Nucleotide diversity in Silene latifolia autosomal and sex-linked genes

Citation for published version:

Digital Object Identifier (DOI):
10.1098/rspb.2010.0606

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the Royal Society B-Biological Sciences

Publisher Rights Statement:
Freely available via Pub Med.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes

Suo Qiu$^{1,2}$, Roberta Bergero$^1$, Alan Forrest$^1$, Vera B. Kaiser$^1$ and Deborah Charlesworth$^{1,*}$

$^1$Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh, Ashworth Laboratory, King’s Buildings, West Mains Road, Edinburgh EH9 3JT, UK

$^2$State Key Laboratory of Biocontrol and Key Laboratory of Gene Engineering of the Ministry of Education, Sun Yat-Sen University, Guangzhou 510275, People’s Republic of China

The plant *Silene latifolia* has separate sexes and sex chromosomes, and is of interest for studying the early stages of sex chromosome evolution, especially the evolution of non-recombining regions on the Y chromosome. Hitch-hiking processes associated with ongoing genetic degeneration of the non-recombining Y chromosome are predicted to reduce Y-linked genes’ effective population sizes, and *S. latifolia* Y-linked genes indeed have lower diversity than X-linked ones. We tested whether this represents a true diversity reduction on the Y, versus the alternative possibility, elevated diversity at X-linked genes, by collecting new data on nucleotide diversity for autosomal genes, which had previously been little studied. We find clear evidence that Y-linked genes have reduced diversity. However, another alternative explanation for a low Y effective size is a high variance in reproductive success. Autosomal genes should then also have lower diversity than expected, relative to the X, but this is not found in our loci. Taking into account the higher mutation rate of Y-linked genes, their low sequence diversity indicates a strong effect of within-population hitch-hiking on the Y chromosome.

**Keywords:** *Silene latifolia*; nucleotide diversity; autosomal genes; sex chromosomes; genetic degeneration

1. INTRODUCTION

The plant *Silene latifolia* is of interest for studying the evolution of sex chromosomes from chromosomes of hermaphrodite ancestors (Bergero et al. 2007; Marais et al. 2008). Support for predictions that selection during the evolution of Y chromosomes will lower the effective population size ($N_e$) for Y-linked genes (Charlesworth & Charlesworth 2000), beyond the effect of the lower numbers of Y chromosomes than X chromosomes in populations, has been the finding that several Y-linked genes in *S. latifolia* have low diversity, compared with the homologous X-linked genes (Filatov et al. 2000, 2001; Ironside & Filatov 2005; Laporte et al. 2005). In non-recombining Y chromosome regions, selection will reduce $N_e$ through hitch-hiking processes, including selective sweeps during the spread of advantageous alleles (causing selective sweeps that would lower diversity at linked silent sites, e.g. Rice 1987; Bachrorg 2004; Bachrorg et al. 2009), selective elimination of deleterious mutations (Charlesworth et al. 1993; Kaiser & Charlesworth 2009) and Muller’s ratchet (Gordo & Charlesworth 2000; Gordo et al. 2002; Engelstädter 2008).

In evaluating the X–Y diversity difference as evidence for selection affecting Y chromosomes, other possible processes affecting diversity must be considered. The data so far collected cannot distinguish whether the difference is due to the predicted reduction in Y $N_e$, or to some process increasing high species-wide diversity for X-linked genes. Possibilities include a higher X than Y mutation rate (although this is the opposite of the expectation for any difference between the mutation rates of the two chromosomes, and the data so far available support that expectation, see Filatov & Charlesworth 2002; Filatov 2005), sexual selection causing a high variance in male reproductive success (Laporte & Charlesworth 2002) or introgression of sequences from another species; hybridization is well known to occur between *S. latifolia* and *S. dioica* (Baker 1948; Minder & Widmer 2007) and seems to have led to sequences within each species having recombined into X-linked genes of the other (Atanassov et al. 2001; Laporte et al. 2005). Population size changes also affect the relative diversity levels of genome regions with different $N_e$ (Fay & Wu 1999; Pool & Nielsen 2007).

A way to test whether the diversity is reduced for Y-linked genes, or elevated for X-linked ones, is to compare diversity with that at autosomal loci. A previous study of a single autosomal gene (*CCL537.1*) found higher diversity than for a Y-linked gene (*SIY*), but considerably lower than for an X-linked gene studied in the same set of plants, leaving the possibility open that X-linked genes might have unusually high diversity; however, this gene may have experienced a recent selective sweep (Filatov et al. 2001). Sequences are available for only one other *S. latifolia* autosomal gene, *Slop*, whose silent site diversity is 0.65 per cent (Muir & Filatov 2007). A third estimate, from an *Ap3* gene (Matsunaga et al. 2003), is no longer relevant, because the gene is probably X-linked (Ishii et al. 2008).

We therefore obtained new data on DNA sequence diversity in *S. latifolia* autosomal loci, and also increased...
Table 1. Genes studied in this paper and the primers used for PCR amplification.

<table>
<thead>
<tr>
<th>gene</th>
<th>locus name</th>
<th>accession number</th>
<th>primers</th>
<th>regions studied</th>
<th>length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E72</td>
<td>pentatricopeptide repeat-containing protein</td>
<td>AT1G02150</td>
<td>E72-f: ATTTAATGGCTAAAGTTCCAGTAATCC</td>
<td>exon2</td>
<td>674</td>
</tr>
<tr>
<td>E72</td>
<td>zinc ion-binding protein (ATVPS11)</td>
<td>AT2G05170</td>
<td>E72-r: CATTAGAAGGTTTCCGGTTAGGTC</td>
<td>exon4</td>
<td>656</td>
</tr>
<tr>
<td>E106</td>
<td>vacuolar ATP synthase subunit D (VATD)</td>
<td>AT3G58730</td>
<td>E106-f: ATACTCCAGGCTTCATATATTGAC</td>
<td>exon1</td>
<td>410</td>
</tr>
<tr>
<td>E265</td>
<td>FCLY (farnesylcysteine lyase) or prenyllycysteine oxidase</td>
<td>AT5G63910</td>
<td>E265-f: GGCATGCTTTAATGTCTGGTATG</td>
<td>intron2</td>
<td>464</td>
</tr>
<tr>
<td>E304</td>
<td>putative transaldolase</td>
<td>AT5G13420</td>
<td>E304-f: CTTTATCAACCAGGAGTACGT GG</td>
<td>intron3</td>
<td>210</td>
</tr>
<tr>
<td>E314</td>
<td>rhodanese-like domain-containing protein</td>
<td>AT1G34000</td>
<td>E314-f: AAACATCGATTTCACGAGATG</td>
<td>intron2</td>
<td>392–397</td>
</tr>
<tr>
<td>E391</td>
<td>proton motive force-dependent protein transporter</td>
<td>AT5G54770</td>
<td>E391-f: GGCAGCAGCACATCTTATCCAGT</td>
<td>exon4</td>
<td>344</td>
</tr>
<tr>
<td>E393</td>
<td>photosystem I subunit L (PSAL)</td>
<td>AT4G12800</td>
<td>E393-f: CAACAGGCTTTTATTTGTCACACC</td>
<td>intron1</td>
<td>13</td>
</tr>
</tbody>
</table>

2. MATERIAL AND METHODS

(a) Genes studied

To obtain autosomal genes, a large number of EST sequences were obtained by 454 sequencing of *S. latifolia* cDNAs (R. Bergero, S. Qiu, H. Borthwick, A. Forrest & D. Charlesworth 2008–2010, unpublished data). The Arabidopsis thaliana genome was used to identify likely orthologues, and genes belonging to large gene families (in *A. thaliana*) were discarded, as these are unlikely to provide simple genetic segregation data. Translated cDNA sequences were aligned with *A. thaliana* coding sequences (Swarbreck et al. 2008), using the software FASTA (Pearson 1990). Single genes with e-value < 10^-5 were retained and annotated in Sequencher v.4.5 (GeneCodes, Ann Arbor, MI, USA), using information from *A. thaliana* orthologues to suggest intron positions and functions. The genes studied, and the polymerase chain reaction (PCR) primers for each gene are given in table 1.

(b) Silene latifolia families and population samples

To identify autosomal (and pseudoautosomal) genes, we studied segregation of length or sequence variants detected in families (following the approach previously used in this species, see Bergero et al. 2007). To study diversity, we sequenced eight autosomal genes, using a set of *S. latifolia* male plants (electronic supplementary material, table S2) from across Europe, including much of the likely native distribution of this species. These plants were grown from seeds collected from naturally pollinated females, and one male was sampled per locality, to provide a sample suitable for estimating effective population sizes from nucleotide diversity, and in which variant frequencies should follow a neutral coalescent unless selection occurs (whereas samples with multiple individuals per population are expected to have a different variant frequency spectrum, see Wakeley & Aliacar 2001). We also estimated diversity for two recently characterized sex-linked genes, *SlCyP* and *Sl7* (Bergero et al. 2007), and a newly discovered gene, *SlXY9* (Kaiser et al. in press). Our new data, with more loci, and also from more *S. latifolia* populations than previous studies, show that Y diversity is considerably reduced, supporting the ongoing genetic degeneration hypothesis.
(c) DNA extraction, PCR reaction and cloning
For the segregation study, seeds were germinated on Murashige and Skoog (Murashige & Skoog 1962) medium. DNA was extracted from 20-day-old seedlings using the FastDNA kit (Q-biogene) following the manufacturer’s instructions. Seedlings’ sexes were determined using a 293-bp indel fixed in intron 2 of all Y-linked copies of the SlCyp gene (Bergero et al. 2007). For the sequence diversity study, DNA was extracted from leaves collected from mature plants grown in greenhouse conditions, using the same method.

Primers for the genes newly sequenced in our study were designed using the EST sequences of S. latifolia (table 1). The primers for SlXY7 were SIX7F3 (GAATGGCAGAATGTTGCAATGCAAC) and SIX7F2 (GGAAA GCTCTTTTCCAAGTGCAC), and, for SlCypXY, cypXF3 (AATTCTCTGTAACATCCTGATCCGACAG) and cypR1 (AGTTGATCTCTGTTCTGATCATGAAACC). Amplifications used a Finzymes’ Piko Thermal Cycler with Phire hot-start DNA polymerase (Finzyme), under the following PCR conditions: one cycle of initial denaturation at 98°C for 30 s, 10 cycles at 98°C for 5 s, 65°C for 5 s, and 72°C for 50 s, and then 25 cycles of 98°C for 5 s, 60°C for 5 s and 72°C for 1 min, and finally one cycle at 72°C for 5 min.

The PCR amplicons were cleaned with ExoSAP-IT (Amersham Biosciences, Tokyo) and sequenced on an ABI 3730 capillary sequencer (Applied Biosystems). When no indel variants were present, the single nucleotide polymorphisms (SNPs) were verified by direct sequencing from both strands. Most such variants were found in multiple individuals. Amplicons containing intron regions were sometimes heterozygous for indels. In such individuals, and for plants with singleton variants, the DNA was re-amplified with a proofreading DNA polymerase (Phusion, Finzyme) using the PCR conditions given above, and cloned into a T-tailed pBSKS+ vector (Stratagene) before sequencing.

As described below, two autosomal genes have high diversity. Our genetic results suggest that they are single-copy genes. However, these data apply only to the parents of the families used, and furthermore the genetic results for E393 are based on a single SNP, while for E391 we used an intron region, not the region whose diversity was estimated. To test for the existence of paralogous copies of these genes among the plants used in the diversity survey, we sequenced plants in family H2005-1 (electronic supplementary material, table S1), with the primers used in the survey. For E393, both the F1 parental plants are heterozygous for 13 SNPs, and 10 progeny sequenced include heterozygotes and both homozygotes, as expected for segregation of a single gene. For E391, the parents are heterozygous for a single SNP, which also segregated in the progeny. Therefore, in this family, these genes do not appear to be duplicated.

(d) Sequence analyses
The newly obtained sequences of each gene were first aligned with Sequencher v.4.5 and manually adjusted using the sequence alignment editor Se-Al v.2.0a11 (Se-Al: Sequence Alignment Editor, http://evolve.zoo.ox.ac.uk/). The alignments were submitted to GenBank, under accession numbers HM188567–HM189181. Polymorphism and haplotype analyses (including estimates of nucleotide diversity and population subdivision) were done using DNASP v.5.00.06 (Librador & Rozas 2009) and MEGA 4 (Tamura et al. 2007). We estimated diversity for silent sites, because this relates closely to Ne values (Loew et al. 2006). Hudson-Kreitman-Aguadé (HKA) tests for differences in diversity between sets of genes (see §3) were done with a maximum-likelihood implementation of the test, multi-locus HKA (mHKA) (Wright & Charlesworth 2004). These tests used the S. vulgaris sequences mentioned above, and divergence from the S. latifolia orthologues was estimated using DNASP. Silent site divergence values for these two species range from about 9 to 22 per cent. We also did Tajima’s (Tajima 1989a) and Fay and Wu’s tests for selection, which depend on variant frequencies (see §3), and the DH test, which is designed to be less sensitive to changes in recent demographic history (Zeng et al. 2006). For these tests, we used variants of all site types, and S. vulgaris as the outgroup.

In addition, the diversity data were compared with published estimates from four other sex-linked genes, SIXY1 (Filatov et al. 2001) and SIXY4 and SIDD44 (Ironside & Filatov 2006; Laporte et al. 2005), which used smaller samples of European S. latifolia, and SlSs (Filatov 2008), which used samples similar to those used here.

We tested for mutation rate differences in four new X–Y gene pairs, three genes used in our diversity analyses (SIXY7, Cyp-XY, SIXY9) and XY6a, which was not included in the diversity survey, because a duplicate gene exists (Bergero et al. 2007). We extracted non-coding and fourfold degenerate sites using DNASP v.5.00.06 (Librador & Rozas 2009), and used the baseML program of the PAML v.3e package (Yang 2001) to obtain maximum-likelihood estimates of the relative substitution rates and perform likelihood ratio tests of the significance of silent substitution rate differences. We estimated the likelihood under two models (i) with three evolutionary rate parameters, one estimated for each of the X-linked, Y-linked and non-sex-linked (S. vulgaris), branches, and (ii) with the same rate for X- and Y-linked genes. Significance of differences between the X- and Y-linked genes was tested using a χ² distribution with 1 d.f.

3. RESULTS
(a) Discovery of autosomal genes
The eight autosomal genes in our diversity survey (table 1) were found to be single-copy and not sex-linked by studying their segregation ratios in families. Electronic supplementary material, table S1 shows that these genes almost always show the expected ratios, with no heterogeneity in ratios between the two sexes of progeny (other than for locus E163, which deviated significantly from the expected ratio in the female progeny, but with no significant heterogeneity between the sexes). Two further genes, E284 and E241, appear to be pseudoautosomal, and will be described in detail elsewhere.

Six of the eight autosomal genes also map to autosomal linkage groups with multiple other genes (R. Bergero, A. Forrest, H. Borthwick, S. Qiu & D. Charlesworth, unpublished data). For two genes, E163 and E265, pseudo-autosomal locations are not ruled out from these data. However, the genotypes in family G2008-3 (electronic supplementary material, table S1) allowed us to test E265 for linkage to the pseudoautosomal gene E241; no linkage was detected.

The haplotypes from males sampled from the natural populations further support autosomal inheritance for
all eight genes. If a locus is sex-linked, we expect variants differentiating the Y-linked alleles from all X sequences. Thus males should always be heterozygous (unless the Y chromosome has lost the gene, in which case all males will appear to be homozygotes). Finding both heterozygous and homozygous males, as we found for all loci, including $E_163$ and $E_265$ (table 2) suggests that a gene is in a recombining region of the genome (it is unlikely that they carry an allele that does not amplify with our primers, because no null homozygotes were seen for any locus in our study).

(b) Autosomal locus diversity results and comparisons with X and Y

Figure 1 summarizes the currently available nucleotide diversity estimates based on silent sites ($\pi_{\text{silent}}$) for all genes studied so far, including previously published results, and three newly sequenced sex-linked genes, $SIXY_7$, $SIXY_9$ and $SICypXY$. The Y-linked genes have much lower diversity values than X-linked or autosomal genes. The X–Y difference is significant by a multi-locus HKA test, using divergence from the homologous $S. vulgaris$ sequences ($p = 1.06 \times 10^{-5}$), excluding $SlSsX$, where no Y sequences are available, and either including or excluding $SlAp3$ (where X-linkage is not firmly established, see Ishii et al. 2008); the Y-autosomal significance value was $p = 2.35 \times 10^{-4}$. Overall, $\pi_{\text{silent}}$ for the eight autosomal loci is 2.88, versus 2.15 per cent for the seven X-linked genes, and 0.14 per cent for their Y-linked homologues. Although the diversity estimates vary widely among loci, the X-autosomal difference is significant ($p = 0.0017$) based on all the X-linked genes, but not if we exclude $SlSsX$ ($p = 0.92$), which may have undergone a selective sweep (Filatov 2008), and differs significantly from the other loci.

The overall $X/A$ diversity ratio is 0.747, very close to the value (0.75) expected with a 1:1 sex ratio and the same variance in male and female reproductive success, so that the effective size for X-linked genes is 0.75 that of autosomal genes (e.g. Laporte & Charlesworth 2002; Keinan et al. 2009). These assumptions predict an effective size for Y-linked genes one-fourth that of autosomal genes, but the $Y/A$ silent site diversity ratio is 0.049 (one-twentieth). Together, these results suggest hitch-hiking processes caused by selection on the non-recombining Y chromosome, lowering the effective size for Y-linked genes below the expected value (see §1).

Table 2. Diversity statistics for $S. latifolia$ autosomal genes.

<table>
<thead>
<tr>
<th>gene</th>
<th>$n^a$</th>
<th>sites in alignment</th>
<th>number of polymorphic sites$^b$</th>
<th>number of different haplotypes$^c$</th>
<th>number of homozygotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E72$</td>
<td>46</td>
<td>674</td>
<td>39</td>
<td>44</td>
<td>1</td>
</tr>
<tr>
<td>$E106$</td>
<td>48</td>
<td>656</td>
<td>6</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>$E163$</td>
<td>48</td>
<td>410</td>
<td>9</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>$E265$</td>
<td>48</td>
<td>674</td>
<td>5</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>$E304$</td>
<td>47</td>
<td>676</td>
<td>44</td>
<td>33</td>
<td>8</td>
</tr>
<tr>
<td>$E314$</td>
<td>46</td>
<td>574</td>
<td>40</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>$E391$</td>
<td>48</td>
<td>413</td>
<td>27</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>$E393$</td>
<td>48</td>
<td>432</td>
<td>38</td>
<td>34</td>
<td>8</td>
</tr>
</tbody>
</table>

$^a$Number of alleles sequenced.
$^b$Excluding indels.
$^c$Including indels.

Figure 1. Summary of nucleotide diversity estimates using silent sites for X, A and PAR genes of $S. latifolia$ (grey bars) and for the Y-linked homologues of the X-linked loci (black bars). The numbers of sites analysed, and references for the data sources are in the electronic supplementary material, table S3.

Two autosomal genes ($E393$ and $E391$) have very high diversity values, but mHKA tests within the set of autosomal genes detect no statistically significant differences for these genes (see below). Our genetic results suggest that the $E393$ and $E391$ genes do not have duplicate copies, and neighbour-joining trees of the sequences contain no separate clusters of $S. latifolia$ sequences (electronic supplementary material, figure S1). The high diversity values are not due to high mutation rates for these genes, because divergence estimates from...
S. latifolia autosomal diversity  S. Qiu et al.  3287

Table 3. Tests for selection for autosomal and X-linked genes, using all site types. Statistically significant values ($p < 0.05$) are shown in bold.

<table>
<thead>
<tr>
<th>gene</th>
<th>Tajima’s test</th>
<th>Fay and Wu’s test</th>
<th>DH test value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_T$</td>
<td>$p$-value</td>
<td>$H$</td>
</tr>
<tr>
<td>autosomal genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E72</td>
<td>-0.941</td>
<td>0.179</td>
<td>0.373</td>
</tr>
<tr>
<td>E106</td>
<td>-1.44</td>
<td>0.058</td>
<td>0.392</td>
</tr>
<tr>
<td>E163</td>
<td>-0.167</td>
<td>0.499</td>
<td>-1.99</td>
</tr>
<tr>
<td>E265</td>
<td>-0.654</td>
<td>0.300</td>
<td>-1.353</td>
</tr>
<tr>
<td>E304</td>
<td>-1.77</td>
<td>0.0158</td>
<td>-1.885</td>
</tr>
<tr>
<td>E314</td>
<td>-0.051</td>
<td>0.552</td>
<td>-0.762</td>
</tr>
<tr>
<td>E391</td>
<td>0.070</td>
<td>0.6021</td>
<td>-1.088</td>
</tr>
<tr>
<td>E393</td>
<td>0.305</td>
<td>0.693</td>
<td></td>
</tr>
<tr>
<td>X-linked genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SlSsX</td>
<td>-1.94</td>
<td>0.007</td>
<td>-1.722</td>
</tr>
<tr>
<td>SlXY4</td>
<td>-2.58</td>
<td>0.461</td>
<td>-1.571</td>
</tr>
<tr>
<td>SlD44X</td>
<td>0.033</td>
<td>0.553</td>
<td>0.470</td>
</tr>
<tr>
<td>SlXY7</td>
<td>-1.362</td>
<td>0.0775</td>
<td>-3.093</td>
</tr>
<tr>
<td>Sl9X</td>
<td>0.320</td>
<td>0.695</td>
<td>-0.480</td>
</tr>
<tr>
<td>SlXC</td>
<td>-0.318</td>
<td>0.425</td>
<td>0.168</td>
</tr>
<tr>
<td>SlX1</td>
<td>0.705</td>
<td>0.0354</td>
<td>0.807</td>
</tr>
<tr>
<td>SlSsX</td>
<td>-2.58</td>
<td>0.033</td>
<td>-0.268</td>
</tr>
<tr>
<td>SlXY6a</td>
<td>0.033</td>
<td>0.553</td>
<td>0.470</td>
</tr>
<tr>
<td>SlX8</td>
<td>0.258</td>
<td>0.461</td>
<td>-1.571</td>
</tr>
<tr>
<td>SlCypX</td>
<td>0.320</td>
<td>0.695</td>
<td>-0.480</td>
</tr>
<tr>
<td>SlX7</td>
<td>-1.362</td>
<td>0.0775</td>
<td>-3.093</td>
</tr>
<tr>
<td>Sl9X</td>
<td>0.320</td>
<td>0.695</td>
<td>-0.480</td>
</tr>
<tr>
<td>SlSsX</td>
<td>-2.58</td>
<td>0.033</td>
<td>-0.268</td>
</tr>
<tr>
<td>SlXY6b</td>
<td>0.033</td>
<td>0.553</td>
<td>0.470</td>
</tr>
<tr>
<td>SlX8</td>
<td>0.258</td>
<td>0.461</td>
<td>-1.571</td>
</tr>
<tr>
<td>SlSsX</td>
<td>-2.58</td>
<td>0.033</td>
<td>-0.268</td>
</tr>
<tr>
<td>SlX8</td>
<td>0.258</td>
<td>0.461</td>
<td>-1.571</td>
</tr>
<tr>
<td>SlXY6b</td>
<td>0.033</td>
<td>0.553</td>
<td>0.470</td>
</tr>
</tbody>
</table>

S. vulgaris and S. dioica sequences are similar to the values for other loci. Finally, we tested for introgression in our S. latifolia populations, using sequences from S. dioica sampled from Southern Finland, Scotland, England, France and Italy (electronic supplementary material, table S2 and figure 1; both species co-occur in all regions except Finland). We found no fixed differences between the species (excluding a formal analysis of introgression), and, even with our non-exhaustive sampling, many polymorphisms are shared (24 for E393 and 7 for E391, versus 18 and 12 polymorphisms, respectively, exclusive to S. latifolia). Silene dioica sequences are also diverse (silent site $\pi > 5\%$ for both loci), but the haplotypes are mostly distinct from those of S. latifolia plants. The few clusters with sequences from both species have very low bootstrap support, and the sequences are mostly not extremely similar (electronic supplementary material, figure S1); for E391, no S. latifolia sequences are among the main S. dioica cluster. For E393, excluding two S. latifolia alleles that cluster with S. dioica sequences reduces diversity only slightly (from 0.077 to 0.075). Introgression from S. dioica thus seems unlikely to be the sole reason for high diversity of these genes, but we cannot exclude past introgression from other species not currently sympatric with S. latifolia.

We also tested for a higher Y than X mutation rate in the four new X–Y genes. The ratio of Y- to X-linked substitution rates, $R_{Y/X}$, exceeds 1 for all four genes, two of them significantly ($SlCypX$, with $p = 0.039$ and $R_{Y/X} = 2.03$, and $SIXY6a$, with $p < 0.001$ and $R_{Y/X} = 2.80$), while two short sequences (approx. 100 bp analysed) were non-significant ($SIXY7$, with $R_{Y/X} = 1.08$, and $SIXY9$, with $R_{Y/X} = 1.42$).

(c) Tests for selection
The overall X/A ratio of diversity estimates is close to the expected value of 0.75. However, diversity varies among both the X-linked and the autosomal genes. The X/A ratio would be affected if some genes in either type of location had been subject to recent selective sweeps, or have high diversity owing to long-term balancing selection, and it is thus possible that the agreement is coincidental. As well as the two autosomal genes with high silent site diversity discussed above, two (E106 and E265) have diversity values even lower than CCLS37.1 (which, as mentioned above, may have undergone a selective sweep). The recent selective sweep at the X-linked $SlSsX$ gene (Filatov 2008) is supported by mHKA tests among the X-linked loci (see above). Only the two low-diversity autosomal genes (E106 and E265) differ significantly from the other autosomal loci by mHKA tests. These may thus have undergone recent selective sweeps, though the test is highly significant only for E265 ($p = 2 \times 10^{-3}$; for E106, $p = 0.026$).

We also detect an unexpectedly high proportion of low-frequency variants, compared with the expectation at equilibrium under neutrality. Tajima’s $D$-values for autosomal genes (using either silent sites or all site types) are mostly negative, often strongly so, but only E304 has an individually significantly negative value (table 3). This overall pattern could be caused by an extreme bottleneck (Tajima 1989a,b), but is more likely to be due to population expansion (Tajima 1993; Innan & Stephan 2000; Ray et al. 2003), which is plausible for this species, which probably expanded after the last ice age. Subdivision may also contribute, since sampling few individuals from many localities can produce negative $D$-values (Ptak & Przeworski 2002). However, D becomes even more negative when we use a single sequence per population, which produces a standard coalescent process if geographical structure is unimportant (Wakeley & Aliacar 2001; De & Durrett 2007). Geographical structure often produces positive Tajima’s $D$-values, owing to local adaptation, but sampling mostly from a common subpopulation or race,
with few individuals from a different race, could produce an unexpected frequency of rare variants. However, the clustering program STRUCTURE2 (Pritchard et al. 2000) identifies no geographical structure in S. latifolia, using six X-linked genes (G. Muir, R. Bergero, D. Charlesworth & D. A. Filatov 2007–2010, unpublished data).

We also applied the DH test for selection, which is less sensitive than other tests to changes in populations’ recent demographic history (Zeng et al. 2006). The E304 gene again yielded a significant test result, and two other genes gave marginally significant Fay and Wu’s tests (table 3). None of the tests were significant for any Y-linked gene (results not shown). Of the five X-linked genes, selection at the SISsX locus is confirmed, and is inferred also for SISX (table 3). Overall, consistent with the HKA test results described above, there is clearly no convincing evidence for selection in or near most of the loci studied, and thus for excluding loci in estimating the X/A ratio; without the low diversity SISsX gene, this ratio increases to 0.866, but excluding E265 also, it becomes 0.688.

Finally, we applied haplotype tests, to test for balancing selection (for the two loci with high diversity, E391 and E393, such tests might also detect evidence for paralogous copies). We used Dnasp to run neutral coalescent simulations, using the estimated recombination rate for each locus, and obtain 95% confidence intervals for the expected numbers of haplotypes. The observed values for all loci were within these intervals.

(d) Population subdivision and comparisons between the X and Y

Another possible explanation for high silent site diversity values is population subdivision, but this seems not to be the case. We analysed diversity separately for the samples from individual populations or from four geographical regions (see §2 and electronic supplementary material, table S2). Despite the large distances between the locations where the plants were sampled, KST values for the autosomal genes (electronic supplementary material, table S4) are not high (the average is 0.125 for silent sites, similar to the value for human populations), with similar values for the different loci. For the sex-linked genes, KST is higher (the X and Y averages are 0.177 and 0.245, reflecting the fact that they both have lower N0). Estimates using individual localities are higher (electronic supplementary material, table S4), but our sample size of two alleles per locality must considerably overestimate subdivision, because such samples often contain only the commonest allele, and monomorphism elevates KST values (Charlesworth 1998).

4. CONCLUSIONS AND DISCUSSION

Our results establish that diversity is low for Y-linked genes, rather than being unusually high for X-linked ones, since the mean X diversity is slightly lower than the autosomal loci, and close to the value expected with a 1:1 sex ratio and the same variance in male and female reproductive success. Hitch-hiking processes on the non-recombining Y chromosome are predicted to lower Y diversity, but a possible alternative reason for low diversity of Y-linked genes is a high variance in male reproductive success. The autosomes should then also have lower diversity than expected, relative to the X, i.e. the X/A ratio should be increased above 0.75 (Laporte & Charlesworth 2002; Ellegren 2009). We therefore conclude that there is no evidence that sexual selection is important in populations of this plant.

Another possibility, however, is a recent change in the demographic history of the population. Changes in population size affect the X and Y chromosomes differently from the autosomes, and can increase or reduce X/A and Y/A diversity ratios for many generations (Fay & Wu 1999; Pool & Nielsen 2007). Rapid growth in population size leads to increased X/A ratios, lasting for considerable times after the size change, whereas a reduction in population size decreases the ratio (the decrease can be severe, but not for long, see Pool & Nielsen 2007). A temporary bottleneck generally lowers X/A ratios, but a very severe bottleneck, followed by growth as the population recovers, can have the opposite effect. The Y/A diversity ratio (and the mitochondrial/autosomal one) are affected in similar ways, but the low effective size of Y-linked genes causes their diversity to recover much faster than for X-linked genes, so that prolonged effects of a bottleneck are less likely to be found.

We therefore compared diversity estimates from two geographical regions, since the size change models predict different ratios in populations whose sizes differ. Although the regions are not strongly isolated (see above), the samples from northern populations (as defined in §2), when analysed separately, tend to have slightly higher diversity than southern ones (14 of 17 genes for which data are currently available), with a south/north diversity ratio for silent sites of 0.78 for autosomal genes (but no difference for the Y-linked sequences). The difference is more pronounced for the X-linked genes: southern populations’ silent site diversity is only 56 per cent of that in the northern sample, so that their X/A ratio is much lower (0.35) than the overall value, consistent with a greatly reduced southern population in the recent past. However, Tajima’s D-values are not more negative in the southern sample.

A low population size in the recent past, from which the population has not yet recovered, could temporarily greatly reduce diversity for Y-linked genes, relative to X-linked ones, and is thus a possible alternative to hitch-hiking to explain the observed low Y sequence diversity. Figure 2 illustrates this. A larger decrease in size could give an even larger X/Y ratio. On the other hand, any such difference will be lessened if Y genes have a higher mutation rate than X ones (as seems to be the case in S. latifolia, see above). It therefore seems unlikely that a population size change can explain the ratios observed in S. latifolia.

Our observed higher KST values for Y-linked genes than for other loci, regardless of the spatial scale of the samples used, suggest that the reduction in Y diversity is mainly caused by within-population processes, such as elimination of deleterious mutations, or local adaptations involving Y-linked genes, and is not principally an effect of a species-wide selective sweep owing to an advantageous mutation fixing in a Y-linked gene (which should lead to lower KST for the Y, until equilibrium is re-established). The highest within-population X/Y diversity ratios average 31 for silent sites (electronic supplementary material,


REFERENCES


Keinan, A., Mullikin, J., Patterson, N. & Reich, D. 2009 Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat. Genet. 41, 66–70. (doi:10.1038/ng.303)


