Long-term balancing selection and the genetic load linked to the self-incompatibility locus in Arabidopsis halleri and A. lyrata

Audrey Le Veve, Nicolas Burghgraeve, Mathieu Genete, Christelle Lepers-Blassiau, Margarita Takou, Juliette de Meaux, Barbara K Mable, Eléonore Durand, Xavier Vekemans, Vincent Castric

Article

Discoveries

Long-term balancing selection and the genetic load linked to the self-incompatibility locus in *Arabidopsis halleri* and *A. lyrata.*

Audrey Le Veve[1], Nicolas Burghgraeve[1], Mathieu Genete[1], Christelle Lepers-Blassiau[1], Margarita Takou[2,4], Juliette De Meaux[2], Barbara K. Mable[3], Eléonore Durand[1], Xavier Vekemans[1], Vincent Castric[1]

[1] *Univ. Lille, CNRS, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France*

[2] *Institute of Botany, University of Cologne, Cologne, Germany*

[3] *Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow, UK*

[4] *Department of Biology, Pennsylvania State University, PA, United States of America*

Author for correspondence : vincent.castric@univ-lille.fr

Abstract

Balancing selection is a form of natural selection maintaining diversity at the sites it targets and at linked nucleotide sites. It is generally expected to facilitate the accumulation of a "sheltered" genetic load of linked deleterious mutations, but the overall extent of its genomic impact remains poorly documented. Taking advantage of plant self-incompatibility as one of the best-understood and most intense examples of long-term balancing selection, we provide the most highly resolved picture of the effect of balancing selection on the sheltered genetic load in any plant genome. We used targeted genome resequencing to evaluate the intensity of indirect selection on the genomic region flanking the self-incompatibility locus in three sample sets in each of the two closely related plant species *Arabidopsis halleri* and *A. lyrata,* and used 100 control regions from throughout the genome to factor out differences in demographic histories and/or sample structure. Nucleotide polymorphism increased strongly around the *S*-locus in all sample sets, but only over a limited genomic extent, as it became indistinguishable from the genomic background beyond the first 25kb. Genes around the *S*-locus carried more mutations than those in the control regions, but in contrast to the classical model for the accumulation of the sheltered load we observed no relaxation of the efficacy of purifying selection. Overall, our results challenge our understanding of how natural selection in one genomic region affects the evolution of the adjacent genomic regions, and have consequences for how systems of mating that enforce outcrossing in populations are maintained.

Keywords: balancing selection, linked selection, sheltered genetic load, deleterious mutations, S-locus, Arabidopsis, self-incompatibility.

Introduction

Balancing selection refers to a variety of selective regimes maintaining advantageous genetic diversity within populations (Delph and Kelly 2014). Notable examples include heterozygote advantage, negative frequency-dependent selection, and spatial heterogeneity. The implication of balancing selection in maintaining genetic diversity across the genome has been widely debated (*e.g.* Asthana et al. 2005). In contrast to genetic linkage to genomic sites subject to either positive or negative selection that generally tends to eliminate surrounding genetic variation (Smith and Haigh 1974, for hitchhiking effect; Charlesworth et al. 1993 and Loewe and Charlesworth 2007, for background selection effects), linkage to loci under balancing selection is expected to locally promote the long-term persistence of variation in surrounding sites (Charlesworth 2006). Theoretical studies by Takahata and Satta (1998), Schierup et al. (2000) and Wiuf et al. (2004) showed that besides the strength of balancing selection and the local rate of recombination, the magnitude of the local diversity increase and its extent along the chromosome critically depend on details of the exact form of balancing selection (Llaurens et al. 2017 for a review). An important result from these studies is that the extended time over which the balanced allelic lineages are maintained (Takahata and Nei 1990; Vekemans and Slatkin 1994) also means more time for recombination to decouple them from their linked sites, such that the extent of the region affected may end up being quite narrow (Hudson and Kaplan 1988; Schierup et al. 2001). In addition to the sheer increase of polymorphism, several balancing selection processes promote heterozygosity. They are thus expected to mask recessive deleterious mutations (Maruyama and Nei 1981), such that linkage to a locus under balancing selection can negatively interfere with purifying selection, diminishing its efficacy and facilitating the local accumulation of a potentially strong genetic load, referred to as the "sheltered load" (Uyenoyama 1997, 2005; Hartfield and Otto 2011). This phenomenon has been considered as the "evolutionary cost" of balancing selection (van Oosterhout et al. 2009; Lenz et al. 2016), and in humans a large number of diseases are indeed associated with variants at genes linked to one of the classical examples of balancing selection in the human genome, the Major Histocompatibility Complex (*MHC*; e.g. Lenz et al. 2016; Matzaraki et al. 2017).

Evaluating the importance of balancing selection and determining its evolutionary consequences has been the focus of sustained interest in the field (Llaurens et al. 2017). Genome resequencing studies have revealed that balancing selection can be a potent force throughout the genome (DeGiorgio et al. 2014), but because of the inherent technical challenges related to the high levels of polymorphism in the genomic regions affected (Vekemans et al. 2021) it is still unclear how widespread the various forms of balancing selection actually are (see e.g. Fijarczyk and Babik, 2015). In support of the sheltered load hypothesis, Lenz et al. (2016) observed a specific accumulation of putatively deleterious mutations (missense variants) in genes that are located inside the human *MHC* region but have no function in immunity and just happen to be linked to the *MHC* alleles. Interestingly, this sheltered load was mostly due to an increase in the mean population frequency of deleterious mutations as compared to genes in a series of "control" regions, but not to an elevation of their overall number, suggesting that the balancing selection process at play for the human *MHC* region elevates polymorphism locally by distorting the frequency of deleterious mutations, rather than by increasing their density. Whether this observation can be generalised to other biological systems under balancing selection is not known.

Self-incompatibility (SI) in plants is perhaps the best understood case of long-term balancing selection (Castric and Vekemans, 2004). SI is a genetic mechanism allowing recognition and rejection of self-pollen, thereby preventing inbreeding and promoting outcrossing in hermaphroditic plants

(Nettancourt, 2001). Pollination between partners expressing identical haplotypes at the *S*-locus leads to rejection of the pollen. This genetic system enforces outcrossing and promotes higher heterozygosity than expected under random mating. In addition, as noted by Wright (1939), pollen produced by individuals carrying rare *S*-alleles will more rarely land on incompatible pistils than pollen produced by individuals carrying *S*-alleles that are more frequent. The action of natural selection on the *S*-locus is thus well elucidated and corresponds to an intense form of negative frequency-dependent selection that allows the stable maintenance of a large number of *S*-alleles within populations. The *S*-alleles are maintained over very long evolutionary times (Vekemans and Slatkin, 1994), and theoretical models predict that a local increase of nucleotide polymorphism should be observed in the linked genomic region (Uyenoyama, 1997 ; Schierup et al. 2000). In gametophytic SI (GSI), pollen SI specificity is determined by its own haploid genome (as found e.g. in Solanaceae), whereas in sporophytic SI (SSI), the pollen recognition phenotype is determined by the male diploid parent (as found e.g. in Brassicaceae). In the Brassicaceae, SSI is controlled by a single genomic region, the *S*-locus (Schopfer, Nasrallah, and Nasrallah 1999 ; Kusaba et al. 2001), composed of two linked genes, *SCR* (encoding the *S*-locus cysteine-rich protein) and *SRK* (encoding the *S*-locus receptor kinase protein), encoding the male and female specificity determinants, respectively.

The phenotypic effect of the sheltered genetic load linked to the *S*-locus can be revealed by controlled crosses to experimentally enforce homozygosity at the S-locus and isolate the specific effect of this homozygosity on proxies of fitness. To the best of our knowledge, such experiments have been performed in five plant species only: *Papaver rhoea*s (Lane and Lawrence 1995), *Solanum carolinense* (Stone 2004), *Arabidopsis halleri* (Llaurens et al. 2009), *A. lyrata* (Stift et al. 2013) and *Rosa* (Vieira et al. 2021). In these species, a detectable genetic load linked to the *S*-locus could be revealed, although its magnitude varied among the *S*-alleles that were brought to the homozygous state. The fact that this load was detectable at the phenotypic level in spite of the inherently limited experimental power of these studies, suggests that the *S*-locus does indeed shelter a substantial load of deleterious mutations. These studies focused on phenotypic characterization of the load, and thus provided no indication about its genomic architecture. At this stage, the nature of this load therefore remains elusive. In *A. halleri*, the *S*-locus has been sequenced entirely in multiple haplotypes, revealing that the non-recombining *S*-locus region contains no protein-coding genes besides the ones controlling the SI machinery itself (*SCR* and *SRK*; Goubet et al. 2012). The load detected phenotypically is therefore likely caused by mutations in the partially linked flanking regions rather than in the non-recombining *S*-locus region itself. A series of studies have set out to determine the genomic extent of the flanking region over which polymorphism was altered by linkage to the *S*-locus in *A. halleri* and *A. lyrata* (Kamau and Charlesworth 2005; Kamau et al. 2007; Ruggiero et al. 2008; Roux et al. 2013). These studies sequenced short fragments of a small subset of the genes immediately flanking the *S*-locus, as well as more distant genes, and compared their polymorphism to that of a handful of "control" coding sequences from across the genome. In both species, the increase of polymorphism was limited to the genes immediately flanking the *S*-locus only, but the very sparse sampling of genes and the sequencing of small gene fragments only did not allow these previous studies to reach solid conclusions on the true genomic extent of this increase, and to precisely quantify the accumulation of deleterious mutations.

In this study, we combined whole genome sequencing data with a targeted resequencing approach to comprehensively analyze all genes and intergenic sequences within 75kb on either side of the *S*-locus in three sample sets each of *A. halleri* and *A. lyrata.* We compared the observed patterns of polymorphism in these regions with those of 100 unlinked randomly chosen regions used as genomic controls. The use of internal genomic controls provides a powerful way to factor out differences in

demographic histories and/or sample structure. We consistently observed an increase of polymorphism within the first 25-kb region immediately flanking the *S*-locus only, with no detectable effect further along the chromosome. Contrary to predictions from classical models of sheltered genetic load, the putatively deleterious mutations that this narrow region carries do not segregate at higher population frequencies than the overall genomic background, and the relative rate of accumulation of non-synonymous to synonymous variants is also not elevated. These patterns are remarkably consistent across the different sequencing methods we employed and also across the different sample sets we studied in spite of differences in their specific demographic histories. Hence, our data suggest that linkage to one of the strongest known balanced polymorphisms does indeed result in elevated polymorphism, but is not associated with a detectable reduction of selection efficacy.

Results

Sequencing the S-locus flanking regions and control regions in large sample sets

To evaluate the impact of balancing selection on the genomic regions flanking the *S*-locus, we focused on the region where previous studies indicated that the signature of balancing selection was most likely encompassed; i.e. over a maximum of 75 kb on each side of the *S*-locus (Kamau and Charlesworth 2005; Kamau et al. 2007; Ruggiero et al. 2008; Roux et al. 2013). We divided this region in three consecutive non-overlapping windows of 25kb (-25, -50 and -75 kb one one side and +25, +50 and +75 kb on the other side; Fig. 1). Together, the two 25kb windows closest to the *S*-locus contain a total of 11 annotated genes in the *A. lyrata* genome, the next upstream and downstream 25-50kb windows together contain 9 genes, and the most distant 25kb regions contain 13 genes (Hu et al. 2011, Fig 1). To compare these regions to the background level of nucleotide polymorphism, we also included in the analysis one hundred 25kb "control" regions unlinked to the *S*-locus. These control regions were randomly chosen across the *A. halleri* genome and selected to closely match the density of protein-coding sequences and transposable elements found at the *S*-locus flanking regions (proportion of CDS within the interval = 0.23 +/-0.0023; proportion of TEs = 0.28 +/-0.0028, Fig. 1). Because the extreme level of sequence divergence of the non-recombining interval containing the *S*-locus itself precludes mapping of short reads among *S*-haplotypes (Goubet et al. 2012), we excluded this region from further analysis and focused on the flanking regions only (Fig.1).

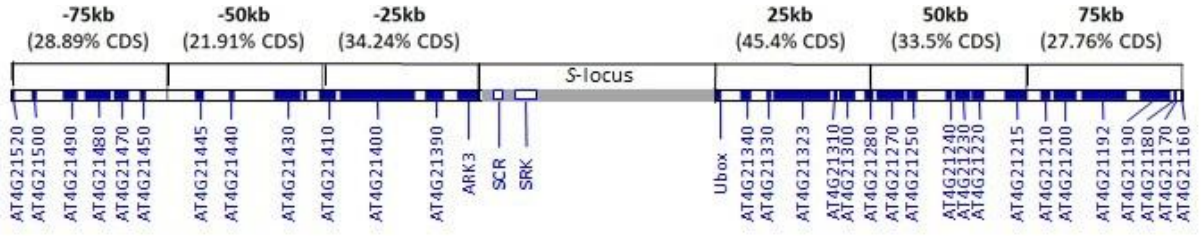

Figure 1: Schematic representation of the S-locus and its flanking regions. *Protein-coding genes in the flanking regions are represented as filled blue rectangles. The genomic regions studied are distributed between positions 9,264,458 and 9,451,731 along chromosome 7 of the A. lyrata genome assembly (Hu et al. 2011). The S-locus (in grey) contains two protein-coding genes only, SRK and SCR (white rectangles) and is flanked by the ARK3 and Ubox genes (in the 5' and 3' directions, respectively). The S-locus region itself was not analysed in the present study. The percentage of CDS in each 25kb window is given on top of the figure.*

To provide a comprehensive picture of the indirect effects of balancing selection, we analysed two closely related species that share the same orthologous SI system and show extensive trans-specific polymorphism at the *S*-locus, *A. halleri* and *A. lyrata* (Castric et al. 2008). In order to evaluate the robustness of our conclusions to different demographic histories, we analysed nucleotide sequence polymorphism data from two natural populations of *A. halleri* (Nivelle, n=25 and Mortagne, n=27 that have been recently introduced in the North of France in association with industrial activities) and two natural populations of *A. lyrata* (Plech, n=18 from the core of the species range and Spiterstulen, n=23 from the edge of the species range, Table S1). To evaluate the robustness of our conclusions to different sampling strategies, we also included samples from more extended geographic regions of *A. lyrata* (North America, n=27 distributed across three distinct populations) and *A. halleri* ssp. *gemmifera* (Japan, n=47 distributed across six distinct populations). For the Nivelle, Mortagne and North American samples, we developed a dedicated sequence capture protocol specifically targeting the control and *S*-locus flanking regions. For the Japan, Plech and Spiterstulen samples, we took

advantage of published whole-genome resequencing datasets, but analysed only polymorphism of the regions included in the capture protocol.

We obtained an average of 59 million reads mapped for the samples sequenced by sequence capture and 1,310 million reads for the WGS samples. After stringent filtering, we were able to interrogate with confidence an average of 960,368 positions in control regions, 28,432 of which were variable and biallelic (3%). As expected, the number of variable sites across the control regions differed among sample sets, reflecting their different demographic histories (Table 1). The *A. halleri* Japan sample set was the least polymorphic of all, with observed heterozygosity *Ho*=0.00096, nucleotide polymorphism π=0.00128 and a proportion of polymorphic sites equal to 0.0088. At the other extreme, the *A. lyrata* Plech population was the most polymorphic, with *Ho*=0.00646, π=0.00758 and a proportion of polymorphic sites equal to 0.0299. These estimations of the background level of nucleotide polymorphism in each sample set were used as internal genomic controls for the study of polymorphism in *S*-flanking regions, where we were able to interrogate an average of 74,866 sites, containing 3,225 variable biallelic positions (4%).

Detection of the footprints of ancient balancing selection on the S-linked regions

Based on these comprehensive polymorphism data, we combined different approaches to characterise the impact of balancing selection. As a first step, we excluded the potential confounding effect that would arise if mutation rates were higher in the *S*-flanking regions than in the control regions. Comparison of the mean levels of divergence between *A. thaliana* and *A. lyrata r*eference genomes showed no evidence for increased divergence in the windows flanking the *S*-locus, as would be expected if they tended to accumulate more mutations per unit time. Instead, these two regions tended to indicate a slight reduction rather than an increase of divergence, albeit not a significant one (Fig. S1).

Next, we followed a multilocus Hudson-Kreitman-Agade (HKA) approach to compare nucleotide polymorphism within *A. lyrata* or *A. halleri* sample sets, taking into account divergence from the outgroup *A. thaliana* between the 33 *S*-flanking genes and 67 randomly chosen control genes. The multilocus HKA test showed a highly significant departure from neutral expectation (mean X^2=821, *P*=0, df=33; Table S3), indicating that the two categories of loci differed in their relative patterns of polymorphism. The mean estimate of the selection parameter for the 33 *S*-flanking genes was above one (mean *k*=1.46 ; Fig. 2, Table S4), indicating higher polymorphism of the *S*-flanking genes compared with the control loci. *k* also tended to increase toward the S-locus, although the magnitude of this pattern varied across samples from different regions and differed between the 5’ and 3’ flanking regions.

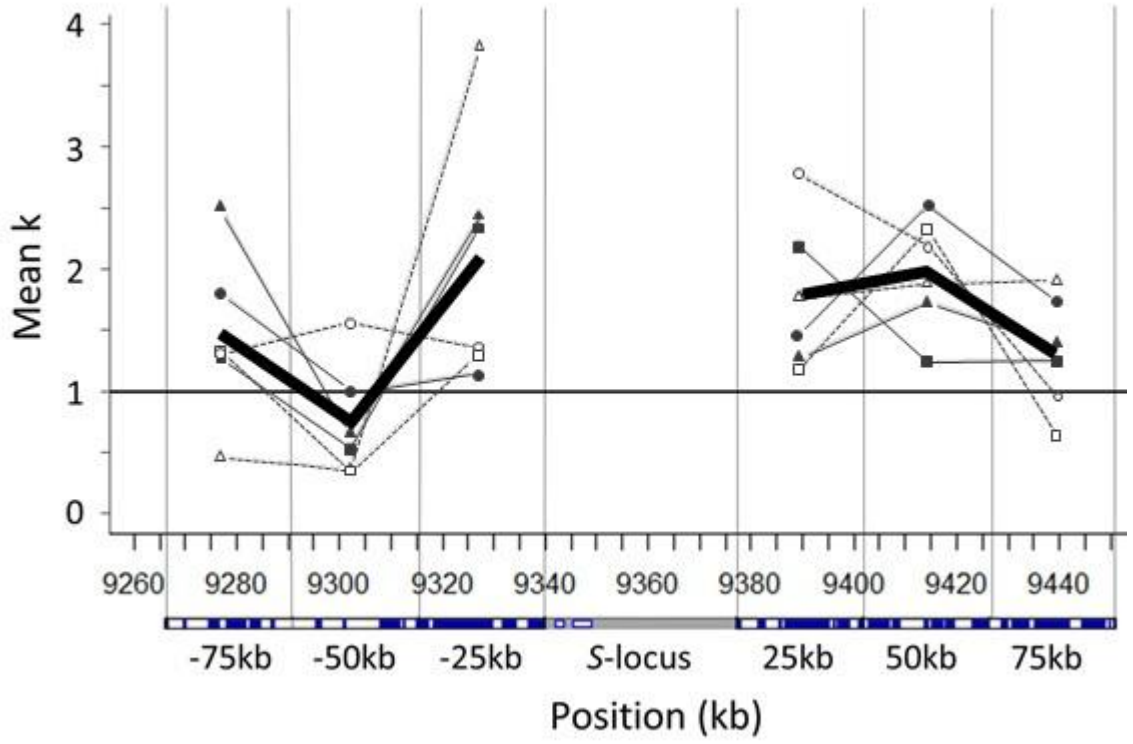

Figure 2: Variation of the mean selection parameter (k) obtained for genes in the S-flanking regions. *The thick solid line represents the mean value of k obtained across the six sample sets. The thin solid lines represent the A. lyrata sample sets (square=Plech, circle=Spiterstulen, triangle=North America). The dashed lines represent the A. halleri sample sets (open square=Japan, open circle=Nivelle, open triangle=Mortagne). The threshold value of 1 (no selection) is represented by the horizontal black line.*

We then used the new powerful approach of Cheng and Degiorgio (2020) that is robust to demographic variations to detect distortions of the site frequency spectrum along the chromosomal fragments and determine the maximum likelihood position of putative targets of balancing selection. We found strong signals of balancing selection in some of the control regions, specifically on chromosomes 3 and 4, but in most cases they were not consistent across all sample sets (Fig. 3). In contrast, the *S*-locus flanking regions carried consistent signals of balancing selection that were detected across all sample sets, even though they were not always the most extreme (Fig. 3). The exact position of the peak detected in the 25kb windows around the *S*-locus varied between sample sets (Fig. S2). Overall, these results provide evidence for a strong and consistent footprint of balancing selection on the regions flanking the *S*-locus.

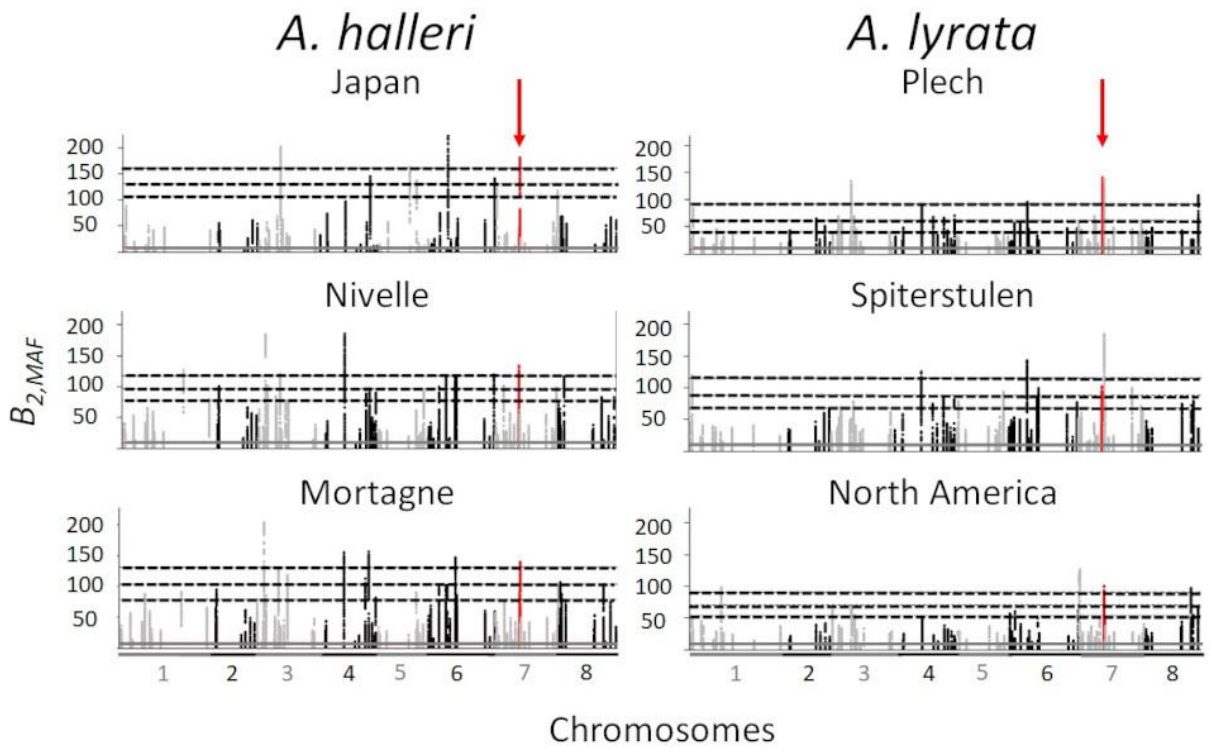

Figure 3: Manhattan plots for signals of balancing selection ($B_{2,MAF}$ scores) in the 100 control regions and the S-locus region (red dots and arrows) along the A. lyrata genome. *Chromosomes 1-8 of the A. lyrata genome, with the 100 control regions distributed on successive chromosomes represented by an alternation of grey and black dots. The horizontal grey solid line represents the median $B_{2,MAF}$ scores across SNPs in the genome, and the black horizontal dashed lines represent the top 5, 2.5 and 1% percentiles.*

The increased polymorphism of the S-locus flanking regions is mostly caused by an increase of the proportion of polymorphic sites

Then we sought to describe in detail how polymorphism of the *S*-locus flanking regions compared with the genomic background. To do so, we compared the values of several summary statistics of polymorphism from the *S*-locus flanking regions to their distribution across the 100 control regions. Specifically, we compared the nucleotide polymorphism (π), the observed heterozygosity (*Ho*), the mean frequency of the minor allele (*MAF*) and the proportion of polymorphic sites (number of observed polymorphic sites divided by the total number of sites considered). Significant excess of polymorphism statistics as compared to control regions was found for almost all sample sets in the two 25kb windows immediately flanking the *S*-locus for *Ho* (by a factor 1.7-fold in Plech to 6.4-fold in Japan, Fig. 4), π (by a factor 1.6-fold in Plech to 5.8-fold in Japan, Fig. 5), and the proportion of polymorphic sites (by a factor 1.6-fold in Plech to 3.9-fold in Japan, Fig. 6, Table S5). We observed only two exceptions to this pattern. For the +25kb window in Plech and the -25kb window in Spiterstulen, *Ho* was not significantly higher than in control regions. In stark contrast, the second and third consecutive 25kb windows on either sides of the *S*-locus generally showed no excess polymorphism as compared to control regions in any sample set, with the exception of the Spiterstulen population, where the -75kb and +50kb windows had a slightly higher proportion of polymorphic sites (Fig. 6).

To verify that the effect we observed was not specific to the particular window size we chose, we used Linear Models to test whether *Ho*, π and *MAF* of individual sites of the *S*-flanking regions

(considered as response variables) declined when physical distance increased away from the *S*-locus. A highly significant negative effect of the distance to the S-locus was observed overall (Table S6), confirming the effect of proximity to the *S*-locus on polymorphism of sites in the flanking regions.

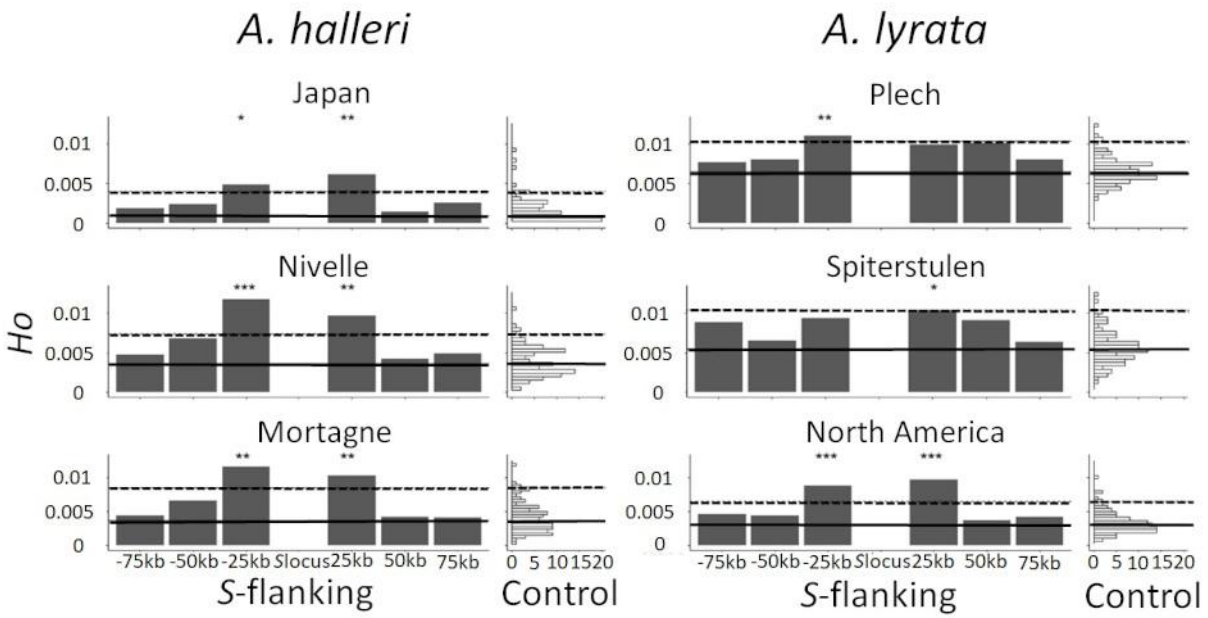

Figure 4: Mean Ho around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of Ho obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of Ho mean in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

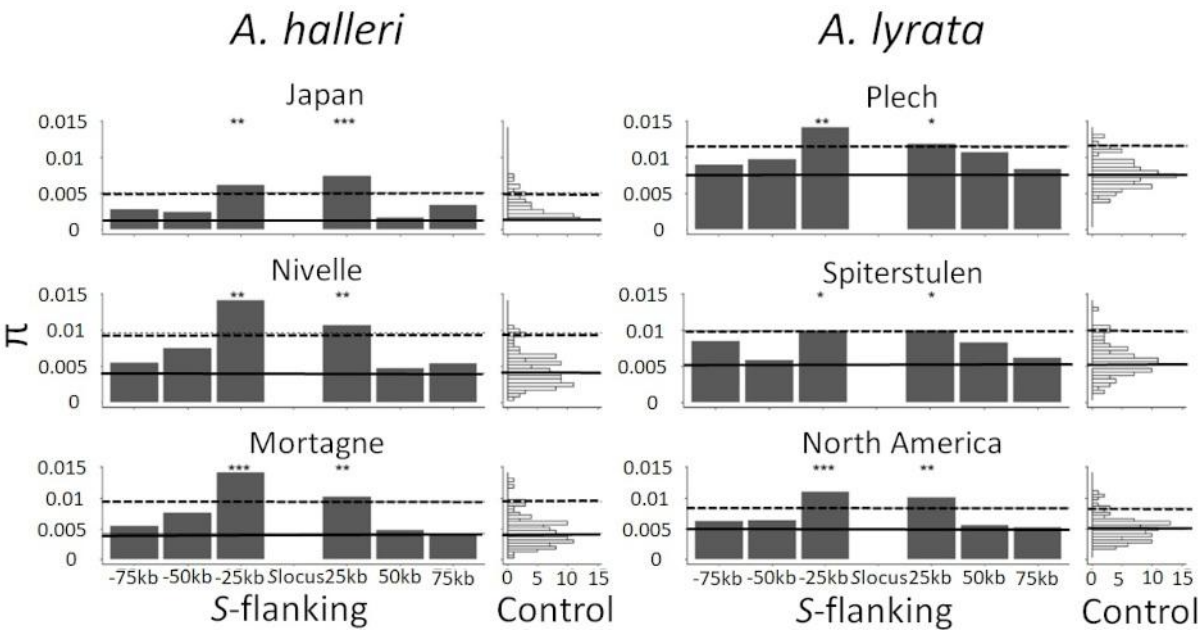

Figure 5: Mean π around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of π obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of π mean in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

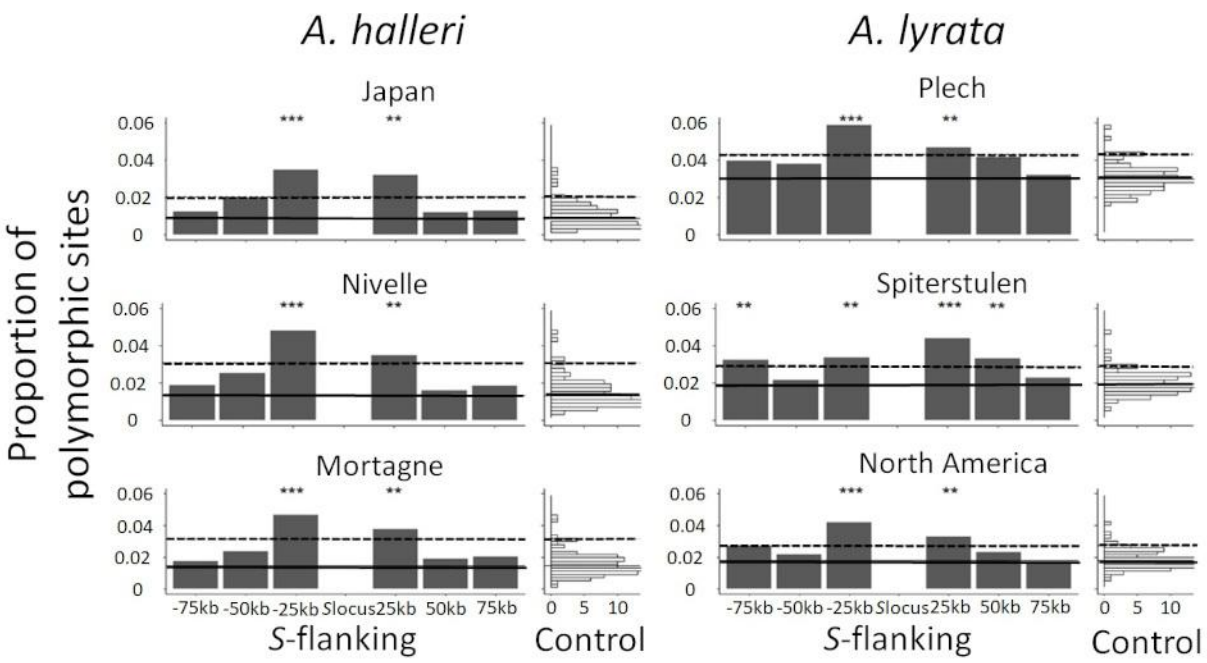

Figure 6: Proportion of polymorphic sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of polymorphic sites in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

An increase of polymorphism can be explained by 1) an increase of the frequency of mutations at polymorphic sites within each sampling region, 2) an increase of the proportion of polymorphic sites, or 3) a combination of both. Noting that the proportion of polymorphic sites increased in the *S*-locus flanking region compared with the control regions, we wanted to test whether the allele frequencies at those polymorphic loci were also affected. To do that, we reiterated the analyses above, but considering the polymorphic sites only. We found no difference with respect to control regions when computing the *Ho*, π and *MAF* statistics on polymorphic sites only, with a single exception for *Ho*, which showed a higher value in the +25kb window flanking the *S*-locus in the North American sample set from *A. lyrata* (Fig. S3). Hence, the higher polymorphism detected in the *S*-locus flanking region is essentially due to an increase of the proportion of polymorphic sites rather than to a shift in the allele frequency spectrum. This is confirmed by the absence of deviation of the Tajima's D statistic compared with control regions (Fig. 7). Overall, our results thus show elevated nucleotide polymorphism at the *S*-locus region, which is mostly caused by a larger number of polymorphic sites rather than by an increased frequency at which the polymorphic sites segregate.

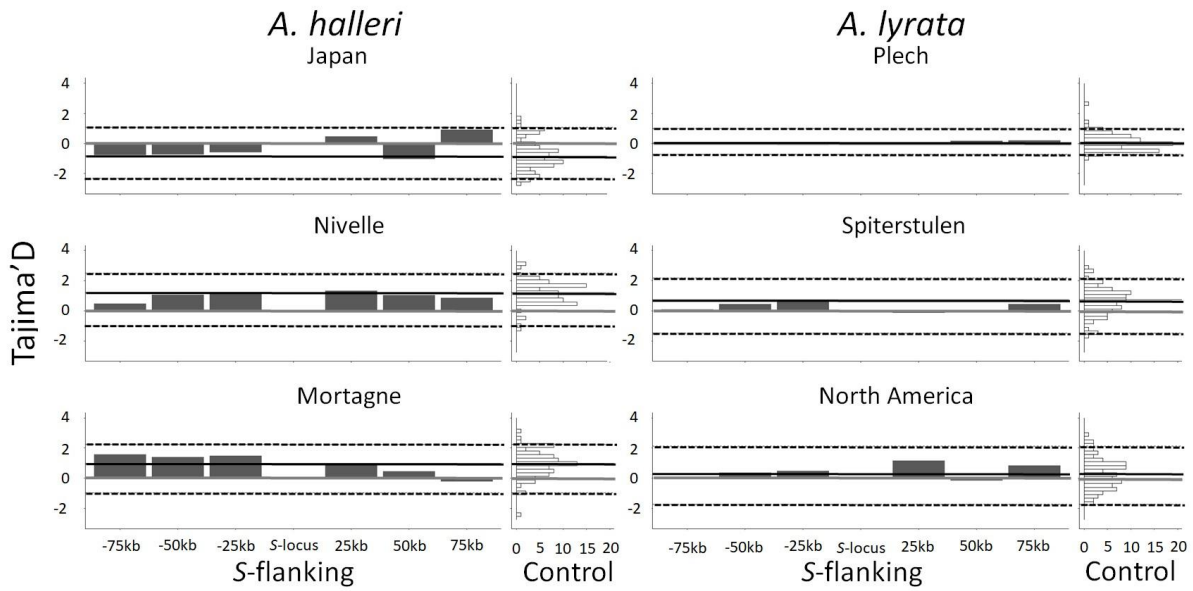

Figure 7: Tajima's D around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean of Tajima's D obtained in S-flanking windows of 25kb. The distributions (count) in the 100 control regions are represented by a histogram (right). The 97.5% and 2.5% percentiles of the distributions are represented by dashed lines. The median value of the distribution in control regions is represented by black lines.*

Higher density of putative deleterious mutations in the S-locus flanking region

To determine if the indirect effect of balancing selection described above is associated with the accumulation of a "sheltered" genetic load, we examined the accumulation of 0-fold degenerate sites only, assuming that the majority of amino-acid polymorphisms are deleterious to some extent (Eyre-Walker and Keightley. 2007). Like for total polymorphism above, we observed an increase of polymorphism at 0-fold degenerate sites in the *S*-locus flanking regions as measured by *Ho*, π or *MAF* when compared to control regions (Fig. S4 and S5 for *Ho* and π respectively), which is mostly due to an increased proportion of polymorphic sites in the first 25kb surrounding the *S*-locus (Fig. 8, Table S7, Fig. S6). The magnitude of this increase as compared to the genomic background ranged from 1.92 to 3.66-fold across sample sets (Table S7). A GLM restricted to 0-fold sites confirmed the effect of proximity to the *S*-locus (Table S6).

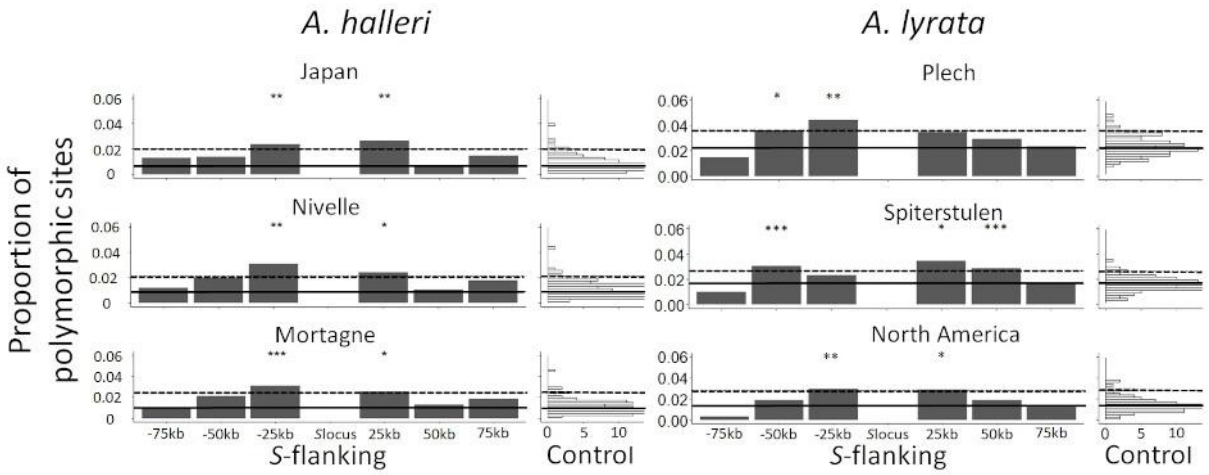

Figure 8: Proportion of polymorphic sites among 0-fold degenerate sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of polymorphic sites in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

We further compared the ratio of π between 0-fold and 4-fold degenerate sites. If balancing selection decreased the efficacy of the purge of deleterious mutations, we expect an elevation of the $\pi_{0\text{-fold}}/\pi_{4\text{-fold}}$ ratio in the *S*-locus flanking regions. However, we found no evidence for such an increase in the *S*-locus flanking regions as compared to the control regions (Fig. 9), with the exception of the -50kb window in the North American sample set of *A. lyrata*.

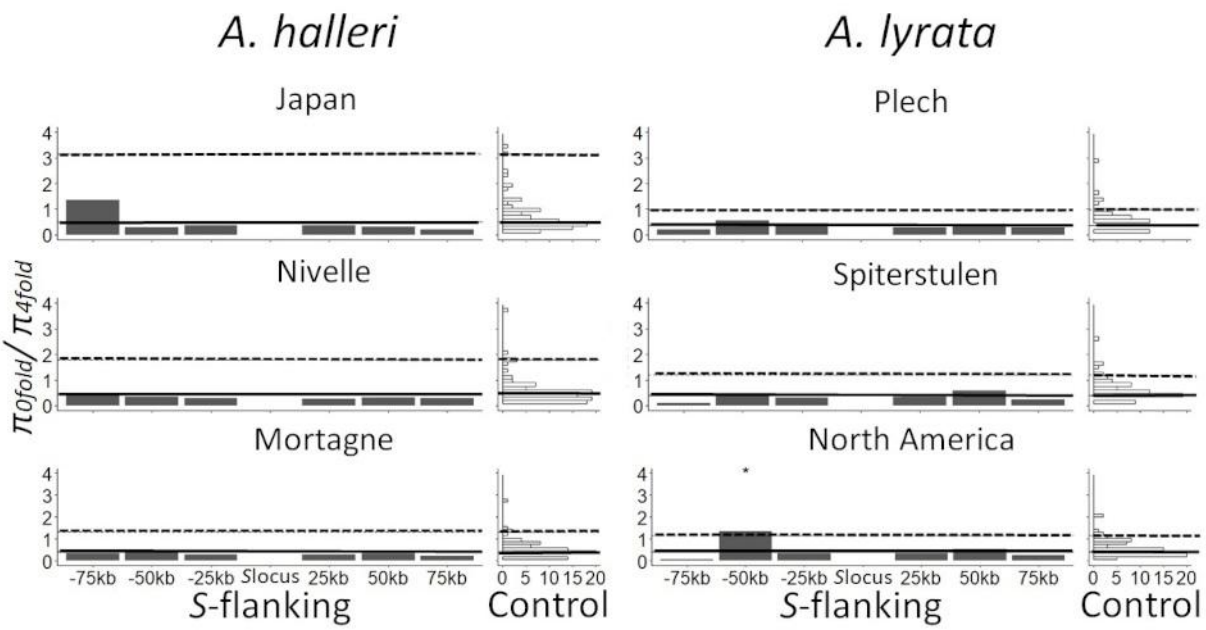

Figure 9: $\pi_{0\text{-fold}}/\pi_{4\text{-fold}}$ ratio around the S-locus and across the control regions from throughout the genome. *The bars represent the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of polymorphic sites in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

Finally, to define more precisely the genes that are affected by the sheltered genetic load, we explored the functional annotations of the genes contained in the *S*-flanking regions (Table S9). The -25kb and +25kb regions, where we found an effect of linkage to the *S*-locus, contained only eleven annotated genes in the *A. lyrata* genome. Four of these genes are receptor-like serine/threonine-protein kinases (*AT4G21410*, *AT4G21400*, *AT4G21390*, *AT4G21380*/*ARK3*), one is an ubiquitination protein (*AT4G21350*/*Ubox*), two are transcription factors (*AT4G21340*, *AT4G21330*), one a peptidase (*AT4G21323*), one a transmembrane protein (*AT4G21310*), one a tetratricopeptide repeat (TPR)-like superfamily protein (*AT4G21300*), and a last one is a subunit of Photosystem II (*AT4G21280*).

Discussion

Elevated polymorphism but no decreased efficacy of purifying selection in the linked region

The segregation of deleterious mutations linked to the S-locus is expected to have important consequences for mating system evolution, affecting the maintenance of SI itself (Porcher & Lande 2005, Gervais et al. 2014) and the conditions for diversification of SI lineages (Uyenoyama 2003). Our analysis across several replicate geographic regions in two closely related species is the first comprehensive genomic study to reveal the extent of the *S*-locus associated sheltered genetic load in a plant genome. In line with theoretical predictions from Schierup et al. (2000), we show that the genomic region directly adjacent to the *S*-locus presents consistent signals of linked balancing selection and that polymorphism is elevated as compared to the genomic background, by a factor up to 5.8. This observation is important because Contrary to this expectation, however, we found no evidence that the efficacy of purifying selection is decreased for the mutations linked to the *S*-locus. The linked regions do accumulate more mutations than the control regions, but they segregate at population frequencies that are indistinguishable from those at control genes, and the $\pi_{0\text{-}fold}/\pi_{4\text{-}fold}$ ratio is also unchanged. This accumulation of mutations is not due to an increased mutation rate in these genes because divergence from *A. thaliana* rather tends to be reduced compared with the genomic background. This latter observation is in line with the repeated introgression observed at the *S*-locus between *A. halleri* and *A. lyrata* (Castric et al. 2008), causing divergence between the two species to be more recent for the *S*-alleles than for the genomic background. Hence, the main factor causing the elevated polymorphism seems to be the deeper coalescence time among allelic lineages, allowing the accumulation of both neutral and deleterious mutations.

One possible explanation for the fact that the distribution of allele frequencies in the linked regions are indistinguishable from those at the control regions can be related to the model proposed by Takahata (1990). This model shows that the genealogical relationships among distinct *S*-allele lineages under gametophytic SI are expected to be identical to those of neutral genes, except they are expanded by a scaling factor f_S. If the number of *S*-alleles in the studied sample sets was extremely large and every chromosome we sampled corresponded to a different *S*-allele lineage, then the only difference between the genealogies of sequences around the *S*-locus and those in the control regions would lie in the different time scales. Testing this hypothesis would require phasing the flanking sequences with each *S*-allele to obtain haplotypes. However, based on the published estimates of *S*-allele frequencies in four of the population samples studied here (the Nivelle population of *A. halleri*, Llaurens et al. 2008; the Japanese *A. halleri* sample set, Genete et al. 2020; and the Plech and Spiterstulen populations of *A. lyrata*, Takou et al. 2021), it is already clear that randomly sampled chromosomes would be very unlikely to systematically correspond to distinct *S*-allele lineages. More importantly, the Takahata (1990) model is based on a gametophytic SI, while the *S*-locus of the Brassicaceae functions as a sporophytic SI, such that dominance/recessivity interactions modulate the selective effect between *S*-alleles (Schierup et al. 1997 ; Billiard et al. 2006). Whereas in gametophytic SI all *S*-alleles are expected to segregate at equal population frequencies and hence sharply depart from the neutral site frequency spectrum, in sporophytic SI, the recessive *S*-alleles are driven to high population frequencies, whereas dominant *S*-alleles remain relatively rarer. This asymmetry may diminish the contrast between the *S*-locus and the genomic background. Finally, in sporophytic SI the recessive *S*-alleles can form homozygous combinations, allowing some purging of deleterious variants, which could help to explain why we detected no apparent decrease of the efficacy of purifying selection. Developing new theoretical models taking into account the structure of the dominance hierarchy between *S*-alleles will be necessary to fully understand the effect of linkage

to a SSI locus. Strikingly, our results are almost the mirror image of the pattern seen at the human *MHC* by Lenz et al (2016), where the elevated polymorphism of the genes in the linked genomic region is not due to deleterious mutations being more abundant, but to the fact that each of them tends to segregate at higher population frequency than the genomic background. For the *MHC*, balancing selection is believed to be driven by pathogen mediated selection (although the exact mechanism remains controversial ; see Spurgin and Richardson 2010), which is very different from the negative frequency dependent selection maintaining diversity at the *S*-locus. Dominance interactions between the balanced allelic lineages are also not expected at the *MHC*. It is currently not clear which specific feature of the balancing selection mechanism acting at the *S*-locus and at the *MHC* causes these sharply distinct genomic signatures.

A limited extent of the footprint along the chromosome

Because SI is arguably one of the most intense forms of long-term balancing selection, it could *a priori* be considered a favourable case to detect the footprints of its genomic signature. Yet, a salient feature of our results is the limited extent of the genomic region over which the effect of linkage to the *S*-locus can be detected, essentially spanning over the immediately 25kb flanking regions only. As compared to other strongly balanced polymorphisms such as sex-determining regions of sex-chromosomes or mating-type loci in fungi, which typically occupy large chromosomal portions, the *S*-locus itself occupies only 30-110 kb in *A. halleri* and *A. lyrata*, and only includes the genes involved directly in the SI recognition machinery (Guo et al. 2011, Goubet et al. 2012). The large chromosomal regions associated with sex-determining regions of sex-chromosomes are believed to result from the progressive extension of successive inversions that can ultimately capture a large number of genes, eventually expanding across most of the length of a chromosome (Charlesworth et al. 2005 ; Otto et al. 2011). The classical models for this process entailed sexual antagonism, whereby the inversions selectively fix mutations that are favorable in one sex in the appropriate genetic combination, and the accumulation of deleterious mutations follows from the action of Müller's ratchet once recombination has ceased (Rice, 1987). However, recent models have shown that the successive fixation of inversions can still take place even in the absence of sexual antagonism, as a result of the effective masking of recessive deleterious mutations accumulated in the flanking regions of the sex-determining loci (Jay et al. 2021, Lenormand and Roze, 2022). The reason why the *S*-locus flanking regions do not undergo this process of inversion, recombination arrest and degeneration in spite of the strong balancing selection it experiences, may be due to the limited size of the region upon which polymorphism is affected and the lack of effective sheltering, preventing the region from extending further efficiently. It would be interesting to investigate genomic patterns of sheltered load in other SI systems, in particular in systems with a collaborative non-self recognition mechanism, such as observed in Solanaceae, as the size of the S-locus is much larger in those systems and apparently includes functional genes unrelated to SI (Wu et al. 2020).

The peak of polymorphism is robust to sample heterogeneity

Overall, in spite of the different demographic histories, sampling structures and sequencing technologies used, we find qualitatively very similar results across species and sampling regions. This provides strong support for the idea that the contrasts we observed between the *S*-locus and the genomic background are robust to these factors. The difference of demographic histories between populations was expected to modify the levels of polymorphism for control and *S*-flanking regions (see e.g. Fijarczyk and Babik, 2015). The *A. lyrata* Plech population presented the highest level of polymorphism in the genomic background, in line with this population being at the core of the range of European *A. lyrata* (Takou et al. 2021). The *A. lyrata* Spiterstulen population had lower

polymorphism in the genomic background, in line with the strong reduction of effective population size it experienced during the colonisation of Norway within the past 100,000 years (Takou et al. 2021). The Nivelle and Mortagne populations of *A. halleri* have recently colonised the north of France during the last century from ancestral German populations (Pauwels et al. 2005). In spite of their composite origin (multiple populations), the North American *A. lyrata* and the Japanese *A. halleri* sample sets had the lowest polymorphism, possibly as the results of major demographic bottlenecks they experienced in the course of post glacial colonisations, at least in *A. lyrata* (Clauss and Mitchell-Olds, 2006; Ross-Ibarra et al. 2008). Although our strategy was not designed to interpret the quantitative differences observed among sample sets, we note that the relative elevation of polymorphism at the *S*-locus seems to be more pronounced in sample sets with lower baseline levels of diversity across the genome. This is consistent with the observation that the *S*-locus appears to be less sensitive to demographic effects than the genomic background (Takou et al. 2021).

The strongest increase of polymorphism at the *S*-locus as compared to the genomic background across all samples was found in the two sample sets composed of multiple populations (*A. halleri* from Japan and *A. lyrata* from North America). Thus, even though population stratification is expected to modify the site frequency spectrum, it did not prevent the detection of increased polymorphism in the flanking regions due to balancing selection at the S-locus. We note that Ruggiero et al. (2008) and Roux et al. (2013) similarly used regional samples, and also detected an excess of polymorphism in genes flanking the *S*-locus, albeit with a much lower level of resolution. Some North American populations have experienced a loss of SI, and have shifted to partial selfing. Selfing is generally expected to reduce the effective rate of recombination, and might thus expand the footprint of balancing selection (Wright et al. 2008). On the other hand, however, selfing may have been expected to relax the intensity of balancing selection on the *S*-locus. The populations considered here were specifically chosen because they are predominantly outcrossing (Foxe et al. 2010), so we expect this effect to be minor.

The nature and number of mutations causing the load

The absence of protein-coding genes within the *S*-locus region itself beyond those directly involved in the SI machinery (Goubet et al. 2012) suggests that mutations causing the sheltered load are likely to lie in the flanking region rather than in the *S*-locus region itself. While Stone (2004), Stift et al. (2013) and Llaurens et al. (2009) showed that a sheltered load linked to the *S*-locus could be detected at the phenotypic level, our study provides the first genomic demonstration of an accumulation of potentially deleterious mutations in the *S*-linked regions. Identifying more precisely the mutations causing the sheltered load would still require fine mapping, but our work suggests that they are most likely to be found in very close proximity to the *S*-locus (<25kb). The phenotypic traits on which the load was documented varied across the different studies (seed dormancy in *Papaver rhoea*s, Lane and Lawrence 1995; seed survival in *Solanum carolinense,* Stone 2004; leaf development and juvenile survival in *A. halleri,* Llaurens et al. 2009 ; juvenile survival in Stift et al. 2013; horticultural traits in *Rosa,* Vieira et al. 2021), as would be expected given that the *S*-locus lies in different genomic environments in distant species, and given that the deleterious mutations are expected to hit the different flanking genes in a random manner.

Kawabe et al (2006) speculated that the low number of genes in the *S*-genomic region is probably not high enough for a large sheltered load to have an impact on fitness compared to the overall genomic load. Here we show that the genomic interval whose polymorphism is affected by linkage with the *S*-locus comprises eleven genes. Several of these genes were previously shown to be associated with deleterious phenotypic traits in *A. thaliana*, or to be ancestral genes shared with the sister family

Cleomaceae *(Ubox, ARK3* and *AT4G21390*, Cheng et al. 2013). For instance, mutants of the transcription factor *AT4G21330* exhibit abnormal anther morphology at the beginning of stage 4 (Zhang et al. 2006) and the gene *ARK3* has been shown to be implicated in root development (Dwyer et al. 1994). Hence, it is clear that some of these genes have important functions, and are obvious candidates for the future dissection of the genetic architecture of the mutation load sheltered by the *S*-locus.

A limitation of our population genetics approach is that it was designed to detect the collective accumulation of mutations rather than individual high-impact mutations. However, it is possible that a low number of high-impact mutations, rather than a collection of small-effect mutations, are causing the load. Indeed, the selective dynamics of lethal mutations *vs.* slightly deleterious mutations can be sharply different (e.g. Clo et al., 2020), and in the latter case finely dissecting the load at the genetic level will remain challenging. In addition, while our sequence capture approach also includes the intergenic sequences, we quantified the load based on coding sequences only. Previous studies demonstrated that polymorphism on intergenic regions could be under purifying selection (Mattila et al. 2019 for an example in *A. lyrata*), so it is also possible that besides the coding sequences, mutations in intergenic sequences contribute to the load, hence making our estimation of the sheltered load an underestimate. Another limitation of our work is the focus on SNPs, while structural variants may also have strong deleterious effects. Long-read sequencing would now be required to achieve a more detailed analysis of these types of polymorphisms. A final limitation of our work is that theoretical models of the effect of SI on the flanking regions have assumed a gametophytic SI system (Schierup et al. 2000), while the SI system in Arabidopsis is sporophytic. An exciting next step will be to compare the number and identity of deleterious mutations associated with dominant *vs.* recessive alleles, a task that will require phasing polymorphisms and is beyond the scope of the present study.

Material and methods

Source plant material

We worked on natural accessions from two closely related species, *A. halleri* and *A. lyrata*, each represented by samples from three regions named Japan, Mortagne (France) and Nivelle (France) for *A. halleri*, and Plech (Germany), Spiterstulen (Norway) and North America for *A. lyrata* (Table S1). For the Japan, Spiterstulen and Plech samples, we used available whole genome sequencing (WGS) data obtained by Kubota et al. (2015) and Takou et al. (2021). The Japan sample set was composed of 47 individuals of *A. halleri*, subsp. *gemmifera* originating from six different populations (17 individuals from Fujiwara, 17 from Ibuki, 2 from Inotani, 3 from Itamuro, 4 from Minoo and 4 from Okunikkawa; Kubota et al. 2015), the Spiterstulen (26 individuals) and Plech (23 individuals) sample sets were from single locations (Takou et al. 2021). For the three other sample sets, we collected individuals and developed a dedicated targeted enrichment capture approach to sequence the genomic regions of interest. The North american sample set of *A. lyrata* was composed of 26 individuals from three highly outcrossing populations from the Great Lakes region, named IND (Indiana Dunes National Lakeshore in Michigan, n=8), PIN (Pinery Provincial Park in Ontario, n=10) and TSS (Tobermory Provincial Park in Ontario, n=8) (Foxe et al. 2010). We collected 25 individuals from the Nivelle population (50°47'N, 3°47'E, France) and 27 individuals from the closely related Mortagne population (50°47'N, 3°47'E, France). In total, we complemented the 88 individuals with whole genome data with an additional 78 newly sampled individuals that we sequenced with the targeted sequence capture approach described below.

S-locus flanking regions and control regions

To evaluate the effect of balancing selection on the *S*-locus, we developed an original approach based on the comparison between the patterns of polymorphism of the two flanking regions on either side of the *S*-locus to those of a set of 100 randomly chosen control regions. The *S*-locus region can be poorly represented in whole-genome assemblies, so we first sequenced them using two *A. halleri* BAC clones that we newly obtained following the approach of Goubet et al. (2008) from a BAC library constructed from a mixture of several *A. halleri* individuals from Italy. These two BAC clones were chosen so as to cover entirely the 5' and 3' regions on either side of the *S*-locus (BAC clones 37G17 and 21E5, respectively ; 10.6084/m9.figshare.16438908). We computed the proportion of coding sequences (CDS) and transposable elements (TEs) on the first 75kb sequences immediately flanking the *S*-locus on these two BAC clones (but excluding the non-recombining region within the *S*-locus itself), and we used these two statistics to select a set of matched control regions from across the *A. halleri* genome (Legrand et al. 2019). To do this, we used bedtools (Quinlan and Hall, 2010) to randomly select 25-kb contiguous genomic intervals. The genomic intervals were retained if their density of CDS and TEs closely matched that of the actual *S*-locus flanking regions (within 10%). If the proportions of CDS and/or TEs departed from those values, the region was discarded and a new region was picked until a total of 100 genomic intervals was included. The genomic coordinates of the control regions are given in Supplementary Table S9, and their sequences in fasta format are available at 10.6084/m9.figshare.16438908. Because the control regions were defined initially on the *A. halleri* reference, we used sequence similarity (using YASS, Noé and Kucherov 2005) to identify orthologous regions along the *A. lyrata* genome.

Library preparation, sequence capture and sequencing

For the 78 newly sequenced individuals, we purified DNA from 15 mg of dried leaves of each sample with Chemagic beads (PerkinElmer) following Holtz et al (2016), using the manufacturer's instructions but with an additional Agencourt AMPure beads (Beckman) purification. DNA was quantified by Qubit and 50 ng of DNA was fragmented mechanically with Bioruptor (Diagenode) to obtain fragments of around 300bp, which we verified using a BioAnalyzer (Agilent) with a DNA HS chip. We prepared indexed genomic libraries using the Nextflex Rapid DNA Seq kit V2.0 (PerkinElmer) using the manufacturer's instructions. Briefly, extremities of sequences were repaired and tailed, ligated with universal adaptors P5/P7 containing multiplexing unique dual index (PerkinElmer), and amplified by five cycles of PCR. We then selected fragments between 150 and 300pb with AMPures beads and pooled all libraries in equimolar proportions.

The pooled libraries then proceeded to a sequence capture protocol using the MyBaits v3 (Ann Arbor, Michigan, USA) approach. Briefly, 120bp RNA probes were designed by MyBaits and synthesised to target the complete set of one hundred 25kb control regions as well as the 75kb regions flanking the *S*-locus on either side, with an average tiling density of 2X (a total of 48,127 probes). In addition to the *S*-locus flanking regions and the control region, the capture array also contained a set of additional probes that were not used in the frame of the present project but are detailed in Supplementary Information. The indexed genomic libraries were hybridised to the probes overnight at a temperature of 65°C, and were finally sequenced by Illumina MiSeq (300pb, paired-end) by the LIGAN-MP Genomics platform (Lille, France).

Read mapping and variant calling

Raw reads from sequence-capture or WGS datasets (see Supplementary table S1) were mapped onto the complete *A. lyrata* reference genome (V1.0.23, Hu et al. 2011) using Bowtie2 v2.4.1 (Langmead and Salzberg, 2012). File formats were then converted to BAM using samtools v1.3.1 (Li et al. 2009) and duplicated reads were removed with the MarkDuplicates program of picard-tools v1.119 (http://broadinstitute.github.io/picard). These steps were performed by the custom Python script sequencing_genome_vcf.py available in https://github.com/leveveaudrey/analysis-of-polymorphism-*S*-locus. We retained only reads which mapped to the *S*-locus flanking or control regions. For the sake of consistency, we followed the same procedure for samples sequenced by WGS. Biallelic SNPs in these regions were called using the Genome Analysis Toolkit v. 3.8 (GATK, DePristo et al. 2011) with the option GVCF and a quality score threshold of 60 using vcftool v0.1.15 (Danecek et al. 2011). For each sample independently, we computed the distribution of coverage depth across control regions using samtools depth (Li et al. 2009). We excluded sites with either less than 15 reads aligned or coverage depth above the 97.5 % percentile, as the latter are likely to correspond to repeated sequences (e.g. transposable elements or paralogs). Sites covered by at least 15 reads but containing no SNP were considered as monomorphic. To exclude the possibility of spurious heterozygosity induced by the presence of paralogs, we verified that no SNP was systematically heterozygous across all individuals. The final number of sites in each sample set is summarised in Tables 1 and 2.

Footprints of balancing selection

For each sample set, we first evaluated the distribution of the $B_{2,MAF}$ statistic across all SNPs, which was designed to capture the distortion of the site frequency spectrum along chromosomes caused by linkage to a site under balancing selection (Cheng and Degiorgio 2020). We then compared the $B_{2,MAF}$

distribution in control regions with the *S*-flanking regions, and considered a significant difference when the mean $B_{2,MAF}$ value was outside the 95% percentile of the distribution in control regions.

To control for a possible difference in mutation rates between genes in the *S*-locus flanking regions and genes in the control regions, we then compared their pattern of molecular divergence between *A. lyrata* and *A. thaliana* (TAIR10 genome) at the sites retained for the polymorphism analysis (i.e. having passed the coverage filter). We identified orthologs as the best hits in the *A. thaliana* genome using YASS, retaining alignments with a minimum e-value of 0.01 and an identity above 70%. Pairs of sequences were then aligned with clustalOmega (Sievers et al. 2011) and the proportion of divergent sites was determined using a custom Python script (https://github.com/leveveaudrey/analysis-of-polymorphism-*S*-locus).

We further compared the ratio of within-species polymorphism to between-species divergence (Hudson et al. 1987) using the multilocus maximum likelihood HKA framework developed by Wright and Charlesworth (2004) and available at https://github.com/rossibarra/MLHKA. The algorithm is currently limited to only one hundred genes, so we tested the 33 *S*-locus flanking genes and a randomly chosen subset of 67 control genes. Specifically, we compared a model with free mutation at each locus and no selection against a model with free mutation but where each of the 33 *S*-locus flanking genes are allowed to have their own selection coefficient (*k*). This parameter corresponds to the relative increase of polymorphism of the *S*-linked genes compared to genes in the control regions, taking into account differences in divergence between *A. lyrata* and *A. thaliana* across loci. We used a log-likelihood ratio test with 33 degrees of freedom to compare the likelihood of these two nested models. Chain length was set to 100,000 and separate analyses were performed for each sample set independently.

Decomposing the signals of balancing selection

We then decomposed the signal of balancing selection across the *S*-locus flanking regions into a series of elementary statistics. For each site, we estimated the observed heterozygosity (*Ho*) as the number of observed heterozygous genotypes divided by the number of individuals in the dataset, and the minor allele frequency (*MAF*). We calculated π at each position using the vcftools –site-pi option (Danecek et al. 2011). When a position of the *A. lyrata* genome was covered but not polymorphic, the *Ho*, *MAF* and π statistics were set to 0. For each statistic, we binned SNPs flanking the *S*-locus into 25kb intervals and compared the distribution of the mean value obtained for sites within non-overlapping windows of 25kb in the *S*-locus flanking regions with the distribution of the mean obtained across the 100 control regions. Finally, we used Linear Models on all the samples cumulated to test for a linear correlation between the exact distance of each SNP to the *S*-locus along the chromosome and each of the polymorphism statistics listed above with the populations as a random effect in LM. Finally, deviation from neutrality was also tested using Tajima's D for each region of 25kb around the S-locus, for which an excess of intermediate frequency polymorphisms suggests the presence of balancing selection (positive values of D), using the vcftools option –TajimaD (Danecek et al. 2011).

Quantifying the sheltered load of deleterious mutations

To determine the extent to which the *S*-locus flanking regions accumulate deleterious mutations, we first reiterated the same analysis with the previous parameters (*Ho*, *MAF*, π), but for the 0-fold degenerate sites only (determined using the script NewAnnotateRef.py; Williamson et al. 2014). We assumed that all nonsynonymous changes are deleterious. Because all mutations at 0-fold degenerate

sites alter the sequence of the encoded protein, we assumed that these mutations are deleterious (neglecting the rare cases where balancing selection could favour amino acid changes). In contrast, mutations at the 4-fold degenerate sites never alter the encoded amino acid, so we used them as neutral references. For the sake of simplicity, we discarded mutations on the 2- or 3-fold degenerate sites.

Acknowledgements

This work was supported by a grant from the France-Berkeley Fund (to VC and Rasmus Nielsen); the European Research Council (NOVEL project, grant number 648321); and the Agence Nationale de la Recherche (TE-MoMa project, grant number ANR-18-CE02-0020-01). AL thanks the ERC and the University of Lille for funding her PhD project. The authors thank the UMR 8199 LIGAN-MP Genomics platform (Lille, France), which belongs to the 'Federation de Recherche' 3508 Labex EGID (European Genomics Institute for Diabetes; ANR-10-LABX-46) and was supported by the ANR Equipex 2010 session (ANR-10-EQPX-07-01; 'LIGAN-MP'). The LIGAN-PM Genomics platform (Lille, France) is also supported by the FEDER and the Region des Hauts-de-France. We thank Stephen I. Wright, Rasmus Nielsen, Violaine Llaurens, Sylvain Glémin and Camille Roux for helpful discussions. The authors thank the Région Hauts-de-France, and the Ministère de l'Enseignement Supérieur et de la Recherche (CPER Climibio), and the European Fund for Regional Economic Development for their financial support for the molecular facilities in Lille. This work has been performed using infrastructure and technical support of the Plateforme Serre, cultures et terrains expérimentaux – Université de Lille for the greenhouse/field facilities.

Supplementary data include Fasta and Bed files of *A. halleri* regions and probes used for the sequence capture available online in figshare database at 10.6084/m9.figshare.16438908. All original sequence data are available in the NCBI Short Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) with accession codes: PRJNA744343. All scripts developed are available in Github (https://github.com/leveveaudrey/analysis-of-polymorphism-S-locus).

Bibliography

Asthana S, Schmidt S, and Sunyaev S. 2005. A limited role for balancing selection. Trends in Genetics. 21: 30–32.

Billiard S, Castric V, Vekemans, X. 2006. A general model to explore complex dominance patterns in plant sporophytic self-incompatibility systems. Genetics. 175: 1351–1369.

Castric V, Vekemans X. 2004. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Molecular Ecology. 13: 2873–2889.

Castric V, Bechsgaard J, Schierup M.H, Vekemans X. 2008. Repeated adaptive introgression at a gene under multiallelic balancing selection. PLOS Genetics. 4, e1000168.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–1303.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity. 95: 118–128.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLOS Genetics. 2, e64.

Cheng S, van den Bergh E, Zeng P, Zhong X, Xu J, Liu X, Hofberger J, de Bruijn S, Bhide A.S, Kuelahoglu C et al. 2013. The Tarenaya hassleriana Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers. The Plant Cell. 25: 2813–2830.

Cheng X, DeGiorgio M. 2020. Flexible mixture model approaches that accommodate footprint size variability for robust detection of balancing selection. Mol Biol Evol. 37: 3267–3291.

Clauss M.J, Mitchell-Olds T. 2006. Population genetic structure of *Arabidopsis lyrata* in Europe. Molecular Ecology. 15: 2753–2766.

Clo J, Ronfort J, Awad D.A. 2020. Hidden genetic variance contributes to increase the short-term adaptive potential of selfing populations. Journal of Evolutionary Biology. 33: 1203–1215.

Danecek P, Auton A, Abecasis G, Albers C.A, Banks E, DePristo M.A, Handsaker R.E, Lunter G, Marth G.T, Sherry S.T, et al. 2011. The variant call format and VCFtools. Bioinformatics. 27: 2156–2158.

DeGiorgio M, Lohmueller K.E, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genetics. 10, e1004561.

Delph L.F, Kelly J.K. 2014. On the importance of balancing selection in plants. New Phytologist. 201: 45–56.

DePristo M.A, Banks E, Poplin R.E, Garimella K.V, Maguire J.R, Hartl C, Philippakis A.A, del Angel G, Rivas M.A, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43: 491–498.

Dwyer K.G, Kandasamy M.K, Mahosky D.I, Acciai J, Kudish B.I, Miller J.E, Nasrallah M.E, Nasrallah J.B. 1994. A superfamily of S locus-related sequences in Arabidopsis: diverse structures and expression patterns. The Plant Cell. 6: 1829–1843.

Eyre-Walker A, Keightley P.D. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8: 610–618.

Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. Molecular Ecology. 24: 3529–3545.

Foxe J.P, Stift M, Tedder A, Haudry A, Wright S.I, Mable B.K. 2010. Reconstructing origins of loss of self-incompatibility and selfing in North american *Arabidopsis lyrata*: a population genetic context. Evolution. 64: 3495–3510.

Genete M, Castric V and Vekemans X. 2020. Genotyping and De Novo Discovery of Allelic Variants at the Brassicaceae Self-Incompatibility Locus from Short-Read Sequencing Data. Molecular Biology and Evolution. 37: 1193–1201.

Goubet P.M, Bergès H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl A.-C, Fobis-Loisy I, Vekemans X, et al. 2012. Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. PLoS Genetics. 8, e1002495.

Guo Y.-L, Zhao X, Lanz C, Weigel D. 2011. Evolution of the S-locus region in Arabidopsis relatives. Plant Physiology. 157: 937–946.

Hartfield M, Otto S.P. 2011. Recombination and hitchhiking of deleterious alleles. Evolution. 65: 2421–2434.

Holtz Y, Ardisson M, Ranwez V, Besnard A, Leroy P, Poux G, Roumet P, Viader V, Santoni S, David J. 2016. Genotyping by sequencing using specific allelic capture to build a high-density genetic map of *Durum Wheat*. PLOS ONE. 11, e0154609.

Hu T.T, Pattyn P, Bakker E.G, Cao J, Cheng J.-F, Clark R.M, Fahlgren N, Fawcett J.A, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43: 476–481.

Hudson R.R, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics. 116: 153–159.

Hudson R.R, Kaplan N.L. 1988. The coalescent process in models with selection and recombination. Genetics. 120: 831–840.

Jay P, Tezenas E, Giraud T. 2021. A deleterious mutation-sheltering theory for the evolution of sex chromosomes and supergenes. BioRxiv 2021.05.17.444504.

Kamau E, Charlesworth D. 2005. Balancing selection and low recombination affect diversity near the self-incompatibility loci of the plant *Arabidopsis lyrata*. Current Biology. 15: 1773–1778.

Kamau E, Charlesworth B, Charlesworth D. 2007. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of *Arabidopsis lyrata*. Genetics. 176: 2357–2369.

Kawabe A, Hansson B, Forrest A, Hagenblad J, Charlesworth D. 2006. Comparative gene mapping in *Arabidopsis lyrata* chromosomes 6 and 7 and *A. thaliana* chromosome IV: evolutionary history, rearrangements and local recombination rates. Genetics Research. 88: 45–56.

Kubota S, Iwasaki T, Hanada K, Nagano A.J, Fujiyama A, Toyoda A, Sugano S, Suzuki Y, Hikosak, K, Ito M, et al. 2015. A genome scan for genes underlying microgeographic-scale local adaptation in a wild arabidopsis species. PLOS Genetics. 11, e1005361.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah J.B, Nasrallah M.E. 2001. Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. Plant Cell. 13: 627–643.

Lane M.D, Lawrence M.J. 1995. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. X. An association between incompatibility genotype and seed dormancy. Heredity. 75: 92–97.

Langmead B, Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9: 357–359.

Legrand S, Caron T, Maumus F, Schvartzman S, Quadrana L, Durand E, Gallina S, Pauwels M, Mazoyer C, Huyghe L, et al. 2019. Differential retention of transposable element-derived sequences in outcrossing Arabidopsis genomes. Mobile DNA. 10: 30.

Lenormand T, Roze D. 2022. Y recombination arrest and degeneration in the absence of sexual dimorphism. Science. 375: 663–666.

Lenz T.L, Spirin V, Jordan D.M, Sunyaev S.R. 2016. Excess of deleterious mutations around *HLA* genes reveals evolutionary cost of balancing selection. Mol Biol Evol. 33: 2555–2564.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25: 2078–2079.

Llaurens V, Billiard S, Leducq J.-B, Castric V, Klein E.K, Vekemans X. 2008. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri* ? Evolution. 62: 2545–2557.

Llaurens V, Gonthier L, Billiard S. 2009. The sheltered genetic load linked to the S locus in plants: new insights from theoretical and empirical approaches in sporophytic self-incompatibility. Genetics. 183: 1105–1118.

Llaurens V, Whibley A, Joron M. 2017. Genetic architecture and balancing selection: the life and death of differentiated variants. Molecular Ecology. 26: 2430–2448.

Loewe L, Charlesworth B. 2007. Background selection in single genes may explain patterns of codon bias. Genetics. 175:1381–1393.

Mattila T.M, Laenen B, Horvath R, Hämälä T, Savolainen O, Slotte T. 2019. Impact of demography on linked selection in two outcrossing Brassicaceae species. Ecology and Evolution. 9: 9532–9545.

Maruyama T, Nei M. 1981. Genetic variability maintained by mutation and overdominant selection in finite populations. Genetics. 98:441–459.

Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. 2017. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. Genome Biology. 18: 76.

Nettancourt D. 2001. Incompatibility and incongruity in wild and cultivated plants. Berlin Heidelberg: Springer-Verlag.

Noé L, Kucherov G. 2005. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research. 33: W540–W543.

van Oosterhout, C. 2009. A new theory of MHC evolution: beyond selection on the immune genes. Proceedings of the Royal Society B: Biological Sciences. 276: 657–665.

Otto S.P, Pannell J.R, Peichel C.L, Ashman T.-L, Charlesworth D, Chippindale A.K, Delph L.F, Guerrero R.F, Scarpino S.V, McAllister B.F. 2011. About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. Trends in Genetics. 27: 358–367.

Pauwels M, Saumitou-Laprade P, Holl A.C, Petit D, Bonnin I. 2005. Multiple origin of metallicolous populations of the pseudometallophyte *Arabidopsis halleri* (Brassicaceae) in central Europe: the cpDNA testimony. Molecular Ecology. 14: 4403–4414.

Quinlan A.R, Hall I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26: 841–842.

Rice W.R. 1987. The Accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution 41, 911–914.

Ross-Ibarra J, Wright S.I, Foxe J.P, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, and Gaut B.S. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. Plos One 3, e2411.

Roux C, Pauwels M, Ruggiero M.-V, Charlesworth D, Castric V, Vekemans X. 2013. Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A. lyrata*. Molecular Biology and Evolution. 30: 435–447.

Ruggiero M.V, Jacquemin B, Castric V, Vekemans X. 2008. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in *Arabidopsis halleri*. Genet Res (Camb). 90: 37–46.

Schierup M.H, Vekemans X and Christiansen F.B. 1997. Evolutionary dynamics of sporophytic self-incompatibility alleles in plants. Genetics. 147: 835–846.

Schierup M.H, Vekemans X, Charlesworth D. 2000. The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. Genetics Research. 76: 63–73.

Schierup M.H, Mikkelsen A.M, Hein J. 2001. Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics. 159: 1833–1844.

Schopfer C.R, Nasrallah M.E, Nasrallah J.B. 1999. The male determinant of self-incompatibility in Brassica. Science. 286: 1697–1700.

Siever, F, Wilm A, Dineen D, Gibson T.J, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7: 539.

Smith JM, Haigh J. 1974. The hitchhiking effect of a favorable gene. Genet Res. 23:23–35.

Spurgin L.G, Richardson D.S. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proceedings of the Royal Society B: Biological Sciences. 277: 979–988.

Stift M, Hunter B.D, Shaw B, Adam A, Hoebe P.N, Mable B.K. 2013. Inbreeding depression in self-incompatible North-American *Arabidopsis lyrata*: disentangling genomic and S-locus-specific genetic load. Heredity. 110: 19–28.

Stone J.L. 2004. Sheltered load associated with S-alleles in *Solanum carolinense*. Heredity. 92: 335–342.

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proceedings of the National Academy of Sciences. 87: 2419–2423.

Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics. 124: 967–978.

Takahata N, Satta Y. 1998. Footprints of intragenic recombination at HLA loci. Immunogenetics. 47: 430–441.

Takou M, Hämälä T, Koch E.M, Steige K.A, Dittberner H, Yant L, Genete M, Sunyaev S, Castric V, Vekemans X, et al. 2021. Maintenance of adaptive dynamics and no detectable load in a range-edge outcrossing plant population. Molecular Biology and Evolution. 38:1820–1836

Uyenoyama M.K. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics. 147: 1389–1400.

Uyenoyama, M.K. 2003. Genealogy-dependent variation in viability among self-incompatibility genotypes. Theoretical Population Biology. 63: 281–293.

Uyenoyama M.K. 2005. Evolution under tight linkage to mating type. New Phytol. 165: 63–70

Vekemans X, Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics. 137:1157–1165

Vekemans X, Castric V, Hipperson H, Müller N.A, Westerdahl H, Cronk Q. 2021. Whole-genome sequencing and genome regions of special interest: Lessons from major histocompatibility complex, sex determination, and plant self-incompatibility. Mol Ecol. 30: 6072–6086.

Vieira J, Pimenta J, Gomes A, Laia J, Rocha S, Heitzler P, Vieira C.P. 2021. The identification of the Rosa S-locus and implications on the evolution of the Rosaceae gametophytic self-incompatibility systems. Sci Rep. 11: 3710.

Williamson R.J, Josephs E.B, Platts A.E, Hazzouri K.M, Haudry A, Blanchette M, Wright S.I. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. PLOS Genetics 10, e1004622.

Wiuf C, Zhao K, Innan H, Nordborg M. 2004. The probability and chromosomal extent of trans-specific polymorphism. Genetics. 168: 2363–2372.

Wright S. 1939. The distribution of self-sterility alleles in populations. Genetics 24: 538–552.

Wright S.I., Charlesworth B. 2004. The HKA test revisited: a maximum likelihood ratio test of the standard neutral model. Genetics. 168: 1071-1076.

Wright S.I, Ness R.W, Foxe J.P, and Barrett S.C.H. 2008. Genomic consequences of outcrossing and selfing in Plants. International Journal of Plant Sciences. 169: 105–118.

Wu L, Williams J.S, Sun L and Kao T.-H. 2020. Sequence analysis of the Petunia inflata S-locus region containing 17 S-Locus F-Box genes and the S-RNase gene involved in self-incompatibility. The Plant Journal. 104: 1348–1368.

Zhang W, Sun Y, Timofejeva L, Chen C, Grossniklaus U, Ma H. 2006. Regulation of Arabidopsis tapetum development and function by DYSFUNCTIONAL TAPETUM1 (DYT1) encoding a putative bHLH transcription factor. Development. 133: 3085–3095.

Figure legend

Figure 1: Schematic representation of the S-locus and its flanking regions. *Protein-coding genes in the flanking regions are represented as filled blue rectangles. The genomic regions studied are distributed between positions 9,264,458 and 9,451,731 along chromosome 7 of the A. lyrata genome assembly (Hu et al. 2011). The S-locus (in grey) contains two protein-coding genes only, SRK and SCR (white rectangles) and is flanked by the ARK3 and Ubox genes (in the 5' and 3' directions, respectively). The S-locus region itself was not analysed in the present study. The percentage of CDS in each 25kb window is given on top of the figure.*

Figure 2: Variation of the mean selection parameter (k) obtained for genes in the S-flanking regions. *The solid large black line represents the mean value of k obtained across the six sample sets. The black solid lines represent the A. lyrata sample sets (square=Plech, circle=Spiterstulen, triangle=North America). The black dashed lines represent the A. halleri sample sets (open square=Japan, open circle=Nivelle, open triangle=Mortagne). The threshold value of 1 (no selection) is represented by the horizontal black line.*

Figure 3: Manhattan plots for signals of balancing selection ($B_{2,MAF}$ scores) in the 100 control regions and the S-locus region (red dots and arrows) along the A. lyrata genome. *Chromosomes 1-8 of the A. lyrata genome, with the 100 control regions distributed on successive chromosomes represented by an alternation of grey and black dots. The horizontal grey solid line represents the median $B_{2,MAF}$ scores across SNPs in the genome, and the black horizontal dashed lines represent the top 5, 2.5 and 1% percentiles.*

Figure 4: Mean Ho around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of Ho obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of Ho mean in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

Figure 5: Mean π around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of π obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of π mean in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

Figure 6: Proportion of polymorphic sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of polymorphic sites in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

Figure 7: Tajima's D around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean of Tajima's D obtained in S-flanking windows of 25kb. The distributions (count) in the 100 control regions are represented by a histogram (right). The 97.5% and 2.5% percentiles of the distributions are represented by dashed lines. The median value of the distribution in control regions is represented by black lines.*

Figure 8: Proportion of polymorphic sites among 0-fold degenerate sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of polymorphic sites in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

Figure 9: $\pi_{0\text{-fold}}/\pi_{4\text{-fold}}$ ratio around the S-locus and across the control regions from throughout the genome. *The bars represent the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of polymorphic sites in the 100 control regions are represented by a vertical histogram on the right. The 95% percentile of the distributions is represented by dashed lines. The median value of the distribution in control regions is represented by black lines. *** = observed value above the 99% percentile of control regions, ** = observed value above the 97,5% of control regions, * = observed value above the 95% of control regions.*

Tables

Table 1: Variation of the median Ho, π, MAF and proportion of polymorphic sites in control regions in each dataset.

Species	Sample names	Sequencing method [a]	Number of populations	Number of positions considered	*Ho* x10^{-3}	*π* x10^{-3}	*MAF* x10^{-3}	Proportion of polymorphic sites x10^{-3}
A. halleri	Japan	WGS	6	953,242	0.96	1.28	0.9	8.8
	Nivelle	Capture	1	1,037,607	3.59	4.08	2.79	13.1
	Mortagne	Capture	1	1,059,569	3.71	4.24	2.93	14.6
A. lyrata	Plech	WGS	1	1,190,287	6.46	7.58	5.11	29.9
	Spiterstulen	WGS	1	1,017,504	5.47	5.33	3.74	19.2
	North America	Capture	3	503,976	3.15	4.97	3.53	17.9

[a] *The nucleotide sequence polymorphism data obtained by whole genome sequencing (WGS) came from published datasets.*