbp) [15]. Thus, genome sizes differ by a factor of two between *Dianthus* and *S. vulgaris* and by a factor of four between *Dianthus* and *S. latifolia*.

To further develop genomic resources for the genus *Silene*, and especially for the dioecious species related to *S. latifolia*, we performed comparative high-throughput transcriptome sequencing using 454 pyrosequencing technology. This method is increasingly used for EST sequencing in both animals [16-18] and plants [19-21]. Advantages over conventional Sanger sequencing based EST projects are the large amount of data generated per run and the fact that cloning is not required as an initial step, factors which substantially reduce the time, labor and cost involved [22, 23].

Here we present the results of comparative transcriptome sequencing in seven *Silene* individuals representing four species, and one *Dianthus* outgroup. These species are closely related and include species with and without sex chromosomes, also differing substantially in genome size. A total of 1,041,122 EST reads, totaling 242,341,741 bp, were obtained from two complete 454 pyrosequencing runs and processed and assembled in the SiESTa database. These ESTs provide a unique and novel resource for ecological and evolutionary studies in *Silene* and *Dianthus*.

Results and Discussion

SiESTa database characteristics

454 pyrosequencing of eight individual cDNA libraries derived from one *Silene latifolia* male (SlM) and two females (SlF, SlFf), one *S. dioica* male (SdM) and female (SdF), one *S. marizii* male (SmM), and one individual of the each of the hermaphroditic species *S. vulgaris* (SvH) and *Dianthus superbus* (Ds) lead to a total of 1,041,122 EST reads. The number of nucleotides sequenced per library varied between 25 million and 46 million in SlFf and Ds respectively (Table 1). In contrast to studies using normalized libraries [17, 19, 21, 24], we used non-normalized libraries with the advantage of searches not identifying weakly expressed genes and a reduced chance of finding alternative splicing variants [25]. However, these factors may negatively impact upon the ability to build contigs. Our reads were assembled into 93,627 contigs (38,256,084 bp) and 309,074 singletons (69,524,702 bp), with an overall total of 402,701 unigenes (107,780,786 bp) that were deposited in a newly developed database called SiESTa (*Silene* EST annotations) (Table 1).

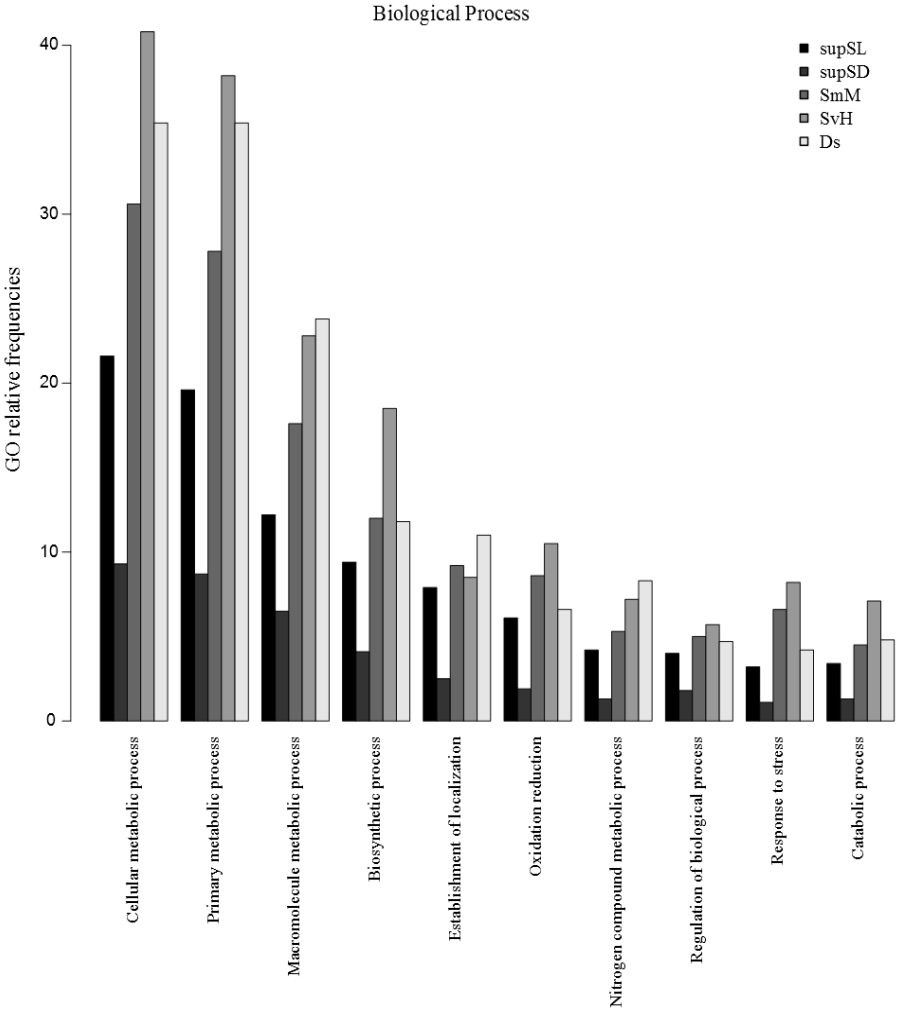
Table 1: SiESTa sequence content. (#) Units in thousands of sequences.

	Library									
	SlM	SlF	SlFf	SdM	SdF	SmM	SvH	Ds	supSL	supSD
# ESTs	119	136	110	113	115	127	123	198	347	228
Nucleotides (Mbp)	28	32	25	27	27	29	29	46	85	54
# Unigenes	61	40	49	71	69	51	32	30	129	129
% Contigs	17	34	17	15	17	28	36	43	24	18
% ESTs in contigs	57	81	63	47	50	71	83	91	72	54
Avg. EST length (bp)	235	232	225	235	233	230	234	232	230	233
Avg. contig length (bp)	403	430	413	395	385	392	401	463	422	396

All reads may be accessed under the accession number ERP000371 in the NCBI Sequence Read Archive and all contigs are available in Genbank Transcriptome Shotgun Assembly (TSA) under the accession numbers JL382689 - JL473671. The unigenes were sorted into eight individual libraries with an average size of 130,140 ESTs and 11,703 contigs per library. Two super-libraries, supSL and supSD, containing the sequences of *S. latifolia* and *S. dioica* individuals, respectively, were also created. Their sizes are 129,456 and 129,252 superunigenes respectively. As reported elsewhere in recent studies [18, 21, 24] short EST reads from 454 sequencing runs may be assembled and annotated to effectively characterize the gene space of non-model organisms. Average read length in our study was 232 bp, close to the lengths obtained in other recent studies that used the GS-FLX platform for sequencing [17, 18, 21], but substantially longer than early studies that used the GS-20 platform where read lengths were 100 to 110 bp [16, 19, 21]. Between 47% and 91% of EST reads were assembled into contigs (for SdM and Ds respectively), while the remainder were kept as singletons (Table 1). Similar percentages of reads assembled into contigs were found in other studies, ranging from 40% to 48% [19, 25] to 88% and 90% [16-18, 21] in both plants and animals. The frequency distribution of ESTs per unigene showed a hyperbolic distribution (Additional File 1), with a single EST read available for most unigenes (singletons), whereas only a small proportion of unigenes include a large number of EST reads. Given that our libraries were not normalized, one can use

the number of ESTs per unigene as an estimate of expression level [26]. This implies that the unigenes composed of many EST reads are highly expressed. An analysis of the ten most strongly expressed genes (i.e. the unigenes with the highest numbers of EST reads) in each library revealed that these correspond to only fifteen different genes (Additional File 2).

Figure 1: Relative frequencies of the most represented Biological Process GO sub-classes across libraries. Figure 1 shows the ten most frequent biological process GO terms at level 3 in the five species *Silene latifolia*, *S. dioica*, *S. marizii*, *S. vulgaris* and *Dianthus superbus*.



Of these, two were found in more than four out of the eight libraries analyzed. Our results indicate that one of these genes codes for an alpha-tubulin homologue of *Arabidopsis thaliana* (present in SlM, SlFf, SdM, SdF, SvH, Ds) and the second for a homologue of a predicted ORF in *Pinus koraiensis* (present in SlM, SdM, SdF, Ds) (Additional File 2). Most of these genes are housekeeping genes that are known to be highly expressed [27-30]. GO annotations revealed that a large number of contigs had a term assigned to them. Of the 93,048 contigs tested (from supSL, supSD, SmM, SvH and Ds), 46,217 were annotated with a GO term. The large number of GO terms annotated in the libraries (53%) further confirms the quality of the contigs of our database.

A comparison of the ten most represented GO annotations reveals substantial homogeneity in the composition of our libraries (Figure 1 and Additional File 3). In addition, the analysis of the ten most represented gene groups, based on the most expressed GO Slim for plants [31], confirmed the homogeneity of gene expression in the buds of the different species studied (Table 2). Not surprisingly, genes involved in cellular component organization translation and transcription are highly expressed in all our individuals.

Table 2: Expression differences among all eight libraries for the ten most frequently represented GO Slim terms. (Biological processes and Molecular functions). For each library, the expression percentage is calculated as the number of reads included in contigs matching a term divided by the total number of reads included in all contigs. Terms are sorted by the total expression common to all libraries in descending order.

GO Slim term	Expression percentage							
	SlM	SlF	SlFf	SdM	SdF	SmM	SvH	Ds
Response to stress	2.3%	8.1%	2.3%	1.5%	2.8%	6.6%	8.2%	4.2%
Cellular component organization	2.6%	2.9%	2.0%	1.2%	2.0%	3.7%	3.5%	7.5%
Translation	1.3%	7.3%	1.8%	1.0%	4.7%	2.2%	3.4%	2.5%
Photosynthesis	2.2%	6.2%	2.3%	0.8%	2.7%	4.6%	3.8%	0.4%
Kinase activity	2.8%	2.7%	2.5%	0.8%	1.1%	3.2%	3.2%	4.6%
Cell communication	2.3%	2.2%	2.7%	1.2%	1.7%	2.1%	1.9%	2.2%
Signal transduction	2.2%	2.2%	2.6%	1.1%	1.6%	2.1%	1.9%	2.1%
Response to abiotic stimulus	1.0%	3.5%	1.2%	0.5%	1.2%	2.5%	3.1%	2.2%
Transcription	1.2%	1.8%	1.1%	0.8%	1.8%	1.3%	1.4%	1.4%
Response to biotic stimulus	0.4%	2.1%	0.4%	0.3%	0.5%	1.2%	1.4%	0.2%

Ninety-nine percent of the unigenes have an ORF predicted by prot4EST (Table 3). About 45% of the predictions are based on Blast similarities, 28% are predicted by ESTScan and the remaining 27% correspond to the longest reading frames of the sequences.

Table 3: Prot4EST ORF prediction results. ORF prediction based on similarity with BLAST results, ESTScan prediction and longest reading frame.

Library	Predicted ORFs				
	Similarity	ESTScan	Longest ORF	Average length	Total
supSL	39%	31%	31%	208	129251
supSD	29%	36%	35%	193	129154
SmM	36%	28%	35%	205	50798
SvH	60%	24%	16%	245	32131
Ds	59%	23%	18%	259	29668

The *Silene* genome is known to include a large number of repeated elements [32-34] and we had to filter out such elements because they contribute to assembly problems. On average, 23,000 reads per library matched repeated elements (data not shown). Numerous repeated elements have recently been identified in *S. latifolia* [32], which make easier contig construction even a large diversity of elements still remains to be characterized.

Our newly developed EST resources for *Silene* and *Dianthus*, with 130,140 ESTs on average, are comparable to the resources available for *Helianthus annuus* (133,684 ESTs) and for *Populus trichocarpa* (89,943 ESTs) [NCBI EST database of October 1, 2010].

The SiESTa database

The newly developed SiESTa database provides several tools that facilitate data and information extraction. The first tool is the unigene search engine (Unisearch), which allows entering a list of unigene or superunigene names. From these, the user directly obtains the link to the sequence annotations and can download all sequences in fasta format. The second tab called "Libraries" allows users to navigate the database. Information about the different libraries, including species identity and used tissue, sex of the individual and the total number of unigenes/superunigenes in the library are presented in a table. By selecting the link on the unigene number, the user may download the complete set of unigenes from each library. The link attached to the library name enables the user to access the unigene

table which lists unigenes, their lengths, the number of ESTs per unigene and the best hit with Uniprot. Selecting any unigene provides access to the unigene sequence, a picture of the EST alignment that is linked to the alignment in fasta format and a table with the five best hits with the *A. thaliana* proteome and Uniprot. In the case of superunigenes, additional information is available, including ORF prediction and the list of unigenes that are part of the superunigene. The third tool, "Query", allows users to search for genes using their annotations. The fourth tab provides a link to a Gene Onthology formatted browser interface from which it is possible to obtain GO annotation for contigs of each species included in SiESTa. The fifth tool is a BLAST search engine that allows users to search for nucleic or protein sequence homology within the eight SiESTa libraries using BLASTN, TBLASTX or TBLASTN searches. The sixth and seventh tabs "Tools" and "FAQ" provide all this information on the web-site.

Homology with plant model species

In order to annotate and evaluate the quality of our reads and of our assemblies, we performed BLASTX searches to align both contigs and singletons from each library with *A. thaliana*, *Vitis vinifera* and *Populus trichocarpa* proteomes and Uniprot (Table 4). On average, 49% of the contigs and 27% of the singletons had a significant hit to the *A. thaliana* proteome. We evaluated the redundancy of the hits and found that on average 32% and 18% of contigs and singletons respectively match strictly different *A. thaliana* protein sequence. These non-redundant protein sequences (noted 'unique' in Table 4) revealed that some of our unigenes could come from distinct regions of the same gene. Compared to the proportions of hits with *A. thaliana*, we noticed an increase of the average percentage of matches for both contigs and singletons respectively, with *V. vinifera* (+0.2% and +1.9%), *P. trichocarpa* (+1.9% and +5.7%) and with Uniprot (+8.8% and +14.7%). Nevertheless, even though most of the *Silene* genes have a match with the three model species, across all libraries, an average of 62, 87 and 189 hits are exclusive to the proteomes of *A. thaliana*, *V. vinifera* or *P. trichocarpa*, respectively (Table 5). Such differences among the investigated proteomes might suggest that *P. trichocarpa* is more closely related to *Silene* than *A. thaliana* and *V. vinifera*. However, the phylogeny of angiosperms compiled by Bremer and coworkers [35] reveals that *Silene* (Caryophyllales) is phylogenetically equally distant from *Vitis* (Vitales), *Populus* (Malpighiales) and *Arabidopsis* (Brassicales). The causes of these observed differences are currently unknown, but a possible explanation may be differential gene loss during the evolution of these plant lineages as observed in other plants [36, 37].

Table 4: BLASTX hits of contigs and singletons in the eight individual libraries with different proteomes. Table 4 shows the number of hits for both contigs and singletons. Non-redundant accessions are recorded in the '% unique' column. A cut-off E-value of 1E-4 was used for each database.

Library	**Contigs**							
	A. thaliana		*V. vinifera*		*P. trichocarpa*		Uniprot	
	%hit	% unique	%hit	% unique	%hit	% unique	%hit	% unique
SlM	31%	21%	31%	20%	33%	23%	41%	33%
SlF	76%	49%	76%	46%	78%	53%	78%	66%
SlFf	35%	24%	35%	24%	36%	26%	56%	47%
SdM	18%	13%	19%	13%	22%	14%	31%	22%
SdF	31%	22%	31%	21%	32%	24%	49%	43%
SmM	55%	35%	54%	33%	56%	39%	57%	48%
SvH	74%	48%	74%	45%	76%	52%	76%	64%
Ds	73%	47%	73%	43%	74%	51%	75%	63%

Library	**Singletons**							
	A. thaliana		*V. vinifera*		*P. trichocarpa*		Uniprot	
	%hit	% unique	%hit	% unique	%hit	% unique	%hit	% unique
SlM	14%	9%	16%	9%	21%	10%	32%	19%
SlF	42%	27%	44%	26%	46%	31%	47%	38%
SlFf	17%	11%	18%	11%	22%	12%	45%	30%
SdM	10%	6%	13%	6%	21%	7%	32%	14%
SdF	16%	10%	18%	10%	22%	11%	39%	25%
SmM	26%	17%	28%	17%	30%	20%	32%	25%
SvH	44%	30%	47%	29%	50%	33%	55%	41%
Ds	46%	33%	46%	32%	48%	36%	49%	42%

Table 5: *Silene* contigs with hits that are exclusive to the *A. thaliana, V. vinifera*, and *P. trichocarpa* proteomes. In the second column are numbers of contigs with hits occurring in all three species; the following columns give the numbers of contigs with hits exclusively to *A. thaliana (*At*)* (3rd column) (these sequences do not have significant matches with either *V. vinifera* or *P. trichocarpa*), *V. vinifera (*Vv*)* (4th column) and *P. trichocarpa (*Pt*)* (5th column).

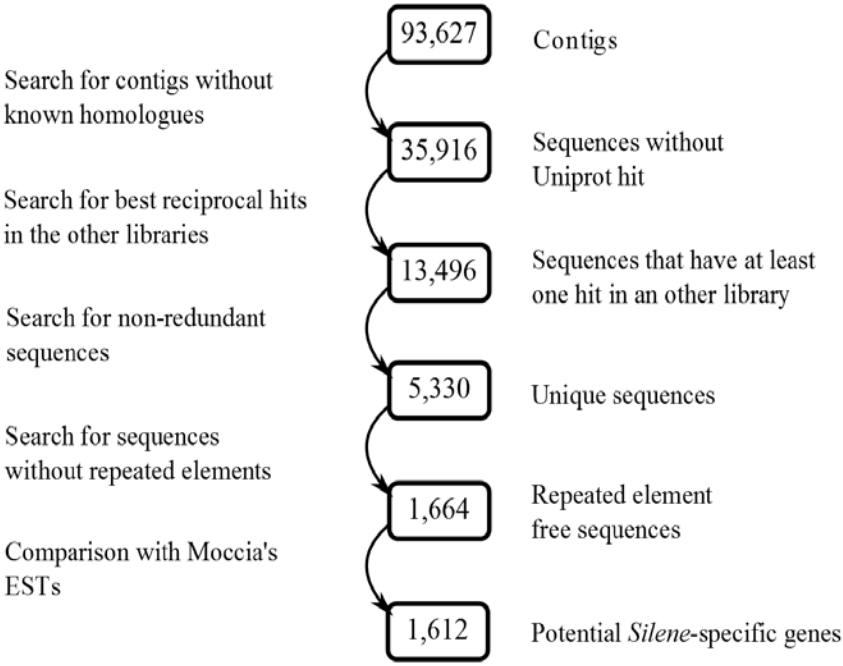
Library	# hits in At,Vv,Pt	# hits At	# hits Vv	# hits Pt
SlM	2886	74	78	220
SlF	9887	49	102	152
SlFf	2732	69	104	105
SdM	1758	32	77	391
SdF	3339	99	70	212
SmM	7418	66	107	187
SvH	8270	44	71	130
Ds	8885	64	93	120
Mean	5646	62	87	189

Contigs lacking known homologs

For 34,848 contigs out of the 93,627 contigs assembled in the present study, homologous genes could not be identified through BLAST searches against several databases (Additional File 4). Of these 34,848 contigs, 22,365 were found only in a single library, whereas 12,483 contigs correspond to sequences found in more than one library. After removing redundancies, 4,931 unigenes remained that had no significant hit in BLASTX searches against Uniprot and were found in at least two libraries. A substantial proportion of these unigenes (69 %) had similarities with additional repeated elements identified from *S. latifolia* (J. Macas, unpublished results) and was removed. The remaining 1,467 contigs were compared with the EST library reported by Moccia et al. [2], and 14% of these contigs had a significant hit. For some of the corresponding ESTs of Moccia et al. [2] there is a significant hit with Uniprot, most likely because these sequences were longer and contained coding sequences, and we were able to infer homology for 56 contigs. After removing one further transposon homologue and 40 sequences consisting of UTR regions, only 15 sequences had a good homology with gene coding regions, but 7 of them had undetermined functions. Of the remaining contigs, 1,411 sequences looked like potential Caryophyllaceae-specific genes (Figure 2). Yang and coworkers [38] recently investigated species-specific genes in the *A. thaliana*, *P. trichocarpa* and

Oryza sativa proteomes. Inter-proteome comparisons revealed that 165 of 26,784 proteins (0.6%) are exclusive to *A. thaliana*, as these proteins have no homologue in either *P. trichocarpa* or *O. sativa* and also in *Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea* (similar results are indicated in *P. trichocarpa and O. sativa*). Similarly, we searched our libraries for genes that are specific for *Silene or Dianthus*. We have identified 1,411 sequences from our studied species that may correspond to Caryophyllaceae-specific genes. These sequences represent about 1.5% of all contigs. The proportions of species-specific proteins identified by Yang et al. [38] in *A. thaliana*, *P. trichocarpa* and *O. sativa* are 0.6%, 0.2% and 1.1% respectively. Our estimate is also less than 2%, but we do not have a sequenced genome available, and consequently, some genes are certainly missing in the calculation and some may have been counted more than once.

Figure 2: Identification of potential Caryophyllaceae-specific genes. The first step identifies sequences without known homologues in reference species; the second and the third steps select sequences that are found in at least two SiESTa libraries. Sequences that partially match repeated elements are removed. In our final step we compared the remaining sequences with the *Silene* EST library of Moccia et al. [2] to identify potential Caryophyllaceae-specific genes.



Possible biases introduced in our estimates include that 1) we used contigs built from cDNA sequences. These are different from full-length protein sequences because they are oftentimes only fragments of coding sequences and it is possible that different contigs contain non-overlapping regions of the same gene as revealed by Table 4. 2) Singletons were not included in this analysis because their quantity prevented computation. 3) The lack of well-annotated genome sequences of species closely related to *Silene* reduced chances to find more homologous sequences. 4) Our EST libraries were non-normalized, and it is thus possible that further Caryophyllaceae-specific genes were missed because they were not sufficiently expressed to be represented in our database. Points 2 and 4 might increase the proportion of Caryophyllaceae-specific genes while points 1 and 3 might decrease it. Further studies will reveal whether these sequences are indeed specific to Silene and what their functions are. For this purpose, our SiESTa database provides a valuable resource.

SNP detection, validation and heterozygosity estimates

QualitySNP predicted between 4,500 and 12,000 polymorphic sites in our eight libraries, the results of SNP analysis for each individual library being presented in Table 6. There are on average 31 SNPs per 10,000 bp of expressed sequences in *Silene* and *Dianthus* and most SNPs are substitutions (78.6%).

Table 6: Library SNP content. * Only contigs assembled from at least 4 reads were considered. The total length of these contigs was used to calculate the percentage of heterozygous positions. All SNPs that are not due to substitutions are indels.

Library	SlM	SlF	SlFf	SdM	SdF	SmM	SvH	Ds
Contigs*	2909	5486	2287	2993	2912	4982	5028	6094
Contigs with SNPs	1221	1517	976	1333	1078	1709	1619	2513
SNPs	6648	6308	5576	6307	4653	7361	7381	12282
Substitutions	4847	5165	3873	5356	3516	5681	6402	9927
% Transitions / transversions	52/48	61/39	47/53	63/37	56/44	57/43	60/40	61/39
% heterozygous positions	0.39	0.19	0.43	0.38	0.28	0.26	0.27	0.32

Of the 48 polymorphic positions predicted by qualitySNP that were selected for validation, 32 (67%) of SNPs were confirmed by Sanger sequencing of PCR products. Polymorphic positions that are not associated with single-nucleotide repeats were

selected for validation, because pyrosequencing is known to experience difficulties in sequencing these regions [39]. We observed that such regions often induce incorrect predictions of SNPs by qualitySNP. From our SNP data we cannot directly estimate nucleotide diversity, because our SNP estimates are based on reads from single individuals. However, the detected polymorphisms allow estimating heterozygosity in the different species. Polymorphism varies between 19 and 43 SNPs per 10kb of expressed sequences for SlF and SlFf respectively (Table 6). Similar values were reported in maize [20], with between 33 and 47 SNPs per 10kb, in *Oryza sativa* [40], with around 30 SNPs per 10kb, and in *A. thaliana* [41], with around 40 SNPs per 10kb. By analyzing 27 genes in *Silene latifolia*, a recent study estimated a polymorphism rate of about 551 SNPs per 10kb [42], which is ten times higher than in other plants. The origin and the large number of individuals sampled in that study is probably the reason for these high estimates. Our results suggested that there is no difference in the proportion of heterozygous positions between the dioecious species and the gynodioecious species in the genus *Silene* (mean of 0.32 and 0.3 respectively). However, a lower level of heterozygosity was detected in the *S. latifolia* female library SlF (0.19) compared to other libraries. SIF belongs to an inbred line, which explains the low polymorphism exhibited by this individual. On the contrary, an increase of polymorphism was detected in the F1 individual SlFf, which was obtained by crossing the two other *S. latifolia* plants, SlM and SlF. Polymorphisms detected in this individual provide valuable markers for the development of a linkage map for *S. latifolia* and its sex chromosomes.

Conclusions

The high quality EST database SiESTa provides valuable resources for molecular ecologists studying Caryophyllaceae, particularly for the genus *Silene*. It provides the necessary molecular resources to develop microsatellite and SNP markers for linkage mapping and population genetic analyses, provides access to candidate genes for specific traits, such as heavy metal tolerance or flower color variation, and enables identification of X and Y-linked gene copies. Moreover this online database (www.siesta.ethz.ch) provides access to sequences and annotations of four *Silene* and one *Dianthus* species lacking fully-sequenced genomes. The two 454 sequencing runs described in this study generated more than one million sequencing reads, allowing for the identification of about 74,000 genes and about 56,000 SNPs. We hope that the availability of these resources will encourage further investigations into the genomics and evolutionary biology of *Silene* and related species.

Methods

RNA extraction & cDNA sequencing

We extracted RNA from one flower bud belonging to eight individuals of five closely related species; three dioecious species: *Silene latifolia, S. dioica* and *S. marizii* and two gynodioecious species: *S. vulgaris* and *Dianthus superbus*. For *S. latifolia* and *S. dioica*, both sexes were included in this study, whereas for *S. marizii*, only a male individual was used. For the two gynodioecious species, we used flowers from hermaphrodite individuals.

Flower buds prior to anthesis were collected from plants grown in a greenhouse under long day conditions at the ETH Zurich and were immediately frozen in liquid nitrogen. Flower buds of *Dianthus superbus* were collected in the field (Davos, Switzerland) and immediately placed in RNALater (Ambion) and stored at room temperature for three days. Total RNA was isolated using TriFast (PeqLab), stored in liquid nitrogen, and sent to GATC Biotech (Konstanz, Germany) for library construction. cDNA was prepared using the SMART™ PCR cDNA Synthesis Kit (Clontech), concatenated by ligation, nebulized, tagged and sequenced using the GS FLX protocol (Roche). Two tagged libraries were combined in half a picotiter plate for sequencing.

EST processing

All sequences were generated in two complete runs on a Roche GS-FLX 454 pyrosequencing machine and eight fasta files containing trimmed reads were extracted from the sff files. Short reads (<50 nt), as well as reads derived from mitochondria and plastids were removed using SeqClean [43]. Repeated elements were then removed using RepeatMasker [44] with a Viridiplantae database compiled in RepBase (01/08/2008 version) [45] to which we added the *Silene latifolia* – specific repeated elements identified by Cermak and coworkers [32]. EST reads were then clustered and contigs built using TGICL [46] with the default parameters (95% of identity and 40 bp minimum for sequence overlap). In addition to the resulting contigs, the remaining singletons (unique EST reads) were then added to the database as unigenes.

We constructed separate EST libraries from all eight individuals used in this project: a *Silene latifolia* male (SlM), two *S. latifolia* females, which are “mother” and “daughter” (respectively SlF and SlFf), one male and female each of *S. dioica* (SdM and SdF, respectively), a *S. marizii* male (SmM), as well as one individual each of *S. vulgaris* (SvH) and *Dianthus superbus* (Ds). Additionally, two super-libraries were constructed that combine sequences from the three *S. latifolia* (called supSL) and the

two *S. dioica* individuals (supSD), respectively. Due to large demand placed on CPU use, ESTs from chloroplast and mitochondrial genes were removed from the assembly process for supSL, thereby reducing the number of reads used from 365,089 to 347,047.

Unigene annotation

Similarity searches were carried out in two steps, the first of which involved BLAST similarity searches [47] of the contig sequences versus Uniprot (UniProt Rel. 13 = SWISS-PROT 55 + TrEMBL 38, 29 April 2008) and added the annotation results to the database. Because of the large number of contigs being searched for similarities, we used PC clusters at the French National Institute of Nuclear Physics and Physics of Particles located in Lyon (IN2P3). In our second step, BLASTX searches against the *Arabidopsis thaliana* proteome were then performed and the results included in the database. In both steps, the E-value cut-off used was 1E-04 and the five best hits were included in the SiESTa database.

Prot4EST [48] was then used to predict open reading frames (ORFs) using the following criteria: 1) if a unigene had a significant BLAST hit with Uniprot, the ORF from the best hit was used as template for ORF prediction; 2) if the unigene had no BLAST hit, ESTScan [49] predicted peptides that were used to predict ORFs. We used *Arabidopsis thaliana* as 'training model' for ESTScan prediction (codon usage matrix from May 2009 [50]); 3) if ESTScan failed to predict any peptides, the longest ORF from the 6-frame was retained. In addition to prot4EST predictions, we retained all ORFs that were at least 180 bp long when they were in a different reading frame than the prediction done at step 2) and 3). We ran prot4EST on individual libraries Ds, SmM and SvH, and on the super-libraries supSL and supSD. Predicted ORFs were added to the database. Gene ontology (GO) annotation was carried out using Blast2GO [51]. In the mapping step, a pool of candidate GO terms was obtained for each unigene by retrieving GO terms associated with the 20 first BLAST hits (BLASTx against NCBI nr: E-value cut-off: 1e-3; HSP coverage percentage: 0.33). In the annotation step, reliable GO terms were then selected from the pool of candidate GO terms by applying the core annotation function of Blast2GO. Default parameters were used (GO weights: 5; score threshold: 55; Evidence code weights: default). In order to complete the functional annotation (based on BLAST) with protein domain information, InterproScan [52] was run (based on the longest unigene's ORF) and GO terms associated with protein domains were merged with the GO terms kept at the annotation step.

We used the tools provided by the GO consortium to build our own GO database dedicated to our species by loading the 'unigene products - GO' association files found with Blast2GO [http://www.geneontology.org/godatabase/ archive/full/ 2009-03-01/]. To search and browse the gene ontology and visualize the gene products associated with a particular GO term, we implemented an instance of the Amigo browser [http://www.geneontology.org/GO.tools.browsers.shtml].

Homology investigation

Annotation of the unigenes led to the identification of two major groups of unigenes: the first one with matches to Uniprot and the second without matches. We used the contigs in the first group to estimate the proportion of homologous genes shared with the plant model species *Arabidopsis thaliana*, *Vitis vinifera* and *Populus trichocarpa*.

We then used unigenes of the second group, i.e. unigenes for which no hits with Uniprot were obtained, to assess whether potentially new Caryophyllaceae-specific genes could be found. To avoid spurious results, we removed all unigenes in this group that were found only in a single individual. To do so, we performed pairwise BLASTN searches between libraries and removed all sequences without hits in other libraries. From the sequences with hits in other libraries, we kept only one sequence for further analysis. Finally, using the database recently developed by Macas and coworkers (unpublished results) of newly identified *S. latifolia* repeated elements, we tested whether the remaining contigs contained repeated elements that were not removed by RepeatMasker. When more than 20% of the contig length resulted from repeated elements, contigs were discarded. We then compared the remaining sequences with the EST library developed by Moccia and coworkers [2]. This EST library was established by standard Sanger sequencing of a normalized cDNA library. It contains only 3,105 unigenes, but these are on average longer than our 454-based unigenes. Moreover, these unigenes were used in the construction of a custom cDNA microarray that has been used in expression analyses by Aria Minder (unpublished results). This comparison firstly allowed us to identify homology that we missed due to the commutative property of homologies and secondly, to assess the proportion of genes that lack annotation which are expressed in *S. latifolia*.

SNP detection, validation and heterozygosity estimation

SNPs were identified using qualitySNP [53], a haplotype-based SNP finder that groups sequences sharing the same nucleotides at each polymorphic site using the resulting clusters defined with CAP3 [54] and predicts if the SNP position is supported. We used the CAP3 clusters build during TGICL assembly. We ran

qualitySNP with the default parameters on the contigs of each individual library. The software searched for polymorphisms in contigs formed by at least four ESTs and identified all potential SNPs that occurred at least twice. Tips of each sequence (30bp in 5' end and 20% of sequence length in 3' end) were set as low quality (LQ) regions and the rest as high quality regions following the method used by Tang and coworkers [53]. Only high quality SNPs with a minimum score of two were retained. Since we did not search for polymorphisms common to individuals but rather within individuals, SNPs identified by qualitySNP were used to estimate heterozygosity within each individual.

In order to confirm their quality, substitutional SNPs identified by qualitySNP were selected for validation. Specifically, we selected unigenes present in both libraries, SlM and SlFf that contained SNPs (i.e. heterozygous positions in either SlM or SlFf). SNP Primers were designed using Primer3 [55] to be located at least 40bp from the SNP and PCR amplify fragments of at least 200bp in length. Seventeen primer pairs were designed in order to validate 48 SNPs.

The SiESTa database may be accessed using the login and password below at http://www.siesta.ethz.ch
login: 5!LeN3
password: 4cent5ante4

Author contributions

NB collected and analyzed data and drafted the manuscript. DC designed the database, analyzed the data and assisted in writing. CO participated in data capture. GABM participated in coordinating the study and assisted with data analysis, interpretation and preparation of the manuscript. AW conceived and coordinated the study and assisted with drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank M. Gnesotto for assistance with SNP analysis, T. Torossi, B. Blattmann and C. Michel for their assistance in the lab and A. Minder, M.D. Moccia, M. Pâris and N. Zemp for valuable comments on the manuscript, and M. Scarborough for is help with English writing. Thanks also to J. Macas and E. Kejnovsky for providing the sequences of repeated elements. IN2P3 (CNRS) kindly granted us access to their computer resources. S. Delmotte, L. Humblot, S. Penel, B. Spataro and G. Perrière assisted in using the PBIL server (BBE - UMR CNRS 5558). Financial support for this study was provided by the Genetic Diversity Centre (GDC)

at the ETH Zurich and an ETH Zurich grant (TH-07 06-3) to AW. The work done in Lyon was supported by Agence Nationale de la Recherche (ANR) to GABM (ANR-08-JCJC-0109).

References

1. Bernasconi G, Antonovics J, Biere A, Charlesworth D, Delph LF, Filatov D, Giraud T, Hood ME, Marais GAB, McCauley D, Pannell JR, Shykoff JA, Vyskot B, Wolfe LM, Widmer A: ***Silene*** **as a model system in ecology and evolution**. *Heredity* 2009, **103**(1):5-14.

2. Moccia MD, Oger-Desfeux C, Marais GAB, Widmer A: **A White Campion (*Silene latifolia*) floral expressed sequence tag (EST) library: annotation, EST-SSR characterization, transferability, and utility for comparative mapping**. *BMC Genomics* 2009, **10**.

3. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Gruzner F, Deakin JE, Whittington CM, Schatzkamer K, Kremitzki CL, Graves T, Ferguson-Smith MA, Warren W, Graves JAM: **Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes**. *Genome Res* 2008, **18**(6):965-973.

4. Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Moneger F: **A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants**. *Plos Biology* 2005, **3**(1):47-56.

5. Minder AM, Rothenbuehler C, Widmer A: **Genetic structure of hybrid zones between *Silene latifolia* and *Silene dioica* (Caryophyllaceae): evidence for introgressive hybridization**. *Molecular Ecology* 2007, **16**(12):2504-2516.

6. Karrenberg S, Favre A: **Genetic and ecological differentiation in the hybridizing campions *Silene dioica* and *S. latifolia***. *Evolution* 2008, **62**(4):763-773.

7. Waelti MO, Muhlemann JK, Widmer A, Schiestl FP: **Floral odour and reproductive isolation in two species of *Silene***. *Journal of Evolutionary Biology* 2008, **21**(1):111-121.

8. Widmer A, Lexer C, Cozzolino S: **Evolution of reproductive isolation in plants**. *Heredity* 2009, **102**(1):31-38.

9. Rieseberg LH, Willis JH: **Plant speciation**. *Science* 2007, **317**(5840):910-914.

10. Lexer C, Widmer A: **The genic view of plant speciation: recent progress and emerging questions**. *Philosophical Transactions of the Royal Society B-Biological Sciences* 2008, **363**(1506):3023-3036.

11. Minder AM, Widmer A: **A population genomic analysis of species boundaries: neutral processes, adaptive divergence and introgression between two hybridizing plant species**. *Molecular Ecology* 2008, **17**(6):1552-1563.

12. Costich D, Meagher T, Yurkow E: **A rapid means of sex identification in *Silene latifolia* by use of flow cytometry**. *Plant Molecular Biology Reporter* 1991, **9**(4):359-370.

13. Bennett MD, Smith JB: **Nuclear-DNA amounts in angiosperms**. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 1991, **334**(1271):309-345.

14. Široký J, Lysak MA, Doležel J, Kejnovský E, Vyskot B: **Heterogeneity of rDNA distribution and genome size in *Silene* spp**. *Chromosome Research* 2001, **9**(5):387-393.

15. Figueira A, Janick J, Goldsbrough P: **Genome size and DNA polymorphism in *Theobroma cacao***. *Journal of the American Society for Horticultural Science* 1992, **117**(4):673-677.

16. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing**. *Molecular Ecology* 2008, **17**(7):1636-1647.

17. Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL: **Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis***. *BMC Genomics* 2009, **10**.

18. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GS-FLX**. *BMC Genomics* 2009, **10**.

19. Cheung F, Haas BJ, Goldberg SMD, May GD, Xiao YL, Town CD: **Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology**. *BMC Genomics* 2006, **7**.

20. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing**. *Plant Journal* 2007, **51**(5):910-918.

21. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome**. *BMC Genomics* 2008, **9**.

22. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

23. Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB: **Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing**. *Plant Physiol* 2007, **144**(1):32-42.

24. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA: **Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery**. *BMC Genomics* 2010, **11**.

25. Wheat CW: **Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing**. *Genetica* 2010, **138**(4):433-451.

26. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila, Arabidopsis***. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(8):4482-4487.

27. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E: **Housekeeping genes as internal standards: use and limits**. *Journal of Biotechnology* 1999, **75**(2-3):291-295.

28. Martins RP, Ostermeier GC, Krawetz SA: **Nuclear matrix interactions at the human protamine domain - A working model of potentiation**. *Journal of Biological Chemistry* 2004, **279**(50):51862-51868.

29. Xu WY, Bak S, Decker A, Paquette SM, Feyereisen R, Galbraith DW: **Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana***. *Gene* 2001, **272**(1-2):61-74.

30. Micheli F, Holliger C, Goldberg R, Richard L: **Characterization of the pectin methylesterase-like gene AtPME3: a new member of a gene family comprising at least 12 genes in *Arabidopsis thaliana***. *Gene* 1998, **220**(1-2):13-20.

31. Mohandas TK, Speed RM, Passage MB, Yen PH, Chandley AC, Shapiro LJ: **Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp**. *Am J Hum Genet* 1992, **51**(3):526-533.

32. Čermák T, Kubát Z, Hobza R, Koblížková A, Widmer A, Macas J, Vyskot B, Kejnovský E: **Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes**. *Chromosome Research* 2008, **16**(7):961-976.

33. Kejnovsky E, Kubat Z, Macas J, Hobza R, Mracek J, Vyskot B: **Retand: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat**. *Molecular Genetics and Genomics* 2006, **276**(3):254-263.

34. Matsunaga S, Yagisawa F, Yamamoto M, Uchida W, Nakao S, Kawano S: **LTR retrotransposons in the dioecious plant *Silene latifolia***. *Genome* 2002, **45**(4):745-751.

35. Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, Soltis DE, Soltis PS, Stevens PF, Anderberg AA, Moore MJ, Olmstead RG, Rudall PJ, Sytsma KJ, Tank DC, Wurdack K, Xiang JQY, Zmarzty S, Angiosperm Phylogeny G: **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III**. *Botanical Journal of the Linnean Society* 2009, **161**(2):105-121.

36. Ku H-M, Vision T, Liu J, Tanksley SD: **Comparing sequenced segments of the tomato and Arabidopsis genomes: Large-scale duplication followed by selective gene loss creates a network of synteny**. *Proceedings of the National Academy of Sciences* 2000, **97**(16):9121-9126.

37. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events**. *Nature* 2003, **422**(6930):433-438.

38. Yang XH, Jawdy S, Tschaplinski TJ, Tuskan GA: **Genome-wide identification of lineage-specific genes in *Arabidopsis, Oryza* and *Populus***. *Genomics* 2009, **93**(5):473-480.

39. Shendure J, Ji HL: **Next-generation DNA sequencing**. *Nature Biotechnology* 2008, **26**(10):1135-1145.

40. Yu J, Wang J, Lin W, Li SG, Li H, Zhou J, Ni PX, Dong W, Hu SN, Zeng CQ, Zhang JG, Zhang Y, Li RQ, Xu ZY, Li ST, Li XR, Zheng HK, Cong LJ, Lin L, Yin JN, Geng JN, Li GY, Shi JP, Liu J, Lv H, Li J, Deng YJ, Ran LH, Shi XL, Wang XY *et al*: **The Genomes of *Oryza sativa*: A history of duplications**. *Plos Biology* 2005, **3**(2):266-281.

41. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng HG, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao KY, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J: **The pattern of polymorphism in *Arabidopsis thaliana***. *Plos Biology* 2005, **3**(7):1289-1299.

42. Qiu S, Bergero R, Forrest A, Kaiser VB, Charlesworth D: **Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes**. *P Roy Soc B-Biol Sci* 2010, **277**(1698):3283-3290.

43. **Computational Biology and Functional Genomics Laboratory** *http://compbiodfciharvardedu/tgi/software/*.

44. Smit A, Hubley R, Green P: **RepeatMasker**. *http://repeatmaskerorg*.

45. Jurka J: **Repeats in genomic DNA: mining and meaning**. *Current Opinion in Structural Biology* 1998, **8**(3):333-337.

46. Pertea G, Huang XQ, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J: **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets**. *Bioinformatics* 2003, **19**(5):651-652.

47. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

48. Wasmuth JD, Blaxter ML: **Prot4EST: Translating Expressed Sequence Tags from neglected genomes**. *BMC Bioinformatics* 2004, **5**.

49. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining Hidden Markov models**. *Bioinformatics* 2003, **19**:II103-II112.

50. Bouzidi MF, Franchel J, Tao QZ, Stormo K, Mraz A, Nicolas P, Mouzeyar S: **A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions**. *Theor Appl Genet* 2006, **113**(1):81-89.

51. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

52. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF *et al*: **InterPro: the integrative protein signature database**. *Nucleic Acids Res* 2009, **37**(Database issue):D211-215.

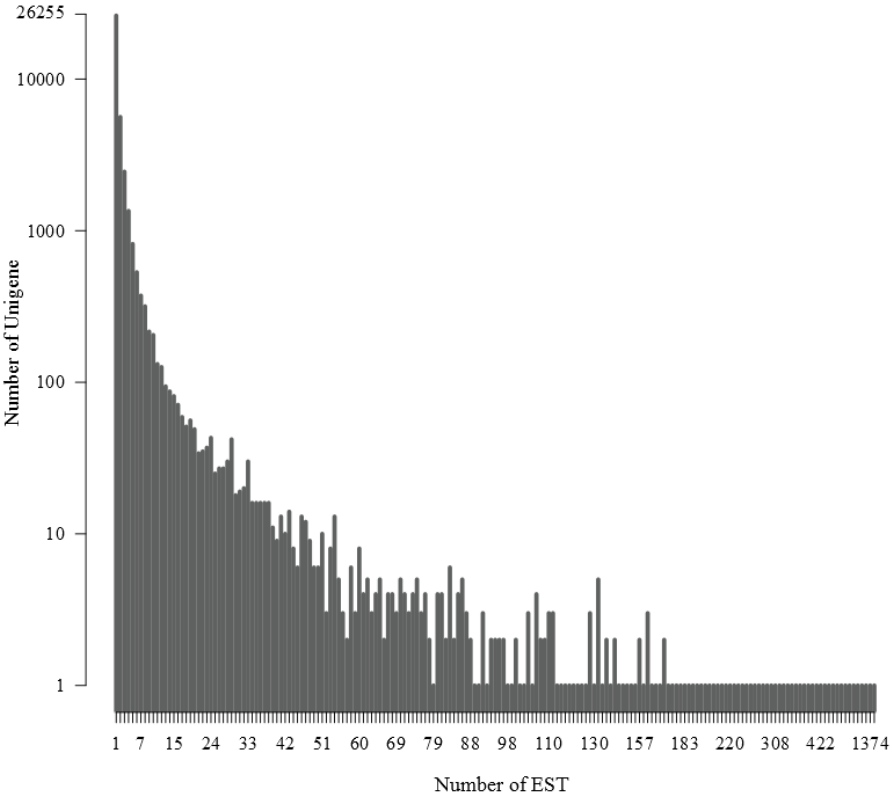
53. Tang JF, Vosman B, Voorrips RE, Van der Linden CG, Leunissen JAM: **QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species**. *BMC Bioinformatics* 2006, **7**.

54. Huang XQ, Madan A: **CAP3: a DNA sequence assembly program**. *Genome Research* 1999, **9**(9):868-877.

55. Rozen S, Skaletsky H: **Primer3 on the WWW for General Users and for Biologist Programmers**. In: *Bioinformatics Methods and Protocols*. Edited by Misener S, Krawetz SA, vol. 132: Humana Press; 1999: 365-386.

Additional material

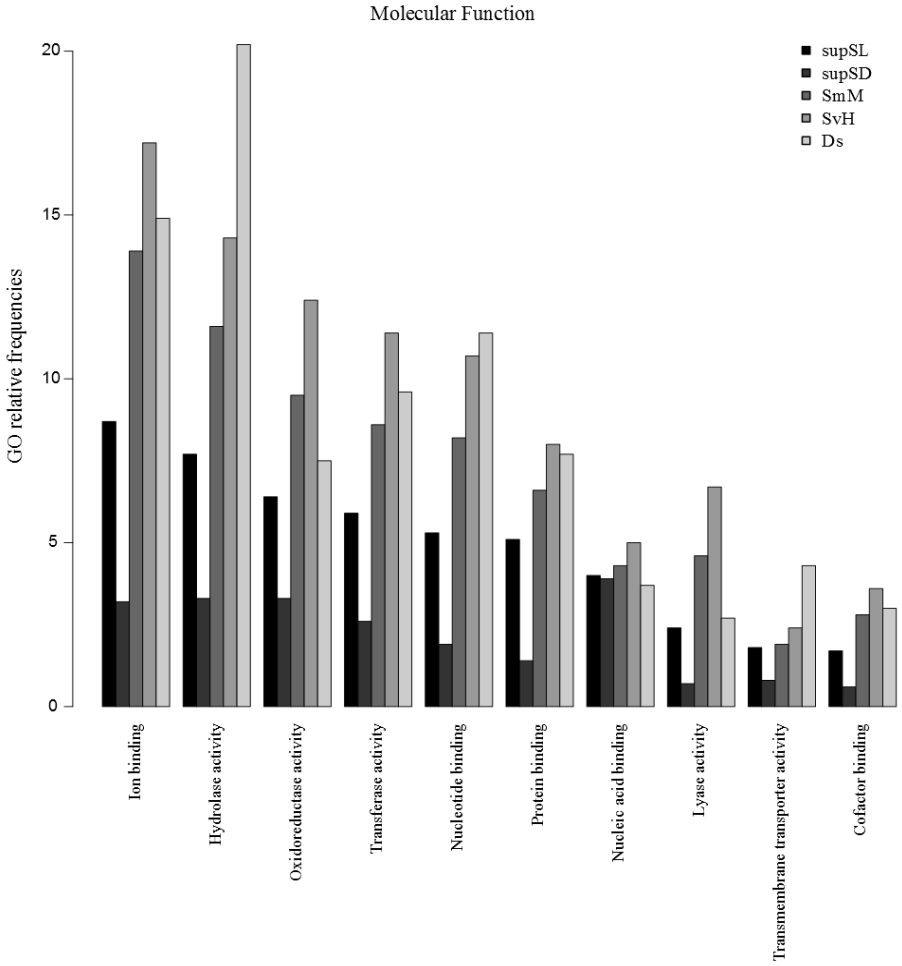
Additional File 1: Distribution of ESTs per unigene. Distribution of EST reads per unigene in the SlF library. The x-axis represents EST reads per unigene and the y-axis the number of unigenes.



Additional File 2: Highly expressed genes. Fifteen genes were identified among the ten most strongly expressed genes in each library. The two first genes coding for homologues of alpha-tubulin and an unknown gene in *Pinus koraiensis* are present in at least four of our libraries. 'Presence' indicates the number of libraries in which a given gene was found among the ten most strongly expressed genes. E-value cut-off 1E-4.

Gene annotation	Accession	Species	Presence
ORF124	A4QMB9	*Pinus koraiensis*	SlM, SlFf, SdM, SdF, SvH, Ds
Alpha tubulin	P29510	*Arabidopsis thaliana*	SlM, SdM, SdF, Ds
TNP2	C7FDI5	*Glycine max*	SlM, SdM, SdF
Cytochrome P450 like TBP	O04892	*Nicotiana tabacum*	SlM, SdF, SvH
Ribulose bisphosphate carboxylase small chain 1A, chloroplastic	P10795	*Arabidopsis thaliana*	SlF, SmM, SvH
Putative uncharacterized proteine	Q3I6J2	*Silene latifolia*	SlM, SlFf
Probable cinnamyl alcohol dehydrogenase 9	P42734	*Arabidopsis thaliana*	SlM, SlFf
Nodulin / glutamate-ammonia ligase-like protein	Q9SCP3	*Arabidopsis thaliana*	SlM, SlFf
Dicarboxylate/tricarboxylate carrier	Q9C5M0	*Arabidopsis thaliana*	SlM, SlFf
Pollen-specific leucine-rich repeat extensin-like protein 1	Q9LJ64	*Arabidopsis thaliana*	SlM, SdM
Non-specific lipid-transfer protein 3	Q9LLR7	*Arabidopsis thaliana*	SlF, SmM
Putative uncharacterized protein	Q8GW53	*Arabidopsis thaliana*	SlF, SvH
50S ribosomal protein L16, chloroplastic	P56793	*Arabidopsis thaliana*	SdM, SdF
Putative pectinesterase/pectinesterase inhibitor 28	Q3E8Z8	*Arabidopsis thaliana*	SmM, Ds
Germin-like protein subfamily 3 member 3	P94072	*Arabidopsis thaliana*	SmM, SvH

Additional File 3: Relative frequencies of the most represented Molecular Function GO sub-classes across libraries. Additional file 3 shows the 10 most frequent molecular function GO terms at level 3 in the five species *Silene latifolia*, *S. dioica*, *S. marizii*, *S. vulgaris* and *Dianthus superbus*.



Additional File 4: BLAST hits of unigenes in the eight individual libraries with different databases. BLAST hits of unigenes in the eight individual libraries with the following databases (E-value cut-off 1E-4):
AT = *Arabidopsis thaliana*, VV = *Vitis vinifera*, PT = *Populus trichocarpa*, SL = *Silene latifolia*.
Protein sequences were downloaded from:
http://plants.ensembl.org/info/data/ftp/index.html: AT proteome = TAIR10.pep 07.02.2011, VV proteome = IGGP12x.pep 07.02.2011, PT proteome = JGI2.0.pep 07.02.2011. AT EST = AGI_release_15, VV EST = VVGI_release_7, PT EST = PPLGI_release_5. SL mtDNA is *S. latifolia* mtDNA described in Sloan *et al.* (2010). Uniprot release 07.2010. *A. thaliana* mtDNA, cpDNA, exon, intron, intergenic, 3' UTR and 5' UTR come from TAIR10.

Database	Library							
	SlM	SlF	SlFf	SdM	SdF	SmM	SvH	Ds
A. thaliana proteome	10,242	21,201	9,749	8,083	12,795	17,411	17,614	17,009
V. vinifera proteome	11,357	21,721	10,326	10,006	13,708	17,866	18,271	17,097
P. trichocarpa proteome	13,802	22,578	11,868	15,111	16,233	19,105	19,203	17,549
Uniprot	20,388	23,026	23,044	22,399	28,175	19,882	20,262	17,797
TAIR10_exon	4,687	8,620	4,948	3,522	5,638	7,005	7,403	3,196
TAIR10_intron	354	400	428	293	430	373	313	292
TAIR10_intergenic	2,938	1,133	3,784	1,694	3,607	675	2,205	631
TAIR10_3_UTR	214	474	179	215	282	408	411	348
TAIR10_5_UTR	163	281	132	169	186	295	236	221
A. thaliana EST	7,521	11,028	8,540	4,859	9,373	8,447	10,533	8,302
V. vinifera EST	9,104	13,322	9,728	7,247	11,250	10,206	12,341	9,589
P. trichocarpa EST	7,955	12,653	9,134	4,917	10,083	9,598	11,861	9,102
SL_mtDNA	989	429	905	560	842	322	907	338
TAIR10_chrM	400	139	698	241	388	132	136	273
TAIR10_chrC	668	233	1,096	346	720	216	146	126

Chapter II

Published article:

Comparative analysis of a plant pseudoautosomal region (PAR) in *Silene latifolia* with the corresponding *S. vulgaris* autosome

Nicolas Blavet[1], Hana Blavet[2], Radim Cegan[2], Niklaus Zemp[1], Jana Zdanska[2], Bohuslav Janoušek[2], Roman Hobza[2,3] and Alex Widmer[1]

BMC Genomics (2012), 13: 226

[1]Institute of Integrative Biology (IBZ), ETH Zürich, Universitätstrasse 16, Zürich, 8092, Switzerland

[2]Institute of Biophysics, Laboratory of Plant Developmental Genetics, Academy of Sciences of the Czech Republic, v.v.i. Kralovopolska 135, Brno, CZ-61200, Czech Republic

[3]Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Sokolovská 6, Olomouc, CZ-77200, Czech Republic

Abstract

Background:

The sex chromosomes of *Silene latifolia* are heteromorphic as in mammals, with females being homogametic (XX) and males heterogametic (XY). While recombination occurs along the entire X chromosome in females, recombination between the X and Y chromosomes in males is restricted to the pseudoautosomal region (PAR). In the few mammals so far studied, PARs are often characterized by elevated recombination and mutation rates and high GC content compared with the rest of the genome. However, PARs have not been studied in plants until now. In this paper we report the construction of a BAC library for *S. latifolia* and the first analysis of a > 100 kb fragment of a *S. latifolia* PAR that we compare to the homologous autosomal region in the closely related gynodioecious species *S. vulgaris.*

Results:

Six new sex-linked genes were identified in the *S. latifolia* PAR, together with numerous transposable elements. The same genes were found on the *S. vulgaris* autosomal segment, with no enlargement of the predicted coding sequences in *S. latifolia*. Intergenic regions were on average 1.6 times longer in *S. latifolia* than in *S. vulgaris*, mainly as a consequence of the insertion of transposable elements. The GC content did not differ significantly between the PAR region in *S. latifolia* and the corresponding autosomal region in *S. vulgaris*.

Conclusions:

Our results demonstrate the usefulness of the BAC library developed here for the analysis of plant sex chromosomes and indicate that the PAR in the evolutionarily young *S. latifolia* sex chromosomes has diverged from the corresponding autosomal region in the gynodioecious *S. vulgaris* mainly with respect to the insertion of transposable elements. Gene order between the PAR and autosomal region investigated is conserved, and the PAR does not have the high GC content observed in evolutionarily much older mammalian sex chromosomes.

Keywords: BAC library, Pseudoautosomal region, PAR, *Silene latifolia*, Sex chromosome, Evolution.

Background

Heteromorphic sex chromosomes (XY/ZW) can often be distinguished from autosomes by the absence of recombination in at least a part of their length and Y/W chromosome degeneration [1, 2]. Plants with sex chromosomes have evolved rarely but repeatedly in many plant lineages, and sex chromosomes have reached various levels of differentiation [3]. In *Asparagus officinalis* and *Carica papaya* for example, X and Y chromosomes have diverged little and recombine along most of their length [4-6] whereas in *Rumex acetosa* and *Silene latifolia*, the sex chromosomes in males are largely non-recombining [7, 8]. In *S. latifolia*, X and Y chromosomes can recombine only in the regions known as pseudoautosomal regions (PARs). Westergaard [9] originally identified one PAR on each of the q-arms of *Silene latifolia* sex chromosomes. Later, Lengerova *et al*. (2003) [8] using fluorescent in situ hybridization (FISH) revealed that the X PAR is located on the p-arm, whereas the PAR on the Y chromosome is located on the q-arm. More recently, Scotti and Delph [10] proposed that PARs exist on both ends of the X and Y chromosomes, similar to the situation in humans [11]. A further similarity to mammalian sex chromosomes is that the *S. latifolia* sex chromosomes diverged gradually [12] , which led to the formation of evolutionary strata. Comparisons between the evolutionarily young *S. latifolia* sex chromosomes (about 10 million years [12, 13]) and those of eutherian mammals (about 110 million years [14]) have revealed that similar processes are involved in the evolution of sex chromosomes in both animals and plants.

The sex chromosomes of *S. latifolia* most likely evolved from a single pair of autosomes as previously shown [12, 13], with one autosome of the gynodioecious relative *S. vulgaris*, a species lacking sex chromosomes, carrying homologues of *S. latifolia* sex-linked genes [12, 13]. *Silene latifolia* and *S. vulgaris* have the same haploid chromosome number (n = 12), but differ substantially in genome size. The *Silene latifolia* haploid genome is 2646 Mbp in females [15], with the X chromosome being about 400 Mbp in length [16], whereas the haploid genome size of *S. vulgaris* is 1103Mbp [17] and autosomes are about 100 Mbp long.

In this study we analyzed a part of the *S. latifolia* PAR located on the p-arm of the X chromosome and on the q-arm of the Y chromosome (henceforth referred to as PAR) and of the corresponding *S. vulgaris* autosome, in order to study collinearity and divergence between these chromosome parts and to assess whether the *S. latifolia* PAR has characteristics in common with animal PARs. Furthermore, we investigate whether the *S. latifolia* size increase relative to *S. vulgaris* reflects the increase in size of the entire X chromosome or more closely resembles the increase seen in *S. latifolia* autosomes.

In mammalian genomes, PARs have several interesting properties including increased GC content, higher mutation rates and a level of recombination higher than in the rest of the genome [18, 19] due to the necessity for crossing over in this region [20]. PARs in mice and the human PAR1 appear to serve a critical function in spermatogenesis, as indicated by the fact that their absence prevents X and Y chromosome segregation during male meiosis, causing male sterility [21-23]. However, PARs differ widely in size among mammals (covering about 4 % of the Y chromosome in humans [24, 25] and mice [26], about 8 % in cattle [27] and about 24 % in dogs [28, 29]), with most eutherians sharing the same genes situated closest to the telomere but having the pseudoautosomal boundary (PAB), separating the PAR from the sex-specific part of the sex chromosomes, at variable positions [27]. In mice, the PAB is located in the gene *Fxy*. Exons 1-3 are located in the X specific part, while exons 4-10 are located in the PAR [30]. The segment of this gene located in the PAR has a higher GC content than its X-specific portion [30, 31].

In order to analyze the *S. latifolia* PAR and the corresponding region on the *S. vulgaris* autosome, we first established and screened a bacterial artificial chromosome (BAC) library of *S. latifolia* with the marker ScOPA09 that has previously been found to be located in the PAR of the closely related dioecious species *S. dioica* [32] and has successfully been identified and used for mapping *S. latifolia* sex chromosomes [12]. The marker ScOPA09 is located in the *S. latifolia* PAR which is known to recombine once per generation in males [33] and makes up about 10% of the Y chromosome [34, 35]. In *S. vulgaris*, the marker OPA is lacking. We therefore first sequenced a clone of the *S. latifolia* BAC library containing the marker ScOPA09. Sequencing was performed by Sanger and 454 pyrosequencing to explore the suitability of different sequencing strategies for BAC assembly. From these sequences we identified new markers and used them to screen the *S. vulgaris* BAC library for a homologous clone. Both BAC sequences were assembled into >100,000 bp-long scaffolds using GS De Novo Assembler (Roche).

Here we present the results of a genomic comparison between an area located in the PAR of the X chromosome p-arm and in the q-arm of the Y chromosome of the dioecious plant species *S. latifolia* and its homologous autosomal area in the closely related gynodioecious species *S. vulgaris*. Our results identify the first physically mapped genes located in the *Silene* PAR and reveal characteristics of a plant pseudoautosomal region.

Results and Discussion

Our study reports the first comparative analysis of a BAC sequence from a plant pseudoautosomal region and the corresponding autosomal area in a related

species that lacks sex chromosomes. Comparative mapping of a limited number of sex-linked genes in *S. latifolia* and autosomal genes in *S. vulgaris* has previously demonstrated large-scale synteny between the X chromosome of *S. latifolia* and one *S. vulgaris* autosome [12, 13]. Our results provide the first evidence for small-scale synteny and strong collinearity at the gene level within a restricted region of the *S. latifolia* sex-chromosomes, the PAR located in the p-arm of the X chromosome and in the q-arm of the Y chromosome, and the corresponding *S. vulgaris* autosomes.

BAC sequencing, assembly and annotation

The 454 paired-end sequencing of both a *S. latifolia* BAC clones containing marker ScOPA09 and of a homologous BAC clone from *S. vulgaris* gave more than 150,000 reads for each BAC clone. These were assembled into 171,870 bp and 116,096 bp-long scaffolds for *S. latifolia* and *S. vulgaris*, respectively.

We found a total of twenty-eight homologous sequences (seventeen different accession numbers) with the *A. thaliana* proteome. Of these, nine were found in both *Silene* species, two were identified only in *S. vulgaris* (one of them twice), and six were found only in *S. latifolia*. A total of 16 out of 28 sequences are most likely transposable or repeated elements as indicated by their annotations extracted from the protein domain family database ProDom [36] and repeat coverage (Table 1). The repeat coverage is based on BLAST hits with a *Silene* repeated elements library [37].

Among the nine sequences shared by the two *Silene* species, the areas matching TAIR accessions AT4G23160 (*CRK8*), AT2G01050, AT1G43760 and ATMG00860 contain repeated elements. Moreover, in *S. vulgaris*, the part matching ATMG00860 is contained in a match with the transposon sequence Q3I6J4_SILLA, which also includes the sequence matching accession number AT3G01410 (Table 1).

Using annotated *Silene* transposable elements [34] we found that the region matching *CRK8* is similar to a *Copia*-like retrotransposon [38], and that the region matching Q3I6J4_SILLA is similar to a *Retand*-like retrotransposon (see Kejnovský *et al.* (2006) for description [39]). Moreover, the sequences matching AT2G01050, AT1G43760 and *CRK8* are found in both scaffolds at different positions, which provide further evidence that these sequences are transposable elements.

Table 1: Putative transposable elements identified in *Silene* BAC clones. Putative transposable elements were identified by BLASTX searches against the *Arabidopsis thaliana* proteome (TAIR10) with an E-value cut-off of 1E-4, ProDom annotations and analysis of the coverage in the *Silene* repeated sequences database [37]. [a]The sequence has been found twice in the BAC clone.

Found in	TAIR accession	Function	ProDom annotation	*Silene* repeats coverage
S. latifolia/ *S. vulgaris*[a]	ATMG00860	Mitochondrion/ hypothetical protein	Transposable element	0%
S. latifolia/ *S. vulgaris*	AT4G23160	Cysteine-rich receptor-like protein kinase 8/ polyprotein	Transposable element	80%
S. latifolia	ATMG00710	Mitochondrion/ hypothetical protein	Transposable element	100%
S. vulgaris[a]	AT3G01410	Putative RNase H	Transposable element	77%
S. latifolia	ATMG00310	Mitochondrion/ hypothetical protein	Transposable element	49%
S. latifolia	AT5G41980	Uncharacterized protein	Transposable element	0%
S. latifolia/ *S. vulgaris*	AT2G01050	Nucleic acid binding / zinc ion binding/ uncharacterized protein	Transposable element	25%
S. latifolia/ *S. vulgaris*	AT1G43760	Uncharacterized protein	Transposable element	20%
S. latifolia	AT4G20520	RNA binding / RNA-directed DNA polymerase/ uncharacterized protein	Transposable element	0%
S. latifolia	ATMG01250	Mitochondrion/ hypothetical protein	Transposable element	23%

The five remaining sequences correspond to new pseudoautosomal genes. They are homologues of the *A. thaliana* genes *ESP1* (AT4G22970), *BIP1* (AT5G28540),

ACBP1 (AT5G53470), and of genes AT5G53500 and AT5G41970. These latter two genes we named PAR1 and PAR2, respectively (Table 2).

Table 2: Putative genes identified in *Silene* BAC clones. Putative genes were identified using BLASTX searches against the *Arabidopsis thaliana* proteome (TAIR10) with an E-value cut-off of 1E-4, ProDom annotations were used to determine function or detect repeats and transposable elements. % identity with *A. thaliana* sequence is given first for *S. latifolia* and then for *S. vulgaris*. * PAR2 is truncated approximately by half in *S. latifolia*.

Gene name	Found in	TAIR accession	Function	ProDom annotation	%identity
ESP1	*S. latifolia*/ *S. vulgaris*	AT4G22970	Separase	Separase	34/36
BIP1	*S. latifolia*/ *S. vulgaris*	AT5G28540	Luminal-binding protein 1	ATP-binding	80/74
ACBP1	*S. latifolia*/ *S. vulgaris*	AT5G53470	Acyl-CoA-binding domain-containing protein 1	Lipid-binding	44/42
PAR1	*S. latifolia*/ *S. vulgaris*	AT5G53500	WD-40 repeat family protein	Hydrolase phosphatase	41/42
PAR2*	*S. latifolia*/ *S. vulgaris*	AT5G41970	Uncharacterized protein	Metal dependent	56/70
PAR3	*S. latifolia*	AT3G15000	Uncharacterized protein	Plastid developmental protein DAG	54/-
SVA1	*S. vulgaris*	AT4G27700	Rhodanese-like domain-containing protein	Rhodanese	-/71

Finally, two other gene sequences, AT3G15000 and AT4G27700, that we named PAR3 and SVA1, were found on the BACs of *S. latifolia* and *S. vulgaris*, respectively (Table 2). However, because the BAC sequences only partly overlap, we do not have the homologous copies of these genes in the other species. The PAR3 gene in *S. latifolia* corresponds to a putative pseudoautosomal gene. The *S. vulgaris* SVA1 sequence is homologous to a gene coding for a rhodanese protein in *A. thaliana* (information collected from TAIR [http://arabidopsis.org/]).

Genes located in PARs close to the PAB (less than 50 cM) often present sex-specific expression [40]. Using RNA-seq data from Muyle *et al.* [41] we found no evidence for sex-biased expression of the genes located in the *S. latifolia* PAR (Additional table S1).

GC content

In mammalian PARs, high recombination associated with biased gene conversion (BGC) [42-44] results in a high GC content [27, 30, 45, 46]. Comparisons of GC and GC3 content between the PAR and non-PAR regions of the human X chromosome revealed a higher GC and GC3 content in the pseudoautosomal region [42]. Further studies of the human PAR revealed that the GC content decreases from 64 % close to the telomeric region to 55 % in the middle of the PAR and is only 38 % close to the pseudoautosomal boundary (PAB) [19]. Similar declines of the GC content were also found in other mammals, including cattle [27] and murine species [47].

A recent analysis of sequence polymorphisms in plants has revealed that the mating system affects GC content, with a higher content in outcrossing compared to selfing taxa being observed, but this effect is significant only in Poaceae that are known to have unusual GC contents [48]. Whether this effect is due to BGC and why it is observed only in Poaceae, however, is not clear. To date, evidence for BGC in plant sex chromosomes is lacking, but given the relatively small size of the investigated PAR (about 10 % of the Y chromosome [34, 35]) in *S. latifolia*, which is comparable to PAR size in mammals, and the fact that recombination occurs during meiosis, a higher recombination rate is expected in this region as compared with other regions of the genome.

In contrast to mammals, the *Silene latifolia* PAR can directly be compared with a homologous autosomal region in a closely related species. If the *S. latifolia* PAR has an increased GC content compared to autosomes, then this should be detectable in a comparative analysis. We then determined the GC and GC3 contents for each gene (Table 3). A comparison between *S. latifolia* and *S. vulgaris* revealed no significant difference (GC content: $t = 0.0638$, p-value = 0.9507; GC3 content: $t = 0.0521$, p-value = 0.9597). Moreover we determined the GC and the GC3 content of nine sex-linked genes that had previously been identified and are located in the sex-specific region of the X chromosome [49-57] (Additional table S2).

Table 3: GC and GC3 content comparison. GC and GC3 contents were computed on gene sequences of identical size in both *S. latifolia* and *S. vulgaris*. PAR2 is truncated in the BAC sequence of *S. latifolia* and *ESP1* is truncated in *S. vulgaris*. In these cases, only partial sequences with coverage in both species were compared. Sl: *Silene latifolia*, Sv: *S. vulgaris*. *Exon and intron sequences were taken into account.

Gene name	***ESP1***	***BIP1***	***ACBP1***	**PAR1**	**PAR2**
Sl sequence length (bp)	5416	2013	1125	2049	507
Sl exon GC content (%)	42.4	47.3	48.3	42.3	48.3
Sl exon GC3 content (%)	41.2	53.5	42.9	38.1	49.7
Sv sequence length (bp)	5416	2013	1125	2049	507
Sv exon GC content (%)	42.4	47.3	47.7	42.4	48.3
Sv exon GC3 content (%)	41.1	53.7	42.9	37.3	49.1
Sl putative gene GC content*	38.2	42.3	38.8	40.4	39.9
Sv putative gene GC content*	37.5	42.8	39.2	40.2	40.6

No difference was found in the GC and GC3 content between the newly identified PAR genes and the genes located in the sex-specific region of the X chromosome ($t = 0.601$, p-value = 0.5656 and $t = 0.6295$, p-value = 0.5437, respectively). These results indicate that the pattern typical for mammalian PARs is not present in the investigated part of the *S. latifolia* PAR. This may indicate that the *S. latifolia* PAR maintains its "autosomal" features, as the sex chromosomes in this species are evolutionarily young. Alternatively, the studied region of the *S. latifolia* PAR might be close to the pseudoautosomal boundary where the GC content is lower than in more distal PAR areas, as has been found to be the case in most mammals [19, 27, 46]. Indeed, the ScOPA09 marker was estimated to be located 15 cM from the pseudoautosomal boundary (PAB) in *S. dioica* [32]. We obtained a very similar estimate of 11 cM for *S. latifolia* in this study (both calculations were done using Kosambi and Haldane mapping functions, for details see Additional table S3). Even though we presently do not know the physical distance between the PAB and the BAC clone studied here, the results of our genetic analysis clearly show that all genes identified in this study are located in the PAR, as evidenced by their cosegregation with marker ScOPA09 (Additional table S4). Furthermore, these genes are recombining with the same recombination frequency of 11 % with the PAB (Additional table S3).

Structure comparison

Large-scale collinearity between the *S. latifolia* X chromosome and *S. vulgaris* autosome has repeatedly been reported in studies of *S. latifolia* sex chromosome evolution [13, 52, 58, 59]. In addition to large-scale collinearity we here report the presence of small-scale collinearity spanning five genes whose linear arrangement is conserved. We also show that the length of these genes is identical between the studied *Silene* species. Indeed, while we assessed whether the investigated *S. latifolia* PAR presents signs of chromosome enlargement, because the X chromosome is about four times the size of a *S. vulgaris* autosome, we analyzed exon and intron lengths of the five genes previously reported (Table 4). We then compared with a Student's t-test whether the average size of both introns and exons of the different genes between both species were similar and we found no significant difference between *S. latifolia* and *S. vulgaris* ($t = -0.0817$, p-value $= 0.9369$ and $t = -0.005$, p-value $= 0.9961$ for intron and exon comparisons respectively).

Table 4: Comparison of gene sizes. Genes PAR2 and ESP1 are truncated in *S. latifolia* and *S. vulgaris* respectively, due to their positions at the end of the investigated BAC clones. For these genes, only the fragment covered in both species was reported below. Sl: *Silene latifolia*, Sv: *S. vulgaris*.

Gene name	***ESP1***	***BIP1***	***ACBP1***	**PAR1**	**PAR2**
Sl sequence length (bp)	15455	3401	4909	3107	1127
Sl exons	24	6	6	7	3
Sl exons length (bp)	5416	2019	1143	2049	510
Sl introns length (bp)	10039	1382	3766	1058	617
Sv sequence length (bp)	16677	3217	5042	3043	1138
Sv exons	24	6	6	7	3
Sv exons length (bp)	5422	2046	1143	2049	507
Sv introns length (bp)	11255	1171	3899	994	631
% identity gene	80.7	85.6	73.6	92.6	92.8
% identity exons	96.5	96.9	92.9	97.1	95.9
Sv/Sl introns	1.1	0.9	1.0	0.9	1.0

The conserved intron size may indicate a functional role in gene regulation. Indeed, introns enlarged by repetitive elements were found to affect gene expression in rice [60]. However, substantial differences in length occur in intergenic regions due to transposable element insertions. Additional figure 1 presents the global alignment of both BAC sequences. Intergenic regions are highly diverged in size and consequently major gaps are visible in the alignment. We considered as intergenic all

regions in-between the five genes described in this paper for which there are copies in both *Silene* species. The total length of the intergenic region in the *S. latifolia* BAC is 115,909 bp, and 71,866 bp in the *S. vulgaris* BAC. This difference of 44,043 bp is highly significant ($\chi^2 = 30823.1$, df = 3, p-value < 2.2e-16) and corresponds to a 61 % increase of the *S. latifolia* chromosome size as compared to *S. vulgaris*. This increase is due to the insertion of transposable elements in the *S. latifolia* PAR region.

Microsatellite comparison

We found 577 and 377 microsatellite loci in *S. latifolia* and *S. vulgaris* respectively. A comparison of the average proportion of microsatellite loci (mono-, di-, tri- and tetranucleotide microsatellite) between both *Silene* species using a Student's t-test revealed no significant difference (t = 0.5461, p-value = 0.5867). A previous analysis of microsatellites in plants revealed a negative correlation between microsatellite frequency and genome size [61]. The *Silene latifolia* X chromosome is about four times larger than the *S. vulgaris* autosome. However, our results revealed that the *S. latifolia* PAR contains a similar density of microsatellite repeats as the *S. vulgaris* autosomes, suggesting that microsatellites play no or only a minor role in the size increase of the *Silene latifolia* PAR. Then, we searched for long-mer microsatellite accumulation in *S. latifolia* and *S. vulgaris* (see Methods), which are expected to be rare in the PAR [62]. We found one $(ATC)_{10}$ microsatellite locus in *S. latifolia* and one occurrence each of $(T)_{35}$, $(ATA)_{16}$, $(AAG)_{19}$ and $(TTA)_{25}$ in *S. vulgaris*. The low density of long-mer microsatellites observed here confirms the previously reported paucity of microsatellite repeats on the *S. latifolia* X chromosome PAR [62].

Transposable element insertion

We found three transposable elements containing long terminal repeats (LTR) in *S. latifolia* and two in *S. vulgaris*. The estimates of the invasion of these elements vary from about 17,200 years ago to 766,000 years ago (Additional table S5). These elements were inserted after the divergence of the *Silene* sex chromosomes approximately $5 \sim 10$ million years ago [12], which may be an indication of highly active transposable elements in the PAR. However, given that we have observed a smaller than expected size increase in the PAR (about 1.6x instead of about 4x), we hypothesize that the enlargement of the X chromosome occurs mainly in the non-PAR areas of the X chromosome and is due to large-scale accumulation of different tandem repeats [63, 64] and retrotransposons [65, 66]. The observed larger size of the studied *S. latifolia* PAR segment in comparison to the *S. vulgaris* autosome is close to the

approximate difference in size between autosomes in *S. latifolia* and *S. vulgaris* and may therefore not reflect the size increase seen in the sex chromosomes.

Conclusions

In this study we present the first analysis of a fragment belonging to the *S. latifolia* pseudoautosomal region located in the p-arm of the X chromosome and the q-arm of the Y chromosome. The analysis of BAC sequences revealed five new pseudoautosomal genes that are conserved in size and linear arrangement between *S. latifolia* and *S. vulgaris*, indicating small-scale gene collinearity between the X chromosome and the corresponding autosomal region. No increase in GC or GC3 content was found in the studied PAR area, indicating that either the evolutionarily young *S. latifolia* PAR is not GC rich or alternatively, that the studied region is close to the pseudoautosomal boundary, where no increase in GC content is expected. A structural comparison revealed that non-coding regions of the *S. latifolia* PAR contain multiple transposable and repeated elements and are overall about 61 % longer than in *S. vulgaris*. This size increase is similar to the size difference between *S. latifolia* and *S. vulgaris* autosomes and may therefore reflect a genome-wide, rather than a sex chromosome-specific trend. Our study reports the first comparative analysis of a partial pseudoautosomal region in a plant which we compare to a closely related species lacking sex chromosomes, thereby providing new insights into genome size and sex chromosome evolution in *Silene latifolia*.

Methods

BAC preparation

S. latifolia and *S. vulgaris* were grown from seeds in a climate chamber. Fresh leaf material was harvested after initiation of flowering. *Silene latifolia* leaves were snap frozen in liquid nitrogen, packaged in dry ice and shipped to Amplicon Express, Pullman, Washington, where the BAC library was constructed from high molecular weight (HMW) genomic DNA following the method described by Tao *et al*. (2002) [67]. The *S. vulgaris* BAC library was assembled in the Institute of Experimental Botany of the AS CR Laboratory of Molecular Cytogenetics and Cytometry, Olomouc. In summary, DNA was partly digested with *Hind*III and inserted into the pECBAC1 vector. Ligations were transformed into DH10B *E. coli* cells (Invitrogen) and plated on LB agar containing appropriate concentrations of chloramphenicol, X-gal and IPTG. Clones were robotically selected with a Genomic Solution G3 and transferred into 384 well plates, grown for 18 h, replicated and frozen at −80°C. In order to identify positions and plate numbers of each clone, they were placed on a grid in duplicate on Hybond N + (Amersham, Biosciences) nitrocellulose

membranes in a 4x4 pattern. The membranes were incubated and processed as described by Bouzidi *et al*. (2006) [68]. The *S. latifolia* BAC library was arrayed on six membranes of 18,432 colonies and one membrane containing 9,216 clones. The average insert-size of the library is 128 kb. The *S. vulgaris* BAC library was arrayed on three membranes with 18,432 colonies each. The average insert-size of the library was 110 kb [50]. Probes for radioactive hybridizations were labeled with α^{32}P using the Prime-It II Random Primer Labelling Kit (Stratagene) according to the manufacturer's protocol. The presence of the marker in positive BACs was verified by PCR. BAC DNA was isolated with the Large Construct Kit (Qiagen). The *S. latifolia* BAC clone containing the marker ScOPA09 was then sequenced. In order to test which method was most suitable for subsequent BAC assembly, we used three sequencing methods: shotgun Sanger sequencing, 3 kb and 8 kb 454 paired-end pyrosequencing on a GS-FLX machine (Roche). Sanger sequencing was performed by the GATC Biotech laboratory in Konstanz, Germany, and 454 sequencing by the Functional Genomics Center Zurich (FGCZ). As the ScOPA09 marker is not present in *S. vulgaris*, we developed suitable markers from neighboring loci using the *S. latifolia* BAC sequence in order to identify a homologous BAC clone in *S. vulgaris*. Primers used for screening the BAC library are presented in Table 5.

Sequencing, assemblies and annotations

Assembly of the initial shotgun sequences of the *S. latifolia* BAC lead to the identification of numerous relatively short contigs because of the presence of multiple repeat regions. We therefore tested two different 454 pair-end sequencing methods (3 kb and 8 kb) in order to overcome these problems. Both approaches provided similar results, but because the 8 kb paired-end sequencing method requires more DNA, we used 3 kb paired-end sequencing for the corresponding *S. vulgaris* BAC clone (Additional table S6).

Table 5: Primers used. The annealing temperature for all primer pairs is 60°C.

Gene name	Forward-primer (5' – 3')	Reverse-primer (5' – 3')	Product size (bp)
ESP1	AAATACCCAGCCCGTAGCTT	TGCTCAATACATGCCTCCAG	414
BIP1	CGAAAGATGAAGCTCCCAAG	CCCTTCTTGTCCAAACCGTA	990
ACBP1	TTAGCCCTGGCAGTCATCTT	AGGAAGTGTGTTCGGTGGAG	252
PAR1	TTTTCCTCAGGCCATAATGC	GGCTACCGAGAACACCATGT	245
PAR2	CTCCAAATTCTCGGGTTCAA	GCTCAAACACTCCACCAACA	176

All 454 sequences were *de novo* assembled using Roche GS De Novo Assembler with 98 % minimum overlap identity and 60 bp as minimum overlap length. We set the expected depth parameter with reference to the estimated size of the BAC (determined by pulse field gel electrophoresis (PFGE)) and the number of reads sequenced. In addition to 454 sequencing and assembly, we used targeted Sanger sequencing to close gaps remaining in the *S. latifolia* scaffold after the assembly process. On *S. vulgaris* we successfully used the Epicentre transposon insertion Kit (EZ1982K) to fill a 987 bp-long gap remaining after assembly. Nevertheless, a few short stretches remain unsequenced after the assembly. Both scaffolds were submitted to GenBank under accession numbers JN574439-JN574440.

The scaffolds were then annotated by similarity using BLASTX [69] with UniProtKB (Swiss-Prot + TrEMBL, 13 July 2010), the *Arabidopsis thaliana* proteome (TAIR10_20100802) and transposable elements (TAIR9_TE) [www.arabidopsis.org] with an E-value cut-off of 1E-4. We used the exon prediction tool Genscan to identify coding sequences with *A. thaliana* as the training model [70] and used ProDom [36] and a *Silene* repeated element database [37] to detect *Silene*-specific repeats and transposable elements. Transposable element annotations were then completed using annotated *S. latifolia* repeats [34].

Genetic analysis

In order to verify the pseudoautosomal location of the BAC clone studied here, we performed segregation analyses using nucleotide polymorphisms in the genes *ESP1*, PAR2, *ACBP1* and ScOPA09. First, PCR products of these genes were amplified and sequenced to look for partially sex-linked restriction polymorphisms segregating in the pseudobackcross population RB1. This population was prepared by crossing a female plant from a Swiss population, with a male plant obtained from a cross between a female of an inbred line, U9 (from Utrecht, kindly provided by Prof. Sarah Grant), and male plant from a Swiss population. Putative restriction polymorphisms (CAPS) were verified by restriction analysis, and genotyping was performed in 76 DNA samples available. For the list of primers and restriction enzymes used, see Additional table S7. The observed recombination frequencies were used for the calculation of map distances using the onemap package of R [71] with both the Kosambi and the Haldane mapping functions [72, 73].

Structure comparison

Using BLAST [69] annotation results we searched for genes and transposable elements, also using Perfect Microsatellite Repeat Finder (http://sgdp.iop.kcl.ac.uk/nikammar/repeatfinder.html) with default parameters

(minimum number of repeats = 3, minimum repeat unit length = 2 and maximum repeat unit length = 100). Mononucleotide repeats were identified manually. Only mononucleotide microsatellites with a minimum repeat number greater than or equal to 12 were counted. We assessed the frequency of all mono-, di-, tri- and tetranucleotide microsatellites and tested whether a smaller density of these microsatellites in *S. latifolia* than in *S. vulgaris* was associated with the observed genome size difference between species [61]. Furthermore, we estimated the frequency of long-mer microsatellite stretches (mononucleotides ≥ 30 repeats, dinucleotides ≥ 15 repeats and trinucleotides ≥ 10 repeats) in both species. We used a combination of *A. thaliana* annotation and Genscan [70] exon prediction to compare the size of intergenic regions and introns, and we measured the percentage identity of exons and introns and the GC content at third codon positions (GC3). In order to take into account the possibility that a gene was truncated due to its position at the beginning or end of a BAC, the gene fragment for which we have coverage in both species was used in calculations, whereas gaps were excluded. Both the GC and GC3 contents of the PAR genes were then compared with the GC and GC3 contents of nine X-linked genes (*DD44X, SlAP3X, SlX1, SlCypX, SlX3, SlX4, SlX7, SlX9* and *SlssX*) located in the non-recombining part of the *S. latifolia* sex chromosomes [49-57]. In order to test whether average GC and GC3 contents differ between species we used Student's t-tests in R [71].

Transposable element analysis

We identified LTRs of transposable elements using LTR_Finder [74] set to the default parameters. We then aligned paired LTR sequences and determined the number of mutations (substitutions and insertions/deletions) between them. We estimated the age of the LTR invasion using the following equation: N/(2*L*K), where N is the number of base substitutions between the two LTRs, L is the length of the LTR and K the base substitution per site per year [75]. We used a value of K = 23E-9 as the average substitution rate per site per year [76].

Author contributions

NB performed data analysis and drafted the manuscript. HB screened the *S. latifolia* BAC library and assisted with writing the manuscript. RC screened the *S. vulgaris* BAC library and helped with transposable element analysis. JZ and BJ performed genetic analyses. BJ also assisted with writing the manuscript. RH participated in the coordination of the study and helped with data analysis, interpretation and drafting of the manuscript. AW conceived the study, coordinated

and supervised all stages and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank L. Poveda, M. Kuenzli and W. Qi from the Functional Genomic Center Zurich (FGCZ) for assistance relating to sequencing, T. Torossi, C. Michel and the Genetic Diversity Center (GDC) at ETH Zürich for technical support, S. Zoller for bioinformatics support and M. Scarborough for help with English writing. We further acknowledge support by J. Macas and E. Kejnovský who provided sequences of repeated elements and J. Bartoš who participated in BAC library analysis. This study was supported by an ETH Zurich grant (TH-07 06–3) to AW, by Czech Science Foundation grants (522/09/0083 to RH and P501/12/G090 to BJ) and Centre of the Region Haná for Biotechnological and Agricultural Research grant (ED0007/01/01) to RH.

References

1. Wilson MA, Makova KD: **Genomic Analyses of Sex Chromosome Evolution**. *Annu Rev Genomics Hum Genet* 2009, **10**:333-354.

2. Bergero R, Charlesworth D: **The evolution of restricted recombination in sex chromosomes**. *Trends in Ecology & Evolution* 2009, **24**(2):94-102.

3. Jamilena M, Mariotti B, Manzano S: **Plant sex chromosomes: molecular structure and function**. *Cytogenetic and Genome Research* 2008, **120**(3-4):255-264.

4. Reamon-Buttner SM, Schondelmaier J, Jung C: **AFLP markers tightly linked to the sex locus in *Asparagus officinalis* L.** *Mol Breeding* 1998, **4**(2):91-98.

5. Reamon-Buttner SM, Jung C: **AFLP-derived STS markers for the identification of sex in *Asparagus officinalis* L.** *Theor Appl Genet* 2000, **100**(3-4):432-438.

6. Liu ZY, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu QY, Pearl HM, Kim MS, Charlton JW, Stiles JI, Zee FT, Paterson AH, Ming R: **A primitive Y chromosome in papaya marks incipient sex chromosome evolution**. *Nature* 2004, **427**(6972):348-352.

7. Lengerova M, Vyskot B: **Sex chromatin and nucleolar analyses in *Rumex acetosa* L.** *Protoplasma* 2001, **217**(4):147-153.

8. Lengerova M, Moore RC, Grant SR, Vyskot B: **The sex chromosomes of *Silene latifolia* revisited and revised**. *Genetics* 2003, **165**(2):935-938.

9. Westergaard M: **Aberrant Y chromosomes and sex expression in *Melandrium album*.** *Hereditas* 1946, **32**(3-4):419-443.

10. Scotti I, Delph LF: **Selective Trade-offs and Sex-Chromosome Evolution in *Silene latifolia*.** *Evolution* 2006, **60**(9):1793-1800.

11. Freije D, Helms C, Watson M, Donis-Keller H: **Identification of a second pseudoautosomal region near the Xq and Yq telomeres**. *Science* 1992, **258**(5089):1784-1787.

12. Nicolas M, Marais G, Hykelova V, Janoušek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Moneger F: **A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants**. *Plos Biol* 2004, **3**(1):47-56.

13. Filatov DA: **Evolutionary history of *Silene latifola* sex chromosomes revealed by genetic mapping of four genes**. *Genetics* 2005, **170**(2):975-979.

14. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Gruzner F, Deakin JE, Whittington CM, Schatzkamer K, Kremitzki CL, Graves T, Ferguson-Smith MA, Warren W, Graves JAM: **Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes**. *Genome Res* 2008, **18**(6):965-973.

15. Costich D, Meagher T, Yurkow E: **A rapid means of sex identification in *Silene latifolia* by use of flow cytometry**. *Plant Molecular Biology Reporter* 1991, **9**(4):359-370.

16. Ming R, Moore PH: **Genomics of sex chromosomes**. *Curr Opin Plant Biol* 2007, **10**(2):123-130.

17. Široký J, Lysak MA, Doležel J, Kejnovský E, Vyskot B: **Heterogeneity of rDNA distribution and genome size in *Silene* spp**. *Chromosome Research* 2001, **9**(5):387-393.

18. Jeffreys AJ, May CA: **Intense and highly localized gene conversion activity in human meiotic crossover hot spots**. *Nat Genet* 2004, **36**(2):151-156.

19. Chen JF, Lu F, Chen SS, Tao SH: **Significant positive correlation between the recombination rate and GC content in the human pseudoautosomal region**. *Genome* 2006, **49**(5):413-419.

20. Burgoyne PS: **Genetic homology and crossing over in the X and Y chromosomes of mammals**. *Hum Genet* 1982, **61**(2):85-90.

21. Mohandas TK, Speed RM, Passage MB, Yen PH, Chandley AC, Shapiro LJ: **Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp**. *Am J Hum Genet* 1992, **51**(3):526-533.

22. Burgoyne PS, Mahadevaiah SK, Sutcliffe MJ, Palmer SJ: **Fertility in Mice Requires X-Y Pairing and a Y-Chromosomal Spermiogenesis Gene-Mapping to the Long Arm**. *Cell* 1992, **71**(3):391-398.

23. Matsuda Y, Moens PB, Chapman VM: **Deficiency of X-Chromosomal and Y-Chromosomal Pairing at Meiotic Prophase in Spermatocytes of Sterile Interspecific Hybrids between Laboratory Mice (*Mus-Domesticus*) and *Mus-Spretus***. *Chromosoma* 1992, **101**(8):483-492.

24. Bachtrog D, Charlesworth B: **Towards a complete sequence of the human Y chromosome**. *Genome Biol* 2001, **2**(5):reviews1016.1- reviews1016.5.

25. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, Frankish A, Lovell FL, Howe KL, Ashurst JL, Fulton RS, Sudbrak R, Wen G, Jones MC, Hurles ME, Andrews TD, Scott CE, Searle S, Ramser J, Whittaker A, Deadman R, Carter NP, Hunt SE, Chen R, Cree A, Gunaratne P *et al*: **The DNA sequence of the human X chromosome**. *Nature* 2005, **434**(7031):325-337.

26. Perry J, Palmer S, Gabriel A, Ashworth A: **A Short Pseudoautosomal Region in Laboratory Mice**. *Genome Res* 2001, **11**(11):1826-1832.

27. Das PJ, Chowdhary BP, Raudsepp T: **Characterization of the Bovine Pseudoautosomal Region and Comparison with Sheep, Goat, and Other Mammalian Pseudoautosomal Regions**. *Cytogenetic and Genome Research* 2009, **126**(1-2):139-147.

28. Langford C, Fischer P, Binns M, Holmes N, Carter N: **Chromosome-specific paints from a high-resolution flow karyotype of the dog**. *Chromosome Research* 1996, **4**(2):115-123.

29. Young A, Kirkness E, Breen M: **Tackling the characterization of canine chromosomal breakpoints with an integrated in-situ/in-silico approach: The canine PAR and PAB**. *Chromosome Research* 2008, **16**(8):1193-1202.

30. Perry J, Ashworth A: **Evolutionary rate of a gene affected by chromosomal position**. *Curr Biol* 1999, **9**(17):987-989.

31. Montoya-Burgos JI, Boursot P, Galtier N: **Recombination explains isochores in mammalian genomes**. *Trends in genetics : TIG* 2003, **19**(3):128-130.

32. Di Stilio VS, Kesseli RV, Mulcahy DL: **A pseudoautosomal random amplified polymorphic DNA marker for the sex chromosomes of *Silene dioica***. *Genetics* 1998, **149**(4):2057-2062.

33. Westergaard M: **The mechanism of sex determination in dioecious flowering plants.** *Advances in genetics* 1958, **9**:217-281.

34. Čermák T, Kubát Z, Hobza R, Koblížková A, Widmer A, Macas J, Vyskot B, Kejnovský E: **Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes**. *Chromosome Research* 2008, **16**(7):961-976.

35. Filatov DA, Howell EC, Groutides C, Armstrong SJ: **Recent Spread of a Retrotransposon in the Silene latifolia Genome, Apart From the Y Chromosome**. *Genetics* 2009, **181**(2):811-817.

36. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons**. *Nucleic Acids Res* 2000, **28**(1):267-269.

37. Macas J, Kejnovský E, Neumann P, Novák P, Koblížková A, Vyskot B: **Next Generation Sequencing-Based Analysis of Repetitive DNA in the Model Dioceous Plant *Silene latifolia***. *PLoS ONE* 2011, **6**(11):e27335.

38. Kalendar R, Flavell AJ, Ellis THN, Sjakste T, Moisy C, Schulman AH: **Analysis of plant diversity with retrotransposon-based molecular markers**. *Heredity* 2011, **106**(4):520-530.

39. Kejnovský E, Kubát Z, Macas J, Hobza R, Mracek J, Vyskot B: ***Retand*: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat**. *Molecular Genetics and Genomics* 2006, **276**(3):254-263.

40. Otto SP, Pannell JR, Peichel CL, Ashman T-L, Charlesworth D, Chippindale AK, Delph LF, Guerrero RF, Scarpino SV, McAllister BF: **About PAR: The distinct evolutionary dynamics of the pseudoautosomal region**. *Trends in genetics : TIG* 2011, **27**(9):358-367.

41. Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais GAB: **Rapid De Novo Evolution of X Chromosome Dosage Compensation in *Silene latifolia*, a Plant with Young Sex Chromosomes**. *Plos Biol* 2012, **10**(4):e1001308.

42. Galtier N, Piganeau G, Mouchiroud D, Duret L: **GC-content evolution in mammalian genomes: The biased gene conversion hypothesis**. *Genetics* 2001, **159**(2):907-911.

43. Gabriel M: **Biased gene conversion: implications for genome and sex evolution**. *Trends in Genetics* 2003, **19**(6):330-338.

44. Duret L, Galtier N: **Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes**. *Annu Rev Genomics Hum Genet* 2009, **10**(1):285-311.

45. Filatov DA: **A gradient of silent substitution rate in the human pseudoautosomal region**. *Mol Biol Evol* 2004, **21**(2):410-417.

46. Raudsepp T, Chowdhary BP: **The horse pseudoautosomal region (PAR): characterization and comparison with the human, chimp and mouse PARs**. *Cytogenetic and Genome Research* 2008, **121**(2):102-109.

47. Huang SW, Friedman R, Yu N, Yu A, Li WH: **How strong is the mutagenicity of recombination in mammals?** *Mol Biol Evol* 2005, **22**(3):426-431.

48. Glémin S, Bazin E, Charlesworth D: **Impact of mating systems on patterns of sequence polymorphism in flowering plants**. *Proceedings of the Royal Society B: Biological Sciences* 2006, **273**(1604):3011-3019.

49. Moore RC, Kozyreva O, Lebel-Hardenack S, Široký J, Hobza R, Vyskot B, Grant SR: **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* 2003, **163**(1):321-334.

50. Čegan R, Marais GAB, Kubeková H, Blavet N, Widmer A, Vyskot B, Doležel J, Šafář J, Hobza R: **Structure and evolution of Apetala3, a sex-linked gene in *Silene latifolia***. *Bmc Plant Biol* 2010, **10**:180.

51. Delichère C, Veuskens J, Hernould M, Barbacar N, Mouras A, Negrutiu I, Monéger F: **SlY1, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein**. *Embo J* 1999, **18**(15):4169-4179.

52. Qiu S, Bergero R, Forrest A, Kaiser VB, Charlesworth D: **Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes**. *P Roy Soc B-Biol Sci* 2010, **277**(1698):3283-3290.

53. Bergero R, Forrest A, Kamau E, Charlesworth D: **Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes**. *Genetics* 2007, **175**(4):1945-1954.

54. Kaiser VB, Bergero R, Charlesworth D: **A new plant sex-linked gene with high sequence diversity and possible introgression of the X copy**. *Heredity* 2011, **106**(2):339-347.

55. Atanassov I, Delichère C, Filatov DA, Charlesworth D, Negrutiu I, Monéger F: **Analysis and Evolution of Two Functional Y-Linked Loci in a Plant Sex Chromosome System**. *Mol Biol Evol* 2001, **18**(12):2162-2168.

56. Marais GAB, Nicolas M, Bergero R, Chambrier P, Kejnovský E, Moneger F, Hobza R, Widmer A, Charlesworth D: **Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia***. *Curr Biol* 2008, **18**(7):545-549.

57. Filatov DA: **Substitution rates in a new *Silene latifolia* sex-linked gene, SlssX/Y**. *Mol Biol Evol* 2005, **22**(3):402-408.

58. Matsunaga S: **Sex chromosome-linked genes in plants**. *Genes Genet Syst* 2006, **81**(4):219-226.

59. Kaiser VB, Bergero R, Charlesworth D: **Slcyt, a Newly Identified Sex-Linked Gene, Has Recently Moved onto the X Chromosome in *Silene latifolia* (Caryophyllaceae)**. *Mol Biol Evol* 2009, **26**(10):2343-2351.

60. Guo X, Wang Y, Keightley P, Fan L: **Patterns of selective constraints in noncoding DNA of rice**. *BMC Evolutionary Biology* 2007, **7**(1):208.

61. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes**. *Nat Genet* 2002, **30**(2):194-200.

62. Kubát Z, Hobza R, Vyskot B, Kejnovský E: **Microsatellite accumulation on the Y chromosome in *Silene latifolia***. *Genome* 2008, **51**(5):350-356.

63. Hobza R, Lengerova M, Svoboda J, Kubeková H, Kejnovský E, Vyskot B: **An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution**. *Chromosoma* 2006, **115**(5):376-382.

64. Hobza R, Kejnovský E, Vyskot B, Widmer A: **The role of chromosomal rearrangements in the evolution of *Silene latifolia* sex chromosomes**. *Mol Genet Genomics* 2007, **278**(6):633-638.

65. Kejnovský E, Hobza R, Čermák T, Kubát Z, Vyskot B: **The role of repetitive DNA in structure and evolution of sex chromosomes in plants**. *Heredity* 2009, **102**(6):533-541.

66. Čegan R, Vyskot B, Kejnovský E, Kubat Z, Blavet H, Šafář J, Doležel J, Blavet N, Hobza R: **Genomic Diversity in Two Related Plant Species with and without Sex Chromosomes - *Silene latifolia* and *S. vulgaris***. *PLoS ONE* 2012, **7**(2):e31898.

67. Tao Q, Wang A, Zhang H-B: **One large-insert plant-transformation-competent BIBAC library and three BAC libraries of Japonica rice for genome research in rice and other grasses**. *TAG Theoretical and Applied Genetics* 2002, **105**(6):1058-1066.

68. Bouzidi MF, Franchel J, Tao QZ, Stormo K, Mraz A, Nicolas P, Mouzeyar S: **A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions**. *Theor Appl Genet* 2006, **113**(1):81-89.

69. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

70. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**(1):78-94.

71. R Development Core Team: **R: A Language and Environment for Statistical Computing. [*http://www.R-project.org*]**. 2011.

72. Kosambi DD: **The estimation of map distances from recombination values**. *Annals of Human Genetics* 1943, **12**(1):172-175.

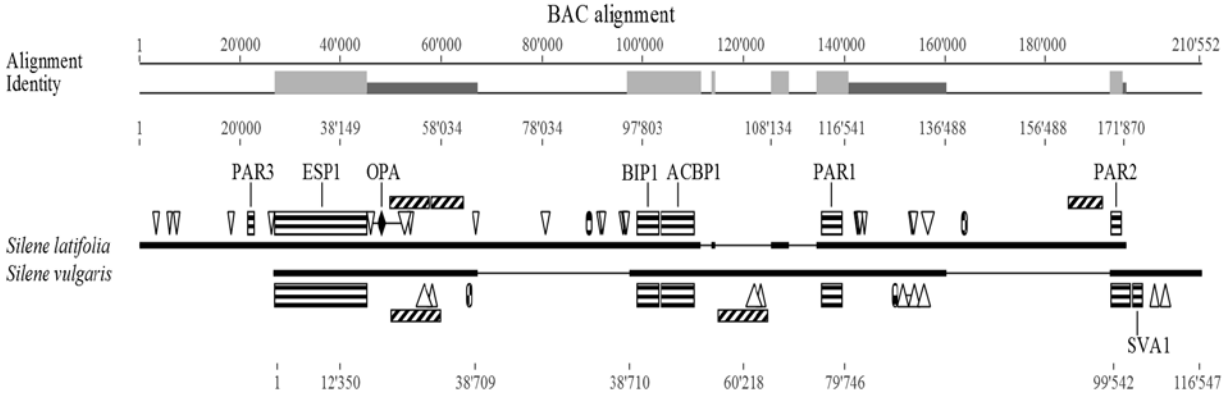
73. Haldane JBS: **The combination of linkage values and the calculation of distance between the loci of linked factors**. *Journal of Genetics* 1919, **8**:299-309.

74. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W265-268.

75. Liu Z, Yue W, Li DY, Wang RRC, Kong XY, Lu K, Wang GX, Dong YS, Jin WW, Zhang XY: **Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres**. *Chromosoma* 2008, **117**(5):445-456.

76. Wolfe KH, Li WH, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs**. *Proceedings of the National Academy of Sciences* 1987, **84**(24):9054-9058.

Additional figure 1: Alignment of *Silene latifolia* and *S. vulgaris* BAC scaffolds. Genes (horizontal strip), transposable elements (white triangle), full LTR transposon sequences (diagonal strip) and uncharacterized nucleotides (checkerboard) are annotated on the BAC sequences (black bold line). Regions of high identity (light gray), low identity (dark gray) and gaps (fine black line) are indicated. The position of marker ScOPA09, a PAR-specific marker used to identify BAC clones located within the *S. latifolia* PAR, is indicated by a black diamond.



Additional table S1: Gene expression. Expression data were extracted from Muyle *et al.* [41].

Name	Normalized expression						Mean		Expression
	U10_37_female	U10_39_female	U10_34_female	U10_49_male	U10_11_male	U10_9_male	Female	Male	
BIP1	2.86E-03	7.47E-03	7.44E-03	6.49E-03	1.19E-02	1.04E-02	5.92E-03	9.60E-03	unbiased
ACBP1	9.26E-04	7.50E-04	6.36E-04	5.77E-04	5.85E-04	5.91E-04	7.71E-04	5.84E-04	unbiased
PAR1	4.50E-04	7.38E-04	7.49E-04	4.09E-04	5.99E-04	5.80E-04	6.46E-04	5.29E-04	unbiased
ESP1	2.05E-04	4.78E-04	5.13E-04	1.42E-04	3.30E-04	4.07E-04	3.99E-04	2.93E-04	unbiased

Additional table S2: Comparison of GC and GC3 contents in *Silene latifolia* sex-linked genes. GC and GC3 contents were calculated for known sex-linked genes located in the non-recombining part of the *S. latifolia* X chromosome. CDS sequences were extracted from GenBank.

Gene name	***DD44***	***SlAP3X***	***X1***	***CypX***	***X7***	***X9***	***X4***	***X3***	***SlssX***
Sequence length (bp)	693	633	1,419	561	738	447	1,118	1,687	839
Exon GC content (%)	45.4	42.2	44.9	42.1	47.1	49.2	46.3	40.5	44.8
Exon GC3 content (%)	49.3	42.6	36.8	36.4	41.9	51	38.2	33.4	54.8

Additional table S3: Genetic analysis of recombination events between the PAB and markers ScOPA09, *ESP1*, PAR2 and *ACBP1*. Standard errors are indicated in brackets.

	ScOPA09	***ESP1***	**PAR2**	***ACBP1***
Recombination fraction (%)	10.5 (3.5)	10.5 (3.5)	10.5 (3.5)	10.5 (3.5)
Distance from PAB - Kosambi function (cM)	10.7 (3.7)	10.7 (3.7)	10.7 (3.7)	10.7 (3.7)
Distance from PAB - Haldane function (cM)	11.8 (4.4)	11.8 (4.4)	11.8 (4.4)	11.8 (4.4)

Additional table S4: Genetic analysis of the linkage between the marker ScOPA09 and each gene *ESP1*, *PAR2* and *ACBP1*.

	ESP1	**PAR2**	***ACBP1***
Recombination fraction (%)	0	0	0

Additional table S5: LTR insertion time estimates

Positions and characteristics of LTRs (Primer binding site (PBS), polypurine tract (PPT), both 5'- and 3'-LTR size) found using LTR_Finder. Time was computed as described by Liu *et al*. (2008) [75]: Nucleotide differences / (2 x LTR length x K). K: substitutions per site per year = 23E-9. Sl = *S.*

Species	Subgroup	Position	Size (bp)	PBS	PPT	5′-LTR size (bp)	3′-LTR size (bp)	Nucleotide differences	Time (years)
Sl	*Copia*	56,822-47,755	9,068	TATCATGAGCCACGGTT	AGGACAAGTGGGAGA	1,264	1,264	1	17,199
Sl	-	56,871-64,160	7,290	CGCCGTTGCCGGGGA	GGGGACAAGCAAATG	1,610	1,612	30	405,077
Sl	-	159,510-167,551	8,042	-	TTAAGGAGGGAAGAT	1,419	1,406	50	766,002
Sv	*Retand*	31,853-20,726	11,127	CGCCGTCTGTGGGGA	AAGCCGACGGAGAAA	605	605	1	35,932
Sv	*Retand*	65,586-54,158	11,429	CGCCGTCTGTGGGGA	AAGCCGACGGAGAAA	735	735	8	236,616

Additional table S6: Results of the different BAC sequencing approaches and assemblies. The same *S. latifolia* BAC clone was sequenced by shotgun Sanger sequencing and 454 pyrosequencing with different paired-end libraries.

	Species			
	S. latifolia			*S. vulgaris*
Sequencing method	454 pair-end 3kb	454 pair-end 8kb	Sanger	454 pair-end 3kb
Assembly software	GS De Novo Assembler	GS De Novo Assembler	Phrap	GS De Novo Assembler
Reads	162,365	129,253	1,154	194,585
Coverage	140	135	19	280
Average read length (bp)	168	230	2,660	180
Assembled reads	88.59%	92.51%	94.71%	90.83%
Assembled bases	93.47%	96.40%	30.03%	96.25%

Additional table S7: Primers and restriction enzymes used for genetic analysis. The annealing temperature for all primer pairs is 60°C.

Marker name	Forward-primer (5' – 3')	Reverse-primer (5' – 3')	Restriction enzyme
ESP1	GACTGGTAAAAAAGAAACGATTTAT	TGGAAGAATCGTGGTTCCAT	CAPS (TaqI)
ScOPA09	GCAATTCACCATCCTCTGCTCCCCAACCAC	ATGGTCTTTGGGCCCTTATC	CAPS (MboI)
ACBP1	CTCGAGATTCCTCCTCGTGA	CCGCTTTCAAAGACGACAAT	CAPS (MwoI)
PAR2	AATGACAACTCTCTACCGTCCA	CAGCTAGAAACAAGCGATGC	CAPS (AccI)

Chapter III

Manuscript:

Plant sex chromosome evolution: a genomic view in *Silene latifolia*.

An unpublished manuscript co-authored with Hana Blavet[2], Jos Käfer[3], Radim Cegan[2], Clothilde Deschamps[4], Aline Muyle[3], Niklaus Zemp[1], Roman Hobza[2,5], Gabriel Marais[3] and Alex Widmer[1]

[1]Institute of Integrative Biology (IBZ), ETH Zürich, Universitätstrasse 16, Zürich, 8092, Switzerland
[2]Institute of Biophysics, Laboratory of Plant Developmental Genetics, Academy of Sciences of the Czech Republic, v.v.i. Kralovopolska 135, Brno, CZ-61200, Czech Republic
[3]Laboratoire de Biométrie et Biologie évolutive, Université Lyon 1, CNRS, UMR5558, Villeurbanne, F-69622 cedex, France
[4]Pôle Rhône-Alpes de Bioiformatique (PRABI), Villeurbanne, F-69622 cedex, France
[5]Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Sokolovská 6, Olomouc, CZ-77200, Czech Republic

Abstract

Background:

Silene latifolia has become a model for the study of plant sex chromosome evolution. This species has a sex chromosome system similar to mammals, with females being homogametic (XX) and males heterogametic (XY). While mammalian sex chromosomes present an evolutionary old system with sex chromosomes that emerged 150 million years ago (mya),, the *S. latifolia* sex chromosomes emerged about 10 mya. *Silene vulgaris* is a closely related species of *S. latifolia* that lacks sex chromosomes. Comparisons between these plant species allow investigating early steps of the evolution of X and Y chromosomes in plants. We present in this study, an analysis of homologous BAC sequences identified on the *S. latifolia* X and Y chromosomes and the corresponding *S. vulgaris* autosomes, and we report the location of several newly identified sex-linked genes.

Results:

We identified 24 BAC clones containing known sex-linked genes and discovered 17 new sex-linked genes and 51 *S. vulgaris* autosomal genes that are potentially *S. latifolia* sex-linked gene homologs. In addition, we localized 59 recently discovered sex-linked genes and estimated positions of 31 others. Moreover, we found 18 new gene triplets (*S. latifolia* X, Y and *S. vulgaris* autosomal homologs). Structural analysis revealed small-scale collinearity conserved between genes close to *S. latifolia SlX6a*, a gene located in the oldest evolutionary stratum. BAC sequence assembly allowed to physically map genes for which only genetic maps were so far available and allowed us to identified the telomeric region of the *S. latifolia* X chromosome q-arm.

Conclusions:

The present study demonstrates the strength of combining BAC library analyses and next generation sequencing in a comparative approach. The notable absence of detectable pseudogenes on the Y chromosome and the homogeneous distribution of genes with reduced expression of the Y-linked allele along the X chromosome suggest that gene loss on the Y chromosome is rare and that inactivation of Y-linked gene copies is random in *Silene latifolia*.

Keywords: BAC library, next generation sequencing, sex chromosomes, *Silene*

Introduction

Silene latifolia sex chromosomes have emerged about 10 million years ago [1, 2]. While *S. latifolia* sex chromosomes are about fifteen times younger than mammalian sex chromosomes [3], several similarities have been found between both systems. First, *Silene latifolia* has heteromorphic sex chromosomes with females being the homogametic sex (XX) and males the heterogametic sex (XY). Degeneration of the *S. latifolia* Y chromosome is evidenced by the lethality of YY and Y0 mutants [4, 5], reduced expression of Y alleles [6], accumulation of transposable elements (TE) [7, 8] and insertion of chloroplastic DNA [9]. As previously found in human sex chromosomes [10], the suppression of recombination between *S. latifolia* X and Y chromosomes is a gradual process led by chromosomal rearrangements such as inversions, which resulted in formation of evolutionary strata [1, 11]. Recently, dosage compensation has been detected in *S. latifolia* [12] showing an equal dosage of X transcripts for many genes in both males and females, similar to the situation in mammals [13].

The major limitation for the analysis of the sex chromosomes in *Silene* was until recently the availability of sex-linked genes. During the last decade, only about ten genes were identified on *S. latifolia* sex chromosomes [6, 14-18]. However, with the expansion of next generation sequencing technologies, new sex-linked genes were recently identified. Indeed, next-generation sequencing methods based on cDNA sequencing facilitated the collection and analysis of large numbers of *Silene* gene sequences [19] and identification of novel putatively sex-linked genes [12, 20, 21]. A major limitation of these approaches for the study of sex chromosome evolution, however, is that the identification of sex-linked genes relies on the identification of male-specific polymorphisms that represent Y-linked alleles. Consequently, these approaches cannot inform us about the frequency with which alleles on the degenerating Y chromosome have been lost or silenced, even that they can give hints as hemizygous X-linked genes can be identified.

We therefore used a different approach based on sequencing selected bacterial artificial chromosome (BAC) clones. While this method is highly demanding in terms of time and resources needed to develop and screen BAC libraries, compared to cDNA sequencing for example with the mRNA-seq approach, it allows to detect also weakly expressed genes and pseudogenes, helps identifying the exon/intron structure and the linear arrangement of genes and allows analyzing intronic and intergenic regions. A combination of both cDNA library and BAC sequencing can further confirm sex-linkage of some of the newly obtained genes. Moreover, with the expression data provided by the cDNA libraries, we can test whether genes with

reduced expression of the Y-allele and dosage compensated genes occur preferentially in the oldest strata that contain the most differentiated genes [1, 6, 11].
Here we report the results of a genomic comparison between sequences of 24 BAC clones coming from *S. latifolia* X and Y chromosomes and *S. vulgaris* homologous autosomes.

Material and Methods

BAC preparation

BACs were prepared following the methods described by Blavet *et al.* (in review) and Čegan *et al.* (2010) [22]. The genes used to screen the BAC library are listed on Table 1. In order to analyze selected regions along the *S. latifolia* sex chromosomes and the corresponding regions on the *S. vulgaris* autosomes, we first screened a BAC library of *S. latifolia* with different markers that were previously found to be located on the X chromosome (see Table 1) and have successfully been used for linkage mapping of the *S. latifolia* sex chromosomes [1, 11]. BAC sequencing was performed by 454 pyrosequencing (Roche). From these sequences we identified new genes and used them to screen a *S. vulgaris* BAC library for homologous clones. The *Silene latifolia* BAC library was then screened again these genes to detect Y copies. All BAC sequences were assembled into contigs using GS De Novo Assembler (Roche).

Table 1: Sex-linked genes used to screen the BAC library.

Gene name	Information	References
SlX1	WD repeat protein	[14]
SlXDD44	Oligomycin sensitivity-conferring protein	[16]
Slss	Spermidine synthase	[17]
SlX3	Putative calcium dependent protein kinase	[6]
SlX4	Fructose-2,6-biphosphatase	[15]
SlX6a	Unknown protein	[18]
SlX6b	Unknown protein	[18]
SlX7	Unknown protein	[18]
SlXCyp	Peptidyl-prolyl cis-trans isomerase	[6]

Sequencing, assemblies and annotations

All 454 sequences were *de novo* assembled using Roche GS De Novo Assembler with 98% minimum overlap identity and 60 bp as minimum overlap length. We set the expected depth parameter with reference to the estimated size of the BAC (determined by pulse field gel electrophoresis (PFGE)) and the number of reads sequenced.

The contigs were annotated by similarity using BLAST [23] with UniProtKB (Swiss-Prot + TrEMBL, 13 July 2010), the *Arabidopsis thaliana* proteome (TAIR10_20100802), *S. latifolia* cDNA contigs [12, 19] and transposable elements (TAIR9_TE) [www.arabidopsis.org] with an E-value cut-off of 1E-4. We also used a *Silene* repeated element database [24] to detect *Silene*-specific repeats and transposable elements. Transposable element annotations were then completed using annotated *S. latifolia* repeats [8]. After BLAST identification of Illumina contigs built by Muyle *et al.* (2012) [12], we used gene expression data to test whether gene expression patterns differ among the evolutionary strata. Statistical analyses were performed with R [25]. Moreover we used both BAC and Illumina sequence read alignments to estimate numbers of both non-synonymous and synonymous substitutions for all genes.

Results

BAC sequencing, assembly and annotations

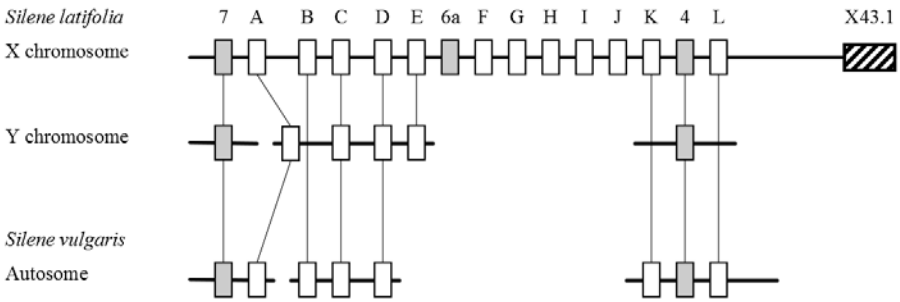
The 454 sequence reads of *S. latifolia* and *S. vulgaris* BAC clones were assembled in contigs and a total of 158 genes were identified through BLAST searches. Among the identified genes, we detected 90 genes that were recently identified as sex-linked genes by cDNA sequencing [12], with 59 and 31 of these being localized on *S. latifolia* and *S. vulgaris* BACs, respectively. In addition we found 17 new sex-linked genes and 51 *S. vulgaris* genes that are potentially sex-linked in *S. latifolia*. Among the identified genes, 18 triplets (homologous gene copies from *S. latifolia* X and Y chromosomes, and *S. vulgaris* autosome) were found. A total of 362 transposable or repeated elements were identified, as indicated by their annotations with Uniprot (*www.uniprot.org*) and repeat coverage. The repeat coverage is based on BLAST hits to a *Silene* repeated elements library [24].

Sex chromosome structure and comparison

Large-scale collinearity between the *S. latifolia* X chromosome and *S. vulgaris* autosome has repeatedly been reported in studies of *S. latifolia* sex chromosome evolution [2, 26-28]. This collinearity was evidenced by genetic

mapping. With the development of genomic resources, small-scale collinearity can be analysed at much higher resolution. In the present study we report both the assembly of the region containing the genes *SlX6a*, *SlX7* and *SlX4* [15, 18] located on the X chromosome and the collinearity of these genes, with copies on the *S. latifolia* Y chromosome and *S. vulgaris* autosomes (Figure 1). We therefore confirmed the proximity of genes *SlX6a*, *SlX7* and *SlX4* that were expected to be closely linked based on earlier genetic mapping results [11] and found that the gene *SlX4* is potentially linked to the telomeric region of the X chromosome (Figure 1). Moreover we have evidenced the absence of neighboring genes close to genes *SlY4* and *SlY7*. These are the only genes found over 150 kb and 205 kb of Y chromosome sequences, respectively, while several transposable elements are present along the sequences (Supplementary Figure 1).

Figure 1: Schematic representation of sex-linked gene arrangement. Gray rectangles indicate known sex-linked genes: 7 = *SlX7*, 6a = *SlX6a* and 4 = *SlX4*. White rectangles indicate newly discovered sex-linked genes. The rectangle with diagonal lines indicates X43.1 repeats that characterize the telomere of the *S. latifolia* X chromosome q-arm.



Pseudogene investigation

The identified X-linked genes were used to search for pseudogenes on the Y chromosome derived BAC clones. For several genes in close proximity to genes *SlX4*, *SlX6a* and *SlX7* on X-derived BAC clones, we have not detected any Y copies on BAC clones containing the genes *SlY4* and *SlY7*. We used protein sequences determined from available cDNA reads [12, 19] and open reading frame (ORF) predictions from prot4EST [29] for each X-linked gene in order to detect partial Y-linked copies, but none were identified.

Table 2: Sex-biased gene expression. Gene expression and dosage compensation are inferred from the results of Muyle *et al.* (2012) [12].

Stratum	Number of genes	Gene with reduced Y expression	Dosage compensated genes
1	7	5	0
2	20	13	3
3	11	7	4

Gene expression and dosage compensation

With the recent results on dosage compensation in *S. latifolia* evidenced by Muyle *et al.* (2012) [12], we assessed both the distribution of expression patterns and dosage compensated expression of the newly identified sex-linked genes with respect to their location in different evolutionary strata along the X chromosome. Table 2 presents the results of the expression analyses for genes located on the investigated BAC clones and indicates the amount of genes that are dosage compensated. We found that the distribution among evolutionary strata along the X chromosome is random for sex-linked genes that have a reduced expression of their Y allele (χ^2 = 0.255, df = 2, p-value = 0.8803) and that are dosage compensated (χ^2 = 4.0929, df = 2, p-value = 0.1292). Moreover, no significant difference exists in the average level of Y allele expression reduction between the genes from the different strata (anova: p-value = 0.8005).

Non-synonymous and synonymous substitutions

Using RNA-seq reads from Muyle and coworkers [12] in combination with our BAC sequences, we analyzed substitution patterns of the genes for which we have copies from the *S. latifolia* X and Y chromosomes and the *S. vulgaris* autosomes. For the 18 gene triplets identified, we counted non-synonymous and synonymous substitutions occurring in both the X and Y alleles using *S. vulgaris* as outgroup (Table 3). Most non-synonymous mutations occurred in the alleles located on the Y chromosomes (2.3 and 2.6 times more on average than on the X for non-synonymous and synonymous mutations respectively) and are most common in genes in stratum 1, significantly for synonymous mutations (χ^2 = 7.2275, df = 2, p-value = 0.02695) but not significantly for non-synonymous mutations (χ^2 = 3.98, df = 2, p-value = 0.1367).

Table 3: Non-synonymous and synonymous substitutions in the different gene triplets. The genes are sorted according to their positions along the X chromosome. Strata are separated by dashed lines. NS = non-synonymous substitutions; SS = synonymous substitutions.

Name	BAC	Length		NS		SS	
		X	Y	X	Y	X	Y
Contig_29617	4	253	253	0	1	0	1
XY4 (Contig_29527)	4	370	360	8	15	15	28
gene_50 (Contig_1767)	6	464	234	1	14	3	56
gene_79 (Contig_62587)	7	784	451	9	50	25	119
XY7 (Contig_1849)	7	363	361	8	11	16	31
XY3 (Contig_49583)	3	158	253	11	6	28	14
Total stratum 1				37	97	87	249
Contig_61876-64482	dd44	354	354	1	0	2	1
XYDD44 (Contig_60039)	dd44	108	107	1	2	1	1
Xyss (Contig_8045)	ss	347	342	2	11	4	8
Contig_3463	cyp	1028	1028	2	5	8	8
Contig_53812	cyp	222	222	0	3	0	0
Contig_53821	cyp	106	106	0	1	0	0
Contig_58571	cyp	450	450	2	4	1	4
Contig_59644	cyp	148	148	0	0	4	6
Contig_63486	cyp	367	367	1	2	0	1
XYCyp (Contig_53905)	cyp	639	639	6	11	10	18
Total stratum 2				15	39	30	47
Contig_22623	1	86	86	0	0	0	0
Contig_44823	1	543	543	0	2	0	2
Contig_48921	1	165	165	0	0	0	2
Contig_53010	1	680	680	1	1	0	2
Contig_53773	1	129	129	0	0	0	1
Contig_59073	1	835	833	2	5	4	18
Contig_6732	1	820	729	3	4	3	7
XY1 (Contig_55531)	1	133	317	5	0	1	3
Total stratum 3				11	12	8	35
Total				63	148	125	331

Discussion

Y chromosome enlargement and weak gene loss

The comparison between the different BAC sequences revealed new evidence for small-scale gene collinearity and enlargement of the Y chromosome by insertion of transposable elements and showed that gene loss on the Y chromosome is reduced in comparison to much older sex chromosomes such as those of humans [10]. We found that both genes, *SlY4* and *SlY7*, are isolated, without any other gene located in close proximity, on their respective BACs (which are more than 150 kb in size), whereas other genes were localized in close proximity to the X copies of these genes (about 2 kb and 1 kb for the closest gene to *SlX4* and *SlX7*, respectively). Moreover, no pseudogenes, resulting from the degeneration of the neighboring X-copy genes, has been detected on the Y chromosome which may indicate that *S. latifolia* Y-linked gene loss is very weak, as suggested from recent RNA-seq analysis [21]. Both results imply that massive insertion of transposable elements in the intergenic regions between Y-linked gene copies contributed to the growth of the Y chromosome and separated neighboring genes (see Additional Figure 1). This scenario is supported by our finding of numerous transposable elements on Y-derived BACs. The insertion of transposable elements is known to be one of the main causes of sex chromosome degeneration and enlargement [7, 30, 31] and we here found a case where intergenic regions have been enlarged more than 70 times, which is more than expected given that the *S. latifolia* Y chromosome is about 7.5 times larger than the *S. vulgaris* autosome [32]. A comparison with the moderate enlargement discovered in the pseudoautosomal region of the sex chromosome [33], the present result suggest that some regions of the *S. latifolia* Y chromosome have been differentially invaded by transposable elements and that TE invasion is strongest in that part of the chromosome where recombination stopped first during the evolution of the sex chromosomes.

Inactivation of Y-linked alleles

While evolutionary strata can readily be identified based on synonymous site divergence between X and Y-linked gene copies and are commonly found in many sex chromosome systems, patterns of divergence in gene expression between gene copies on the different chromosomes are less clear. A recent study in *S. latifolia* estimated that the expression level of Y alleles is about 87% of that of X alleles and concluded that most of the Y-linked gene copies are active [21]. However, the physical arrangement of the studied genes is unknown. Using expression data of the genes located on our BACs in the different strata, we could test the hypothesis that reduced expression of Y-linked alleles is strongest in the oldest stratum. Contrary to expectations, however, our results suggest that genes of the Y chromosome are

randomly inactivated. Thus, our results for the evolutionary young *S. latifolia* Y chromosome resemble these obtained for the *Drosophila miranda* neo-Y chromosome, for which a random model of gene inactivation was inferred [34].

Three models have been proposed for the inactivation of Y-linked alleles : first, the direct model, in which genes that are nonfunctional on the Y (due to frameshift mutations) are targeted for inactivation [35, 36]; second, the random model, in which genes are inactivated randomly with respect to their functionality on the Y, independent of frameshift mutations or transposable element insertions [35, 37, 38]; and third, the large-scale inactivation model, according to which large genomic regions are silenced simultaneously [35, 39]. For *S. latifolia*, we can exclude the direct model, because sequence divergence is most pronounced in the oldest stratum and inactivation would be preferentially found there. In addition, dosage compensation would be expected to occur on this stratum. Secondly, the large-scale inactivation model does not seem to explain the patterns of expression we observed. If *S. latifolia* Y chromosome evolution would follow this model, we should see inactivation, and then dosage compensation happening on neighboring genes, which is not the case. Finally, our results that show no preferential localization of genes with reduced Y expression suggest that the Y chromosome is more probable to follow a random inactivation process such as *Drosophila miranda* one [34].

Mutation rate and relaxed purifying selection

Our analysis of the distribution of non-synonymous and synonymous substitution provides new evidence for degeneration of Y chromosome-linked genes. Previous comparison of dN/dS ratios (nonsynonymous/synonymous substitution rates) between both X and Y alleles of *S. latifolia* sex-linked genes evidenced degeneration of the Y copies and revealed that these genes are evolving under purifying selection [6, 20]. While the analysis conducted by Marais and coworkers was limited to seven well characterized genes [6], the recent study conducted by Chibalina and Filatov was based on about 400 sex-linked genes characterized by RNA-seq, but for which location is still unknown [20]. Our results enlarge the set of sex-linked genes with known locations in the different strata. We found an accumulation of both non-synonymous and synonymous substitutions in Y-linked alleles, preferentially in stratum one. Such a pattern could result from a higher mutation rate on the Y chromosome [17, 40]. However, because purifying selection should eliminate deleterious non-synonymous substitutions, the trend that Y-linked genes accumulate more non-synonymous substitution compared to the X-linked copies may also support the hypothesis that relaxed purifying selection occurs on the Y chromosome [17, 20], which is most pronounced in the oldest stratum one.

Conclusions

In this study we present the analysis of several BAC sequences from *S. latifolia* X and Y chromosomes and *S. vulgaris* autosome. These sequences cover each of the three evolutionary strata identified on *S. latifolia* sex chromosomes. The analysis of the BAC sequences revealed the location of 76 sex-linked genes and showed a new case of collinearity between the different chromosomes of both *Silene* species, as well as a new example of massive transposable element insertion occurring on the Y chromosome. Absence of pseudogenes may indicate a weak process of gene loss on the *S. latifolia* Y chromosome and the presence of Y-linked genes with a reduced expression randomly distributed in the different strata suggests random inactivation of the Y-linked genes. Moreover, the analysis of substitutions confirms both a higher mutation rate on the Y chromosome and the hypothesis of a relaxed purifying selection on *S. latifolia* Y chromosome for which our results show a more important effect on the most diverged stratum one.

Acknowledgements

We thank L. Poveda, M. Kuenzli and W. Qi from the Functional Genomic Center Zurich (FGCZ) for assistance relating to 454 sequencing, T. Torossi, C. Michel and the ETH Zürich Genetic Diversity Center (GDC) for technical support, and S. Zoller for bioinformatics support. We further acknowledge support by J. Macas and E. Kejnovský who provided sequences of repeated elements and J. Bartoš who participated in BAC library analysis. This study was supported by an ETH Zurich grant (TH-07 06-3) to AW, by Czech Science Foundation grants (522/09/0083) to RH and Centre of the Region Haná for Biotechnological and Agricultural Research grant (ED0007/01/01) to RH.

References

1. Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Moneger F: **A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants**. *Plos Biol* 2005, **3**(1):47-56.

2. Filatov DA: **Evolutionary history of *Silene latifola* sex chromosomes revealed by genetic mapping of four genes**. *Genetics* 2005, **170**(2):975-979.

3. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Gruzner F, Deakin JE, Whittington CM, Schatzkamer K, Kremitzki CL,

Graves T, Ferguson-Smith MA, Warren W, Graves JAM: **Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes**. *Genome Res* 2008, **18**(6):965-973.

4. Ye D, Oliveira M, Veuskens J, Wu Y, Installe P, Hinnisdaels S, Truong AT, Brown S, Mouras A, Negrutiu I: **Sex determination in the dioecious Melandrium. The X/Y chromosome system allows complementary cloning strategies**. *Plant Science* 1991, **80**(1–2):93-106.

5. Veuskens J, Ye D, Oliveira M, Ciupercescu DD, Installé P, Verhoeven HA, Negrutiu I: **Sex determination in the dioecious Melandrium album: androgenic embryogenesis requires the presence of the X chromosome**. *Genome* 1992, **35**(1):8-16.

6. Marais GAB, Nicolas M, Bergero R, Chambrier P, Kejnovsky E, Moneger F, Hobza R, Widmer A, Charlesworth D: **Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia***. *Curr Biol* 2008, **18**(7):545-549.

7. Hobza R, Lengerova M, Svoboda J, Kubekova H, Kejnovsky E, Vyskot B: **An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution**. *Chromosoma* 2006, **115**(5):376-382.

8. Čermak T, Kubat Z, Hobza R, Koblizkova A, Widmer A, Macas J, Vyskot B, Kejnovský E: **Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes**. *Chromosome Research* 2008, **16**(7):961-976.

9. Kejnovský E, Kubat Z, Hobza R, Lengerova M, Sato S, Tabata S, Fukui K, Matsunaga S, Vyskot B: **Accumulation of chloroplast DNA sequences on the Y chromosome of <i>Silene latifolia</i&gt**. *Genetica* 2006, **128**(1):167-175.

10. Lahn BT, Page DC: **Four Evolutionary Strata on the Human X Chromosome**. *Science* 1999, **286**(5441):964-967.

11. Bergero R, Charlesworth D, Filatov DA, Moore RC: **Defining regions and rearrangements of the *Silene latifolia* Y chromosome**. *Genetics* 2008, **178**(4):2045-2053.

12. Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais GAB: **Rapid *De Novo* Evolution of X Chromosome Dosage Compensation in**

***Silene latifolia*, a Plant with Young Sex Chromosomes**. *Plos Biol* 2012, **10**(4):e1001308.

13. Vicoso B, Bachtrog D: **Progress and prospects toward our understanding of the evolution of dosage compensation**. *Chromosome Research* 2009, **17**(5):585-602.

14. Delichère C, Veuskens J, Hernould M, Barbacar N, Mouras A, Negrutiu I, Moneger F: **SIY1, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein**. *Embo J* 1999, **18**(15):4169-4179.

15. Atanassov I, Delichère C, Filatov DA, Charlesworth D, Negrutiu I, Monéger F: **Analysis and Evolution of Two Functional Y-Linked Loci in a Plant Sex Chromosome System**. *Mol Biol Evol* 2001, **18**(12):2162-2168.

16. Moore RC, Kozyreva O, Lebel-Hardenack S, Siroky J, Hobza R, Vyskot B, Grant SR: **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* 2003, **163**(1):321-334.

17. Filatov DA: **Substitution Rates in a New Silene latifolia Sex-Linked Gene, SlssX/Y**. *Mol Biol Evol* 2005, **22**(3):402-408.

18. Bergero R, Forrest A, Kamau E, Charlesworth D: **Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: Evidence from new sex-linked genes**. *Genetics* 2007, **175**(4):1945-1954.

19. Blavet N, Charif D, Oger-Desfeux C, Marais G, Widmer A: **Comparative high-throughput transcriptome sequencing and development of SiESTa, the Silene EST annotation database**. *BMC Genomics* 2011, **12**(1):376.

20. Chibalina Margarita V, Filatov Dmitry A: **Plant Y Chromosome Degeneration Is Retarded by Haploid Purifying Selection**. *Current biology : CB* 2011, **21**(17):1475-1479.

21. Bergero R, Charlesworth D: **Preservation of the Y Transcriptome in a 10-Million-Year-Old Plant Sex Chromosome System**. *Current biology : CB* 2011, **21**(17):1470-1474.

22. Čegan R, Marais GAB, Kubeková H, Blavet N, Widmer A, Vyskot B, Doležel J, Šafář J, Hobza R: **Structure and evolution of Apetala3, a sex-linked gene in *Silene latifolia***. *Bmc Plant Biol* 2010, **10**:180.

23. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

24. Macas J, Kejnovský E, Neumann P, Novák P, Koblížková A, Vyskot B: **Next Generation Sequencing-Based Analysis of Repetitive DNA in the Model Dioceous Plant *Silene latifolia***. *PLoS ONE* 2011, **6**(11):e27335.

25. R Development Core Team: **R: A Language and Environment for Statistical Computing. [*http://www.R-project.org*]**. 2011.

26. Matsunaga S: **Sex chromosome-linked genes in plants**. *Genes Genet Syst* 2006, **81**(4):219-226.

27. Kaiser VB, Bergero R, Charlesworth D: **Slcyt, a Newly Identified Sex-Linked Gene, Has Recently Moved onto the X Chromosome in *Silene latifolia* (Caryophyllaceae)**. *Mol Biol Evol* 2009, **26**(10):2343-2351.

28. Qiu S, Bergero R, Forrest A, Kaiser VB, Charlesworth D: **Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes**. *P Roy Soc B-Biol Sci* 2010, **277**(1698):3283-3290.

29. Wasmuth JD, Blaxter ML: **Prot4EST: Translating Expressed Sequence Tags from neglected genomes**. *BMC Bioinformatics* 2004, **5**.

30. Kejnovský E, Hobza R, Čermak T, Kubat Z, Vyskot B: **The role of repetitive DNA in structure and evolution of sex chromosomes in plants**. *Heredity* 2009, **102**(6):533-541.

31. Bergero R, Forrest A, Charlesworth D: **Active miniature transposons from a plant genome and its nonrecombining Y chromosome**. *Genetics* 2008, **178**(2):1085-1092.

32. Široký J, Lysák MA, Doležel J, Kejnovský E, Vyskot B: **Heterogeneity of rDNA distribution and genome size in *Silene* spp**. *Chromosome Research* 2001, **9**(5):387-393.

33. Blavet N, Blavet H, Cegan R, Zemp N, Zdanska J, Janousek B, Hobza R, Widmer A: **Comparative analysis of a plant pseudoautosomal region (PAR) in *Silene latifolia* with the corresponding *S. vulgaris* autosome**. *BMC Genomics* 2012, **13**:226.

34. Bachtrog D: **Expression Profile of a Degenerating Neo-Y Chromosome in Drosophila**. *Curr Biol* 2006, **16**(17):1694-1699.

35. Bachtrog D: **Sex chromosome evolution: Molecular aspects of Y-chromosome degeneration in Drosophila**. *Genome Res* 2005, **15**(10):1393-1401.

36. Orr HA, Kim Y: **An Adaptive Hypothesis for the Evolution of the Y Chromosome**. *Genetics* 1998, **150**(4):1693-1698.

37. Charlesworth B, Charlesworth D: **The degeneration of Y chromosomes**. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 2000, **355**(1403):1563-1572.

38. Charlesworth B: **The evolution of chromosomal sex determination and dosage compensation**. *Curr Biol* 1996, **6**(2):149-162.

39. Steinemann M, Steinemann S: **Enigma of Y chromosome degeneration: Neo-Y and Neo-X chromosomes of Drosophila miranda a model for sex chromosome evolution**. *Genetica* 1998, **102-103**(0):409-420.

40. Filatov DA, Charlesworth D: **Substitution Rates in the X- and Y-Linked Genes of the Plants, *Silene latifolia* and *S. dioica***. *Mol Biol Evol* 2002, **19**(6):898-907.

Supplementary material

Supplementary Figure 1: Gene 4 BAC arrangements. Genes (dark gray) and transposable elements (white), are annotated on the BAC sequences (black bold line). The position of genes *SIX4*, *SIY4* and *Sv4* used to identify BAC clones are indicated by dot filled boxes.

General discussion

The genus *Silene* has become a model system in many research areas, including ecology, evolution and genetics [Bernasconi, et al., 2009] and *Silene latifolia* has become established as a model to study heteromorphic sex chromosomes [Charlesworth and Charlesworth, 2000;Kejnovský and Vyskot, 2010;Nicolas, et al., 2005]. In this thesis I built on what is known about sex chromosome evolution in *Silene latifolia* and made use of next generation sequencing methods to perform analyses that contribute to the understanding of sex chromosome architecture and evolution in *S. latifolia*, and of genome evolution in closely related plant species.

With the rise of next generation sequencing (NGS), several sequencing projects on organisms lacking available genomic resources have been developed (e.g. *Melitaea cinxia* [Vera, et al., 2008], *Sarcophaga crassipalpis* [Hahn, et al., 2009], *Laodelphax striatellus* [Zhang, et al., 2010]). Our approach to sequence transcriptomes in several related species has demonstrated that comparative parallel transcriptome sequencing is an efficient approach for developing genomic resources in groups of non-model organisms. Indeed, we assembled about 74'000 genes in the studied species and identified around 1400 genes that are currently known only in the Caryophyllaceae family. Moreover we detected thousands of single nucleotide polymorphisms (SNPs) that provide the necessary molecular resources for linkage mapping and population genetic analyses.

For the first time, we analyzed a large fragment of a plant pseudoautosomal region (PAR) based on genomic sequences. Using bacterial artificial chromosome (BAC) libraries from both *Silene latifolia* and *S. vulgaris* that have been screened for the pseudoautosomal marker OPA [Di Stilio, et al., 1998], we compared over 100 kb of homologous sequences. We found five new *S. latifolia* pseudoautosomal genes and their *S. vulgaris* autosomal homologs. This allowed us to check whether characteristics of PARs seen in the old and highly differentiated mammalian sex chromosome system are also found in evolutionary young plant sex chromosomes. We found that, unlike in mammals, the *S. latifolia* PAR has no increased GC content, which is in line with the idea that in each lineage, PARs follow their own evolutionary dynamics [Otto, et al., 2011]. The comparative analysis of BAC sequences from both *Silene* species, the dioecious *S. latifolia* and the gynodioecious *S. vulgaris*, allowed us to address the question why *S. latifolia* sex chromosomes are enlarged in size. Contrary to expectations we found no evidence for intron size increase, in contrast to

what has previously been found in several sex-linked genes in *S. latifolia* [Marais, et al., 2008]. On the other hand, we observed that insertions of transposable elements occurred in the studied intergenic regions and contribute to the enlargement of the *S. latifolia* sex chromosomes, as previously shown for both X and Y chromosomes [Hobza, et al., 2007;Hobza, et al., 2006;Kejnovský, et al., 2009]. However, our findings indicate that the observed size increase in the *S. latifolia* X chromosome is lower than expected, given the large size of the entire X chromosome, and is closer to the size difference between the autosomes of *S. latifolia* and *S. vulgaris*. Thus, the reasons for the substantial increase in size of the *S. latifolia* X chromosome must be sought in the non-recombining region. The usefulness of BAC libraries for genome sequencing (e.g. hemp [van Bakel, et al., 2011], woodland strawberry [Shulaev, et al., 2011]) as well as for comparative analysis (e.g. bovine with other mammalian pseudoautosomal region comparison [Das, et al., 2009]) has been confirmed with our study.

The subsequent comparative analysis of the sex-specific region of *S. latifolia* sex chromosomes with *S. vulgaris* autosome sequences allowed the localization of 76 sex-linked genes with 18 triplets (*S. latifolia* X, Y and *S. vulgaris* autosomal homologs). Absence of pseudogenes from the Y chromosome BACs investigated suggests that gene loss occurs rarely in the *S. latifolia* Y chromosome, unlike the situation in the human Y chromosome [Lahn and Page, 1999]. From our results we further conclude that *S. latifolia* Y-linked genes are randomly inactivated, as found in the *Drosophila miranda* neo-Y chromosome [Bachtrog, 2006], and confirmed both a higher mutation rate and a relaxed purifying selection of the Y-linked genes, as previously suggested based on a much smaller number of genes investigated [Chibalina and Filatov, 2011;Filatov, 2005;Filatov and Charlesworth, 2002]. Moreover, we found that the recently documented compensation by the X for the reduced expression of alleles on the Y chromosome [Muyle, et al., 2012], occurs randomly among evolutionary strata on the *S. latifolia* sex chromosomes.

This study shows how comparative analyses of closely-related species either carrying or lacking sex chromosomes can contribute to our understanding of early stages of sex chromosome evolution.

References

Bachtrog, D. (2006) Expression Profile of a Degenerating Neo-Y Chromosome in Drosophila. *Current Biology* **16**, 1694-1699.

Bernasconi, G. *et al.* (2009) *Silene* as a model system in ecology and evolution. *Heredity* **103**, 5-14.

Charlesworth, B. & Charlesworth, D. (2000) The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **355**, 1563-1572.

Chibalina, Margarita V. & Filatov, Dmitry A. (2011) Plant Y Chromosome Degeneration Is Retarded by Haploid Purifying Selection. *Current biology : CB* **21**, 1475-1479.

Das, P. J., Chowdhary, B. P. & Raudsepp, T. (2009) Characterization of the Bovine Pseudoautosomal Region and Comparison with Sheep, Goat, and Other Mammalian Pseudoautosomal Regions. *Cytogenetic and Genome Research* **126**, 139-147.

Di Stilio, V. S., Kesseli, R. V. & Mulcahy, D. L. (1998) A pseudoautosomal random amplified polymorphic DNA marker for the sex chromosomes of *Silene dioica*. *Genetics* **149**, 2057-2062.

Filatov, D. A. (2005) Substitution rates in a new Silene latifolia sex-linked gene, *SlssX/Y*. *Molecular Biology and Evolution* **22**, 402-408.

Filatov, D. A. & Charlesworth, D. (2002) Substitution Rates in the X- and Y-Linked Genes of the Plants, *Silene latifolia* and *S. dioica*. *Molecular Biology and Evolution* **19**, 898-907.

Hahn, D., Ragland, G., Shoemaker, D. & Denlinger, D. (2009) Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics* **10**, 234.

Hobza, R., Kejnovsky, E., Vyskot, B. & Widmer, A. (2007) The role of chromosomal rearrangements in the evolution of *Silene latifolia* sex chromosomes. *Molecular Genetics and Genomics* **278**, 633-638.

Hobza, R. *et al.* (2006) An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution. *Chromosoma* **115**, 376-382.

Kejnovský, E., Hobza, R., Čermak, T., Kubat, Z. & Vyskot, B. (2009) The role of repetitive DNA in structure and evolution of sex chromosomes in plants. *Heredity* **102**, 533-541.

Kejnovský, E. & Vyskot, B. (2010) *Silene latifolia*: The Classical Model to Study Heteromorphic Sex Chromosomes. *Cytogenetics and Genome Research* **129**, 250-262.

Lahn, B. T. & Page, D. C. (1999) Four Evolutionary Strata on the Human X Chromosome. *Science* **286**, 964-967.

Marais, G. A. B. *et al.* (2008) Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. *Current Biology* **18**, 545-549.

Muyle, A. *et al.* (2012) Rapid *De Novo* Evolution of X Chromosome Dosage Compensation in *Silene latifolia*, a Plant with Young Sex Chromosomes. *Plos Biology*

Nicolas, M. *et al.* (2005) A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *Plos Biology* **3**, 47-56.

Otto, S. P. *et al.* (2011) About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. *Trends in Genetics* **27**, 358-367.

Shulaev, V. *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics* **43**, 109-116.

van Bakel, H. *et al.* (2011) The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology* **12**, R102.

Vera, J. C. *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**, 1636-1647.

Zhang, F. *et al.* (2010) Massively parallel pyrosequencing-based transcriptome analyses of small brown planthopper (*Laodelphax striatellus*), a vector insect transmitting rice stripe virus (RSV). *BMC Genomics* **11**, 303.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Dr. Alex Widmer for his knowledge and support. I feel very lucky to have had the opportunity to work with him. I also would like to thank Dr. Gabriel Marais and Prof. Dr. Paul Schmid-Hempel for being my co-examiners.

I would like to thank Dr. Roman Hobza for all his help and advices.

All my gratitude to Claudia Michel, Beatrice Blattmann and Tania Torossi for all the work in the lab, and Stephan Zoller for his bioinformatics support.

Many thanks go to Kathrin Rentsch for her help in administrative matters (I would not find a place to sleep in Zurich without her) and her constant smile.

This thesis would have not been possible without my friends and colleagues Nicolas Quèbre, Adrien Favre, Joelle Rahmé, Maria Domenica Del Prete, Massimiliano Gnesotto, Margot Pâris, Aria Minder, Radim Čegan, Bohuslav Janoušek, Delphine Charif and Christine Oger-Desfeux.

A big thank you is addressed to all the past and current members of the institute for the great and stimulating ambiance, and very special thanks go to my office mates, Ana Marcela Florez-Rueda, Sonja Hassold, Daniela Keller Léonie Suter and Olivier Putallaz for their nice company.

Last but not least, I would like to thank my wife Hana and our daughter Valérie, my parents and family in France and my new family in Czech Republic for their love and support.

Curriculum Vitae

Nicolas Blavet
Born in March 6, 1983 in Villefranche-sur-Saône, France
Citizen of France

2007-2012 **Ph. D. thesis** at the Institute of Integrative Biology, Plant Ecological Genetics, Environmental Sciences, ETH Zurich: "*Silene latifolia* (Caryophyllaceae) sex chromosome evolution" under the supervision of Prof. Dr. Alex Widmer (ETH Zurich)

2005-2007 **Master of Science** in Mathematical and Informatics Analysis of Life (aMIV), speciality: Bioinformatics at the Université Claude Bernard Lyon 1 (UCBL), France.

2006-2007 **Trainee** at the Laboratory of Biometrics and Evolutionary Biology (LBBE): "Study of the diazotroph Bacteria and Archaea phylogeny to determine horizontal transfers" under the supervision of Prof. Dr. Guy Perrière.

2005-2006 **Trainee** at the LBBE: "Search for conserved introns in all Metazoans to design primers for polymorphism analysis" under the supervision of Dr. Sylvain Mousset.

2005 **Licence** in Biochemistry (bachelor's degree equivalent) at the UCBL.

2004 **DEUG** in Biochemistry (two-year university diploma) at the UCBL.

2001 **Baccalauréat** S SVT, speciality mathemathics (A-level equivalent), Lycée Claude Bernard, Villefranche-sur-Saône, France

Selected Publications

Blavet N, Blavet H, Čegan R, Zemp N, Zdanska J, Janoušek B, Hobza R, and Widmer A: Comparative analysis of a plant pseudoautosomal region (PAR) in *Silene latifolia* with the corresponding *S. vulgaris* autosome. *BMC Genomics* 2012, **13**: 226.

Blavet N, Charif D, Oger-Desfeux C, Marais G, Widmer A: Comparative high-throughput transcriptome sequencing and development of SiESTa, the *Silene* EST annotation database. *BMC Genomics* 2011, **12**(1):376.

Čegan R, Vyskot B, Kejnovský E, Kubat Z, Blavet H, Šafář J, Doležel J, Blavet N, Hobza R: Genomic Diversity in Two Related Plant Species with and without Sex Chromosomes - *Silene latifolia* and *S. vulgaris*. *PLoS ONE* 2012, **7**(2):e31898.

Chenuil A, Hoareau T, Egea E, Penant G, Rocher C, Aurelle D, Mokhtar-Jamai K, Bishop J, Boissin E, Diaz A, Krakau M, Luttikhuizen P, Patti F, Blavet N, Mousset S: An efficient method to find potentially universal population genetic markers, applied to metazoans. *BMC Evolutionary Biology* 2010, **10**(1):276.

Čegan R, Marais GAB, Kubeková H, Blavet N, Widmer A, Vyskot B, Doležel J, Šafář J, Hobza R: Structure and evolution of *Apetala3*, a sex-linked gene in *Silene latifolia*. *Bmc Plant Biol* 2010, **10**:180.