doi:10.1017/S0016672307009056 Printed in the United Kingdom

A selective sweep in or near the *Silene latifolia* X-linked gene *SlssX*

D. A. FILATOV*
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

(*Received 1 August 2007 and in revised form 15 October 2007*)

Summary

The most prominent feature of Y chromosomes is that they do not recombine and are usually genetically degenerate, containing only a few genes. White campion *Silene latifolia* has evolved sex chromosomes relatively recently, probably within the last 10–15 million years. Perhaps due to its recent origin, the Y chromosome in this species has not completely degenerated and most isolated X-linked genes have intact Y-linked homologues. A gene encoding a protein with strong homology to spermidine synthases, *Slss*, is the exception to this rule, as the Y-linked copy of this gene has apparently lost its function. Here I report evidence for a recent selective sweep in the X-linked copy of this gene (*SlssX*) that could reflect compensatory evolution in an X-linked gene that has lost a functional Y-linked homologue. The spread and fixation of an advantageous mutation in *SlssX* has resulted in a dramatic loss of genetic diversity and an excess of high-frequency derived polymorphisms in this gene. As the sweep has not affected the closely linked *DD44X* gene, the selective advantage of the mutation that has driven the sweep in the *SlssX* gene might have been less than 1%.

1. Introduction

Despite their independent evolution, sex chromosomes in different organismal groups have similar properties (Bull, 1983): recombination is restricted between X and Y chromosomes, and the male-specific non-recombining Y chromosome (or female-specific W chromosome in the case of female heterogamety) exhibits genetic degeneration via the loss of functional genes and the accumulation of repetitive DNA (reviewed in Charlesworth & Charlesworth, 2000). In most mammals and *Drosophila* species Y chromosomes are small and heterochromatic, and contain only a few genes (reviewed in Lahn *et al.*, 2001). Even plant Y chromosomes, which are typically young (only few million years old), show signs of genetic degeneration (Guttman & Charlesworth, 1998; Liu *et al.*, 2004; Filatov, 2005*b*).

Y chromosome degeneration is thought to occur due to 'sheltering' of the Y-linked genes with X-linked homologues and complete linkage of genes on the non-recombining Y chromosome. Lack of recombination exacerbates such population genetic processes as hitch-hiking (fixation of deleterious mutations due to linkage to favourable mutations spreading in the population; Rice, 1987) and background selection (selection against deleterious mutations at linked genes, resulting in reduced effective population size and stochastic accumulation of mildly deleterious mutations; Charlesworth *et al.*, 1995), leading to reduced effective population size of the Y-linked genes and accumulation of deleterious mutations by Muller's ratchet (Charlesworth & Charlesworth, 1997). The loss of genetic diversity in the non-recombining regions is consistent with the results of the studies of DNA polymorphism in various organisms (reviewed in Charlesworth & Charlesworth, 2000), including the white campion (Filatov *et al.*, 2000, 2001; Ironside & Filatov, 2005), which have found low levels of DNA diversity at Y-linked loci compared with X-linked and autosomal genes in most organisms studied.

The X chromosome (or Z in case of female heterogamety) continues to recombine and does not degenerate, which can compensate to some extent for the loss of functional genes from the Y (or W) chromosome. Dosage compensation has independently

* E-mail: dmitry.filatov@plants.ox.ac.uk

evolved in mammals and *Drosophila* to correct for the differences in the amount of product of sex-linked genes in males and females (reviewed by Straub & Becker, 2007). Even the recently evolved neo-sex chromosomes of *Drosophila miranda* show signs of evolving dosage compensation (Martin *et al*., 1996; Bone & Kuroda, 1996). Once the Y-linked copy of a gene is lost, selection may favour an elevated expression level of the homologous X-linked genes in males, leading to partial dosage compensation. Thus, one may expect that advantageous mutations frequently spread in X-linked genes (a process termed 'selective sweep' by Maynard Smith & Haigh, 1974) once they are under additional selective pressure to compensate for degenerate Y-linked homologues. However, no cases of recent selective sweeps, or other evidence of positive selection in X-linked genes associated with the degeneration of Y-linked homologues, have been reported previously.

The plant genus *Silene* has evolved sex chromosomes in a small cluster of dioecious *Silene* species (section *Elisanthe*: *S. latifolia*, *S. dioica*, *S. diclinis*, *S. heuffelii* and *S. marizii*), while the rest of the genus is non-dioecious (exept *S. otites*, which apparently evolved dioecy independently from *Elisanthe*) and lacks sex chromosomes. Silent divergence between dioecious *S. latifolia* and hermaphroditic *S. conica* is approximately 15%, suggesting that *S. latifolia* sex chromosomes are probably not older than 15 million years (Filatov & Charlesworth, 2002) and most likely evolved from a pair of autosomes (Filatov, 2005*a*). As YY plants (having no X) are usually inviable (Ye *et al*., 1990), the *S. latifolia* Y chromosome has probably degenerated to some extent. However, most genes isolated from *S. latifolia* X chromosome have intact Y-linked homologues, suggesting that genetic degeneration has not proceeded too far and most Y-linked genes are still intact. The first gene isolated from the *S. latifolia* X chromosome, *MROS3X*, was reported to have a dysfunctional Y-linked copy (Guttman & Charlesworth, 1998). However, more recently it was demonstrated that *MROS3X*/*Y* genes are members of a multicopy gene family with several copies on the autosomes and sex chromosomes (Kejnovsky *et al*., 2001). Thus, the Y-linked degenerate copy of *MROS3* is not necessarily the remnant of the Y-linked homologue of *MROS3X*, but it may be a paralogue translocated from the autosomes. High *MROS3X*/*Y* divergence (>30%) also supports this view. All the X-linked genes isolated more recently do have intact Y-linked homologues (Delichère *et al*., 1999; Atanassov *et al*., 2001; Moore *et al*., 2003; Filatov, 2005*b*; Bergero *et al*., 2007). Furthermore, all these Y-linked genes except *SlssY* show no signs of genetic degeneration.

The pair of sex-linked genes that do show signs of degeneration in the Y-linked copy, the *SlssX* and *SlssY* genes, encode proteins with a strong similarity to spermidine synthases and silent divergence between these genes is 8% (Filatov, 2005*b*). Although the Y-linked copy of this gene is transcribed, it contains several mutations that change conserved amino acid positions known to be important for spermidine synthase activity and are expected to affect functionality of the protein (Filatov, 2005*b*). The X-linked copy of this gene is intact and the loss of function of the Y-linked copy might have imposed an additional burden on the X-linked copy to produce enough protein in males to 'shelter' the loss of function of the Y-linked copy.

The previous work focused on genetic mapping of the X-linked genes (Filatov, 2005*a*) revealed unusually low diversity in the *SlssX* gene, suggesting a recent selective sweep in this gene. This study aims to test the hypothesis of the sweep in the *SlssX* gene comparing patterns of DNA polymorphism in the *SlssX* gene and other X-linked genes: *SlX1*, *SlX4* and *DD44X*. The first two genes are located relatively far away from the *SlssX* (7·4 and 17·9 cM, respectively), while *DD44X* is very closely (<1 cM) linked to *SlssX* (Filatov, 2005*a*). As previous analyses of polymorphism in *SlX1*, *SlX4* and *DD44X* genes revealed no indication that these genes may have been affected by a recent selective sweep (Filatov *et al*., 2000; Laporte *et al*., 2005; Ironside & Filatov, 2005), they can be used as useful reference loci to be compared with *SlssX*.

2. Materials and methods

To compare amounts and patterns of DNA polymorphism in *S. latifolia* X-linked genes, I used a set of *S. latifolia* individuals collected from across Eurasia (Table 1). Sampling was designed to cover the entire range of the species. A small quantity of leaf material was taken from each plant and frozen. DNA was extracted using the magnetic beads-based protocol of the ChargeSwitch Plant DNA kit (Invitrogen). Fragments of *DD44X* and *SlssX* genes were amplified from the genomic DNA of male plants (sequences for all PCR and sequencing primers are provided in Table 2). The *DD44X* was amplified using the primers DD44X+1 and DD44XY2.1R. The *SlssX* gene was amplified with the primers Slss+6X and Slss-18X. Other primers listed in Table 2 were used for sequencing of internal regions of the PCR products amplified from *DD44X* and *SlssX* genes. Amplifications were performed under the following PCR conditions: initial denaturation at 94 °C for 2 min followed by 38 cycles of 93 °C for 20 s, 53 °C for 30 s, and 68 °C for 4 min. Expand high-fidelity DNA polymerase (Roche) was used for both *DD44X* and *SlssX* genes. Because males possess only a single X chromosome, *DD44X* and *SlssX* PCR products amplified from male genomic DNA were hemizygous and were sequenced

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

Table 1. *The list of* Silene *samples used in this study*

Sample	Country	GPS coordinates		*SlssX*	*DD44X*
Silene vulgaris	UK	na	na	+	+
Sa_IL9A_Czech	Czech Republic	na	na	+	+
Sa_IL78D_UK	UK	N52 16.207′	E0 12.135′	+	+
Sa_IL72A_Austria	Austria	N48 18.638′	E16 24.225′	+	−
Sa_IL58C_Rus	Russia	N56 11.480′	E60 28.330′	+	−
Sa_IL57B_Rus	Russia	N56 35.774′	E61 2.149′	+	−
Sa_IL54E_Rus	Russia	N53 8.537′	E92 53.530′	+	−
Sa_IL53A_Rus	Russia	N53 8.537′	E92 53.530′	+	−
Sa_IL3B_Belgium	Belgium	N50 41.648′	E4 46.681′	+	+
Sa_IL36F_UK	UK	N52 24.888′	W2 7.217′	+	−
Sa_IL35A_UK	UK	N52 25.129′	W2 6.354′	+	−
Sa_IL34E_UK	UK	N52 24.888′	W2 7.217′	+	−
Sa_IL33E_UK	UK	N52 24.297′	W2 9.134′	+	−
Sa_IL32G_UK	UK	N52 23.753′	W2 8.285′	+	+
Sa_IL30D_UK	UK	N40.61587	W7.74018	+	+
Sa_IL29C_France	France	N47 45.723′	E3 53.496′	+	−
Sa_IL28C_France	France	N47 39.869′	E4 13.968′	+	−
Sa_IL25H_UK	UK	N52 23.77	W2 13.06	+	+
Sa_IL24H_France	France	N47 42.433′	E4 14.020′	+	+
Sa_IL21F_Belgium	Belgium	N50 41.648′	E4 46.681′	+	−
Sa_IL1B_UK	UK	N51 46.293′	W0 11.071′	+	−
Sa_IL18F_France	France	N48 43.319′	E3 1.906′	+	−
Sa_IL12A_Czech	Czech Republic	na	na	+	+
Sa_IL122C_Portugal	Portugal	na	na	+	+
Sa_IL11G_Spain	Spain	na	na	+	+
Sa_IL10G_UK	UK	N52 27.028′	W1 56.091′	+	+
Sa_708_Rus	Russia	na	na	+	+
Sa_706_Rus	Russia	na	na	+	+
Sa_703_Rus	Russia	na	na	+	+
Sa_699_Rus	Russia	na	na	+	+
SaWF	Romania	na	na	+	+
SaVM1	Romania	na	na	+	+
SaVL1	Romania	na	na	+	+
SaVI	Romania	na	na	−	+
SaSM2	Spain	na	na	−	+
SaSM1	Spain	na	na	−	+
SaMa1	UK	N52 5.21′	W2 20.04′	+	+
SaMB3	Romania	na	na	−	+
SaMB2	Romania	na	na	+	+
SaLo1	UK	na	na	+	+
SaIE	Romania	na	na	+	+
SaGh1	Romania	na	na	−	+
SaCz2	Czech Republic	na	na	−	+
SaCB	Romania	na	na	−	+
SaBR3	Spain	na	na	+	+
SaBR1	Spain	na	na	−	+
SaBC3	Spain	na	na	−	+
SaBC1	Spain	na	na	−	+
Sa702	Russia	N57 50.227	E34 50.645	+	+
Sa700	Russia	N57 43.412	E34 46.525	+	+
Sa582	Poland	na	na	−	+
Sa538F	Austria	N47 34.058	E15 18.910	−	+
Sa534f	Austria	N47 33.967	E15 18.898	−	+
Sa527	Austria	N48 15.948	E15 21.508	−	+
Sa516	Austria	na	na	−	+
Sa267	UK	N56 3.83′	W2 46.84′	+	+
Sa205	UK	N55 53.52′	W3 3.53′	−	+
Sa357	UK	N52 10′ 58″	E0 11′ 9″	−	+
Sa763	UK	na	na	+	+
IL4C	Romania	na	na	−	+
IL17H	France	N48 38.936′	E3 56.285′	+	+
IL37	UK	N52 9.695′	E0 6.331′	+	+

Plus and minus signs indicate whether the *SlssX* and *DD44X* genes were sequenced for a particular sample.

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

Table 2. *PCR and sequencing primers*

Name	Sequence
Slss+6X	AGTGTTGTAGGCTATAATTTGGTACAC
Slss+9	GTAATCATTTTGCCATCATCTCTT
Slss+15	GGAGAAGCACATTCCCTGAAAG
Slss-13	CTCTTGGTATGCACACTCATCC
Slss-12	AAAGTGTTGGATAGAGATTCCATAT
Slss-5	CTAGGGTATGTTGGAACTGTAGTCC
Slss+1	GTCCGTTGCAAAGGCTCTTC
Slss-18X	AGCCGCTGAATGGATCTGCA
DD44X+1	ATGTCAATGGCGAACCGCAT
DD44X251F	GTGGTTTGGGAACTCGTAGG
DD44X3F2	TTTCTAAACATGTGGAGCTCAGG
DD44X3F11	TGTCATGCATAGGTGTTCATCATAG
DD44XY+4	TCATTGGTATTAGGTGCCTGTGG
DD44xhr1	TAGCAGGTTCAGATCGACCC
DD44XY3.2F	CTTTGCTACCAAGGCTCCTG
DD44XYr2.1	CTCCATCTGTCTTGCCCTGG

directly using a BigDye v3.1 sequencing kit and a 3700 automated sequencer (Applied Biosystems). All polymorphic sites were double-checked manually. Chromatograms were checked and corrected, contigs assembled and alignments constructed using ProSeq3 software (Filatov, 2002). The sequences have been submitted to GenBank under accession numbers EU265839–EU265926.

The program ProSeq3 was used to calculate two measures of nucleotide diversity (π (Nei, 1987) and Watterson's (1975) θ_w) for *DD44X* and *SlssX*. The sequence diversity of *SlssX* was compared with that of *DD44X* and two other X-linked genes using the HKA test (Hudson *et al*., 1987), implemented in DnaSP (Rozas *et al*., 2003). A single *DD44X* and *SlssX* sequence from the hermaphroditic species *S. vulgaris* was used as an outgroup. Bias of the site-frequency distribution in the *SlssX* and *DD44X* sequences was tested using Tajima's test (Tajima, 1989) as implemented in ProSeq3. The excess of high-frequency derived polymorphisms was estimated using Fay & Wu's (2000) *H* statistic implemented in DnaSP. The deviation of the number of distinct haplotypes in the sample from neutral expectation was tested using Fu's (1997) F_s statistic implemented in DnaSP. The significance of all the neutrality tests was calculated using coalescent simulations without recombination, which is a conservative approach. The coalescent simulations were conducted using DnaSP and ms (Hudson, 2002) programs. The population-scaled recombination rate, ρ, was estimated using the composite likelihood method of McVean *et al.* (2002). The minimum number of recombination events, R_m (Hudson & Kaplan, 1985) was conducted using DnaSP.

Estimation of the time since the selective sweep was estimated using the approach proposed by Przeworski (2003). A program *msHH* implementing this approach was kindly provided by M. Przeworski. This program was used to generate samples from the posterior distribution of the time since the selective sweep, as well as the selective coefficient for the mutation that caused the sweep. The location of the target of positive selection, as well as the selective coefficient of the adaptive mutation, was estimated using the composite likelihood method of Kim & Stephan (2002). The program *clsw* (Kim, 2005) was used to find the maximum composite likelihood estimates of the population-scaled selective coefficient and the position of the beneficial mutation as well as the likelihood ratio of the sweep versus neutrality models.

3. Results

(i) *DNA polymorphism*

Sequencing of a 3·5 kb long fragment of the *SlssX* gene from 42 individuals revealed only 39 nucleotide polymorphisms (Table 3), which corresponds to a per-nucleotide population-scaled mutation rate of $\theta_w = 0{\cdot}0029$ (all sites) when measured using Watterson's (1975) estimator. This level of polymorphism is surprisingly low compared with the previously studied X-linked genes *SlX1*, *SlX4* and *DD44X* ($\theta_w = 0{\cdot}02$, 0·024 and 0·026, respectively), and is similar to the level of polymorphism observed in *S. latifolia* Y-linked genes, which is significantly reduced compared with polymorphism in X-linked genes (Filatov *et al*., 2000, 2001; Laporte *et al*., 2005; Ironside & Filatov, 2005).

Estimates of DNA polymorphism in a gene may be affected by sampling, local mutation rate, demography (of the populations), and non-neutral evolution of the gene or in the flanking regions nearby. The sampling used in the current study to measure DNA polymorphism in the *SlssX* gene differed from the

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

Table 3. *DNA polymorphism in four* S. latifolia *X-linked genes*

	Total								Silent				Non-coding			Synonymous			Non-synonymous		
	N^a	L^b	S^c	π^d	Tajima's D	Fs^e	Rm^f	ρ^g	L^b	S^c	π^d	$K(JC)^h$	L^b	S^c	π^d	L^b	S^c	π^d	L^b	S^c	π^d
SlX1	26	1607	106	0·015	−0·5	−4·065	17	0·0123	1228	99	0·019	0·09	1119	88	0·0186	108·6	11	0·021	383·4	8	0·0024
SlX4	39	754	56	0·0144	−0·36	−1·73	6	0·0001	224	41	0·015	0·16	58	8	0·019	166	33	0·044	530	15	0·0045
DD44X	46	1104	156	0·028	−0·32	−4·2	23	0·012	1104	156	0·028	0·06	1104	156	0·028	0	–	–	0	–	–
SlssX	42	3069	39	0·0013	−2·07	−9·86**	8	0·0004	2469	37	0·001	0·104	2295	32	0·0014	174·3	3	0·0007	599·7	2	0·0003

[a] Number of individuals analysed.
[b] Number of alignment positions analysed.
[c] Observed number of polymorphic sites.
[d] Average heterozygosity per nucleotide (Nei, 1987).
[e] Fu's (1997) *Fs* statistic.
[f] Minimal number of recombination events (Hudson & Kaplan, 1985).
[g] Population-scaled recombination rate (Hudson, 2001).
[h] Jukes–Cantor (Jukes & Cantor, 1969) divergence from outgroup *S. vulgaris*.

Table 4. *Pairwise HKA test results for four X-linked genes from* S. latifolia. S. vulgaris *was used as an outgroup for all pairwise HKA tests*

S. latifolia genes	*SlX1*[a]	*SlX4*[b]	*DD44X*[c]
SlX1[a]	–		
SlX4[b]	$P=0{\cdot}28$	–	
DD44X[c]	$P=0{\cdot}39$	$P=0{\cdot}07$	–
SlssX[c]	$\mathbf{P<0{\cdot}0001}$	$\mathbf{P<0{\cdot}0001}$	$\mathbf{P<0{\cdot}0001}$

[a] Data from Filatov *et al.* (2001).
[b] Data from Laporte *et al.* (2005).
[c] Data from this study.

previous studies (e.g. see Table 1 in this study and in Ironside & Filatov, 2005). Thus comparisons of *SlssX* data with other genes have to be treated with some caution. To exclude the effects of sampling differences on estimates of DNA polymorphism, I sequenced a 3·5 kb long region from the *DD44X* gene using the same samples as for the *SlssX* gene. Regions containing polymorphic insertions/deletions (indels) were excluded from the analyses, which reduced the number of alignment positions that can be analysed for DNA polymorphism to 1104 (Table 3). The observed per-nucleotide average heterozygosity in the new *DD44X* dataset, $\pi=0{\cdot}028$ was similar to that observed previously ($\pi=0{\cdot}023$; Ironside & Filatov, 2005). Thus, the sampling difference between this and the previous studies is unlikely to be the cause of the unusually low DNA diversity in the *SlssX* gene.

Any difference in mutation rates between the *SlssX* and *DD44X* genes may also be a cause of variation in the level of polymorphism among the genes. If the *SlssX* gene is a mutational cold-spot, then one would expect to observe less polymorphism in that gene. Differences in mutation rate between different genomic regions can be corrected for by comparing homologous regions with an outgroup (in this case the non-dioecious species *S. vulgaris*). Silent mutations should be predominantly neutral, so the rate of divergence between species will be proportional to the mutation rate. When the diversity differences between *SlssX* and *DD44X* were corrected using the divergence from homologues in *S. vulgaris*, the difference remained significant (as did comparisons for three other X-linked genes; Table 4). The comparison of DNA diversity in the *SlssX* and *DD44X* genes with two other previously studied *S.latifolia* X-linked genes demonstrates that *DD44X* contains similar amounts of polymorphism as the *SlX1* and *SlX4* genes, while there is a significant lack of diversity in the *SlssX* gene, compared with the three other X-linked genes (Table 4).

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

(ii) *Recombination rates*

Although the physical distance between the *SlssX* and *DD44X* genes is not known, it has been demonstrated that these genes are tightly linked on the X chromosome – no recombinants between these genes have been observed previously (Filatov, 2005*a*). However, from the linkage disequilibrium pattern, the region containing *SlssX* and *DD44X* genes is not a recombinational cold-spot. Both genes contain pairs of polymorphisms with all four combinations of alleles, which could have ocurred either due to multiple mutations at the same site, or due to recombination. The minimal number of recombination events, estimated using the four-gamete test of Hudson & Kaplan (1985) is 8 and 23 in the *SlssX* and *DD44X* genes, respectively. In accordance with the view that the *DD44X*–*SlssX* region is not a recombinational cold-spot, there is no significant linkage disequilibrium between the two genes despite their tight genetic linkage.

Although it is clear that recombination does occur in the *DD44X*–*SlssX* region, the population-scaled recombination rate per nucleotide, ρ $(=3N_e r$ for X-linked genes, where r is the recombination rate and N_e is the effective population size), estimated using the composite likelihood approach (Hudson, 2001; McVean *et al*., 2002), is more than an order of magnitude lower for *SlssX* compared with *DD44X* (Table 3). The population recombination rate, however, depends on the effective population size (which in turn is scaled to polymorphism). Hence, the lower ρ in *SlssX* may be due to a lower effective population size in the region and not a lower recombination rate. Normalization of ρ by the population-scaled mutation rate, θ $(=3N_e\mu$, where μ is the mutation rate), provides a way to correct for differences in the amounts of polymorphism in different genes. Using this correction, the difference between the *SlssX* and *DD44X* genes remains (0·092 and 0·746, respectively). Thus, although the four-gamete test (Hudson & Kaplan, 1985) and the composite-likelihood estimator of ρ detect recombination in the *SlssX* gene, recombination rate at this locus is probably lower than in *DD44X*. Nevertheless, if genetic diversity in the *SlssX* gene were affected by a recent selective sweep (see below), then linkage disequilibrium may not reflect the true recombination rate in this gene, as selective sweeps can generate fairly strong linkage disequilibrium (Stephan *et al*., 2006; McVean, 2007), resulting in an underestimation of the local recombination rate.

(iii) *Frequency-spectrum-based neutrality tests*

The significantly reduced DNA diversity in the *SlssX* gene, compared with the other three X-linked genes, suggests that diversity may have been reduced by a selective sweep in or near the *SlssX* gene. A selective sweep is expected to bias the frequency spectrum of mutations in regions adjacent to the selected mutation (Braverman *et al*., 1995), as the spread of the adaptive allele should wipe out genetic diversity and all new mutations that arise after the sweep will be at low frequency. Indeed, there is a significant excess of mutations present only once in the sample (singletons) in *SlssX*, but not in *DD44X*. Tajima's *D* (Tajima, 1989), a neutrality test devised to detect this kind of bias in the frequency spectrum, is significant for *SlssX*, but not for *DD44X* (Table 3). Although the bias in the frequency spectrum can be caused by demographic factors, such as population expansion, demography cannot specifically affect *SlssX* alone.

The spread of an advantageous allele in a population is expected to drag some closely linked polymorphisms to high frequency, which is a characteristic footprint of positive selection (Fay & Wu, 2000; Przeworski, 2002). Using the *S. vulgaris Slss* gene sequence as an outgroup to establish the ancestral state of the polymorphic sites, nine of 42 polymorphisms in the *S. latifolia SlssX* dataset are high-frequency derived polymorphisms. Fay & Wu's (2000) *H* statistic demonstrates that there is a significant excess of such polymorphisms in the *SlssX* ($H=-10{\cdot}91$, $P<0{\cdot}01$), but not in *DD44X* ($H=1{\cdot}33$, NS).

(iv) *Timing of the selective sweep in* SlssX *gene*

The significantly reduced DNA diversity in the *SlssX* gene compared with other X-linked genes in *S. latifolia* and the excess of high- and low-frequency polymorphisms in the frequency spectrum of the *SlssX* gene strongly suggest a recent selective sweep in or near the *SlssX* gene. The recovery of genetic diversity after a selective sweep takes about $4N_e$ generations. The bias in frequency spectrum by Tajimas' *D* statistic is almost undetectable after approximately $2N_e$ generations, while the excess of high-frequency polymorphisms detectable by Fay & Wu's *H* statistic disappears after $0{\cdot}5$–$1N_e$ generations (Przeworski, 2002). As the signature of the sweep is still detectable by both *D* and *H* statistics the sweep might have occurred less than $1N_e$ generations ago.

To estimate the time of the sweep more precisely, I used the approach proposed by Przeworski (2003). This method summarizes polymorphism data using three summary statistics: the number of polymorphic sites, the value of Tajimas' *D* and the number of observed haplotypes. Given the observed values of these statistics, the method produces a sample from the posterior distribution of the time since the selective sweep, as well as the selective coefficient for the mutation that caused the sweep (*s*). The sample of size 2000 from the posterior distribution of the number of

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

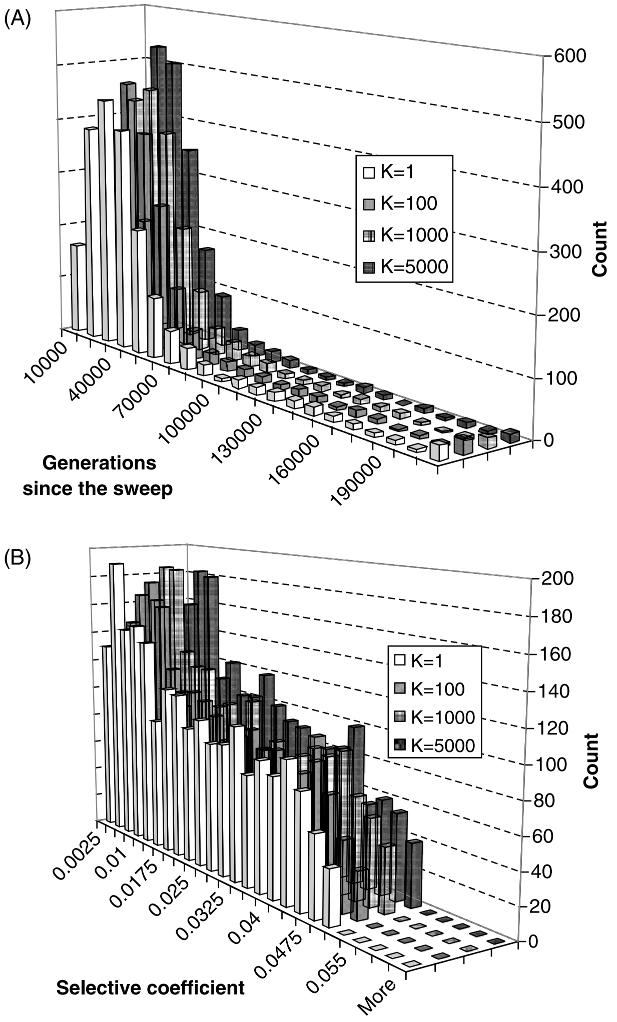

Fig. 1. Estimation of time since the selective sweep and the strength of selection using the method of Przeworski (2003). Shown are 2000 samples from the posterior distributions of the time since the selective sweep (*A*) and the selective coefficient for the mutation that caused the sweep (*B*). The four series correspond to different positions of the selected mutation (K) in the *SlssX* at the beginning of the gene, 100 bp, 1000 bp and 5000 bp upstream of the gene.

generations since the selective sweep (T_{gen}) forms a relatively narrow peak centred at 30 000 generations (Fig. 1*A*). On the other hand, the sample from the posterior distribution of the selective coefficient forms a fairly wide peak between 0 and 0·05, with the maximum at 0·005 (Fig. 1*B*), demonstrating that the selective sweep might have been caused by a beneficial mutation with advantage of less than 5%; however, there is not enough power to obtain a more precise estimate of the strength of selection using this method. These estimates are robust to the choice of the prior values for population size (data not shown), and to the choice of the position of the adaptive mutation in the sequence (*K*; see Fig. 1).

To compare the posterior distributions of T_{gen} for different *S. latifolia* X-linked genes, I applied the same analysis to the *DD44X*, *SlX1* and *SlX4* datasets. The distributions of T_{gen} for these genes are fairly wide and are centred at around 650 000 generations, while the distribution for the *SlssX* gene is in sharp contrast with the three other X-linked genes (Fig. 2). The wide peaks of the *DD44X*, *SlX1* and *SlX4* genes may reflect an ancient demographic event, such as population expansion (e.g. Taylor & Keller, 2007), as they give very similar estimates of T_{gen} for the three genes, while the narrow peak in the *SlssX* gene corresponds to a gene-specific event, providing further support for the recent selective sweep in *SlssX*.

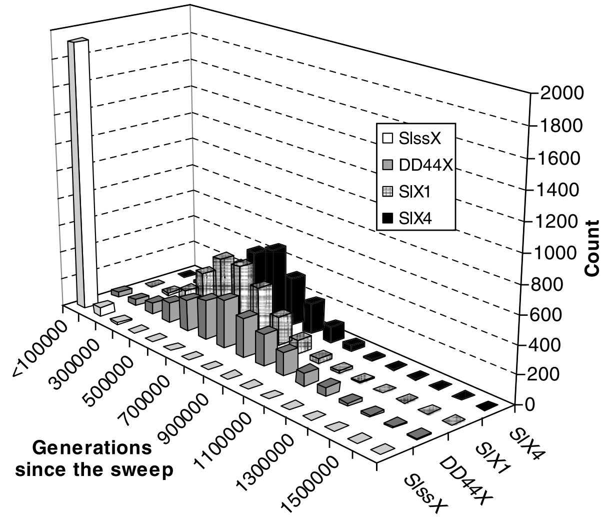

Fig. 2. Samples from posterior distributions of the time since the selective sweep or population expansion for *SlssX*, *DD44X*, *SlX1* and *SlX4 S. latifolia* datasets.

(v) *Selective coefficient and location of the target of the sweep*

A selective sweep is expected to influence a neutral linked locus if recombination between the selected and neutral loci is less than half the selection coefficient, *s*/2 (Maynard Smith & Haigh, 1974; Stephan *et al.*, 1992). As the *DD44X* and *SlssX* genes are very closely linked (<1 cM; Filatov, 2005*a*), and the sweep in or near the *SlssX* gene apparently does not affect the *DD44X* gene, the selective coefficient of the positively selected mutation might have been relatively low (probably less than 1–2%). To estimate the strength of selection in or near the *SlssX* more precisely and to predict the location of the target of the sweep, I used the composite likelihood method of Kim & Stephan (2002). This method compares the observed patterns of nucleotide variation under neutrality and the selective sweep models and allows for a composite likelihood ratio test (LR) of the selective sweep scenario. It also estimates the most likely position of the target of positive selection as well as the population-scaled selective coefficient

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

Table 5. *Composite likelihood ratio analysis of positive selection in* SlssX *gene using the method of Kim & Stephan (2002)*

SlssX	Unfolded frequency spectrum				Folded frequency spectrum			
ρ[a]	LR1[b]	P(LR1)[c]	$1{\cdot}5N_es$[d]	Position[e]	LR2[b]	P(LR2)[c]	$1{\cdot}5N_es$[d]	Position[e]
0·0001	−89·08	0·03	6·23	2760	−59·16	0·015	6·59	3466
0·0006	−10·99	0·03	5·26	2826	3·305	0·009	5·67	2665
0·006	4·98	0·05	25·18	2445	10·249	<0·001	46·06	2455
0·06	4·99	0·05	437·6	2439	10·267	<0·001	793·5	2454

[a] Population-scaled recombination rate (Hudson, 2001).
[b] Likelihood ratio test comparing nested models with and without positive selection.
[c] Probability of rejecting neutrality in the likelihood ratio test.
[d] Estimated population-scaled selective coefficient.
[e] Inferred position of the target of positive selection.

($1{\cdot}5N_es$ for an X-linked gene and $2N_es$ for an autosomal gene).

The results of this analysis for the *SlssX* gene are presented in Table 5. As was noted above, the estimate of recombination rate in this gene could have been biased by the selective sweep, thus the analysis was conducted with a range of different values of the per nucleotide population-scaled recombination rates ($\rho = 3N_er$). Regardless of the recombination rate used or whether data with folded or unfolded frequency spectra (i.e. not distinguishing or distinguishing ancestral and derived alleles, respectively) were employed, the composite likelihood ratio tests reject a neutral model in favour of the selective sweep model. The estimate of the strength of selection that might have caused the sweep depends on the amount of recombination assumed in this analysis. With the 'observed' value of $\rho = 0{\cdot}0006$ (estimated for the *SlssX* gene using the composite likelihood method of McVean *et al.*, 2002), the population-scaled selective coefficient ($1{\cdot}5N_es$) is about 5. Assuming the effective population size of *S. latifolia* is $\sim 10^6$ (e.g. Taylor & Keller, 2007), the selective coefficient of the mutation that has caused the selective sweep in the *SlssX* gene is of the order of 3×10^{-6}, which is much lower than the estimate of the selective coefficient obtained by the method of Przeworski (2003), which yielded a posterior distribution with the maximum at $s = 0{\cdot}005$. Higher recombination rates assumed for *SlssX* give higher estimates of $1{\cdot}5N_es$ (Table 5). Given the difficulty of obtaining a reliable estimate of recombination rate in the *SlssX* gene and the width of the peak for s (Fig. 1*B*) obtained by the method of Przeworski (2003), it is not possible to conclude that the two methods produce estimates that are significantly different from each other. Both methods agree that the selective advantage of the mutation that might have caused the selective sweep in the *SlssX* gene is lower than 1% (and probably much lower).

The inferred position of the target of positive selection varies only slightly with the different frequencies of recombination assumed (Table 5). In most cases the target of selection is located between nucleotide position 2400 and 2850 of the dataset, with one case at position 3466 when a very low recombination rate was assumed. The composite likelihood profile for different candidate positions of the target of positive selection is shown in Fig. 3. This profile was calculated for unfolded frequency spectrum data assuming the 'observed' $\rho = 0{\cdot}0006$. Profiles for other recombination rates and for the folded frequency spectra are similar (data not shown). With this profile the highest composite likelihood values are obtained for a selection target around nucleotide position 2700 of the dataset, which corresponds to exon 7 of the *SlssX* gene (exon numbering according to Filatov, 2005*b*). However, inspection of polymorphisms in exon 7 revealed only a silent singleton at position 2767 of the dataset and no fixed differences between *S. latifolia* and *S. vulgaris*. Thus, although there is only one significant ($P < 0{\cdot}05$) major peak of the likelihood ratio identified by the Kim & Stephan (2002) approach, it does not correspond to any apparent amino acid replacement that could have been a target of positive selection. Although the second highest peak at around nucleotide position 3500 is not significant, it may suggest that the target of selection lies downstream of the studied region of the *SlssX* gene. Estimates of the selective coefficient ($1{\cdot}5N_es$) that might have caused the selective sweep in or near the *SlssX* gene depend on the recombination rate assumed for the *SlssX* gene and vary between 794 and 5 (Table 5).

4. Discussion

In this paper I argue that *Silene latifolia* X-linked gene *SlssX* has undergone a selective sweep in the recent past. Although the causes of this sweep remain unknown, the loss of function of the Y-linked

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

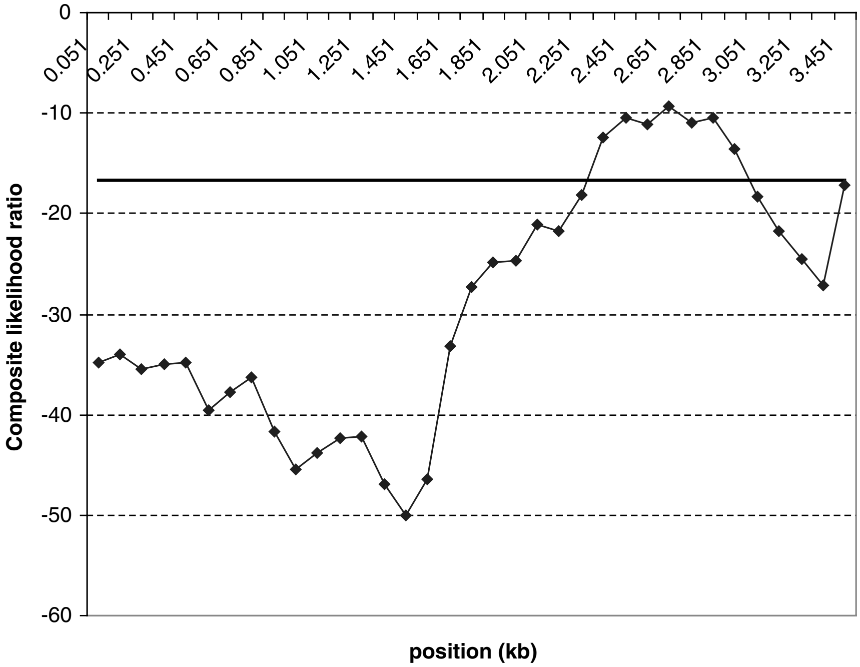

Fig. 3. Composite likelihood ratio profile for *S. latifolia SlssX* gene calculated by the method of Kim & Stephan (2002). Unbroken horizontal line shows the 95% percentile for the distribution of composite likelihood ratio under the neutral model.

homologue, *SlssY* (Filatov, 2005*b*), suggests that the X-linked copy may be under selection to compensate for a dysfunctional copy on the Y chromosome. As the Y-linked copy is still transcribed and does not contain any stop codons and frame shifts, dysfunctionalization of the *SlssY* protein might have occurred very recently. This might have imposed an additional burden on the X-linked *SlssX* copy to produce enough protein in males to 'shelter' the loss of function of the Y-linked copy. Although this is by no means direct evidence for selection for dosage compensation, the finding of a selective sweep in an X-linked gene with a degenerate Y-linked copy may be suggestive of selection to compensate for the loss of function of the Y-linked gene.

The spread of an adaptive allele in a population is expected to result in a loss of genetic variation at linked sites around the selected locus, as the fixation of the advantageous allele will result in the fixation of neutral linked variants (Maynard Smith & Haigh, 1974). In the presence of recombination, not all sites on the chromosome are completely linked. These sites may recombine in or out of the background of the spreading advantageous allele. As a result, sites that are further away (in terms of recombination distance) from the selected site will be less affected by the selective sweep. On average, this action results in the formation of a 'V-shaped valley' of polymorphism centred on the selected locus (e.g. Kim & Stephan, 2000). The width of such a valley depends on how quickly the sweep occurs (i.e. on the magnitude of the selective coefficient), as well as on the recombination rate in the region. The new mutations ocurring in the region of reduced polymorphism after the selective sweep are initially present in the population at low frequency, hence the biased frequency spectrum is another characteristic 'signature' of the selective sweep. In this paper I reported reduced DNA diversity and biased frequency spectrum in the *S. latifolia* X-linked gene *SlssX*, which contrasts with other X-linked genes, including the closely linked *DD44X*. The observed pattern of polymorphism is strongly suggestive of a relatively recent selective sweep in or near the *SlssX* gene. Other explanations, such as population expansion that can also cause bias in the frequency spectrum, seem unlikely given that the low polymorphism and excess of low-frequency polymorphisms are specific for the *SlssX* gene. Moreover, the significant excess of derived high-frequency polymorphisms in *SlssX* detected by Fay & Wu's (2000) *H* statistic is fairly specific to selective sweeps and not expected under a population expansion scenario. Sometimes population structure may create a significant Fay & Wu's *H* (Przeworski, 2002); however, multiple genes are expected to be affected, while *H* is significant only for *SlssX*.

As the most likely region of the target of selection identified by the method of Kim & Stephan (2002) does not contain any amino acid replacements that could be obvious targets of positive selection, the selective sweep could have been driven by a mutation beyond the studied region. It could be a mutation in

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

an unknown neighbour gene that is closer to *SlssX* than *DD44X*. However, given fairly low gene density in plant genomes, a more likely possibility is a mutaion in *SlssX* or in regulatory sequences in introns or untranslated regions that may affect the level of *SlssX* expression. For example, local chromatin structure may significantly affect gene expression and dosage compensation in mammals and *Drosophila* is known to be acompanied by changes in chromatin structure (Straub & Becker, 2007). The relationships between the specific sequence motifs and chromatin structure are not well understood, so it is difficult to infer what changes in or near the *SlssX* gene could have been advantageous, if this involved changes in local chromatin structure.

Interestingly, polymorphism in the *DD44X* gene closely linked to the *SlssX* apparently is not affected by the sweep in or near *SlssX*. A selective sweep is expected to influence a neutral linked locus if recombination between the selected and neutral loci is less than half the selection coefficient, $s/2$ (Maynard Smith & Haigh, 1974; Stephan *et al*., 1992). No recombinants were observed in genetic crosses between these two genes and recombinational distance between *SlssX* and *DD44X* should be less than 1 cM (Filatov, 2005*a*). Thus, any sweep in *SlssX* that is driven by a mutation with selective advantage of 1–2% or more should also affect *DD44X*. In fact, the selective advantage of the mutation in or near the *SlssX* gene might have been lower than 1%, as the methods of Kim & Stephan (2002) and Przeworski (2003) provide estimates of about 0·0003% and 0·5%, respectively.

The discovery of a selective sweep in the *SlssX* may be a mere coincidence with the fact that this gene may be evolving dosage compensation; nevertheless, it provides a strong incentive to conduct a further study of relative expression levels of the *SlssX* and the spermidine synthase activity in males and females. If expression studies confirm that this gene is evolving dosage compensation, then the evidence for a recent selective sweep in the *SlssX* will be the first report of evolving dosage compensation 'caught in action'.

I thank Graham Muir for correction of the manuscript. Barbara Mable and two anonymous reviewers provided useful suggestions that helped to improve the manuscript. This work was supported by a grant to D.F. from NERC.

References

Atanassov, I., Delichère, C., Filatov, D. A., Charlesworth, D., Negrutiu, I. & Monéger, F. (2001). Analysis and evolution of two functional Y-linked loci in a plant sex chromosome system. *Molecular Biology and Evolution* **18**, 2162–2168.

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

Bone, J. R. & Kuroda, M. I. (1996). Dosage compensation regulatory proteins and the evolution of sex chromosomes in *Drosophila*. *Genetics* **144**, 705–713.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796.

Bull, J. J. (1983). *Evolution of Sex Determining Mechanisms*. Menlo Park, CA: Benjamin/Cummings.

Charlesworth, B. & Charlesworth, D. (1997). Rapid fixation of deleterious alleles can be caused by Muller's ratchet. *Genetical Research* **70**, 63–73.

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

Charlesworth, D., Charlesworth, B. & Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**, 1619–1632.

Delichère, C., Veuskens, J., Hernould, M., Baarbacar, N., Mouras, A., Negrutiu, I. & Moneger, F. (1999). *SlY1*, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein. *EMBO Journal* **18**, 4169–4179.

Fay, J. C. & Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

Filatov, D. A. (2002). PROSEQ: a software for preparation and evolutionary analysis of DNA sequence data sets. *Molecular Ecology Notes* **2**, 621–624.

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

Filatov, D. A. (2005*b*). 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-link and Y-linked genes of the plants, *Silene latifolia* and *S. dioica*. *Molecular Biology and Evolution* **19**, 898–907.

Filatov, D. A., Moneger, F., Negrutiu, I. & Charlesworth, D. (2000). Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. *Nature* **404**, 388–390.

Filatov, D. A., Laporte, V., Vitte, C. & Charlesworth, D. (2001). DNA diversity in sex-linked and autosomal genes of the plant species *Silene latifolia* and *Silene dioica*. *Molecular Biology and Evolution* **18**, 1442–1454.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Guttman, D. S. & Charlesworth, D. (1998). An X-linked gene with a degenerate Y-linked homologue in a dioecious plant. *Nature* **393**, 263–266.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, R. R., Kreitman, M. & Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Ironside, J. E. & Filatov, D. A. (2005). Extreme population structure and high interspecific divergence of the silene Y chromosome. *Genetics* **171**, 705–713.

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. N. H. Munro), pp. 21–132. New York: Academic Press.

Kejnovsky, E., Vrana, J., Matsunaga, S., Soucek, P., Siroky, J., Dolezel, J. & Vyskot, B. (2001). Localization of male-specifically expressed MROS genes of *Silene latifolia* by PCR on flow-sorted sex chromosomes and autosomes. *Genetics* **158**, 1269–1277.

Kim, Y. (2005). Programs for simulation and composite likelihood analysis of selective sweep. University of Rochester, USA.

Kim, Y. & Stephan, W. (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* **155**, 1415–1427.

Kim, Y. & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777.

Lahn, B. T., Pearson, N. M. & Jegalian, K. (2001). The human Y chromosome, in the light of evolution. *Nature Reviews Genetics* **2**, 207–216.

Laporte, V., Filatov, D. A., Kamau, E. & Charlesworth, D. (2005). Indirect evidence from DNA sequence diversity for genetic degeneration of the Y-chromosome in dioecious species of the plant *Silene*: *SlY4*/*SlX4* and *DD44-X*/*DD44-Y* gene pairs. *Journal of Evolutionary Biology* **18**, 337–347.

Liu, Z., Moore, P. H., Ma, H., Ackerman, C. M., Ragiba, M., Yu, Q., Pearl, H. M., Kim, M. S., Charlton, J. W., Stiles, J. I., Zee, F. T., Paterson, A. H. & Ming, R. (2004). A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* **427**, 348–352.

Martin, I., Franke, A., Bashaw, G. & Baker, B. (1996). The dosage compensation system of *Drosophila* is co-opted by newly evolved X chromosomes. *Nature* **383**, 160–163.

Maynard Smith, J. & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35.

McVean, G. (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics* **175**, 1395–1406.

McVean, G., Awadalla, P. & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241.

Moore, R. C., Kozyreva, O., Lebel-Hardenack, S., Siroky, J., Hobza, R., Vyskot, B. & Grant, S. R. (2003). Genetic and functional analysis of *DD44*, a sex-linked gene from the dioecious plant *Silene latifolia*, provides clues to early events in sex chromosome evolution. *Genetics* **163**, 321–334.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189.

Przeworski, M. (2003). Estimating the time since the fixation of a beneficial allele. *Genetics* **164**, 1667–1676.

Rice, W. R. (1987). Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116**, 161–167.

Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497.

Stephan, W., Song, Y. S. & Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**, 2647–2663.

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

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Taylor, D. R. & Keller, S. R. (2007). Historical range expansion determines the phylogenetic diversity introduced during contemporary species invasion. *Evolution* **61**, 334–345.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–275.

Ye, D., Installe, P., Ciupercescu, D., Veuskens, J., Wu, Y., Salesses, G., Jacobs, M. & Negrutiu, I. (1990). Sex determination in the dioecious Melandrium. I. First lessons from androgenic haploids. *Sexual Plant Reproduction* **3**, 179–186.

https://doi.org/10.1017/S0016672307009056 Published online by Cambridge University Press