Balancing selection, genetic load and dominance between self-incompatibility alleles in Arabidopsis : an empirical and theoretical study of this ménage à trois

Audrey Le Vève

Université de Lille – Sciences et Technologies

Ecole Doctorale – 104 -Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITE DE LILLE

Discipline : Sciences agronomiques et écologiques

Spécialité : Biologie des populations et écologie

Balancing selection, genetic load and dominance between self-incompatibility alleles in Arabidopsis : an empirical and theoretical study of this ménage à trois.

Sélection équilibrante, fardeau génétique et dominance entre allèles d'auto-incompatibilité chez Arabidopsis: étude empirique et théorique d'un ménage à trois

Soutenue le 22 mars 2022 par

LE VEVE Audrey

Membres du jury:

Sylvain Glémin, *Directeur de recherche,* ECOBIO-UMR 6553 CNRS/U. Rennes	**Rapporteur**
Dan Schoen, *Professeur,* Department of Biology - Mc Gill U.	**Rapporteur**
Violaine Llaurens, *Directrice de recherche,* ISEB - UMR 7205 CNRS/MNHN Paris	**Examinatrice**
Juliette De Meaux, *Professeure,* CEPLAS - U. Cologne	**Examinatrice**
Vincent Castric, *Directeur de recherche,* EEP - UMR 8198 CNRS/U. Lille	**Directeur de thèse**
Xavier Vekemans, *Professeur,* EEP - UMR 8198 CNRS/U. Lille	**Co-directeur**
Eléonore Durand, *Maître de Conférences,* EEP - UMR 8198 CNRS/U. Lille	**Co-encadrante**

Présidente du Jury: Violaine Llaurens

Contents

General introduction

A) Balancing selection, genetic dominance and the genetic load : a three-way interaction.

I) Why and how to study the diversity of genomes?

Understanding the evolutionary processes impacting the diversity of genomes is a major challenge in evolutionary genomics, with far-reaching implications for biodiversity conservation, agronomic improvement or human health. In many cases, endangered species have small and/or declining populations, and in such populations inbreeding and loss of genetic diversity are unavoidable. This loss of genetic diversity can compromise the evolutionary response to environmental change. Moreover, in small populations, deleterious alleles can increase in frequency, and eventually reduce fitness. Over long evolutionary time, the fixation of deleterious alleles can lead to negative population growth and a further decline towards extinction (Avise, 1989; O'Brien, 1994; Frankham, 2005), a phenomenon also called “mutational meltdown” (Lynch et al., 1995). This factor is central in conservation management (see e.g. Robinson et al., 2019). In agronomy, the relationship between genome diversity and agronomic traits of interest is commonly studied with the aim of improving domesticated species (Hamblin et al., 2011). With climate change, nourishing the human population is a challenge. So, the field of agronomy tries to improve quantitatively and qualitatively domestic species to make them more resistant to new environmental pressures, pathogens but also nutritionally better. In molecular genetics, quantitative traits are first decomposed in their Mendelian components by quantitative trait loci (QTL) analyses, followed by fine-mapping of promising QTLs. Thousands of QTLs associated with agronomic traits were found in crops and represent a reservoir of alleles for breeders to create improved varieties (Nguyen et al., 2019). However, very few QTLs were successfully used in marker-assisted selection because of insufficient precision in their genomic localization. One challenge is to increase the precision of the QTL positions to make the introgressed segment as small as possible and to avoid possible undesirable side effects due to flanking genes. Thus, a precise description of the polymorphism is essential (Das et al., 2017). Finally, the diversity of human genomes is intensively studied for the diagnosis, understanding and treatment of human diseases (Guttmacher and Collins., 2002 for review). To date, the diagnosis of rare Mendelian diseases has been the primary clinical application of sequencing the genomes of individual patients. Genomic sequencing allows reporting of thousands of pathogenic mutations identified in recent years, and novel gene-disease associations are proliferating (Gillissen et al., 2011). Diagnosis by genomic sequencing is indicated for the

detection of genetic variants in patients with suspected monogenic disorders (Krier et al,. 2016).

The DNA sequence ultimately contains the hereditary information. Therefore the ability to measure or infer such sequences is essential to biological research. To do this, we need enough data to detect the maximum genetic diversity in the maximum number of organisms and in the maximum number of populations. For example, the power to detect mutations associated with traits is positively correlated with the number of individuals sequenced for studies (Mills and Rahal., 2019). This collection of genomic data is mainly limited by the cost and time of sequencing (Mardis, 2008). After the development of the very first sequencing technologies in 1973 (Gilbert and Maxam, 1973; Sanger et al., 1977), successive generations of methods developed to sequence more and more portions of genomes for more and more individuals of different species and populations have made it possible to document polymorphism in an increasingly precise manner (Kulski, 2016 for review). For example, the amplification of genes of interest by PCR allowed the analysis of portions of the genome for which primers had been developed for certain species. Unfortunately, this methodology remained expensive for a long time and only allowed the analysis of portions of the genome, generally known genes, for a limited number of individuals (Kulski, 2016). Moreover, the computational capacities of traditional population genetic approaches only allowed their applications to small samples and/or local chromosome regions (Chen, 2015). It was only with the improvement and development of next-generation sequencing technologies at the beginning of the 21st century (Kulski, 2016) that massive production of genomic polymorphism data throughout the tree of Life became possible (Llamoril et al., 2008). In fact, since 2001, improvements in sequencing techniques have reduced the cost of sequencing one megabase from $10,000 to $0.01 (Gloss and Dinger., 2018). In addition, the speed of DNA sequencing has been greatly accelerated, up to 90 times faster (Jain et al., 2018). For example, whereas it took the Human Genome Project initiative more than ten years to sequence the first human genome, it now takes a few hours. This acceleration in the speed of sequencing and the decrease in its cost has allowed to increase considerably the number of reference genomes throughout the tree of Life. Today, on NCBI, we count more than 60,000 reference genomes, partially or totally sequenced. This allows the genomic study of very different organisms. Finally, the improvement of sequencing techniques over the last fifteen years has allowed us to increase the power and precision of detection of genetic variability in various populations of different species. Thus, the ability to rapidly sequence large numbers of individuals from different populations and from different species, enables the powerful analysis of genetic diversity from different natural populations at high resolution, including mutations segregating at low frequency. It is now possible to provide an almost exhaustive description of polymorphism for large sample sizes and chromosomal regions, and for a large number of species.

II) Evolutionary processes impacting polymorphism

Mutation is the fundamental force of evolution because it is the one that initially creates the genetic variability of populations. Although a mutation that appears in a population can come from another population by migration, it ultimately appeared by mutation in the first population (Loewe and Hill, 2010). The variability within genomes will first depend on the rate of appearance of these mutations. Then, it will depend on the different evolutionary forces that will affect them. Several categories of mutations can be distinguished based on how they affect fitness. Mutations having no effect on fitness are called neutral. Advantageous mutations have a positive effect on the selective value of carriers of this mutation, and on the other hand deleterious mutations have a negative effect on the selective value of their carriers. After it appears by mutation or migration, the only force acting on a neutral mutation is genetic drift (Kimura, 1968). This process is defined by the evolution of allele frequencies due to the random sampling of alleles from one generation to the next. This evolutionary force can lead to ultimate fixation or to the loss of the mutation, and its intensity depends chiefly on the size of the population (Fig. 1; Masel, 2011). On the other hand, a non-neutral mutation will have a tendency to become fixed in the population if it is advantageous, in which case it is said to be subject to positive directional selection. Conversely, a deleterious mutation will tend to be lost over the generations. It is under negative directional selection, also called purifying selection (Fig. 1, Loewe and Hill, 2010).

The massive production of sequence data over the past 20 years has enabled in-depth studies of polymorphism at the genome level. A particularly striking feature is the non-independent variation of polymorphic sites along chromosomes (Charlesworth et al., 1993; Barton 1995; Charlesworth et al., 2003; Oleksyk et al., 2010; Slotte, 2014). Indeed, variation at a given site can be influenced by selection on neighbouring linked sites (Charlesworth et al., 2003), a phenomenon recognized early on and now called " indirect selection ". The effect of selection at one site on variation at another site depends on the rate of recombination between the two sites: the higher this rate, the greater the probability of dissociating the polymorphic sites (Fig. 1). When the site under selection is under positive selection, it causes the fixation of surrounding related neutral mutations. This phenomenon is also called the “hitchhiking effect” (Fig. 1, top right; Smith and Haigh, 1974). When the selected mutation is under purifying selection, mutations on the linked sites will tend to be lost. This process is then referred to as “background selection” (Fig. 1, top left);. It should be noted that, in general, the effect of selection and / or drift at a given site will tend to eliminate / fix the surrounding genetic polymorphism with the linked mutation. There is, however, one type of selection, which by definition causes polymorphism to be maintained: balancing selection.

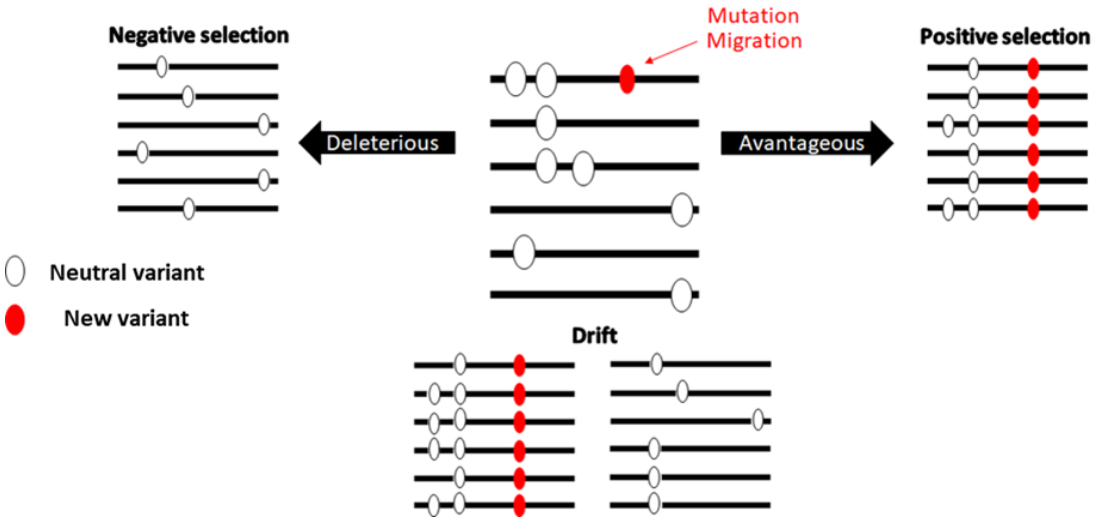


Figure 1: Impacts of different selective processes on polymorphism at the local scale. *The lines symbolise chromosomal regions, the red ovals represent mutations subject to evolutionary forces, the white ovals represent the linked neutral mutations. Each association of a mutation subject to evolutionary forces and neutral mutations on a line constitutes a haplotype. The haplotypes in the centre represent the initial state of a fictitious population. The arrival of a new mutation can be done by migration or mutation (red arrow). The association between each mutation on a haplotype can be broken depending on the rate of recombination, i.e on the genetic distance. The effect of the mutation subjected to evolutionary forces decreases on the most distant neutral mutation because it is more often dissociated by recombination.*

III) Consequences of balancing selection on linked polymorphism

Balancing selection is defined as the set of selective processes leading to the maintenance of allelic diversity at a given locus (Charlesworth, 2006). These diverse processes have the common property of tending to prevent the fixation of individual alleles. The stable frequency equilibrium is then intermediate and the level of heterozygosity is high. The best-known of these selective processes are heterozygous advantage (overdominance model), negative frequency-dependent selection and selection in fluctuating environments in time and / or in space. Many traits involved in sexual recognition are influenced by negative frequency-dependent selection (Llaurens et al., 2017 for review). It is also often observed in traits involved in competition for resources (Benkman, 1996 for an example in mandible orientation in crossbill finches). Many traits involved in resistance to parasites are influenced by overdominance selection because heterozygotes may recognize a larger range of pathogens than homozygotes (Llaurens et al., 2017). When allele fitness fluctuates over time, this may promote the long-term maintenance of polymorphism. For example, time-varying selection is considered one of the drivers of host-parasite coevolution because the varying composition of pathogens in the environment over time forces hosts to maintain an arsenal of defence (Decaestecker et al. 2007). A large number of genes under balancing selection have been identified, including e.g. the *HLA* genes responsible for immunity that are clustered at the major histocompatibility complex (*MHC*) in humans and other mammals,

the loci involved in self-incompatibility in plants, the complementary sex determination genes in Hymenoptera (Gloag et al., 2016).

Balancing selection typically causes an increase in heterozygosity (Navarro and Barton, 2002), an increase in polymorphism (Degiorgio et al., 2014) and a modification of the allelic frequency spectra (Cheng and Degiorgio, 2020) at the locus subject to selection. Balancing selection can maintain many distinct allelic lines over large timescales, up to several million years (e.g., ~ 60 Ma for *HLA*, Klein et al., 2007) causing substantial divergence among the balanced allelic lines. However, within each of the allelic lineages in multiallelic systems, the depth of the pedigrees should be low (Vekemans and Slatkin., 1994 for an exemple in gametophytic self-incompatibility). Thus, the coalescence time between the allelic lines should be greater than that expected under a neutral model, whereas the coalescence time within each line should be less. In addition, Takahata (1990) showed that the overall shape of the coalescent tree of the distinct balanced allelic lines under multiallelic overdominance should be identical to that of a neutral coalescent tree, albeit over a much expanded time scale.

When recombination is limited, this effect of balancing selection should theoretically be reflected in the genomic regions linked to the locus under balancing selection (Fig. 2). Thus, loci under balancing selection are distinguished from other genomic regions by a local increase in polymorphism (DeGiorgio et al., 2014) and heterozygosity (Charlesworth, 2006). This increase in heterozygosity has, for example, been observed around *HLA* genes under balancing selection in several human populations (Hedrick and Thomson, 1983). Finally, each allelic line should accumulate its own unique association of linked mutations, unless recombination decouples this association (Charlesworth et al., 2003).

The chromosomal extent of the indirect effect of balancing selection has been theoretically studied by Takahata and Satta (1998), Schierup et al., (2000) and Wiuf et al., (2004). The effect depends on the strength and form of the balancing selection exerted on the locus / gene undergoing selection, as well as the rate of recombination between this locus / gene and the surrounding regions. The increase in the time during which allelic lines under balancing selection are maintained (Takahata and Nei, 1990; Vekemans and Slatkin, 1994) allows extended time for recombination to decouple allelic lines from their linked genomic regions. Thus, the extent of the chromosomal region affected by balancing selection should be quite narrow (Schierup et al., 2000).

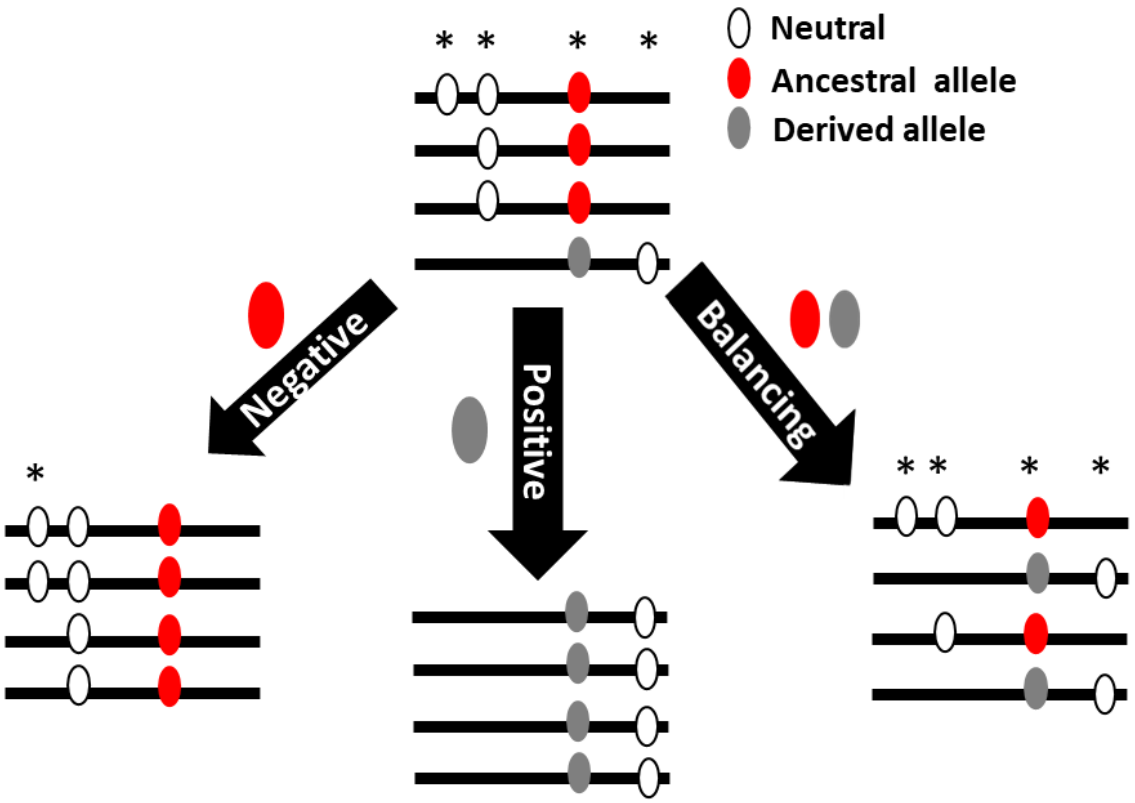


Figure 2: Schematic representation of the predicted impacts of various forms of natural selection on polymorphism at the local scale. *The lines symbolise chromosomal regions, the red / grey ovals represent mutations subject to evolutionary forces, the white ovals represent the linked neutral mutations. Each association composed of a mutation subjected to evolutionary forces and neutral mutations on a trait constitutes a haplotype. The haplotypes at the top centre represent the initial state in a population. The asterisks represent the polymorphic sites. If the derived mutation (grey) is subjected to negative selection (bottom left), the neutral mutations linked to the ancestral allele (red) increase in frequency and the number of polymorphic sites is reduced. If the derived mutation is subjected to positive selection (bottom middle), the neutral mutations linked to the derived allele increase in frequency and the number of polymorphic sites is reduced. If the mutations are subjected to the balancing selection (bottom, right), all neutral mutations are maintained.*

Detection of balancing selection remains a difficult task because the relevant signatures are subtle and narrow and may be masked by other forms of natural selection (Fijarczyk and Babik, 2015 for review). False positives can result from demography, population structure, multiple mutations, and interspecific introgression (Fijarczyk and Babik, 2015). The low power and high incidence of false positives in the tests used to detect balancing selection do not facilitate the study of these consequences on genome diversity. Moreover, the use of different detection criteria might impact findings as different mechanisms of selection can produce different molecular signatures. For example, a lack of spatial genetic structure, estimated by F_{ST}, may indicate the persistence of balanced polymorphisms at a given locus (Fijarczyk and Babik 2015; Llaurens et al., 2017). In fact, a locally depressed F_{ST} compared with neutral markers is expected around the locus under negative frequency-dependent selection. However, an opposite trend is expected when the balanced polymorphism is promoted by spatially variable selection.

In order to confront the many theoretical expectations about the effect of balancing selection on linked polymorphism (Charlesworth, 2006), it is necessary to have genomic data for regions linked to different loci under balancing selection. Also, because of the wide variety of balancing selection processes known for different genes in different species, it is necessary to compare the effects of balancing selection on linked genetic diversity as a function of the type of selection process. But for this, 1) the balancing selection process of the locus under study must be well defined, and 2) the linked region must be completely known (i.e. with a correct and complete reference genome available). Moreover, if we want to define the size of the linked region where we can detect the effects of the balancing selection, it is necessary that this linked region is well defined over a large distance.

IV) Consequences of balancing selection on the genetic load

Deleterious mutations are initially distributed throughout the genome, as a function of the local mutation rate. However, some genomic regions, where loci under balancing selection are present, may be more inclined to accumulate them. Indeed, balancing selection locally enforces heterozygosity in the flanking regions (Kamau et al., 2007), such that linked mutations in these regions that are recessive should be exposed to purifying selection less often than mutations in other parts of the genome (Uyenoyama, 1997; Fig. 3). Balancing selection should thus prevent the elimination of deleterious mutations, resulting in the accumulation of a mutational load "linked" to the locus under balancing selection (see for example van Oosterhout, 2009 for genes linked to *HLA* in humans). In addition, the allele frequencies of mutations in the linked regions could be impacted by balancing selection: we expect an increase in allele frequencies for all mutations, even those that are deleterious (Cheng and DeGiorgio, 2020). Lenz et al., (2013) confirmed that genes within the *MHC* region but not involved in immunity do indeed exhibit an increase in the frequency of mutations predicted to be deleterious as compared to the rest of the human genome. This increase in the frequency of deleterious mutations in these linked regions is possibly responsible for many genetic diseases associated with this locus (Lenz et al., 2013).

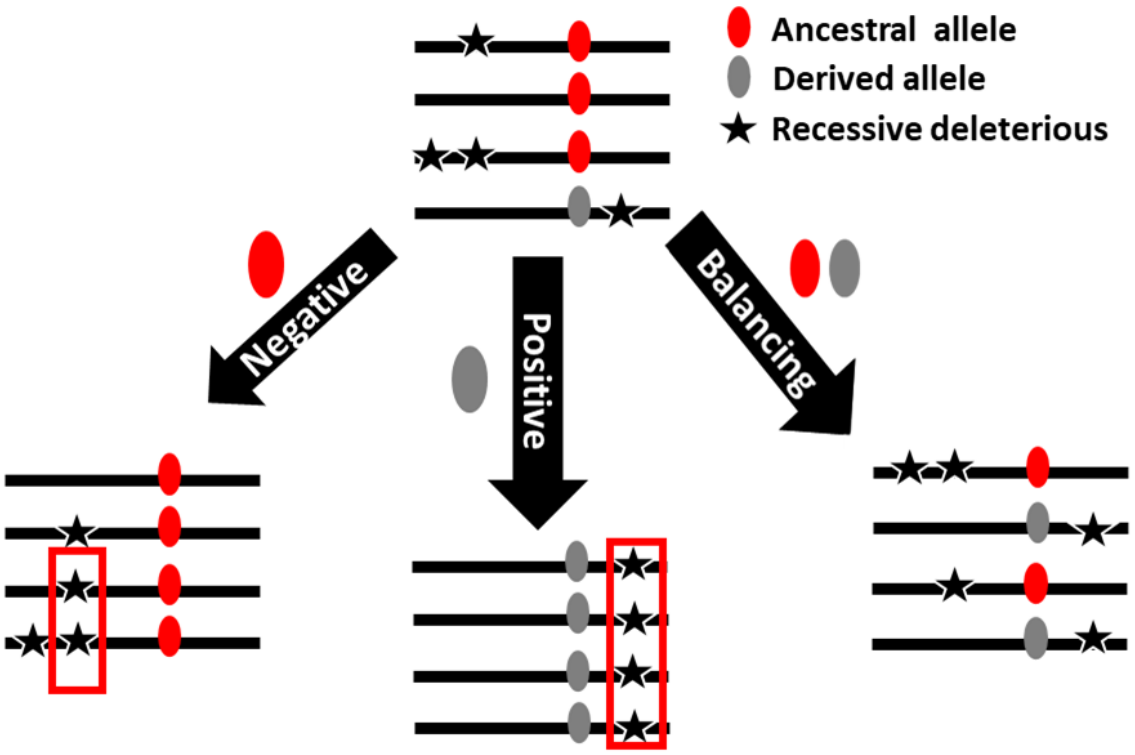


Figure 3: Schematic representation of the predicted impacts of various forms of natural selection on genetic load at the local scale. *The lines symbolise chromosomal regions, the red / grey ovals represent mutations subject to evolutionary forces, the black stars represent the linked recessive deleterious mutations. Each association composed of a mutation subjected to evolutionary forces and deleterious on a trait constitutes a haplotype. The haplotypes at the top centre represent the initial state in a population. If the derived mutation (grey) is subjected to negative selection (bottom left), the deleterious mutations linked to ancestral allele (red) in the homozygous state are expressed (red box) and the individual shows a lower selective value. Elimination of this individual by selection allows the elimination, or purge, of these deleterious mutations. If the derived mutation is subjected to positive selection (bottom middle), the deleterious mutations linked to derived mutation in the homozygous state are expressed (red box) and the individual shows a lower selective value. Selection leads to the purge of these deleterious mutations. If the mutations are subjected to the balancing selection (bottom, right), they are maintained in the heterozygous state, as well as all the deleterious mutations associated. The deleterious recessive mutations in the heterozygous state are not expressed and are maintained.*

A genetic load associated with balancing selection in various situations like the fire ant social supergene has also been suggested (Llaurens et al., 2017 for review). However, the examples with genomic demonstration and a clear comprehension of the architecture of this genetic load linked to a locus under balancing selection remain rare.

Understanding the evolution of the genetic load is important on several accounts. First, deleterious mutations probably play a major role in causing inbreeding depression (Charlesworth and Charlesworth., 1999). Moreover, and more specifically, the genetic load can contribute to the maintenance of balanced polymorphisms. Deleterious mutations associated with some alleles of a locus under balancing selection can drive overdominance,

whereby heterozygotes enjoy better fitness than homozygotes expressing the associated genetic load. This further reinforces the persistence of balanced polymorphism in natural populations (Uyenoyama 2003), and an empirical example in Butterfly supergene mimicry has been recently proposed (Jay et al., 2021). However, the association of adaptive alleles with genetic load might also be expected to contribute to their elimination from the populations, because other alleles without genetic load could benefit from higher fitness. Altogether, however, the accumulation of deleterious mutations can strengthen balancing selection over long evolutionary timescales (Llaurens et al., 2017 for review)

V) The central role of genetic dominance in population genetics

Dominance is one of the basic properties of genetic systems. The concept was already inherent to the work of Mendel in 1866 and can be defined as a deviation from additivity of the phenotypic expression of the two alleles in a heterozygous offspring. Mendel observed that the F1 generation of peas, resulting from the crossing of two pure lines, uniformly manifests one of the two parental characters, known as dominant. The F1 did not express the second parental trait in spite of being present in its genotype (because it was passed on to the F2 where it was expressed in ¼ of the offspring). This trait is said to be recessive (Fig. 4).

Dominance affects the fixation of beneficial alleles because it determines the degree to which they are "visible" by natural selection in heterozygotes and thus affects their fate, especially immediately after they arise by mutation (Charlesworth and Charlesworth, 2010, Fig. 4). Dominance can also affect the dynamics of elimination of deleterious alleles, and in fact the vast majority of segregating deleterious mutations in natural populations seem to be recessive (Charlesworth and Willis., 2009). Haldane (1924) predicted that if a new beneficial allele arises in a population, the probability that it eventually reaches fixation is influenced by the dominance coefficient of the allele. The reason why the dominance coefficient matters is because early in the life of the allele, while it is at low frequency, it is mostly present in the population in heterozygous form. Therefore all else being equal, dominant beneficial alleles can increase in frequency due to selection faster than recessive alleles, increasing their probability of eventual fixation in the population. This effect has become known as “Haldane’s sieve”. Wright (1934) completed this assumption by the prediction of a negative correlation between the effect of selection (*s*) and dominance (*h*) of mutations. Thus, dominance is a central property of population genetic models.

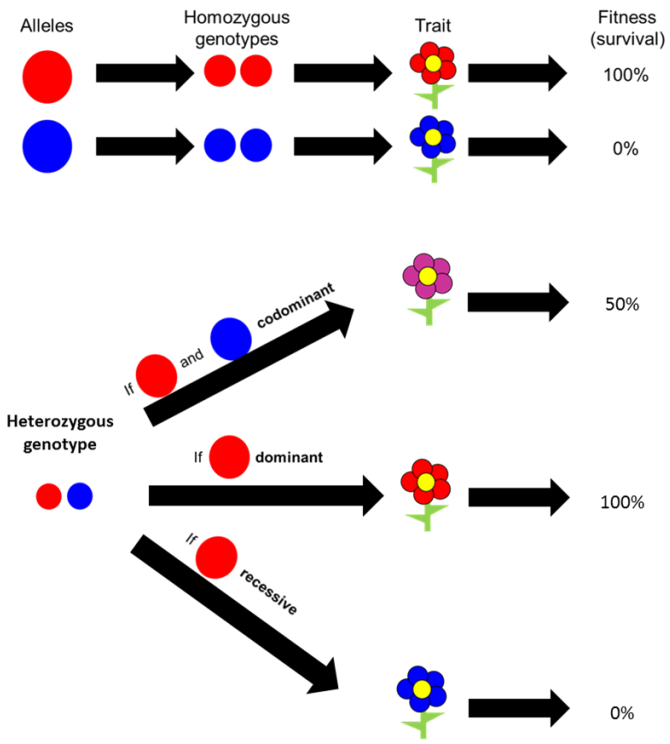


Figure 4: Variation of traits with dominance of two alleles in population. *In a population, we considered two alleles, red and blue rounds. At the homozygous state, each genotype encodes the phenotype red and blue petals respectively. The individuals with the red petals phenotype survive all the time however the blue petals phenotype is lethal. If the two alleles are codominant, the heterozygotes express an intermediate phenotype, the purple petals phenotype. The probability of survival for the heterozygotes is 0.5. If the blue allele is recessive, the heterozygotes express the red petals phenotype. The probability of survival for the heterozygotes is 1. If the blue allele is dominant, the heterozygotes express the blue petals phenotype. The probability of survival for the heterozygotes is null.*

The causes of dominance have been the subject of a heated debate in evolutionary genetics that began in the 20th century between Ronald A. Fisher and Sewall Wright (Billiard and Castric., 2011). For Fisher (1928), dominance interactions between alleles would result from the intervention of genetic elements, referred to as “dominance modifiers”. Without dominance modifiers, heterozygotes would exhibit intermediate fitness relative to the two homozygous genotypes, and would be counter-selected in the case of deleterious alleles. In the presence of a modifier, the heterozygotes would exhibit a fitness equivalent to that of dominant homozygotes (Fisher, 1928), such that these individuals would not express the deleterious allele. Wright (1929) argued that these modifiers are only effective in heterozygous individuals, which are generally present at low frequencies in natural populations in the case of mutations that are deleterious, and showed that there would be insufficient selection pressure for the efficient selection of such modifiers across the genome. Haldane (1930) and Wright (1934) instead proposed that the dominance phenotype of an allele would be derived from a biochemical property, based on enzymatic activity (Haldane, 1930; Wright, 1934). The fitness of heterozygotes would be determined by the relationship between the activity of the gene product and the associated phenotype, plotted on a curve similar to an enzyme saturation curve (Fig. 5).

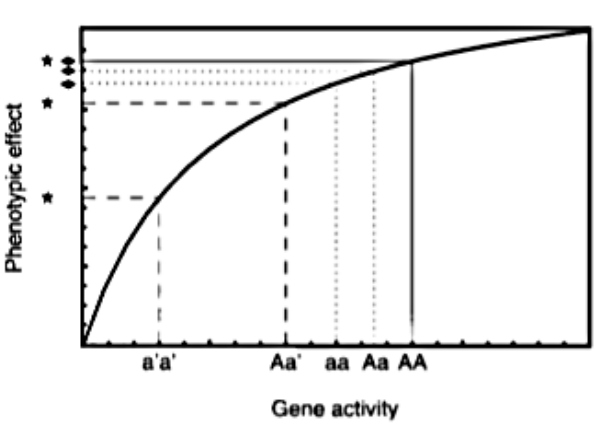


Figure 5 : Relationship between phenotypic effect and gene activity (after Wright 1934a; from Otto and Bourguet, 1999). *Compared with a wild-type genotype (AA; thin solid line), minor effect mutations (a; thin dashed lines) have an additive effect on gene activity (X-axis) and are partially recessive at the phenotypic level (diamonds-on Y-axis). Major effect mutations (a'; dashed lines) show a more pronounced recessivity at the phenotypic level (stars on the y-axis).*

In 1991, a study on Chlamydomonas demonstrated the prevalence of recessive deleterious mutations even though this unicellular alga is predominantly haploid: even in the absence of heterozygous individuals, dominance relationships between alleles were widespread (Orr, 1991). This observation was congruent with the predicted negative correlation between the selection coefficient (*s*) and the dominance (*h*) of mutations in Wright's theory, and was interpreted as discrediting Fisher's theory that a modifier active in heterozygotes is responsible for the evolution of dominance (Orr, 1991). Thus, the existence of dominance modifiers was largely rejected and Wright's physiological theory, later confirmed by the development of the enzymatic theory (Kacser and Burns, 1981), eventually gained the status of a paradigm (Veitia, 2006).

However, Wright conceded that under balancing selection, the high level of heterozygosity could theoretically allow the selection of dominance modifiers. Indeed, under balancing selection, high allelic diversity is maintained and the level of heterozygosity tends to be high. Theoretical work confirmed that it is indeed possible to select for dominance modifiers in overdominance (Otto and Bourguet, 1999) or frequency-dependent models (Peischl and Schneider, 2010), in particular if the dominance modifier is strongly genetically linked to the locus under selection (Fig. 6; Otto and Bourguet, 1999). To date, however, examples of dominance modifiers and their modes of action remain scarce, despite the number of cases of balancing selection that could theoretically allow their evolution (Billiard and Castric, 2011). In fact, the selection for dominance modifiers has been theoretically studied in at least three model systems: sporophytic self-incompatibility in plant (Llaurens et al., 2009a, Schoen and Busch., 2009), Batesian mimicry in butterflies (Charlesworth and Charlesworth., 1975) and two-species models in loci involved in host–parasite interactions (Nuismer and Otto., 2005). In all of the models described above, high levels of heterozygosity might indeed confer the potential for dominance modifier's evolution. However, Billiard and Castric (2011) argued that many other systems maintaining diversity either stably or transiently could potentially share the same favourable properties for dominance evolution, but remain to be investigated. These include temporally varying environments, persistent sexually antagonistic variation caused by intra-locus sexual conflict, mating systems with multiple sexual morphs (gynodioecy, dioecy, heterostyly) and, more generally, all systems with negative frequency-dependent selection.

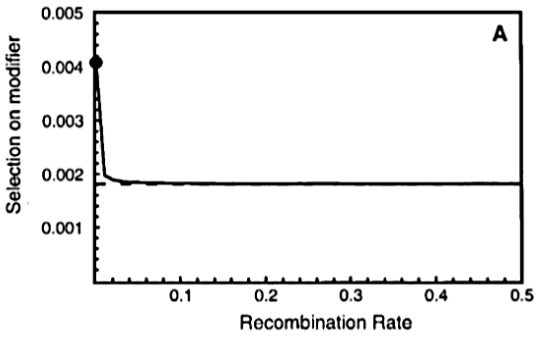


Figure 6 : Selection for a dominance modifier increasing fitness of heterozygotes for an A locus under an overdominant model with recombination rate (Otto and Bourguet, 1999). *In a region with very little recombination with A (r<0.02), under heterozygote selection, a dominance modifier that increases the fitness of A can be selected.*

Overall, this overview shows that interaction between balancing selection, dominance and the genetic load can be complex. The accumulation of the genetic load is impacted by, first, balancing selection because it tends to maintain a high level of heterozygosity, thus masking recessive deleterious mutations. This effect of balancing selection should extend into linked regions depending on the rate of recombination between the locus under balancing selection and these flanking regions. Moreover, we have shown that the genetic load can impact dominance (i.e. relationship between *h* and *s* established by Wright), but we don't know the consequences on dominance evolution if this genetic load is increased by balancing selection. However, genomic studies to support these hypotheses are rare because they require a well-defined and understood balancing selection system for which genomic studies are affordable. We also see that, theoretically, balancing selection should favour dominance evolution involving modifiers, but this relationship between balancing selection and dominance evolution remains poorly studied. To increase our knowledge of this relationship, we need to study a system of balancing selection for which such dominance modifiers are known to exist.

B) The sporophytic self-incompatibility system: a textbook example of balancing selection.

In order to study the interactions between balancing selection, dominance, and the genetic load, it is necessary to study a sufficiently well-understood case of balancing selection, for which dominance relationships between alleles are known. The sporophytic incompatibility system (SSI) is a one-of-a-kind case to address this question because the process of balancing selection is well understood since the work of Wright (1939). Furthermore, as I explain below, the SSI represents to our knowledge the only documented example of dominance relationships between alleles involving modifiers as described by Fisher in 1928 (Billiard and Castric, 2011).

I) The sporophytic self-incompatibility system

The self-incompatibility system in plants is a genetic mechanism allowing the recognition and rejection of self-pollen. This mechanism enforces cross-fertilization and prevents inbreeding depression (Nettancourt, 2001). There are two types of self-incompatibility systems: the gametophytic system, found in more than 60 plant families, including Solanaceae, Rosaceae and Papaveraceae and the sporophytic self-incompatibility system, documented in Brassicaceae, Asteraceae, and Convolvulaceae. In the sporophytic system, pollen expresses the phenotype of the male diploid parent, whereas in the gametophytic system pollen expresses its own haploid genotype. In Brassicaceae, SSI is controlled by the *S*-locus, composed of two genes: *SCR* (*S*-locus cysteine-rich) encoding a pollen surface protein and *SRK* (*S*-locus receptor kinase) encoding its stigma receptor protein (Goubet et al., 2012). If, during pollination, the two proteins form a ligand-receptor complex (Ma et al., 2016), an intracellular signalling cascade inhibits hydration at the surface of the stigmatic papillae, causing pollen rejection (Fig. 7).

This system promotes heterozygosity at the *S*-locus and maintains an important allelic diversity in populations by favouring rare alleles: individuals carrying rare alleles produce pollen less often recognized and rejected by pistils under cross-pollination, increasing their range of compatible partners compared to pollen produced by individuals with alleles that are frequent in the population. This is a case of negative frequency-dependent selection (Wright, 1939; Castric and Vekemans, 2004; Fig. 8).

Arabidopsis halleri and *lyrata* are species belonging to the Brassicaceae family. They are close relatives of the model plant in genetics: *A. thaliana*. However, unlike *A. thaliana*, these species are outcrossers and exhibit a functional SSI system. Thus, they are classic candidates for the study of the impact of balancing selection at the *S*-locus.

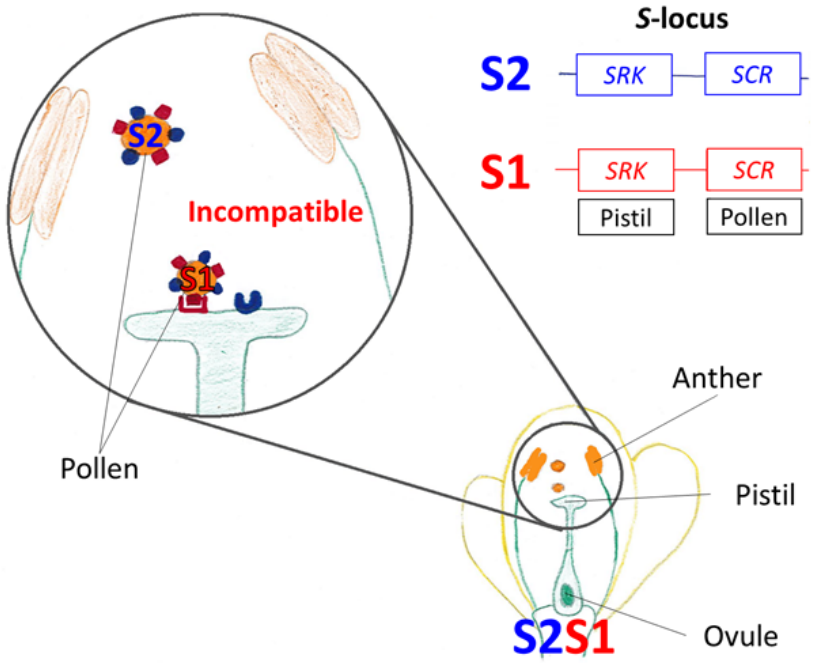


Figure 7: The sporophytic self-incompatibility system. *The system is controlled by the S-locus (top right), which consists of two linked genes SRK and SCR. SRK encodes a receptor on the pistil that recognizes specifically a pollen surface protein encoded by the linked SCR allele. The combination of an SRK allele and the linked SCR allele constitutes a haplotype. At the pollen level, if the two S-alleles are codominant, the two SCR alleles of the diploid antheres are expressed, regardless of the allele at the level of haploid pollen. At the pistil level, if the two S-alleles are codominant, the two SRK alleles of the diploid pistil are expressed. If, during pollination, an SRK receptor on the pistil recognizes a SCR protein from the pollen (top left), the pollen is rejected. The cross is incompatible.*

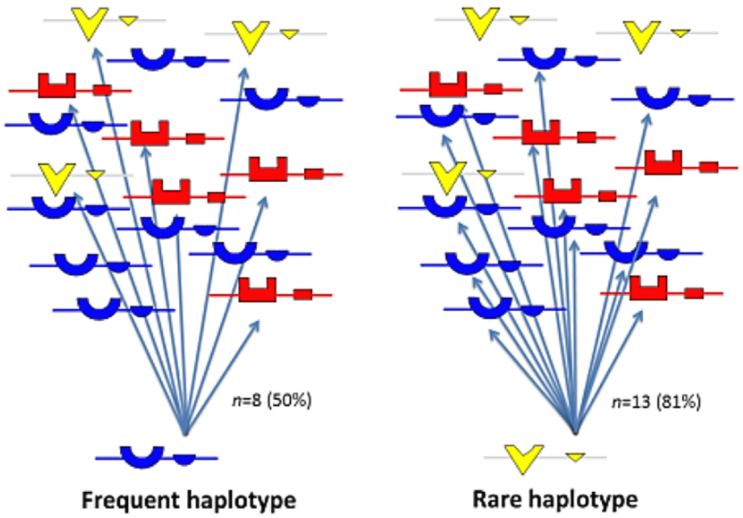


Figure 8: Negative frequency-dependent selection at the S-locus. *We considered one population with three S-alleles represented by the blue, red and yellow colours. Each S-allele is constituted by a combination of one SRK and one SCR gene. The blue allele is frequent (n=8). ollen expressing this allele is therefore recognized and rejected by 50% of the potential partners, and only 50% (n=8) of the potential partners are compatible. The yellow allele is rare (n=3). Pollen expressing this allele is therefore recognized and rejected by only 19% of the potential partners and 81% (n=13) of the potential partners are compatible. The advantage of the yellow allele will continue until its frequency becomes equivalent to that of the other allele. For the sake of simplicity, a single S-allele is expressed by pistils and pollen.*

II) Evolution of polymorphism in the flanking regions of the *S*-locus

Because of the strong negative frequency-dependent selection acting on SSI systems, the *S*-locus is expected to show a local increase in nucleotide polymorphism across the linked regions (Uyenoyama, 1997; Takahata and Satta, 1998; Schierup et al., 2000). Over the recent years, a series of studies have shown that in *A. halleri* and *A. lyrata*, a peak of polymorphism was detected around the *S*-locus over a distance of a few kb (Kamau and Charlesworth, 2005; Kamau et al., 2007; Ruggiero et al., 2008; Roux et al., 2013), although its exact magnitude remained poorly documented due to the small sampling size of previous studies, both in terms of the number of individuals and of genes sequenced (Fig. 9). Indeed, Kamau and Charlesworth (2005) demonstrated, in *A. lyrata*, an excess of polymorphism in two genes flanking the *S*-locus, *AT4G21390* and *Ubox*, compared to two other more distant genes (Fig. 9). In addition to the low number of genes analysed, their study included the offspring of four individuals only from a single population of Iceland. Kamau et al (2007) extended the study to nine populations of *A. lyrata* from Iceland and revealed that the polymorphism in the flanking genes at the *S*-locus is more structured among the different *S*-alleles than among the populations. In this study, the sampling was greatly increased, including from 20 to 54 sequences per gene studied, allowing a better estimation of the impact of the *S*-locus on the polymorphism of the flanking genes. The number of flanking genes analysed was also increased, from four to six, but some of these genes are located far from the *S*-locus: beside the two genes immediately flanking the *S*-locus on either sides (*ARK3* and *Ubox*), the four other genes were located between 189 and 554kb away from the *S*-locus (Fig. 9). This rather long distance probably explains that an increase in polymorphism was again detected only in the two first flanking genes. Ruggiero et al. (2007) focused on the species *A. halleri*, and detected an increase of polymorphism in three of the four genes immediately flanking the *S*-locus (Fig. 9). In addition to having increased the number of genes analysed close to the *S*-locus, the comparison with five genes not linked to the *S*-locus, serving as a control, finally makes it possible to define the increase of polymorphism as an excess over the genome background, as quantified by an Hudson Kreitman Aguade test (Hudson, Kreitman, and Aguadé., 1987). Paradoxically, in these genes, the expected elevation of Tajima's D around a locus under balancing selection was not detected. However, again the number of genes analysed around the *S*-locus remained low: the four genes analysed were only a small subset of the thirty genes present in the 75kb on either sides of the *S*-locus and were unevenly distributed over, at best, 50kb around the *S*-locus. In 2013, Roux et al. more than doubled the number of flanking genes analysed at the *S*-locus, but still comprised only a third of all the genes present in this region (Fig. 9). In this study, for the first time, the two species *A. lyrata* and *A. halleri* were compared, revealing that for both species, the two genes immediately flanking the *S*-locus, *ARK3* and *Ubox*, exhibit an excess of polymorphism in both species. This effect was more pronounced in *A. halleri* than in *A. lyrata*. However, once again, the sampling of only 31 individuals from across 6 different populations was low.

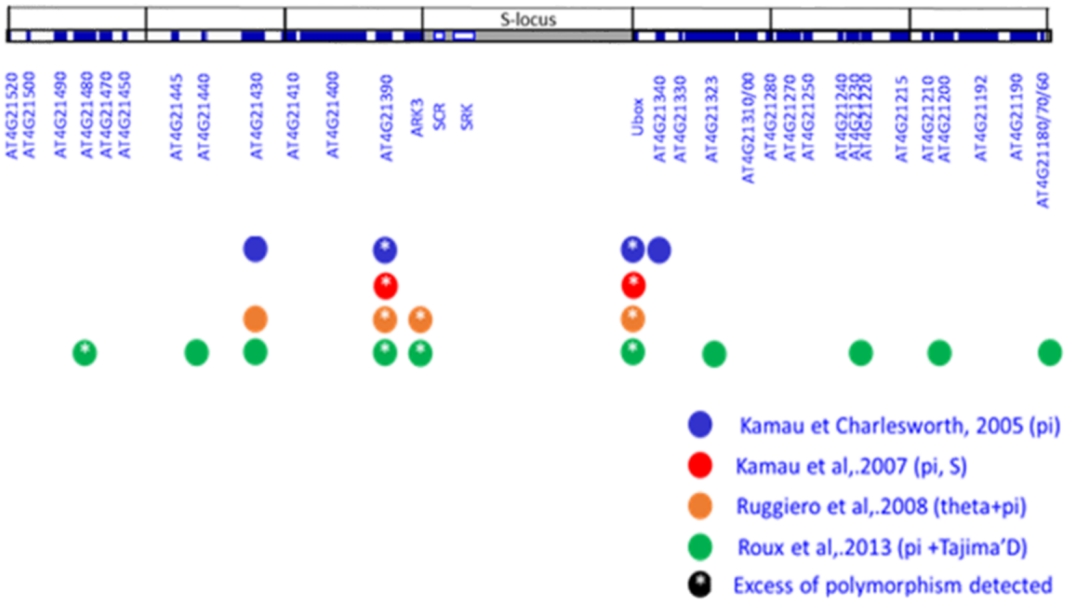


Figure 9: Comparison of gene sampling strategies from previous studies studying the peak of polymorphism around the S-locus in Arabidopsis halleri and A. lyrata. *The genes encoding proteins are represented by blue rectangles. The S-locus (in grey) contains two genes encoding proteins, SRK and SCR. The S-locus is immediately flanked by the genes ARK3 and Ubox (in the 5 'and 3' directions, respectively). Genomic coordinates are given along chromosome 7 of the A. lyrata (Hu et al., 2011). The circles represent the different genes considered in previous studies. Green = Roux et al., 2013, orange = Ruggiero et al., 2008, red = Kamau et al., 2007, blue = Kamau and Charlesworth., 2005. Genes showing a signal of balancing selection signal show a white dot in the corresponding circle. Signals detected are indicated in parentheses after the name of the study.*

While these studies have made clear that the *Ubox*, *ARK3*, and *At4G21390* genes exhibited polymorphism accumulation, they had important limitations regarding the density of genes studied and the size of the samples that preclude definitive conclusions to be drawn regarding the definition of this polymorphism accumulation and the size of the impacted region. With the improvement of sequencing techniques, it is now possible to analyse the genetic composition of many individuals from a large number of populations. In addition, with the development of rapid and reliable sequencing techniques for alleles at the *S*-locus (Genete et al., 2020), it is now possible to analyse a large number of different *S*-alleles. Improved sequencing techniques have also significantly increased the number of genes that can be analysed around the *S*-locus, for example, through the establishment of sequence capture protocols. Indeed, in previous studies, they used PCR amplification to obtain the sequences of the genes for which they had developed primers. This method excluded all other flanking genes around the *S*-locus. Another possibility was to sequence the BAC clones obtained for different individuals (Guo et al., 2011; Goubet et al., 2012). A BAC clone corresponds to a haplotype around the *S*-locus for an individual. However, with this method, the size of the haplotypes obtained was very variable and it was difficult to compare haplotypes. The gene capture approach consisted of specifically amplifying our entire region of interest, even for *A. lyrata*, while the probes were designated on the *A. halleri* genome. It allowed us to consider the sequencing of a large number of individuals for all thirty genes known around the *S*-locus and the intergenic regions, allowing the possibility of detailing the

extent of the size of the region linked to the *S*-locus indirectly impacted by balancing selection.

III) Existence and consequences of the genetic load linked to the *S*-locus

We have seen previously that balancing selection is expected to promote the accumulation of genetic load in linked regions because it maintains a high level of heterozygosity. However, deleterious recessive mutations are only purged by selection in the homozygous state. Because SI locally forces heterozygosity at the *S*-locus, we expect the accumulation of a genetic load linked to this locus (Glémin et al., 2001; Uyenoyama, 2003). The non-recombining region of *S*-locus has a low gene density (Goubet et al., 2012; Durand et al., 2014). Thus, these deleterious mutations, if they exist, are most likely to be found within the flanking genes. Demonstrating the existence and understanding the evolution of this genetic load linked to this locus under balancing selection from a genomic point of view is essential, especially since it may, theoretically, be responsible for a substantial part of inbreeding depression in small populations (Glémin et al., 2001). In addition, Gervais et al. (2014) showed that the linked load should be considered in order to understand the conditions for maintaining a functional SI system. There has been relatively little empirical work to demonstrate the genetic load associated with the *S*-locus. Experimentally, they require distinguishing the contribution of the *S*-locus linked regions to the overall inbreeding depression (Glémin et al., 2001) by comparing the fitness of offspring from a series of controlled autogamous or allogamous crosses, but between individuals who share or do not share identical alleles at the *S*-locus. This component of inbreeding depression attributed to the *S*-locus has been observed in a species with a gametophytic self-incompatibility system, *Solanum carolinense* (Stone, 2004). In *A. halleri* with aSSI system, Llaurens et al (2009a) observed a significant deficit of *S*-locus homozygotes in the progeny of forced incompatible crosses and a decrease of the width and the length of the leaves in these homozygotes. Likewise, a significant deficit of *S*-locus homozygotes in the progeny of forced self crosses in *A. lyrata* (Stift et al., 2013) confirmed the existence of a sheltered genetic load at this locus.

While the theoretical possibility that a genetic load accumulates around a locus under balancing selection is well established, it has been formally tested by a small number of phenotypic studies only, and molecular characterization of this load has remained challenging. Ruggiero et al (2008), did not detect any impact of balancing selection on the efficacy of selection (as measured by π_N / π_S) on the four genes analysed around the *S*-locus. To our knowledge, no other study has investigated the existence of the genetic load around the *S*-locus from a genomic point of view. Hence, although the evolution of the total polymorphism in the flanking genes has been studied, to date, there is no study demonstrating the existence of this genetic load from a genomic point of view.

IV) Consequences of dominance between *S*-alleles on polymorphism and genetic load linked to *S*-locus

The SSI system of the Brassicaceae entails the possibility that dominance/recessivity interactions exist between S-alleles. Such dominance relationships between *S*-alleles have been demonstrated by controlled crosses in several SSI species (Bateman, 1952; Kowyama et al., 1994) including *A. halleri* (Llaurens et al., 2008; Durand et al., 2014). The level of dominance of the *S*-alleles is expected to have an impact on the sheltered genetic load. First, the dominance hierarchy alters the frequency distribution of *S*-alleles at equilibrium: recessive alleles are phenotypically masked by dominant alleles in heterozygous genotypes. They can thus be passively transmitted (Fig. 10) and reach high frequencies in populations (Cope, 1962). Therefore, the frequency of a deleterious mutation linked to the *S*-locus depends on the frequency of the linked *S*-allele, which depends on dominance. Second, dominance negatively impacts the number of gene copies maintained per *S*-allele in the population and the total coalescence time of these copies (Castric et al., 2010). Thus, two copies of a dominant allele should be more similar than two copies of a recessive allele, which should impact the expression of recessive deleterious mutations linked to each allele when in the homozygous state in an individual. Third, recessive deleterious mutations in linked regions are masked in heterozygous genotypes and thus escape purging. The dominance network between *S*-alleles should cause a selection asymmetry between alleles: dominant alleles should be subject to more intense selection than recessive alleles (Billiard et al., 2006). Recessive *S*-alleles are more often homozygotes in natural populations than dominant alleles (Schierup et al., 1997), so recessive alleles might more easily purge recessive deleterious mutations in flanking regions. As a result, dominant *S*-alleles may be expected to exhibit a higher genetic sheltered load than recessive alleles (Fig. 10). Overall, the probability of fixation of linked recessive deleterious mutations should be higher for dominant *S*-alleles than for intermediate or recessive *S*-alleles (Fig. 11; Llaurens et al., 2009a).

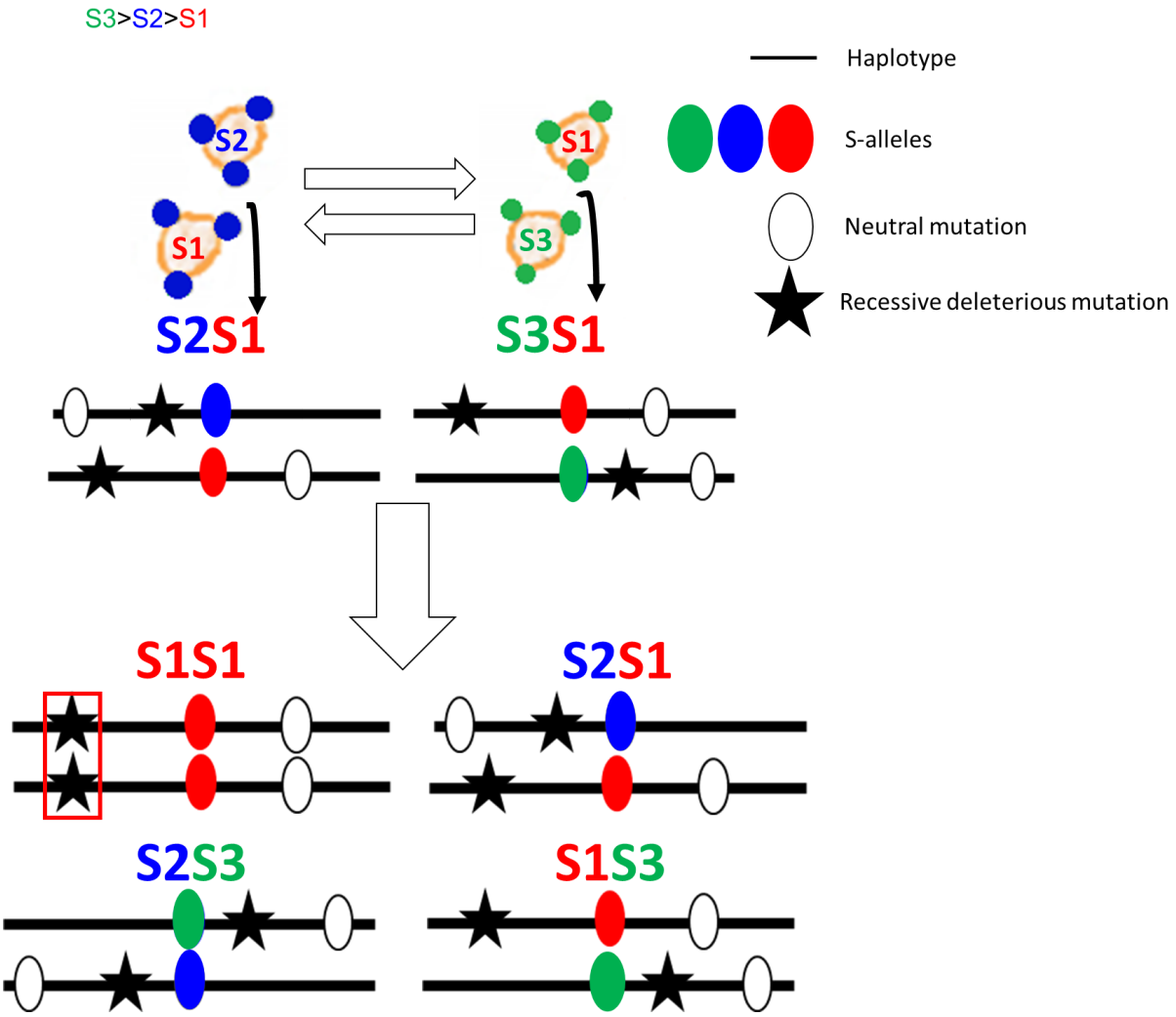


Figure 10 : Diagram of the consequences of dominance relationships between S-alleles on the accumulation of deleterious mutations linked to the S-locus in the SSI system. *We considered two individuals with three S-alleles (top). Each S-allele (red, blue and green circles) is associated to specific haplotypes (black lines) constituted of neutral (white circles) and recessive deleterious mutations (black stars). We observe two compatible crosses (white arrows) and two incompatible crosses (black arrows), including the self crosses. The offspring of the compatible crosses present one homozygous genotype for the recessive S-allele (S1, red circles). The recessive deleterious mutations associated are also homozygotes (red box). This homozygosity of these mutations promotes their exposure to natural selection, and finally their elimination.*

The impact of dominance of *S*-alleles on the genetic load they carry has been investigated by two studies (Llaurens et al., 2009a; Stift et al., 2013), on the species *A. halleri* and *A. lyrata* respectively. In *A. halleri*, the dominance effect was observed by comparing the *S*-linked component of inbreeding depression for several phenotypic traits between homozygotes for two *S*-alleles: the highly recessive Ah01 and an *S*-allele belonging to the most dominant class (Ah15), obtained by enforced incompatible crosses (Llaurens et al., 2009a). The genetic load linked to the dominant *S*-allele was more important than the genetic load linked to the recessive *S*-allele. While the results were consistent with the theoretical predictions (higher load associated with the most dominant *S*-allele), an important limitation was that the

genetic load could be compared between two *S*-alleles only, making it difficult to assess the generality of the observation. Therefore, it is necessary to consider more *S*-alleles, especially in the intermediate classes of dominance. Indeed, Llaurens et al., (2009a) predicted mostly an accumulation of the genetic load in the dominant class relative to the other classes (Fig. 11). Estimating the impact of this genetic load linked to the intermediate alleles would help to clarify whether the genetic load accumulates essentially on the dominant alleles or not. In *A. lyrata*, a significant effect of the *S*-linked component of inbreeding depression for several phenotypic traits was observed for two out of four *S*-alleles tested by forced self-crosses, but these results were not correlated with dominance of the *S*-alleles (Stift et al., 2013). The results were thus in contradiction to the theoretical predictions and did not confirm the results obtained in *A. halleri*. In conclusion, Stift et al. (2013) argue that the stochastic nature of the occurrence of important deleterious mutations in flanking regions of the *S*-locus may be more important than the deterministic effect of *S*-allele dominance. However, an important difference with the Llaurens et al. (2009a) study is that they measured the linked load expressed by individuals carrying two identical *S*-allele copies (obtained by selfing), while in *A. halleri* it was measured in individuals carrying two copies of the same functional *S*-allele obtained from different individuals. This raises the question of the relative importance of fixed deleterious mutations associated with different copies of a given *S*-allele within natural populations, and their contribution to the linked load. Hence, genomic analyses are needed in order to determine the distribution of neutral and non-neutral polymorphisms in flanking regions of the *S*-locus, in association with *S*-allele identity. Overall, while theoretical predictions have been produced regarding the effect of the *S*-locus dominance hierarchy on the linked load, the experimental results that have been published so far have remained limited and have not provided conclusive evidence in favour or against those predictions. An essential limitation is that the experimental design did not allow a large-scale comparison on large numbers of *S*-alleles, and was restricted to phenotypic approaches, hence providing no hint regarding the genomic architecture of the *S*-linked load.

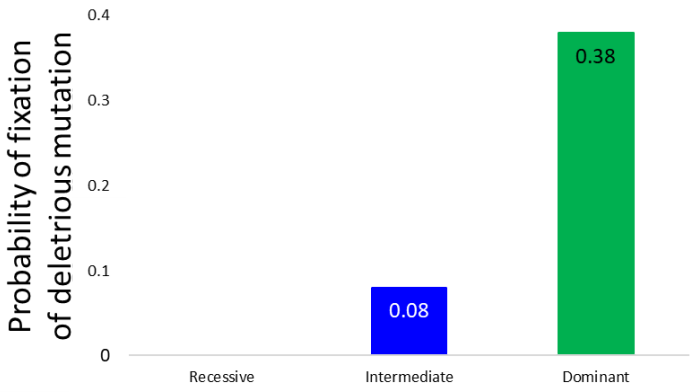


Figure 11 : Fixation probability of a deleterious mutation strictly linked to the S-locus as a function of the dominance class of the S-allele (modified from Llaurens et al., 2009a). *The results were obtained from stochastic simulations. The simulated population had 1000 diploid individuals and 5 different S-alleles. One S1 allele was in the recessive class, two alleles, S2 and S3, were in the intermediate class (blue) and were codominant to each other, and the last two, S4 and S5, were dominant over all others (green) and codominant to each other (S1<[S2=S3]<[S4=S5]). Dominance was only expressed at the pollen level.*

V) Balancing selection and the evolution of the dominance hierarchy between *S*-alleles

Balancing selection is a favourable case for dominance modifier selection because it maintains a high level of heterozygosity. SSI exhibits a case of balancing selection that favours evolution of dominance modifiers linked to the *S*-locus: phenotypic expression of only one of the two *SCR* alleles of the male diploid genotype in a heterozygote or only one of the two *SRK* alleles of the female diploid genotype in a heterozygote increases the range of possible mating partners, while still preventing self-fertilisation (Fig. 12a for an example on pollen; Schoen and Bush, 2009). Consistent with these expectations, allelic dominance relationships are common in Angiosperms (Bateman, 1952). However, these relationships can be very complex for some species, with different dominance hierarchy between pollen and pistil for example (Thompson and Taylor, 1966), which raises questions about the factors impacting the establishment of such relationships.

Based on these observations, two theoretical studies have developed models for the evolution of dominance modifiers associated with an SSI system and demonstrated that selection was sufficiently strong to promote the evolution of such modifiers (Llaurens et al., 2009b; Schoen and Bush, 2009). This evolution is dependent on different factors including allelic diversity at the *S*-locus, selfing rate and inbreeding depression. Allelic diversity was negatively correlated with the time to fix the modifier, but did not modify the final result (Schoen and Bush, 2009), whereas increase of the selfing rate and inbreeding depression prevented the *S*-allele to become more recessive (Llaurens et al., 2009a). Billiard et al (2006) showed that dominance relationships cause an asymmetry of allelic frequencies because recessive alleles are more often masked in heterozygous genotypes than dominant ones. Thus, recessive alleles should be in higher frequencies in the population than dominant alleles (Fig. 12a). Llaurens et al (2009b) find similar results, both when dominance is expressed in pollen and pistils and when it is expressed only in pollen or only in pistils. Schoen and Bush (2009) note that the establishment of dominance relationships between *S*-alleles at the female determinant alone is less advantageous than at the level of the male determinant. Llaurens et al (2009b) also show that the return to a codominant state is disfavored, because codominance limited the number of compatible mates. Overall, these studies showed that codominance is disadvantageous in the SSI system (Llaurens et al., 2009b; Fig. 12b), such that selection for dominance modifiers to escape codominance should be effective. Until recently, however, the existence and molecular nature of these putative modifiers had remained elusive .

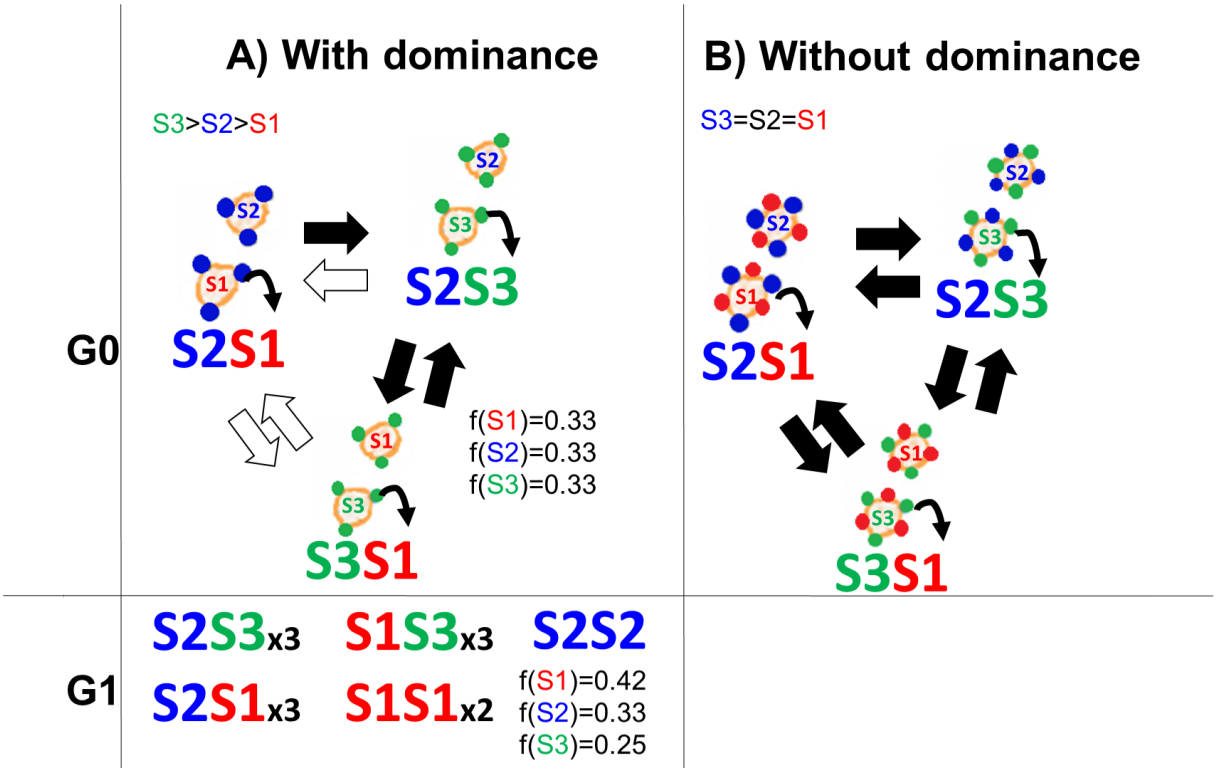


Figure 12 : Consequences of dominance relationships between S-alleles on pollen in the SSI system. *Initially, three S-alleles in egal frequency are distributed between three individuals (top). A) If we consider a linear relationship between the three S-alleles, only the more dominant S-allele is expressed on pollen (blue circle if S2 is the more dominant or green circle if S3 is the more dominant). This system entails three compatible crosses (white arrows) and six incompatible crosses (black arrows), including the self crosses. The offspring present five different genotypes (bottom), including two homozygous genotypes for the recessive (S1) and intermediate (S2) S-alleles. The frequency of recessive S-alleles is increased, while in contrast to the frequency of the dominant S-allele S3 is decreased. B) If the three S-alleles are codominant, the two S-alleles of the parent are expressed on pollen (red circle for S1, blue circle for S2, green circle for S3). We observe only incompatible crosses (black arrows), including the self crosses. Without dominance relationships between the S-alleles, the population thus needs more than three S-alleles to persist.*

VI) Molecular nature of dominance modifiers at the *S*-locus

In *Brassica rapa*, *S*-alleles are classified into two dominance classes in pollen: class I alleles are all dominant over class II alleles (recessive) and are codominant to each other. Yet, they show dominance over the class II alleles, and the latter form a linear dominance hierarchy (Yasuda et al., 2017). In contrast, in stigma, all alleles appear to be codominant. Several studies showed that dominance in pollen of class I over class II alleles in *Brassica rapa* is explained by the loss of expression of the *SCR* transcript of the recessive alleles in the presence of a dominant allele in heterozygotes (Fig. 13a; Kusaba et al., 2002; Kakizaki et al., 2003). Then it was shown that this loss of expression is associated with methylation of the promoter of recessive alleles in the presence of dominant alleles in heterozygotes (Fig. 13b, Shiba et al., 2006). Promoter methylation blocks the production of mRNA from the recessive

alleles. Finally in 2010, it was shown that this methylation is due to the action of a small non-coding RNA, *Smi*, genetically linked to dominant alleles (Tarutani et al., 2010; Finnegan et al., 2011). The linear dominance hierarchy between the class II alleles is explained by an accumulation of mutations on a second sRNA, *Smi2* (Yasuda et al., 2016).

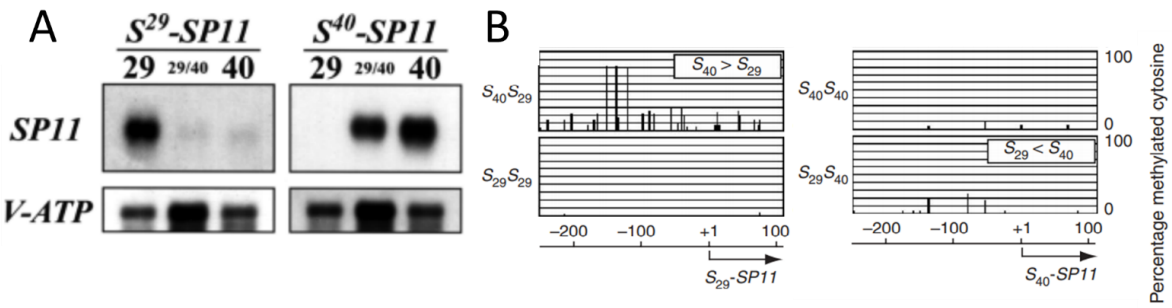


Figure 13 : Loss of expression of the S29 mRNA in S29/S40 heterozygotes by DNA methylation of the S29 promoter. *A: the RNA gel blot of S29 and S40 in Brassica Rapa pollen from S29/S29 (first lane), S29/S40 (second lane), and S40/S40 (third lane; Kakizaki et al., 2003) individuals. The S29 allele is recessive (class II) and S40 is dominant (class I). The RNA of the recessive allele (left) is not expressed in heterozygotes, unlike that of the dominant allele (right). Each allele is expressed in its respective homozygous genotype. B: Methylation profile of the promoter region of S29 (left) and S40 (right) in S29/S40 heterozygotes, S40/S40 and S29/S29 homozygotes (Shiba et al., 2006). Only the S29 promoter is methylated and only in heterozygotes (top, left).*

In *A. halleri,* Durand et al., (2014) characterised the dominance hierarchy among six *S*-alleles and showed that it was mostly linear, with only one case of codominance observed in pollen (Fig. 14a). The regulatory network of this dominance hierarchy involves at least 8 families of sRNAs (Fig. 14b) linked to the different *S*-alleles distributed in four distinct dominance classes (Durand et al., 2014). These sRNAs originate from stem-loop structures like the precursors of miRNAs (Carthew and Sontheimer, 2009). The precise transcriptional repression pathways by which they achieve their function are being investigated in the lab. These sRNA families seem to have appeared successively and independently in the different allelic lineages. This observation raises the question of the mechanisms by which such a complex regulatory network can evolve.

VII) Evolution of dominance interactions by mutations of sRNAs and their targets

Some of these sRNAs, such as *mirS3*, are shared in a large number of phylogenetically distant allelic lineages (Fig. 15a), suggesting an ancient origin. These sRNAs occur on the different allelic lineages with different nucleotide sequences, with potential consequences on the recessive *SCR* alleles that they are able to repress. Burghgraeve et al. (2020, Fig. 15b) showed that the interaction between the sRNAs and their targets entails a threshold model, such that even point mutations either in the sRNA or its targets can dramatically affect the silencing efficiency of the interaction. Beside the accumulation of point mutations on existing

interacting partners, the network can also be modified by the acquisition of new targets by recessive *SCR* alleles. The fact that several sRNAs have different target sites on the *SCR* gene suggests that target sites can be dynamically acquired and lost, each time modifying the regulatory network. Finally, some sRNAs such as *mirS4* or *mir867*, are shared among only a small number of closely related lineages, suggesting a recent origin. The existence of these recent sRNAs raises questions about the molecular mechanisms by which they have emerged. Different origins of sRNAs have been suggested in the plant kingdom, in particular regarding miRNAs, and involve reverse duplication of portions of their target genes, duplication of another sRNA precursors, insertion of transposable elements, or spontaneous formation of a stem-loop structure by a succession of mutations (Nozawa et al., 2012). However, the mutational mechanisms by which the new sRNA at the *S*-locus such as *mirS4* or *mir867* have not been investigated. Overall, the evolution of the network can entail mutations on either the sRNAs or their targets, and can involve the creation of novel genetic elements (either new sRNAs or new target sites). At this stage, the relative importance of these mechanisms remains unclear.

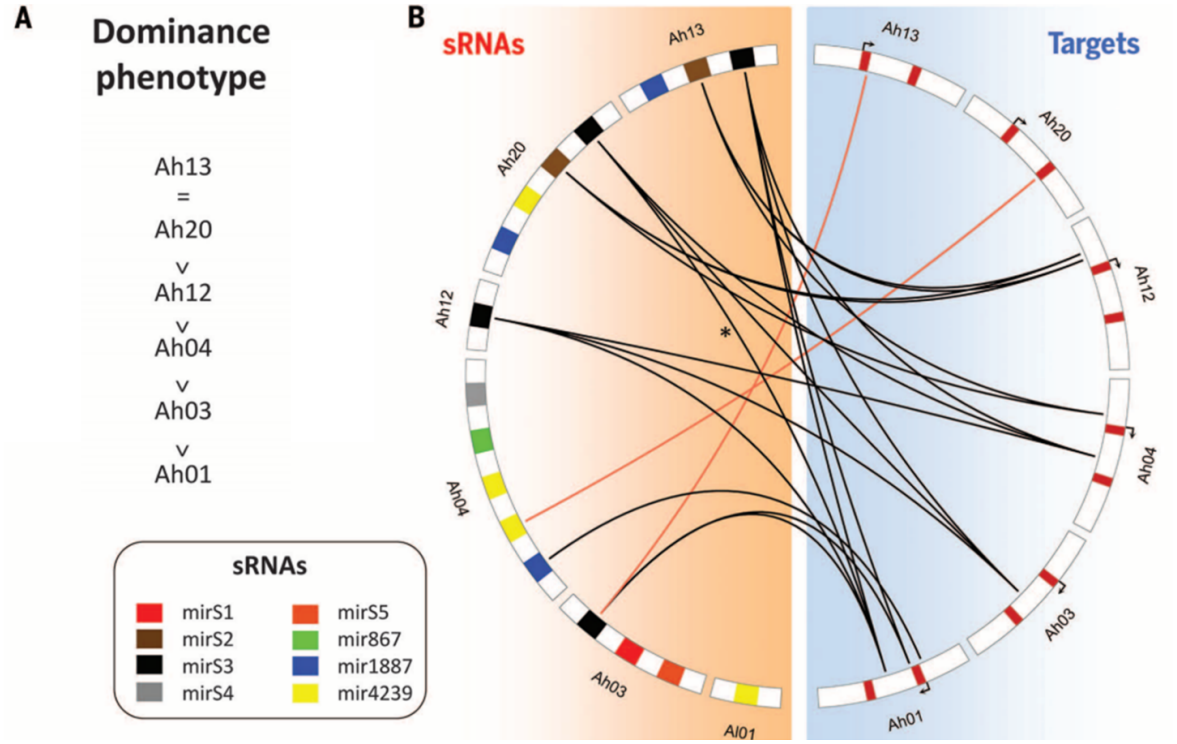


Figure 14 : The observed dominance relationships between 6 S-alleles (A) is partially explained by interactions between 8 sRNA families (B) distributed over these alleles and their targets in A. halleri (Durand et al., 2014). *A) Dominance/recessivity hierarchy between six SCR alleles based on phenotypic evaluation. B) Prediction of sRNA/SCR allele interactions. Each haplotype is represented twice : the left portion of the circle illustrates the repertoire at sRNA-producing loci (e.g. mirS2 is in brown and is present in haplotypes S13 and S20) and the right portion illustrates the small RNA targets at the SCR alleles. SCR alleles (2 exons in red separated by an intron) are shown to expand over 1 kb in the 5' and 3' directions. Predictions in agreement with the dominance phenotype are indicated by black connections, while those opposed to the phenotype are indicated by red connections*

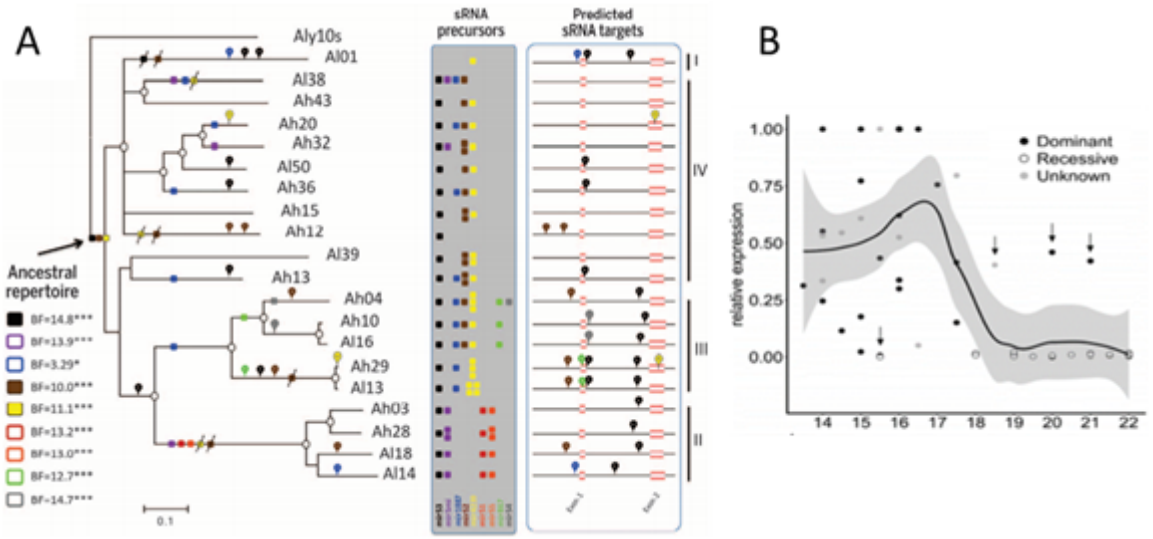


Figure 15 : Repertoire of sRNAs and their targets along the S-allele phylogeny . *A) The phylogeny is based on the complete SRK amino acid sequences of A. halleri and A. lyrata. Nodes with posterior probabilities > 0.95 are represented by white circles. Phylogenetic classes are plotted (I, II, III and IV). Solid and open squares indicate precursors assumed to be present and absent from the ancestral repertoire, respectively. For each precursor, asterisks for Bayes factor (BF) indicate the level of confidence in the presence or absence of the ancestral repertoire (Durand et al., 2014). B) Relative expression of SCR alleles as a function of the alignment score of the "best" interaction between the focal allele (including 2kb of sequence upstream and downstream of SCR) and the population of sRNAs produced by sRNA precursors of the other allele in the genotype. For each allele, expression was normalised relative to the genotype in which the expression was highest. Dots are coloured according to the dominance status of the focal SCR allele in each genotypic context (black: dominant; white: recessive; grey: undetermined). The black line corresponds to a local regression obtained by a smooth function (loess function, span=0.5) and the grey area covers the 95% confidence interval. Vertical arrows point to observations that do not fit the threshold model of transcriptional control (from Burghgraeve et al., 2020).*

Previous models for the evolution of dominance interactions between *S*-alleles assumed dominance modifiers that were single genetic elements : a given *S*-allele was placed along the dominance hierarchy by the identity of the dominance modifier it carried and could be modified freely by mutation. However, it is now clear that dominance is actually controlled by the interaction between pairs of genetic elements : the sRNAs and their target sites. Because a mutated *S*-allele would carry the repertoire of sRNAs and targets of its ancestor, it is now difficult to imagine that an allele could completely change dominance level in a pre-existing hierarchy because this would imply that previous interactions have been eliminated. Thus, the initial level of dominance of *S*-alleles before the appearance of a causal mutation allowing the establishment of a new interaction must be considered. The evolution of an interaction involving two or more alleles can be expected to depend on their initial

places in the hierarchy. This possible importance of the initial level of dominance of the allele in a dominance network has also never been studied in itself. Moreover, since the level of dominance impacts allelic frequencies, one would expect that the nature of the causal mutation, on the sRNA linked to the dominant allele or on the target of the recessive allele, would impact the probability of fixation of this mutation creating a new interaction. Indeed, since recessive alleles are more frequent in the population, we can expect that the mutation that creates an interaction is more advantageous on the target of recessive allele that is more often hidden. Because this notion of interaction between the modifier and its target was not known at the time of the first models, the impact of the nature of the mutation creating a new interaction has not been studied.

Finally, Llaurens et al., (2009b) showed that the genetic load linked to the *S*-locus is predicted to decrease the probability of an allele to become recessive, but also to select for alleles that become dominant, because the latter express their genetic load less often. Thus, in this previous model of dominance evolution, it was inferred that dominance evolved on an "ascending scale", where *S*-alleles would become increasingly dominant. However, as mentioned above, this model did not take into account the actual architecture of the dominance modifiers as separate genetic elements. Whether the predictions formulated by Llaurens et al. (2009a) hold under this genetic architecture has not been investigated.

Objectives of the thesis

The objectives of this thesis were, first, to evaluate the evolutionary consequences of balancing selection and dominance between *S*-alleles on overall polymorphism and the genetic load in genomic regions flanking the *S*-locus, and second, to deepen our knowledge of the factors impacting the evolution of the dominance network between *S*-alleles as observed in *Arabidopsis halleri*. The thesis comprises three chapters:

(1) The objective of the first chapter was to delineate the flanking region linked to the *S*-locus whose polymorphism is indirectly impacted by balancing selection at the *S*-locus and is likely to accumulate the linked genetic load that was observed in *A. halleri* and *A. lyrata*. For this purpose, I performed the first comprehensive polymorphism analysis in the complete genomic regions comprising 75kb on either of the *S*-locus in *A. halleri* and *A. lyrata*, which I compared to 100 matched control regions from across the genome

(2) The objective of the second chapter was to estimate the impact of the dominance hierarchy on the mutations accumulated in the flanking regions of the *S*-locus and the consequences of these mutations on the fitness of individuals. For this, I reconstructed haplotypes for different *S*-alleles from two populations of *A. halleri* and one population of *A. lyrata*, based on sequencing of parents-offspring trios. In each haplotype, I tested for a correlation between the accumulation of deleterious mutations and the dominance level of the *S*-allele associated. I distinguished in particular the mutations that were fixed from those that were segregating within allelic lineages. I completed these analyses with stochastic models to compare the expected distribution of deleterious mutations between lineages of dominant and recessive *S*-alleles. Finally, I measured a series of phenotypes to evaluate the phenotypic impact of the linked load associated with three *S*-alleles of three different dominance levels.

(3) The objective of the third chapter of the thesis was to deepen our knowledge on the factors that can impact the evolution of the dominance network between *S*-alleles. For this purpose, I simulated the evolution of a linear dominance system controlling the expression of *S*-alleles at the pollen level, and then studied the effect of different factors, including the genetic load associated with *S*-alleles. I compared in particular the fate of mutations of different molecular nature that created a new dominance interaction. I complemented these models by a detailed molecular and phenotypic analysis of the dominance network observed in *A. halleri*.

The first chapter is in the form of an article in preparation for the journal "Molecular Biology and Evolution". The next two chapters are written in English, as draft articles.

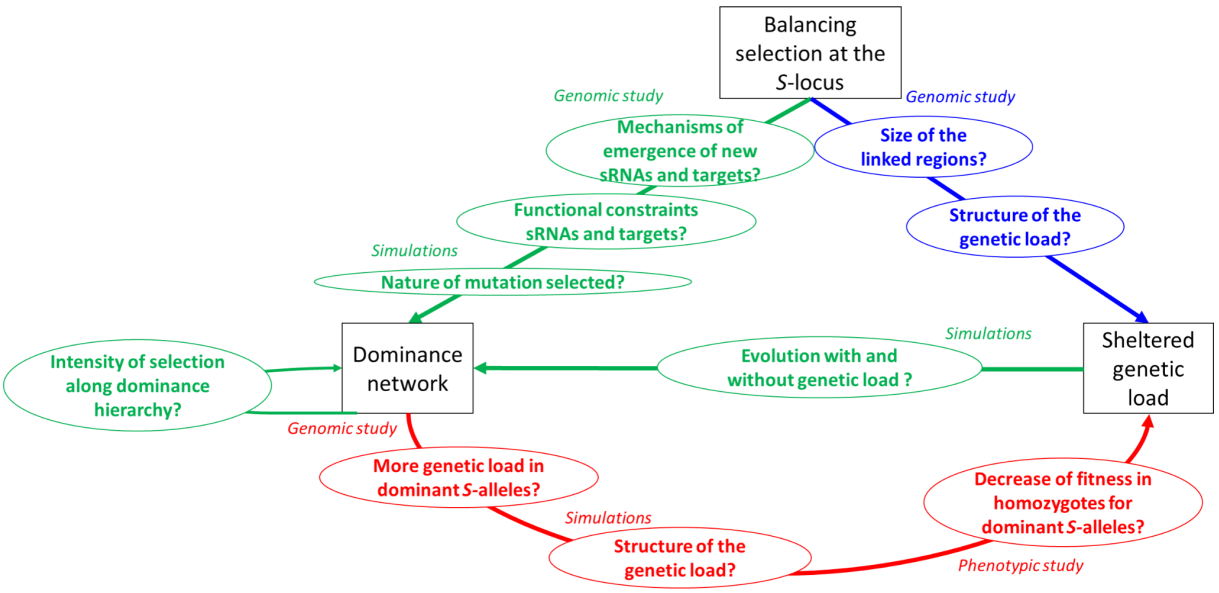


Figure 16: Schematic representation of the organisation of the questions addressed in the thesis. *The lines symbolise the relations between the partners of the ménage à trois (black square) studied. The questions relative to each relation are outlined in the ovals. The methods used to tackle each question are noted in italic. The blue lines represent the relations studied in the first chapter. The red lines represent the relations studied in the second chapter. The green lines represent the relations studied in the last chapter.*

Chapter I

Scientific question:

The first axis of this thesis aims to answer the following questions:

- To what extent is polymorphism in the flanking regions at the *S*-locus impacted by balancing selection?
- Does balancing selection at the *S*-locus cause the accumulation of a detectably increased genetic load in the flanking regions, and is efficiency of purging against deleterious mutations decreased ?

To answer these questions, we compared polymorphism in the complete nucleotide sequences of the 75kb flanking regions on either sides of the *S*-locus to that in 100 randomly chosen regions used as internal genomic controls for three sample sets of *A. halleri* and three sample sets of *A. lyrata.*

Contribution:

Of these six sample sets, three were previously published datasets that I retrieved, and the three others were newly sequenced after gene capture. The *A. lyrata* individuals sequenced by capture were derived from seeds provided by B. Mable and they were grown in the laboratory greenhouse. The *A. halleri* individuals were sequenced by capture and came from two sampling trips in the natural populations carried out by Chloé Ponitzki, Eleonore Durand, Vincent Castric and myself.

The design of the probes for the sequence capture protocol was done by Nicolas Burghgraeve in interaction with the company Mybaits before I started my thesis project. The molecular biology experiments (genomic libraries construction and sequence capture) were carried out by Christelle Lepers-Blassiau and myself. The sequencing was performed by the LIGAN-MP genomics platform (Lille).

The pipeline for read alignment and variant calling was developed by Mathieu Genete and Nicolas Burghgraeve before I started my project. I have developed all the python codes allowing the filtration of variants, the preparation of files for the calculation of B_{2maf}, $\prod$, MAF, Ho, Tajima's D, but also the study of the *A. thaliana* / *A. lyrata* divergence or the use of the MLHKA. Finally, all the statistical analyses in R were done by myself.

This first chapter is in the form of an article in preparation for the journal "Molecular Biology and Evolution".

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], 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*

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 expected to shelter deleterious mutations, facilitating the local accumulation of deleterious mutations, called the “sheltered” load. The impact and extent of balancing selection on polymorphism of the linked regions and on the accumulation of the sheltered genetic load, however, remain poorly documented. Self-incompatibility offers the opportunity to study the indirect effects of long-term balancing selection. Here, we provide the first genomic demonstration of the relation between the balancing selection and the sheltered genetic load in a plant genome. We used targeted genome resequencing to evaluate the intensity of indirect selection on the genomic region flanking the *S*-locus in three sample sets in each of the two closely related plant species, *Arabidopsis halleri* and *A. lyrata.* We found significantly increased polymorphism over the first 25kb around the *S*-locus in all sample sets. In contrast to the classical model for the accumulation of a sheltered load, we found that these genes accumulated more mutations than those in control regions, but the ratio of non-synonymous to synonymous polymorphisms is also unchanged. The comparison between the *S*-flanking regions and control regions provided a powerful way to factor out differences in demographic histories and/or sample structure. Overall, our results reveal that one of the strongest balanced polymorphisms does indeed result in elevated polymorphism of the adjacent genomic regions, but is not associated with a detectable relaxation of the efficacy of purifying selection.

Keywords: balancing selection, sheltered load, deleterious mutations, S-locus, polymorphism.

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 the balancing selection in maintaining genetic diversity on the genome was widely debated (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), partly because of the inherent technical challenges related to high levels of polymorphism in the genomic regions affected (Vekemans et al. 2021). Genome resequencing studies have revealed that balancing selection can be a potent force throughout the genome (DeGiorgio et al. 2014), but 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 three species only: *Solanum carolinense* (Stone 2004), *Arabidopsis halleri* (Llaurens et al. 2009) and *A. lyrata* (Stift et al. 2013). In these three 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 three 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 sequence 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 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 genomic 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).

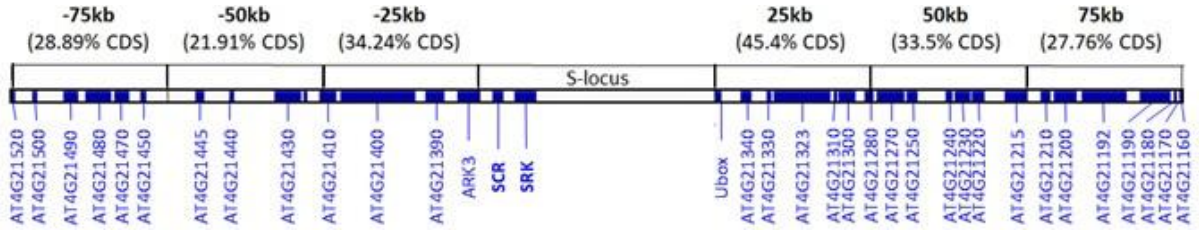


Figure 1: Schematic representation of the S-locus and its flanking region. *Protein-coding genes in the flanking regions are represented as filled blue rectangles. The genomic regions studied are distributed between positions 9,264,458 and 9,451,731 along chromosome 7 of the A. lyrata genome assembly (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* (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 the polymorphism in *S*-flanking regions, where we were able to interrogate an average of 74,866 sites, containing 3,225 variables 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 region.

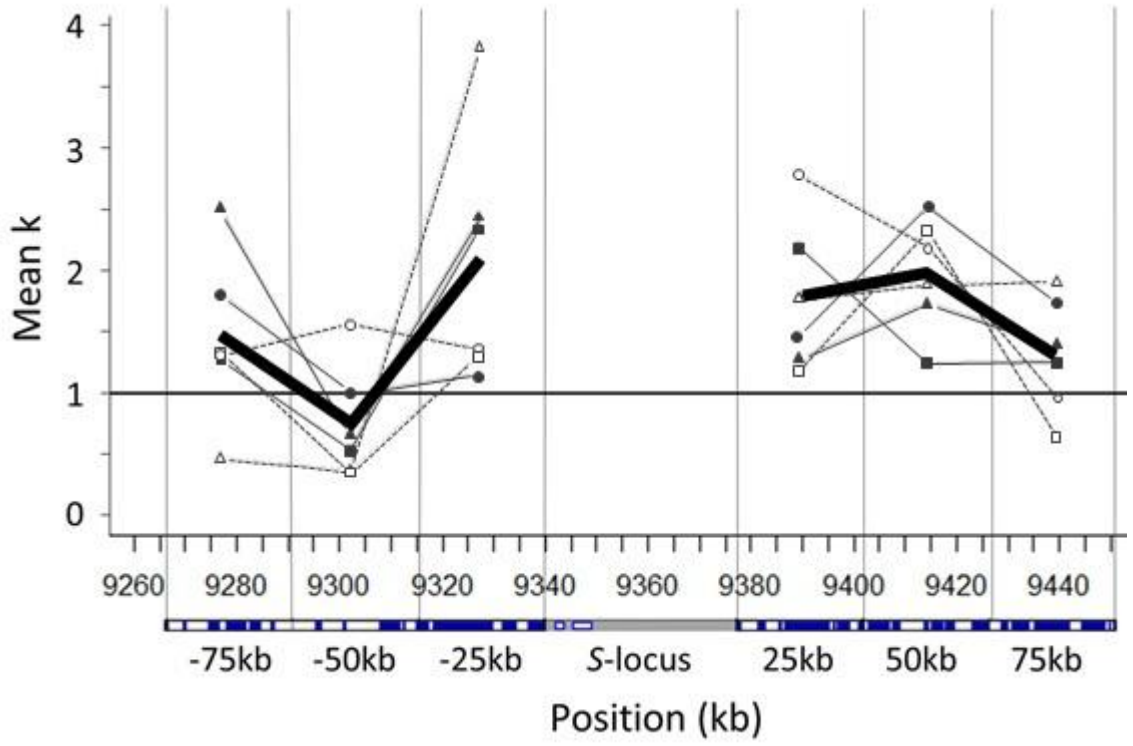


Figure 2: Variation of the mean selection parameter (k) obtained for genes in the S-flanking regions. *The solid large black line represents the mean value of k obtained across the six sample sets. The black solid lines represent the A. lyrata sample sets (square=Plech, circle=Spiterstulen, triangle=North 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.*

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 even though it was not the most extreme, it was the most consistent signal of balancing selection detected across all sample sets (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.

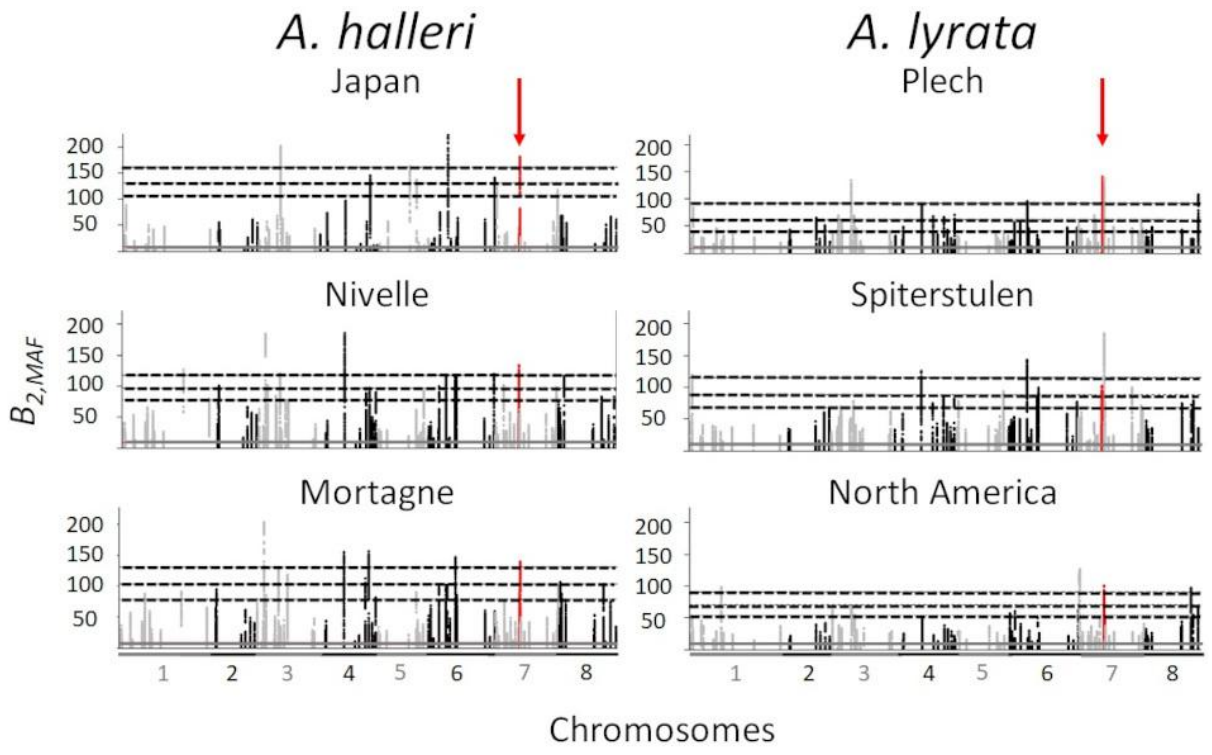


Figure 3: Manhattan plots for signals of balancing selection ($B_{2,MAF}$ scores) in the 100 control regions and the S-locus region (red dots and arrows) along the A. lyrata genome. *Chromosomes 1-8 of the A. lyrata genome, with the 100 control regions distributed on successive chromosomes represented 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 the 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. 7).

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 distance away from the *S*-locus increased. 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.

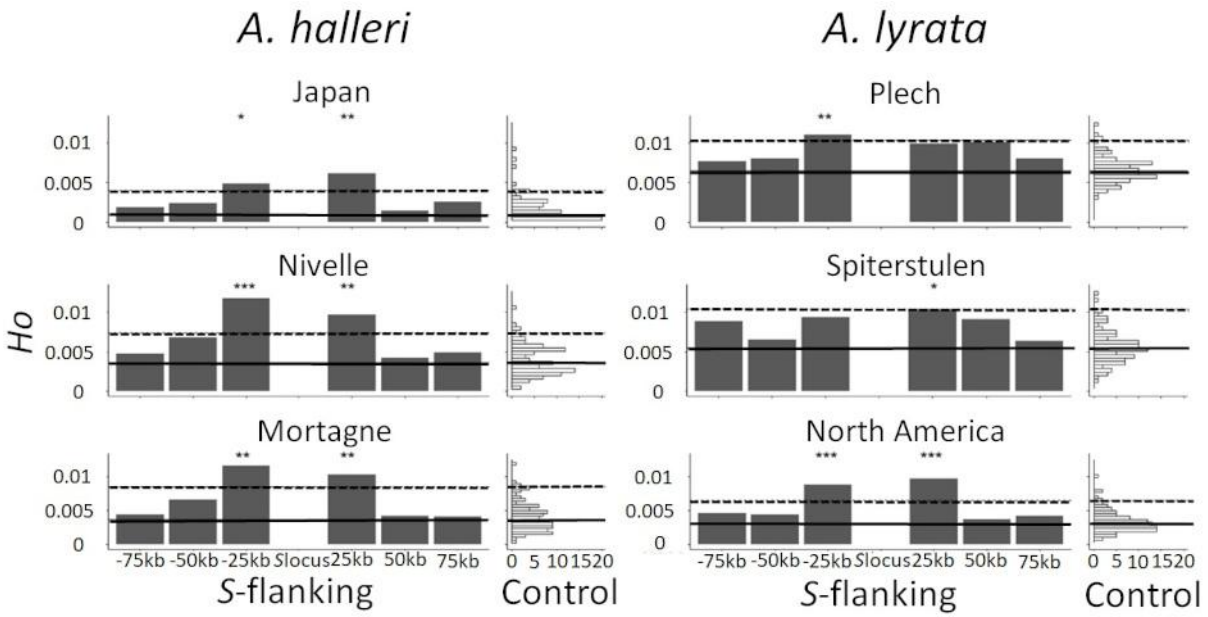


Figure 4: Mean Ho around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of Ho obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of Ho mean in the 100 control regions are represented by a vertical histogram 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.*

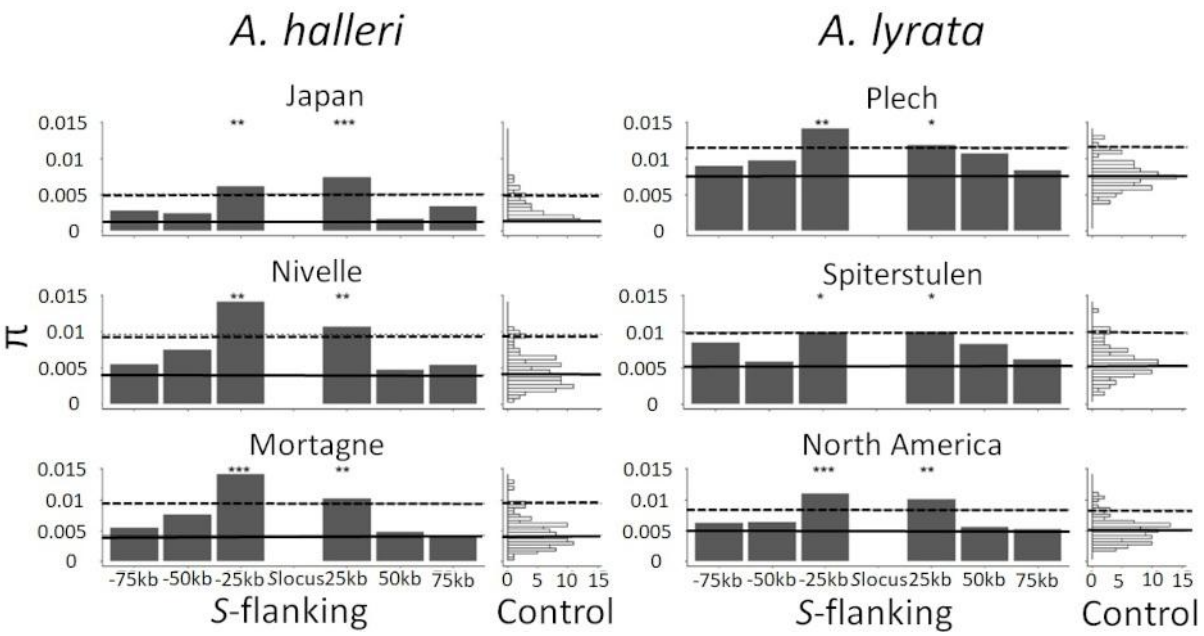


Figure 5: Mean π around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of π obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of π mean in the 100 control regions are represented by a vertical histogram 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.*

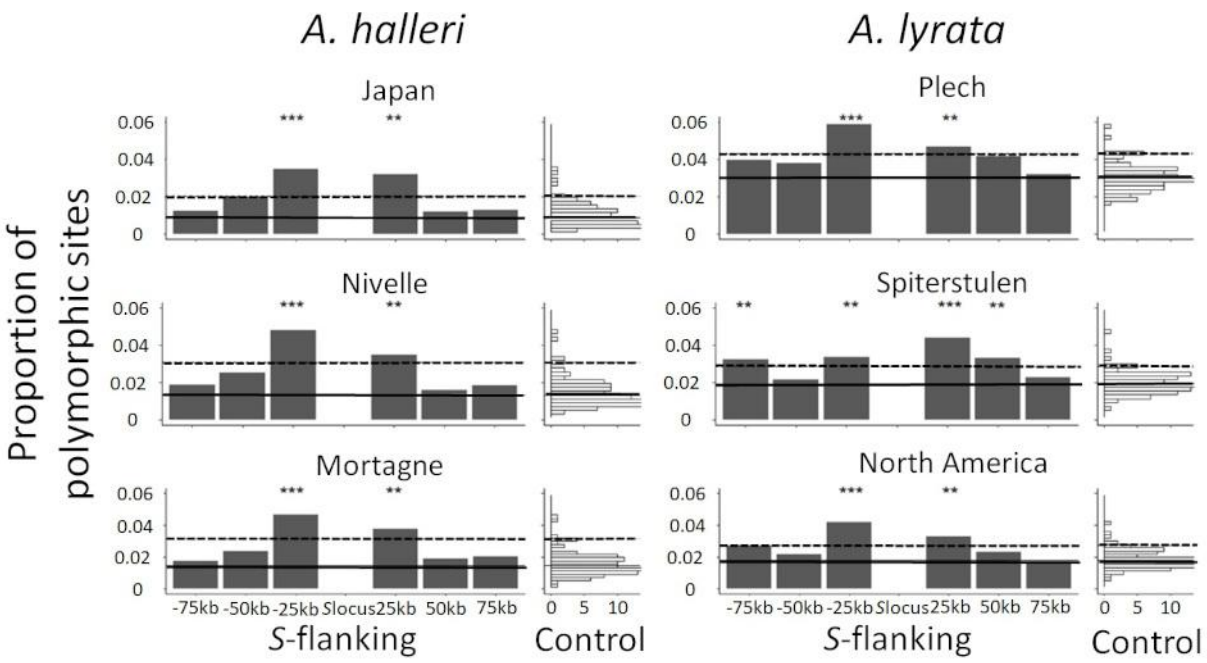


Figure 6: Proportion of polymorphic sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of polymorphic 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 on 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.

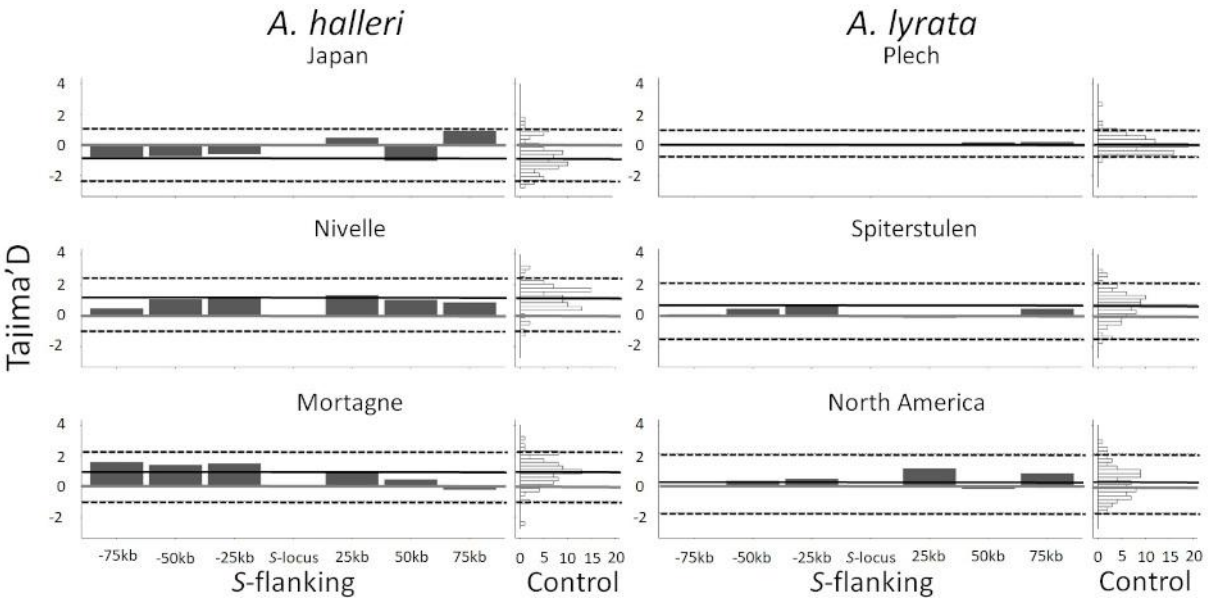


Figure 7: Tajima's D around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean of Tajima's D obtained in S-flanking windows of 25kb. The distributions (count) in the 100 control regions are represented by a histogram (right). The 97.5% and 2.5% percentiles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

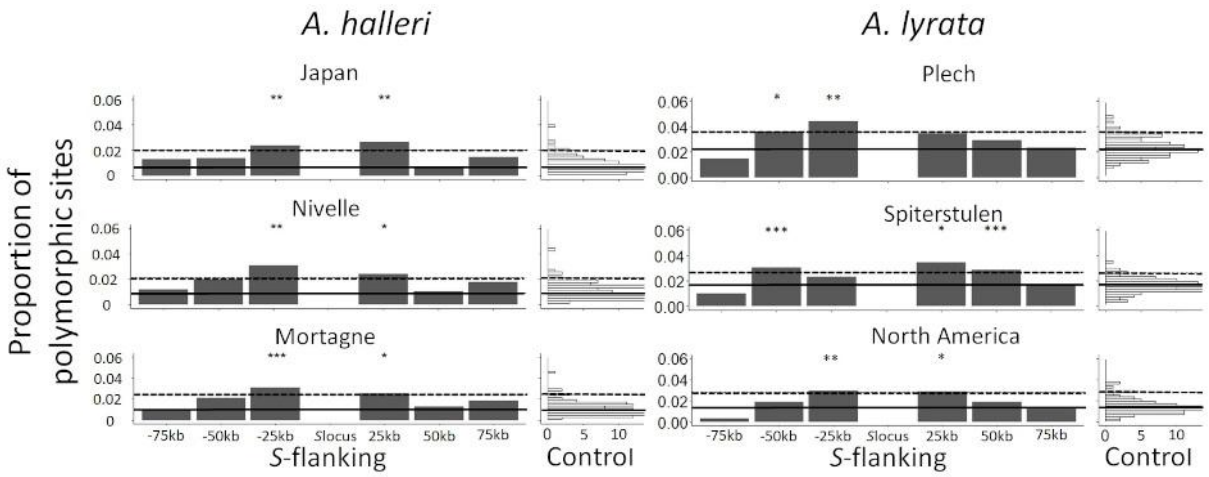


Figure 8: Proportion of polymorphic sites among 0-fold degenerate sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) 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*.

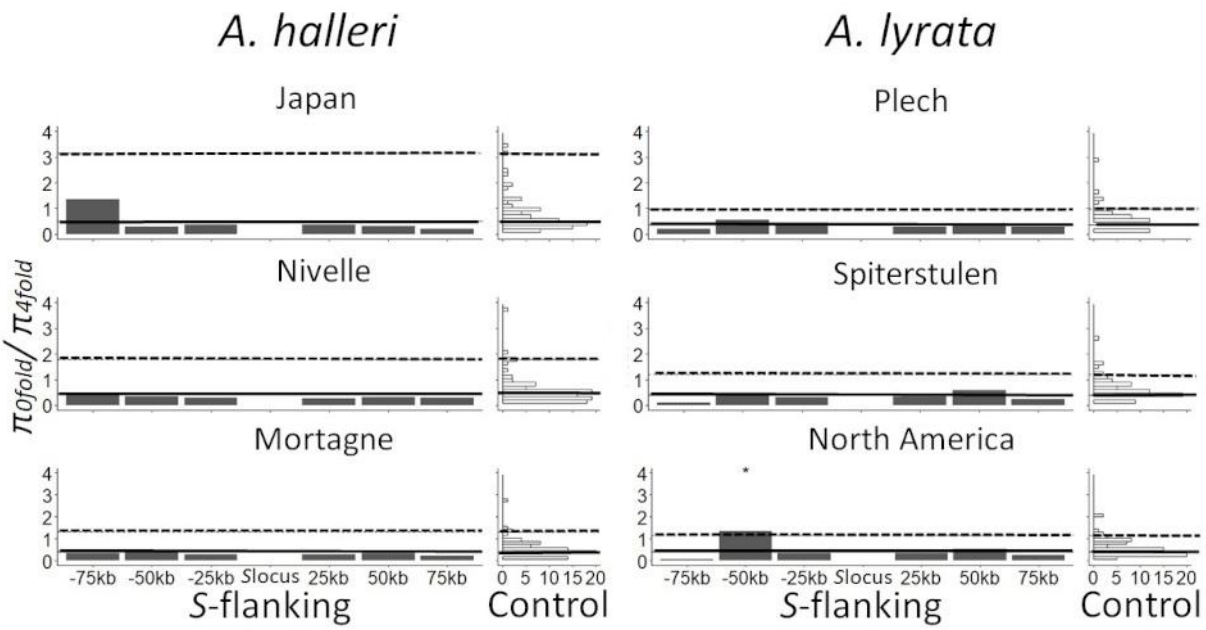


Figure 9: $\pi_{0\text{-fold}}/\pi_{4\text{-fold}}$ ratio around the S-locus and across the control regions from throughout the genome. *The bars represent the proportion of polymorphic sites obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of the proportion of 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

Our analysis across several replicate geographic regions in two closely related species is the first comprehensive genomic study to reveal the extent of the 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. 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 it were the case that actual 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, but based on the published estimates of *S*-allele frequencies in one of the populations studied here (the Nivelle population; Llaurens et al. 2008), it is 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 (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 be expected to 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 sharply 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, at least in *A. halleri* and *A. lyrata*, may be linked 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. Moreover, recombination would also disrupt the functioning of the male-female pairing at the *S*-locus.

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) 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 self-incompatibility, 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 in the *S*-locus region itself. While Stone (2004), Stift et al. (2013) and Llaurens et al. (2009) provided phenotypic evidence for a sheltered load linked to the *S*-locus, our study provides the first genomic demonstration of an accumulation of potentially deleterious mutations in *S*-flanking regions. Identifying more precisely the mutations causing the 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. The phenotypic traits on which the load was documented in the different studies varied (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*. 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 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 (Lynch 2006 ; 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 (Lynch 2006; 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 and Nivelle for *A. halleri*, and Plech, Spiterstulen 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 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 our targeted sequence capture approach.

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 (37G17 and 21E5 respectively ; 10.6084/m9.figshare.16438908). We computed the proportion of CDS and 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 first 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 transposable elements (TE) 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 (based on 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 the 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. The final number of sites in each sample set is summarised in Tables 1 and 2. We assumed that mutations on 0-fold degenerate sites were deleterious, in contrast to mutations on the 4-fold degenerate sites (i.e. positions that did not change amino acid). Mutations on the 2- or 3-fold degenerate sites were not studied for simplification.

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. Mutations on the 2- or 3-fold degenerate sites were not studied for simplification.

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.

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 sequence data are available in 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 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.

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.

Lynch M. 2006. The origins of eukaryotic gene structure. Molecular Biology and Evolution. 23: 450–468.

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, Charlesworth D. 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. Genetics Research. 76: 51–62.

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. 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.

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.

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* [b]	*π* [b]	*MAF* [b]	Proportion of polymorphic sites [b]
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. [b] The values of Ho, π, MAF and of the proportion of polymorphic sites were multiplied by 10^3.

Supplementary informations

The capture array was initially designed to enable the study of a range of genomic regions addressing different scientific questions in our lab. Thus, in addition to the *S*-locus flanking and control regions used in the present study, the capture array also contained a set of probes targeted towards: 1) a library of complete *S*-locus sequences for 36 *S*-alleles obtained from BAC clones ; 2) a library of 123 partial or complete *SRK* sequences from several brassicaceae species (*A. halleri, A. lyrata, A. thaliana, A. kamchatica, Capsella grandiflora, C. rubella, Brassica rapa, B. oleaceae*) ; 3) a set of 185 microRNA genes and their predicted mRNA target sites across the *A. halleri* genome; and 4) a candidate QTL region for heavy-metal tolerance. These additional probes were not used in the frame of the present project, and based on the absence of substantial sequence similarity, they are unlikely to interfere with our results. The complete list and sequences of the probes are available on the figshare database (10.6084/m9.figshare.16438908).

Supplementary data

Tables

Table 1: Summary of the sample sets used.

Species	Populations	Reference	Sample Size	Accession N°
A. halleri	Japan	*Kubota et al. 2015*	47[a]	DRA003268
	Nivelle	*This study*	25	PRJNA744343
	Mortagne	*This study*	27	PRJNA744343
A. lyrata	Plech	*Takou et al. 2021*	18	PRJEB34247, PRJEB33206
	Spiterstulen	*Takou et al. 2021*	23	PRJEB34247, PRJEB33206
	North America	*This study*	26[b]	PRJNA744343

[a] The sample of Japan was represented by 17 individuals from Fujiwara, 17 from Ibuki, 2 from Inotani, 3 from Itamuro, 4 from Minoo and 4 from Okunikkawa. [b] The sample of North America was represented by 8 individuals from IND, 10 from PIN and 8 from TSS.

Table 2: Variation of the median Ho, π, MAF and proportion of polymorphic sites on 0fold sites in control regions in each sample set.

Species	Sample names	Number of positions considered	*Ho*[a]	*π*[a]	*MAF*[a]	Proportion of polymorphic sites[a]
A. halleri	Japan	224,953	0.61	0.9	0.62	7.27
	Nivelle	259,398	2.31	2.78	1.95	9.27
	Mortagne	265,238	2.47	2.71	1.86	10.3
A. lyrata	Plech	210,156	4.89	5.8	4.01	23.02
	Spiterstulen	173,857	4.96	4.72	3.51	16.71
	North America	85,165	2.66	3.82	2.65	13.71

[a] *The values of Ho, π, MAF and the proportion of polymorphic sites have been multiplied by 1000.*

Table 3: Variation of the log likelihood-ratio statistics for the S-flanking genes obtained by the MLHKA test..

Species	Populations	Log likelihood-ratio statistics	P value[a]
A. halleri	Japan	1230	0
	Nivelle	750	0
	Mortagne	598	0
A. lyrata	Plech	422	0
	Spiterstulen	683	0
	North America	1247	0

The results were obtained with the maximum likelihood multilocus HKA framework developed by Wright & Charlesworth (2004). This framework tested for an overall difference in polymorphism between the set of 33 S-locus flanking genes against 67 control genes randomly choose by comparison between a model with free mutation at each locus and no selection against a model with free mutation and selection on the S-locus flanking genes for each dataset. Divergence was estimated by comparison with the A. thaliana genome. [a] *The P values were obtained after a log-likelihood ratio test with 33 degrees of freedom*

Table 4: Variation of the k parameter in each gene in the S-flanking regions of 25kb in A) 5' and B) 3' regions around the S-locus and in each dataset after the MLHKA test.

A)

	genes	AT4G21410		AT4G21400		AT4G21390		AT4G21380	
Pop	**Species**	**theta**	**k**	**theta**	**k**	**theta**	**k**	**theta**	**k**
Japan		0.00492507	0.314132	0.001037	1.85635	0.0057485	**2.34882**	0.00594385	0.722337
Nivelle		0.00786454	0.819052	0.002327	**2.19794**	0.0130924	0.588675	0.00620896	1.86276
Mortagne	*A. halleri*	0.00712319	**7.926**	0.002399	1.62749	0.00795381	1.41259	0.00672362	**4.32222**
Plech		0.010844	**3.34947**	0.002846	1.30877	0.0122983	1.98657	0.00825695	**2.80683**
Spiterstulen		0.006385	1.14679	0.00126	**2.00982**	0.00981867	0.839313	0.00655597	0.631839
North America	*A. lyrata*	0.00343109	0.765859	0.00144436	**2.54048**	0.00814	**4.62788**	0.008937	1.7486

B)

	genes	AT4G21350		AT4G21340		AT4G21330		AT4G21323		AT4G21310		AT4G21300		AT4G21280	
Pop	**Species**	**theta**	**k**	**theta**	**k**	**theta**	**k**	**theta**	**k**	**theta**	**k**	**theta**	**k**	**theta**	**k**
Japan		0.00889017	0.458743	0.00592575	0.870633	0.006253	2.09104	0.00488317	0.973759	0.010381	1.46187	0.00291672	2.51209	0.00589149	0.0310246
Nivelle		0.0102413	2.21437	0.00807714	0.777357	0.00475239	4.49865	0.00265532	2.60695	0.011381	2.45664	0.00682246	1.93612	0.0108968	0.163555
Mortagne	*A. halleri*	0.013252	1.20745	0.00696161	0.975013	0.00840647	4.34776	0.00373788	2.68097	0.0156618	1.17268	0.00648723	3.49745	0.0132466	0.148493
Plech		0.015023	2.28128	0.011355	1.4824	0.00734886	1.32345	0.00444843	2.44496	0.012506	1.08323	0.0111151	1.40248	0.0144089	0.606361
Spiterstulen		0.00638373	2.02058	0.00911717	0.888607	0.00671305	0.488301	0.00244937	3.85729	0.00920909	1.5275	0.00604523	1.60479	0.0132073	0.529902
North America	*A. lyrata*	0.00961609	1.60139	0.00396684	4.70229	0.00813525	1.42165	0.00332678	2.98351	0.0111035	4.32987	0.0027026	4.2177	0.00487777	0.870109

In this model, k measures the degree to which diversity is increased or decreased by the action of selection at each gene. The genetic diversity at each gene was estimated by theta

Table 5: Variation of Ho, Pi, MAF and proportion of polymorphic sites in the 0-fold degenerate sites in S-flanking regions of 25kb in each sample set.

				Ho		Pi		MAF		Proportion of polymorphic sites	
Population	Species	Number of positions considered	*S*-flanking region	Value	*S*-locus region /Median control	Value	*S*-locus region /Median control	Value	*S*-locus region /Median control	Value	*S*-locus region /Median control
Japan	*A. halleri*	65509	-75kb	1.89	1.97	2.79	2.18	2.05	2.27	1.22	1.38
			-50kb	2.35	2.45	2.36	1.84	1.38	1.53	2	2.26
			-25kb	**4.9**	**5.11**	**6.17**	**4.82**	**3.97**	**4.4**	**3.46**	**3.92**
			25kb	**6.17**	**6.44**	**7.41**	**5.78**	**5.15**	**5.71**	**3.2**	**3.62**
			50kb	1.46	1.53	1.61	1.25	1.03	1.15	1.22	1.37
			75kb	2.51	2.62	3.35	2.61	2.56	2.84	1.3	1.46
Nivelle	*A. halleri*	79136	-75kb	4.67	1.3	5.51	1.35	4.13	1.48	1.9	1.46
			-50kb	6.8	1.89	7.48	1.83	5.1	1.83	2.51	1.92
			-25kb	**11.7**	**3.26**	**14.14**	**3.46**	**10.05**	**3.6**	**4.8**	**3.67**
			25kb	**9.66**	**2.69**	**10.71**	**2.62**	**7.58**	**2.72**	**3.51**	**2.68**
			50kb	4.15	1.16	4.71	1.15	3.3	1.18	1.62	1.24
			75kb	4.91	1.37	5.44	1.33	3.69	1.32	1.84	1.41
Mortagne	*A. halleri*	79299	-75kb	4.42	1.19	5.49	1.29	4.06	1.39	1.75	1.19
			-50kb	6.65	1.79	7.58	1.79	5.59	1.91	2.36	1.61
			-25kb	**11.62**	**3.13**	**14.11**	**3.33**	**10.03**	**3.42**	**4.67**	**3.19**
			25kb	**10.32**	**2.78**	**10.21**	**2.41**	**7.07**	**2.41**	**3.75**	**2.56**
			50kb	4.19	1.13	4.72	1.11	3.22	1.1	1.87	1.28
			75kb	4.1	1.11	4.17	0.98	2.87	0.98	2.03	1.39
Plech	*A. lyrata*	85359	-75kb	7.68	1.19	8.97	1.18	6.05	1.18	3.94	1.32
			-50kb	8.01	1.24	9.69	1.28	6.39	1.25	3.81	1.27
			-25kb	**11.02**	**1.71**	**14.1**	**1.86**	**9.45**	**1.85**	**5.9**	**1.97**
			25kb	9.89	1.53	**11.79**	**1.56**	**7.95**	**1.55**	**4.7**	**1.57**
			50kb	10.12	1.57	10.69	1.41	7.05	1.38	4.15	1.39
			75kb	8.02	1.24	8.37	1.1	5.84	1.14	3.22	1.08
Spiterstulen	*A. lyrata*	77834	-75kb	8.84	1.62	8.45	1.59	6.13	1.64	**3.26**	**1.7**
			-50kb	6.53	1.19	5.83	1.09	4.1	1.1	2.17	1.14
			-25kb	9.29	1.7	**9.87**	**1.85**	6.97	1.87	**3.36**	**1.75**
			25kb	**10.38**	**1.9**	**9.98**	**1.87**	6.85	1.83	**4.41**	**2.3**
			50kb	9.09	1.66	8.28	1.56	5.75	1.54	**3.34**	**1.74**
			75kb	6.34	1.16	6.17	1.16	4.38	1.17	2.29	1.19
North America	*A. lyrata*	58060	-75kb	4.66	1.48	6.24	1.26	4.38	1.24	2.71	1.52
			-50kb	4.45	1.41	6.36	1.28	4.52	1.28	2.16	1.21
			-25kb	**8.88**	**2.82**	**10.99**	**2.21**	**7.33**	**2.08**	**4.19**	**2.35**
			25kb	**9.74**	**3.09**	**10.09**	**2.03**	**7.55**	**2.14**	**3.28**	**1.84**
			50kb	3.72	1.18	5.51	1.11	3.86	1.09	2.32	1.3
			75kb	4.21	1.34	5.25	1.06	3.65	1.04	1.73	0.97

For each parameter and each dataset, we compared the value obtained in S-flanking regions of 25kb with the median values obtained in 100 control regions. The values of Ho, π, MAF have been multiplied by 1000. The values of the proportion of polymorphic sites have been multiplied by 100. Values in S-flanking regions greater than 95% of the distribution in control regions were represented in bold.

Table 6: Linear model variation of Ho, MAF and π of all and 0fold degenerate sites with distance to the S-locus.

	Ho		***π***		***MAF***	
	P Value	**Linear effect (by kb)**	**P Value**	**Linear effect (by kb)**	**P Value**	**Linear effect (by kb)**
All sites	<2e-16	-7.06e-5	<2e-16	-8.65e-5	<2e-16	-5.78e-5
0fold sites	4.75e-11	-4.13e-5	2.75e-15	-4.96e-5	5.35e-13	-3.44e-5

Population of origin was included as a random effect.

Table 7: Variation of Ho, Pi, MAF and proportion of polymorphic sites in the 0-fold degenerate sites in S-flanking regions of 25kb in each sample set.

				Ho		Pi		MAF		Proportion of polymorphic sites	
Population	Species	Number of positions considered	S-flanking region	Value	S-locus region /Median control	Value	S-locus region /Median control	Value	S-locus region /Median control	Value	S-locus region /Median control
Japan	*A. halleri*	20739	-75kb	1.74	2.85	2.2	2.43	1.45	2.35	12.95	1.78
			-50kb	0.9	1.47	1.25	1.39	0.66	1.07	13.99	1.93
			-25kb	**3.38**	**5.51**	4.17	4.62	2.74	4.46	**23.92**	**3.29**
			25kb	**4.8**	**7.84**	**5.95**	**6.59**	4.17	6.78	**26.6**	**3.66**
			50kb	1.22	1.99	1.13	1.25	0.8	1.3	6.72	0.92
			75kb	2.39	3.89	3.17	3.51	2.37	3.85	14.63	2.01
Nivelle	*A. halleri*	23856	-75kb	2.58	1.12	3.03	1.09	2.16	1.11	11.95	1.29
			-50kb	6.01	2.6	6.47	2.33	4.44	2.28	20.22	2.18
			-25kb	**7.26**	**3.14**	**8.02**	**2.89**	5.59	2.87	**30.88**	**3.33**
			25kb	**7.02**	**3.04**	**7.45**	**2.68**	5.39	2.77	**24.33**	**2.62**
			50kb	3.04	1.32	3.3	1.19	2.31	1.18	10.74	1.16
			75kb	5.29	2.29	5.51	1.98	3.73	1.91	17.62	1.9
Mortagne	*A. halleri*	23500	-75kb	2.4	0.97	2.9	1.07	2.1	1.13	9.65	0.94
			-50kb	6.7	2.71	**7.16**	**2.64**	**5.49**	**2.96**	21.35	2.07
			-25kb	**7.74**	**3.13**	**8.57**	**3.16**	**5.95**	**3.21**	**31.32**	**3.04**
			25kb	**7.28**	**2.94**	**6.89**	**2.54**	4.85	2.61	**25.67**	**2.49**
			50kb	2.91	1.17	3.31	1.22	2.27	1.22	13.39	1.3
			75kb	3.29	1.33	3.23	1.19	2.06	1.11	18.77	1.82
Plech	*A. lyrata*	18667	-75kb	3.07	0.63	3.28	0.57	2.13	0.53	14.86	0.65
			-50kb	7.79	1.59	**9.56**	**1.65**	**6.59**	**1.64**	**36.24**	**1.57**
			-25kb	**8.94**	**1.83**	**10.69**	**1.84**	**7.23**	**1.8**	**44.16**	**1.92**
			25kb	7.6	1.55	8.84	1.52	6.13	1.53	34.72	1.51
			50kb	6.98	1.43	7.1	1.22	4.8	1.2	29.29	1.27
			75kb	6.04	1.23	6.1	1.05	4.29	1.07	23.58	1.02
Spiterstulen	*A. lyrata*	16939	-75kb	2.35	0.18	2.45	0.52	1.62	0.46	10.21	0.61
			-50kb	7.65	1.63	6.97	1.48	4.58	1.31	**30.5**	**1.83**
			-25kb	6.61	2.45	6.82	1.45	4.93	1.41	23.14	1.39
			25kb	**9.36**	**3.39**	**8.87**	**1.88**	**6.28**	**1.79**	**34.74**	**2.08**
			50kb	**10.8**	**1.31**	**8.6**	**1.82**	**6.57**	**1.87**	**28.71**	**1.72**
			75kb	4.32	1.66	4.18	0.89	2.84	0.81	16.97	1.02
North America	*A. lyrata*	13239	-75kb	0.49	0.18	0.66	0.17	0.38	0.14	3.61	0.26
			-50kb	4.34	1.63	5.23	1.37	3.86	1.46	18.97	1.38
			-25kb	6.5	2.45	**7.87**	**2.06**	5.29	1.99	**29.6**	**2.16**
			25kb	**9.01**	**3.39**	**8.49**	**2.22**	6.09	2.29	**28.89**	**2.11**
			50kb	3.49	1.31	4.74	1.24	3.29	1.24	19.37	1.41
			75kb	4.41	1.66	4.98	1.3	3.91	1.47	12.63	0.92

For each parameter and each dataset, we compared the value obtained in S-flanking regions of 25kb with the median values obtained in 100 control regions. The values of Ho, π, MAF and the proportion of polymorphic sites have been multiplied by 1000. Values in S-flanking regions greater than 95% of the distribution in control regions were represented in bold.

Table 8: Identification of genes in S-flanking regions.

Gene name	S flanking region	Gene ID NCBI	Symbol NCBI	Gene summary NCBI	Genomic coordinates Tha
AT4G21160	+75kb	827864	ZAC	Calcium-dependent ARF-type GTPase activating protein family	NC_003075.7: 11284076-11286767
AT4G21170	+75kb	827865	AT4G21170	Tetratricopeptide repeat (TPR)-like superfamily protein	NC_003075.7: 11286756-11288513
AT4G21180	+75kb	827866	ATERDJ2B	DnaJ / Sec63 Brl domains-containing protein	NC_003075.7: 11288892-11292389
AT4G21190	+75kb	827867	emb1417	Pentatricopeptide repeat (PPR) superfamily protein	NC_003075.7: 11292192-11294014
AT4G21192	+75kb	5008151	AT4G21192	Cytochrome c oxidase biogenesis protein Cmc1-like protein	NC_003075.7: 11294134-11295543
AT4G21200	+75kb	827868	GA2OX8	gibberellin 2-oxidase 8	NC_003075.7: 11302682-11306770
AT4G21210	+75kb	827869	RP1	PPDK regulatory protein	NC_003075.7: 11306945-11308925
AT4G21215	+50/75kb	827870	AT4G21215	transmembrane protein	NC_003075.7: 11310348-11313850
AT4G21220	+50Kb	827871	LpxD2	Trimeric LpxA-like enzymes superfamily protein	NC_003075.7: 11316828-11319137
AT4G21230	+50Kb	827872	CRK27	cysteine-rich RLK (RECEPTOR-like protein kinase) 27	NC_003075.7: 11319139-11321679
AT4G21240	+50Kb	827873	AT4G21240	F-box and associated interaction domains-containing protein	NC_003075.7: 11322411-11323664
AT4G21250	+50Kb	827874	AT4G21250	Sulfite exporter TauE/SafE family protein	NC_003075.7: 11324987-11326958
AT4G21270	+50Kb	827876	ATK1	kinesin 1	NC_003075.7: 11329319-11334168
AT4G21280	**+25Kb/+50Kb**	**827877**	**PSBQA**	**photosystem II subunit QA**	**NC_003075.7: 11334352-11335815**
AT4G21300	**+25Kb**	**827878**	**AT4G21300**	**Tetratricopeptide repeat (TPR)-like superfamily protein**	**NC_003075.7: 11336318-11339052**
AT4G21310	**+25Kb**	**827879**	**AT4G21310**	**transmembrane protein, putative (DUF1218)**	**NC_003075.7: 11339048-11340085**
AT4G21323	**+25Kb**	**827881**	**AT4G21323**	**Subtilase family protein**	**NC_003075.7: 11341963-11345718**
AT4G21330	**+25Kb**	**827883**	**DYT1**	**basic helix-loop-helix (bHLH) DNA-binding superfamily protein**	**NC_003075.7: 11349922-11350694**
AT4G21340	**+25Kb**	**827884**	**B70**	**basic helix-loop-helix (bHLH) DNA-binding superfamily protein**	**NC_003075.7: 11352958-11354824**
AT4G21350	**+25Kb**	**827885**	**PUB8**	**plant U-box 8**	**NC_003075.7: 11356143-11357267**
AT4G21380	**-25Kb**	**827890**	**RK3**	**receptor kinase 3**	**NC_003075.7: 11388827-11393226**
AT4G21390	**-25Kb**	**827891**	**B120**	**S-locus lectin protein kinase family protein**	**NC_003075.7: 11394295-11397726**
AT4G21400	**-25Kb**	**827892**	**CRK28**	**cysteine-rich RLK (RECEPTOR-like protein kinase) 28**	**NC_003075.7: 11398857-11402180**
AT4G21410	**-25Kb/-50Kb**	**827893**	**CRK29**	**cysteine-rich RLK (RECEPTOR-like protein kinase) 29**	**NC_003075.7: 11401711-11405235**
AT4G21430	-50Kb	827895	B160	protein B160	NC_003075.7: 11407804-11412279
AT4G21440	-50Kb	826916	MYB102	MYB-like 102	NC_003075.7: 11418199-11419769
AT4G21445	-50Kb	825893	AT4G21445	receptor-interacting protein	NC_003075.7: 11424666-11425836
AT4G21450	-75Kb	826300	AT4G21450	PapD-like superfamily protein	NC_003075.7: 11425933-11428435
AT4G21470	-75Kb	828232	FMN/FHY	riboflavin kinase/FMN hydrolase	NC_003075.7: 11431104-11433322
AT4G21480	-75Kb	828233	STP12	sugar transporter protein 12	NC_003075.7: 11433303-11435284
AT4G21490	-75Kb	828234	NDB3	NAD(P)H dehydrogenase B3	NC_003075.7: 11436326-11439506
AT4G21500	-75Kb	828235	AT4G21500	transmembrane protein	NC_003075.7: 11440878-11441903
AT4G21520	-75Kb	828237	AT4G21520	Transducin/WD40 repeat-like superfamily protein	NC_003075.7: 11447429-11450571

Genes in the first 25kb around S-locus are represented in bold. The genomic coordinates on chromosome 4 of A. thaliana (orthologous to the S-locus region on chromosome 7 in A. lyrata) in NCBI are in the last column.

Table 9: Genomic location and size of the control regions defined in the A. halleri genome found in the A. lyrata genome by YASS.

Control region studied in *A. lyrata* genome				Corresponding region in *A. halleri* genome			
Chromosome	**First position**	**Last position**	**Size (b)**	**Chromosome**	**First position**	**Last position**	**Size (b)**
1	83447	120180	36733	scaffold200	125357	150357	25000
1	1069475	1082780	13305	scaffold71	293597	298044	4447
1	1418754	1430112	11358	scaffold242	59705	71593	11888
1	4448641	4467582	18941	scaffold42	532332	534566	2234
1	5098531	5118542	20011	scaffold6	832830	834607	1777
1	8784710	8818666	33956	scaffold65	758674	760062	1388
1	9779214	9802443	23229	scaffold81	121244	122596	1352
1	11635934	11651879	15945	scaffold171	218981	237161	18180
1	12266331	12285752	19421	scaffold314	91953	112702	20749
1	15688656	15731735	43079	scaffold281	45642	51281	5639
1	23286221	23293991	7770	scaffold229	67944	81559	13615
1	32869077	32875910	6833	scaffold120	104965	112219	7254
2	3244155	3269796	25641	scaffold166	25681	32122	6441
2	3555776	3569114	13338	scaffold553	11337	14384	3047
2	12848081	12862599	14518	scaffold224	28462	33105	4643
2	14262838	14285032	22194	scaffold3	355622	366086	10464
2	16035207	16049270	14063	scaffold23	339997	352652	12655
2	17654931	17671824	16893	scaffold29	88776	102361	13585

3	939725	967894	28169	scaffold54	221637	225077	3440
3	1082652	1101060	18408	scaffold84	283024	299392	16368
3	2834327	2843749	9422	scaffold47	50432	56219	5787
3	3261529	3277110	15581	scaffold155	166053	178096	12043
3	3614206	3638620	24414	scaffold138	211414	217837	6423
3	4048724	4060778	12054	scaffold46	114384	120992	6608
3	4340346	4364458	24112	scaffold86	201842	211170	9328
3	7536232	7560491	24259	scaffold10	375896	381674	5778
3	7830576	7839059	8483	scaffold60	31634	34804	3170
3	8899593	8920135	20542	scaffold32	553723	568781	15058
3	9289036	9316360	27324	scaffold505	35343	45505	10162
3	9589443	9614893	25450	scaffold136	65513	66259	746
3	11023648	11050966	27318	scaffold358	32810	39527	6717
3	12180930	12186261	5331	scaffold177	211995	214357	2362
3	21611295	21635129	23834	scaffold218	111991	130703	18712
3	22174587	22186666	12079	scaffold225	117783	140486	22703
4	1670520	1677096	6576	scaffold330	70259	77184	6925
4	4120140	4152041	31901	scaffold439	83581	89642	6061
4	4489852	4494894	5042	scaffold361	190619	199250	8631
4	11117870	11132007	14137	scaffold44	29228	45781	16553
4	15130559	15152366	21807	scaffold34	816549	838865	22316
4	16657706	16664114	6408	scaffold330	70259	77184	6925

4	16834544	16844691	10147	scaffold106	226324	239287	12963
4	19035410	19056081	20671	scaffold40	186691	187069	378
4	20412360	20431225	18865	scaffold18	604881	605335	454
4	21525149	21561242	36093	scaffold201	124393	127576	3183
4	22523499	22539888	16389	scaffold129	116213	121581	5368
4	22852372	22871075	18703	scaffold51	393946	397999	4053
5	951841	975828	23987	scaffold48	357354	362171	4817
5	1653956	1664424	10468	scaffold273	72568	81816	9248
5	4003159	4032530	29371	scaffold173	168326	176316	7990
5	12253381	12265905	12524	scaffold150	211653	214281	2628
5	15029455	15051052	21597	scaffold52	332919	347183	14264
5	15605860	15630386	24526	scaffold37	232502	234188	1686
5	16514721	16532625	17904	scaffold137	77095	80599	3504
5	17816355	17837342	20987	scaffold317	16150	30733	14583
5	19398229	19412882	14653	scaffold1	714746	724967	10221
6	437309	469105	31796	scaffold33	17665	21431	3766
6	1147402	1165983	18581	scaffold66	304592	315017	10425
6	2224735	2252345	27610	scaffold64	293559	303151	9592
6	4138280	4163654	25374	scaffold76	328971	340629	11658
6	4249736	4277114	27378	scaffold174	189288	193989	4701
6	6851532	6867611	16079	scaffold121	99101	103370	4269
6	7343155	7365804	22649	scaffold67	295039	320039	25000

6	8241007	8260480	19473	scaffold9	208753	212710	3957
6	10307793	10331203	23410	scaffold612	24347	32762	8415
6	10346324	10381808	35484	scaffold4	104191	107927	3736
6	10797019	10810414	13395	scaffold239	63239	75880	12641
6	21259002	21291721	32719	scaffold444	10560	16306	5746
6	23226457	23244394	17937	scaffold181	161427	180971	19544
6	24844808	24897223	52415	scaffold21	443229	445316	2087
7	433865	457338	23473	scaffold232	119994	144994	25000
7	727463	751667	24204	scaffold116	198360	200426	2066
7	1531420	1570539	39119	scaffold145	199965	215577	15612
7	3634264	3651699	17435	scaffold26	418363	427068	8705
7	4576624	4588042	11418	scaffold93	271013	285571	14558
7	6037403	6066170	28767	scaffold49	432671	453306	20635
7	6691332	6704495	13163	scaffold451	227080	242461	15381
7	7622615	7639343	16728	scaffold28	423509	426510	3001
7	9110002	9121284	11282	scaffold128	21643	30702	9059
7	9568346	9607514	39168	scaffold162	14145	17938	3793
7	9705209	9725734	20525	scaffold316	12224	21232	9008
7	11175491	11196741	21250	scaffold27	415774	440343	24569
7	12975344	12992621	17277	scaffold170	78142	96957	18815
7	19598096	19603720	5624	scaffold599	19858	25961	6103
7	21879403	21897701	18298	scaffold142	5686	17876	12190

7	23035137	23062551	27414	scaffold573	37459	47591	10132
7	23331768	23363142	31374	scaffold235	26308	30432	4124
7	23594987	23610811	15824	scaffold179	168349	177543	9194
7	24099129	24102364	3235	scaffold323	84397	87702	3305
8	639549	668410	28861	scaffold79	484435	509435	25000
8	1193675	1218564	24889	scaffold207	22822	46488	23666
8	1401121	1426726	25605	scaffold194	22610	27336	4726
8	2782403	2805264	22861	scaffold485	56582	58471	1889
8	13458265	13481329	23064	scaffold227	209987	227233	17246
8	13584566	13621779	37213	scaffold653	62306	65454	3148
8	16926540	16961242	34702	scaffold152	206858	209519	2661
8	17573303	17600436	27133	scaffold250	128700	134159	5459
8	19413760	19431480	17720	scaffold11	386718	411718	25000
8	21436030	21456224	20194	scaffold8	358549	363000	4451
8	21802390	21825492	23102	scaffold119	179807	191294	11487
8	22233394	22257401	24007	scaffold36	635729	646885	11156
8	22307313	22313905	6592	scaffold330	70259	77184	6925
scaffold_41	2689	11067	8378	scaffold518	71940	77124	5184

Figures

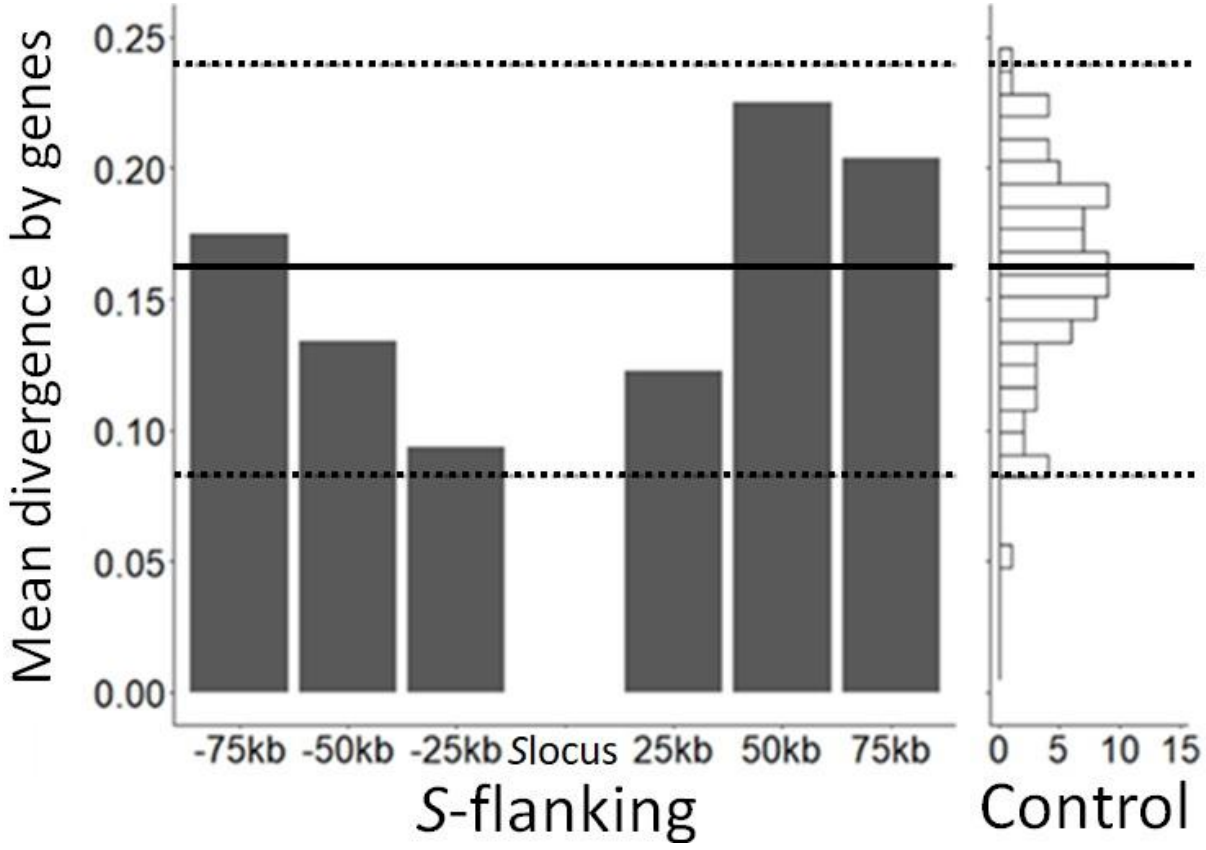


Figure S1 : Mean A. thaliana - A. lyrata divergence in genes distributed around the S-locus and across the control regions from throughout the genome. *Bars represent the mean value of divergence (proportion of divergent sites in genes between the two species) obtained in each region of 25kb around the S-locus. The distribution of the mean divergence in genes in the 100 control regions is represented by the histogram on the right. The median value of the distribution in control regions is represented by the black line and the 97.5% and 2.5% interval by the dashed lines.*

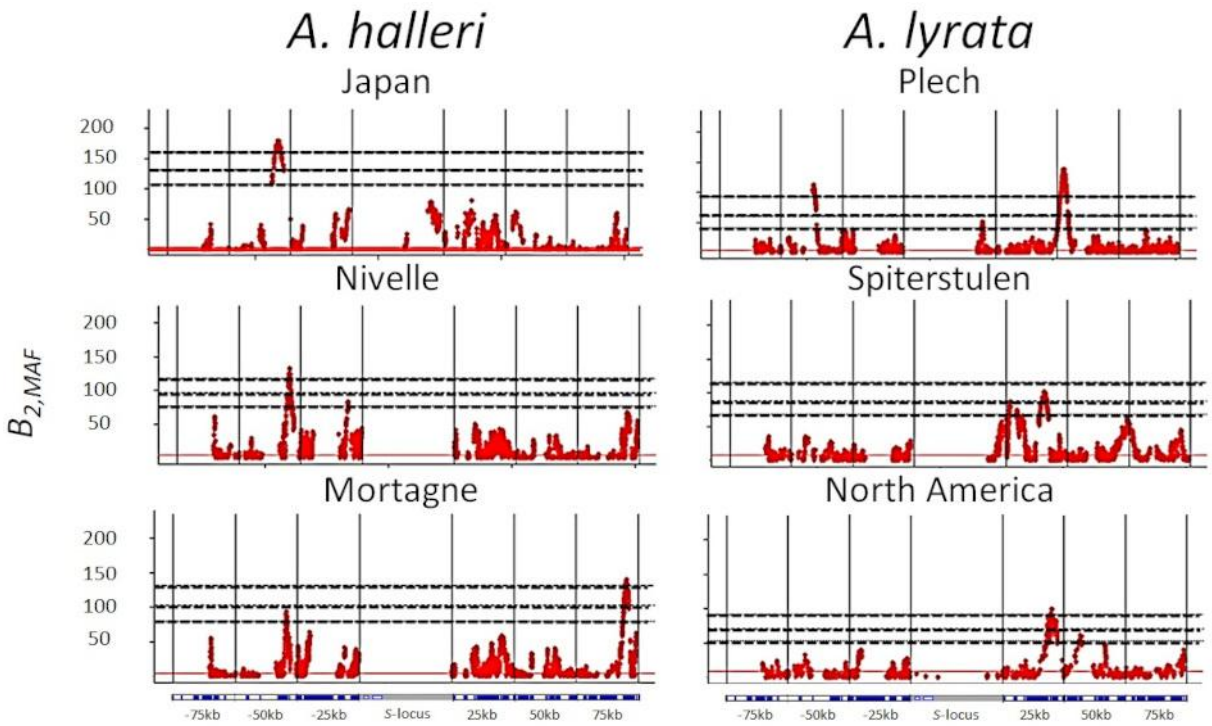


Figure S2 : Signal of balancing selection in the S-flanking regions of the A. lyrata genome. *Manhattan plot displaying $B_{2,MAF}$ scores across the genomic region on chromosome 7 surrounding the S-locus. The S-flanking regions of 25kb are delimited by vertical black lines and the S-locus is represented by a grey box.The horizontal dotted, dashed and medium dashed lines represent cutoff scores for the top 5%, 2.5% and 1% of SNPs across the genome, respectively. The horizontal red solid lines represent the median scores.*

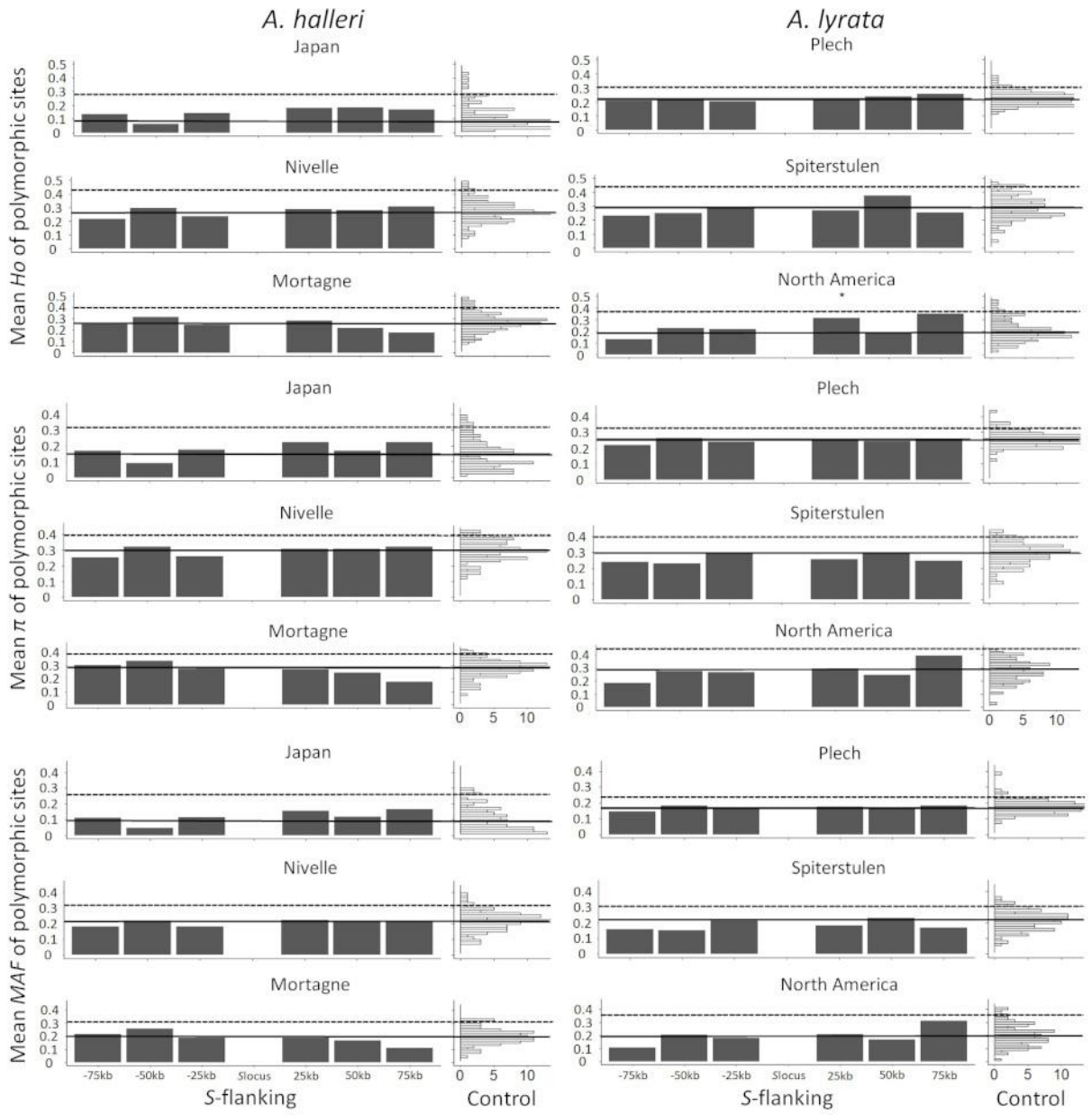


Figure S3 : Mean Ho, π and MAF in polymorphic sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean values obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of means in the 100 control regions are 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.*

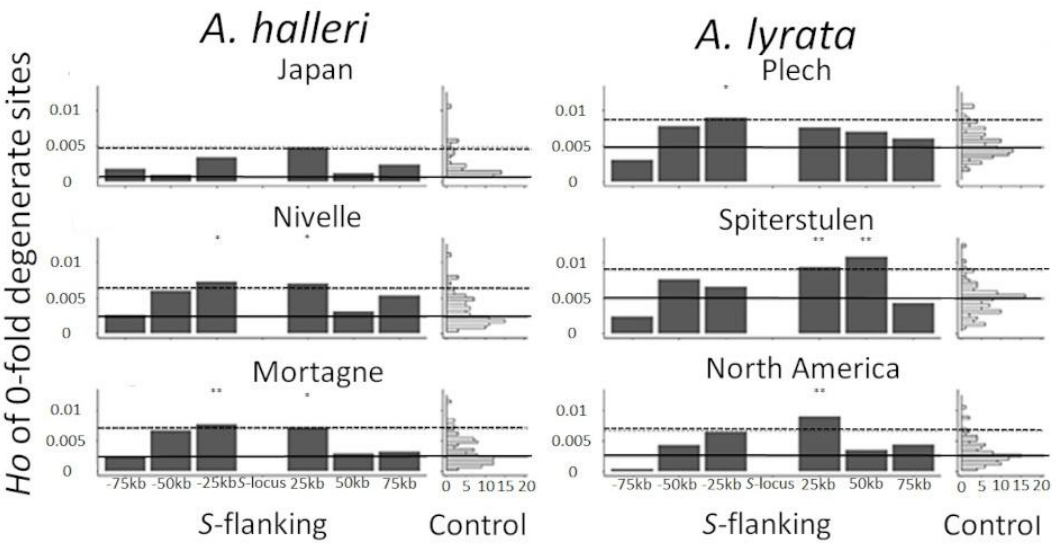


Figure S4: Mean Ho at 0 fold degenerate sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of Ho obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of Ho mean in the 100 control regions are 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 97,5% of control regions, * = observed value above the 95% of control regions.*

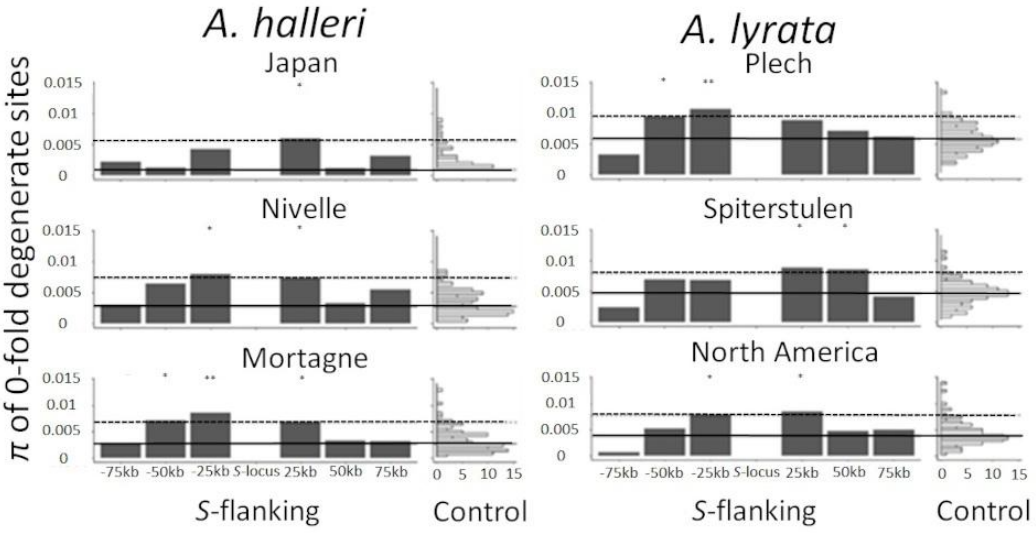


Figure S5: Mean π at 0 fold degenerate sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean value of π obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of π mean in the 100 control regions are represented 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.*

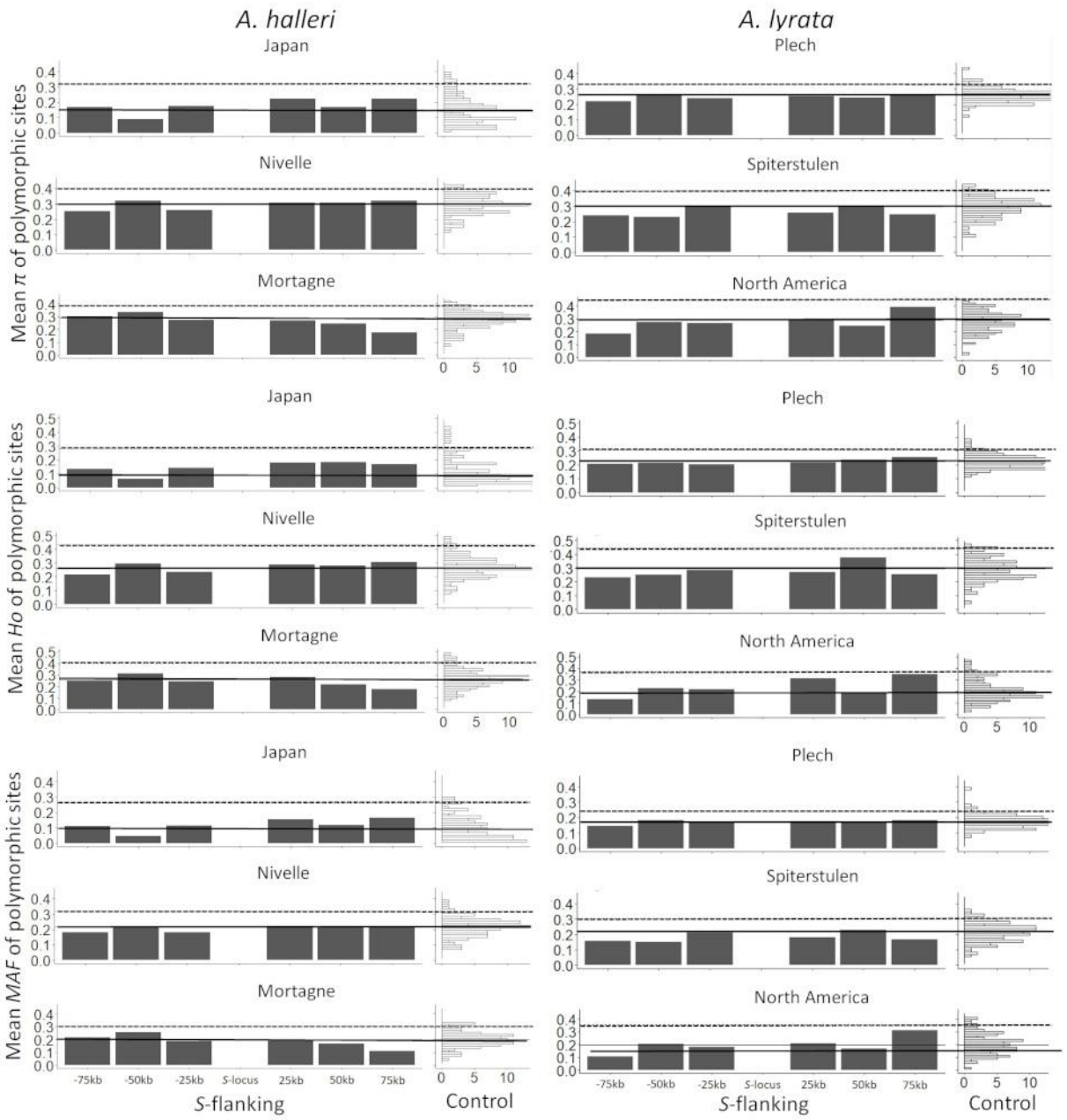


Figure S6 : Mean Ho, π and MAF in polymorphic 0 degenerate sites around the S-locus and across the control regions from throughout the genome. *Each barplot represents the mean values obtained in non-overlapping regions of 25kb around the S-locus. The distributions (count) of means in the 100 control 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.*

Chapter II

Scientific questions:

The second axis of the thesis aims to answer the following questions:

- Is molecular polymorphism in the *S*-locus flanking regions impacted by dominance of the *S*-alleles? If so, how?
- Can these differences in sheltered load be revealed at the phenotypic level ?
- Do dominant and recessive S-alleles differ in the expected proportion of deleterious mutations that are fixed *vs.* segregating within allelic classes ?

To answer these questions, we obtained phased haplotypes of polymorphisms over the 75kb flanking regions of the *S*-locus associated with multiple *S*-alleles from two *A. halleri* and one *A. lyrata* sample set. We then evaluated whether and how the distribution of linked variants varies among S-alleles of different levels of dominance. We complemented this genomic analysis with an analysis of 14 life history traits of individuals from one *A. halleri* population. Finally, we used stochastic models to refine the theoretical predictions on the link between dominance and the sheltered load.

Contribution:

The *A. lyrata* individuals used for haplotype reconstruction were derived from seeds obtained from B. Mable, and were grown in the laboratory greenhouse. The *A. halleri* individuals come from two sampling trips in natural populations carried out by Chloé Ponitzki, Eleonore Durand, Vincent Castric and myself.

Random crosses of individuals from these three populations were performed by myself in May 2019. The plants were maintained by members of the laboratory's experimental platform. The extraction of DNA from the parents as well as the PCRs against the *SRK01* allele were performed in the laboratory by Christelle Lepers-Blassiau and myself.

The design of the probes for the sequence capture protocol was done by Nicolas Burghgraeve in interaction with the company Mybaits. The molecular biology experiments (libraries construction and sequence capture) were carried out by Christelle Lepers-Blassiau and myself. Sequencing was performed by the genomics platforms LIGAN-MP and GenoScreen (Lille).

The pipeline for read alignment and variant calling was developed by Mathieu Genete and Nicolas Burghgraeve before the thesis project. All the python codes allowing the reconstruction of haplotypes, the identification of fixed variants per allele and the calculation of the total number of mutations of each haplotype were developed by myself. The same goes for the codes allowing the calculation of the total number of mutations of each individual. The F_{ST} analysis of the reconstructed haplotypes was performed by myself. Finally, all the statistical analyses in R were done by myself.

Concerning the analysis of life history traits, the controlled crosses allowing the formation of homozygotes for the three alleles were performed by myself in April-May 2020. The seeds

from these crosses were sown by Chloe Ponitzki and myself in September 2020. I measured the biomass related traits mostly by photographic analysis in ImageJ. The photographs were taken by Chloe Ponitzki and myself. The measurements of traits related to reproduction mobilised the entire staff of the experimental platform, a L3 intern, Justine Bertin, and myself, during the period from March 2021 to June 2022. The extraction of parental DNA as well as the PCRs against the expected *S*-alleles in each line allowing genotyping at the *S*-locus of the descendants were carried out in the laboratory by Christelle Lepers-Blassiau and myself. Finally, all the statistical analyses in R were done by myself.

Concerning the stochastic modelling, it was done by myself, according to a previously published model, and previously modified by Sylvain Billiard and myself.

This chapter is written in English, in the form of a draft article.

The structure of the linked genetic load differs between dominant and recessive self-incompatibility alleles in *Arabidopsis halleri* and *A. lyrata*

Le Veve Audrey[1], Genete Mathieu[1], Lepers-Blassiau Christelle[1], Ponitzki Chloé[1], Durand Eleonore[1], Castric Vincent[1], Vekemans Xavier[1]

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

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

Abstract

Evolution of the mating system depends on the inbreeding depression caused by the expression of deleterious mutations in individuals. Further, the accumulation of deleterious mutations can vary across the genome, especially for genes closely linked to loci under balancing selection. Sporophytic self-incompatibility (SSI) is a common genetic mechanism in Angiosperms that enables hermaphrodite plants to avoid selfing and promotes outcrossing. The SSI recognition phenotype is determined by the *S*-locus and entails dominance relationships among alleles. Since natural selection acts asymmetrically on the *S*-alleles according to their level along a dominance hierarchy, it has been suggested that the accumulation of deleterious mutations in genes linked to the *S*-locus depends on the dominance level of the *S*-allele to which they are linked. In this study, we first compared survival and 13 fitness-related traits of homozygote vs heterozygote progenies for three *S*-alleles distributed in three dominance classes. Our analysis revealed a significant sheltered load associated with a relatively recessive *S*-allele (Ah03, class II), but we did not confirm the expected correlation between *S*-allele dominance and the phenotypic impact of their homozygosity. Second, we used a parents-offspring trio approach to phase polymorphisms in the chromosomal regions linked to the *S*-flanking regions for distinct copies of *S*-alleles in different populations of *Arabidopsis halleri* and *Arabidopsis lyrata*. We found that the size of the *S*-flanking regions structured by *S*-allele varies across populations, probably because of variation in demography and local recombination rates. Dominant *S*-alleles showed a higher number of fixed deleterious mutations, but we found no effect of *S*-allele dominance on the total number of putatively deleterious mutations to which they are associated. We extended a previously developed model to demonstrate that the smaller effective population size of the most dominant *S*-alleles the increased fixation of deleterious mutations can compensate for their lower standing variation in the *S*-flanking regions. These observations indicate that the impact of dominance on the genetic sheltered load linked to the *S*-locus is more subtle than previously envisioned, with qualitative rather than quantitative differences of the genetic load along the dominance hierarchy of *S*-alleles.

Keywords: balancing selection, sheltered load, genetic dominance, S-locus, inbreeding depression.

Introduction

Deleterious variation is recurrently introduced in natural populations by new mutations and tends to be eliminated by natural selection. The dynamics of elimination and the resulting mutation-selection balance is modulated by the intensity of genetic drift experienced by the chromosomal segment carrying the mutation across the demographic history of the population (Lynch et al., 2016). Because most deleterious mutations are believed to be recessive, factors affecting homozygosity across the genome are important determinants of their accumulation process. For instance, inbreeding will increase homozygosity genome-wide and therefore allow the expression of deleterious mutations that are usually masked in outbred individuals, leading to the phenomenon of inbreeding depression (Charlesworth and Charlesworth, 1999). Conversely, some forms of long-term balancing selection are rather expected to locally decrease homozygosity (Charlesworth, 2006), allowing the masking of recessive deleterious mutations (Maruyama and Nei, 1981). Linkage to such loci can negatively interfere with purifying selection, diminishing its efficacy and facilitating the local accumulation of a specific genetic load, referred to as the "sheltered load" (Uyenoyama, 1997, 2005; Hartfield and Otto, 2011). This phenomenon has been considered an "evolutionary cost" of balancing selection (van Oosterhout et al., 2009; Lenz et al., 2014), and accordingly a large number of diseases in humans are associated with variants at genes within the Major Histocompatibility Complex (*MHC*), a classical case of balanced polymorphism (Garrigan & Hedrick, 2003). However, at this stage, the generality of this phenomenon remains unclear, and its intensity remains unknown for other balanced polymorphisms.

Homomorphic self-incompatibility (SI) is a genetic mechanism allowing recognition and rejection of self-pollen by hermaphrodite individuals, thereby preventing inbreeding and promoting outcrossing in hermaphroditic plant species (Nettancourt, 2001). Homomorphic SI is one of the most prominent examples of long-term balancing selection (Castric and Vekemans, 2004) and exists in two different versions. In gametophytic SI (GSI, as found e.g. in Solanaceae) pollen specificity is determined by its own haploid genome, while in sporophytic SI (SSI, as found e.g. in Brassicaceae) the pollen recognition phenotype is determined by the male diploid parent. In the Brassicaceae, SI is controlled by a single non-recombining chromosomal region, the *S*-locus (Kusaba et al., 2001 ; Schopfer et al., 1999), composed of two linked genes. *SCR* (*S*-locus Cysteine-Rich) and *SRK* (*S*-locus Receptor Kinase), encode the male and female specificity determinants, respectively. Pollination between partners expressing the same haplotype at the *S*-locus leads to pollen rejection by the pistils.

Deleterious mutations are expected to accumulate in close linkage to the *S*-locus because of the combined effect of (1) enforced heterozygosity that tends to mask recessive deleterious mutations, thus reducing the efficacy of purifying selection and (2) the indirect effect of negative frequency-dependent selection favouring linked mutations introduced on chromosomes carrying rare *S*-alleles regardless of their deleterious effect (Uyenoyama, 1997). Sequencing of multiple *S*-haplotypes in *A. halleri* and *A. lyrata* revealed that the *S*-locus chromosomal segment contains no protein-coding genes other than those controlling the SI machinery itself, but Le Veve et al. (Chapter 1), demonstrated that the two 25-kb regions immediately flanking the *S*-locus on either side indeed present an excess of polymorphism in the *A. lyrata* and *A. halleri* genomes as compared to the genomic background, with potentially deleterious effects. These two regions are partly linked to the *S*-locus and comprise a total of eleven protein coding genes. However, because the sequencing data in Le Veve et al. (Chapter 1) were not phased, it was not possible to determine how the deleterious mutations in this genomic interval were distributed among *S*-alleles in the populations examined.

SSI is characterised by the existence of dominance interactions between *S*-alleles, whereby heterozygous individuals express generally only one of their two *S*-alleles at the phenotypic level, and resulting in a dominance hierarchy among *S*-alleles (Bateman, 1952; Llaurens et al., 2008 for *A.*

halleri). Llaurens et al. (2009a) and Goubet et al. (2012) showed that the dynamics of accumulation of deleterious variation is expected to differ in linkage with dominant versus recessive *S*-alleles. Specifically, recessive *S*-alleles can form homozygous combinations in natural populations more often than dominant *S*-alleles (Schierup et al., 1997), such that recombination may occur occasionally within the S-locus of recessive alleles, allowing them to purge linked recessive deleterious mutations more readily. In addition, because recessive alleles reach higher allele frequencies (Schierup et al., 1997; Billiard et al., 2006), the efficacy of purifying selection on linked variants within the population of allele copies is expected to be higher than for dominant allele copies. As shown by Llaurens et al. (2009a), this is expected to result in a higher fixation probability of deleterious variants linked to the class of dominant *S*-alleles than to the class of recessive *S*-alleles. Based on a series of phenotypic measurements in *A. halleri,* Llaurens et al. 2009a tested this theoretical prediction. They experimentally by-passed SI to obtain selfed progenies in which they compared homozygous *vs.* heterozygous *S*-locus genotypes at different levels of the dominance hierarchy. They revealed several decreased proxies of fitness for homozygotes for one of the most dominant *S*-alleles (Ah15, class IV) as compared to heterozygotes for this *S*-allele, while no such contrast could be detected for the most recessive *S*-allele (Ah01, class I). They concluded that the sheltered load was higher in the most dominant *S*-allele (Ah15) than in the most recessive *S*-allele (Ah01). However, this study compared only two *S*-alleles, which is clearly insufficient to conclusively establish the effect of the dominance hierarchy among *S*-alleles on the sheltered genetic load linked to *S*-alleles.

In this study, we first extended the phenotypic approach of Llaurens et al. (2009a) to evaluate the effect of the *S*-allele dominance level (i.e. the *S*-allele position in the dominance hierarchy) on the sheltered load linked to a series of additional *S*-alleles from the same local population. We then used targeted genome re-sequencing of parents-offspring trios to compare the number of putatively deleterious mutations in the phased chromosomal segments linked to dominant *vs.* recessive *S*-alleles in two *A. halleri* and three *A. lyrata* populations. Finally, we refined the theoretical models by using stochastic simulations to predict the effect of *S*-allele dominance levels on the relative proportion of deleterious mutations that are fixed *vs.* segregating within allelic classes. Overall, our results suggest a more nuanced view of the effect of dominance on the sheltered load, in which recessive *S*-alleles have more segregating but less fixed deleterious mutations, while dominant *S*-alleles eventually compensate for their increased tendency to fix deleterious mutations by the accumulation of a lower number of segregating mutations. The structure of the sheltered load thus differs among *S*-alleles from different dominance classes.

Results

The phenotypic impact of homozygosity varies among S-alleles, but is not correlated with dominance class

To evaluate the effect of *S*-allele dominance on the sheltered load, we modified the phenotypic approach of Llaurens et al. (2009a) to include two additional *S*-alleles from the same local population (Ah03, class II; and Ah04, class III; Nivelle, France). We also included the most recessive *S*-allele (Ah01, class I) in our experiment. Briefly, while Llaurens et al. 2009a used forced selfing to obtain *S*-locus homozygous genotypes, we crossed heterozygous parental individuals sharing a given *S*-allele that was masked by different dominant *S*-alleles (e.g. to obtain Ah_xAh_x homozygotes we deposited pollen from a Ah_xAh_y plant where $Ah_y>Ah_x$ on pistils of a Ah_xAh_z plant where $z\neq y$; see table S1 for more details). This enabled us to obtain full-sib families in which we compared offspring that were homozygous for each of these *S*-alleles with their full sibs that were heterozygous . We obtained 399 offspring from six such crosses.

We first tested whether homozygosity at the *S*-locus affected survival from germination to the reproductive stage. The proportion of Ah01/Ah01 and Ah04/Ah04 homozygotes surviving to the reproductive stage was consistent with Mendelian expectations in their respective families, but only two (7.4%) Ah03/Ah03 homozygotes survived when 6.75 (25%) would have been expected (p=0.02; Table 1). The overall frequency of the Ah03 *S*-allele in the offspring of these families (observed frequency of Ah03 over the 27 individuals in these families = 0.41) did not differ significantly from the 50% mendelian expectation (credible interval 0.37 to 0.63 over 10,000 permutations, p=0.11; Table 1), suggesting that the increased mortality is associated with homozygosity, rather than to lower performance of the Ah03 allele itself. Because an effect was detected on the Ah03 *S*-allele only, which belongs to the intermediate class of dominance and not on the most dominant Ah04 or the most recessive Ah01, these results on transmission ratio distortion do not support a positive relation between *S*-allele dominance and the magnitude of the sheltered load.

Table 1 : Comparison of the proportion of homozygous offspring at the S-locus having reached the reproductive stage with theoretical expectations.

***S*-allele shared by both parents**	**Level of dominance**	**Total number of seedlings reaching the reproductive stage**	**Observed proportion of homozygotes**	**Ratio of the observed / expected proportion of homozygotes (P Value)**	**Observed frequency of the shared *S*-allele in the offspring (P Value)**
Ah01	I	39	0.231 (9/39)	1.1 (0.47)	0.49 (0.46)
Ah03	II	27	0.074 (2/27)	0.29 (0.02)	0.41 (0.11)
Ah04	III	96	0.479 (46/96)	1.04 (0.39)	0.74 (0.37)

The P Values represent the proportions of the distribution equal to or less than the value observed obtained after 10,000 random resamples. The significant values are represented in bold . The effect on the number of homozygotes observed represent the ratio between the expected median value and the observed actual value.

Next, we measured fourteen vegetative and reproductive traits, treating attacks by phytopathogens, phytophages and oxidative stress as random effects if necessary (Table S2). Overall, we found that homozygosity at the *S*-locus did not generally impact the traits analysed. We found only two exceptions to this general pattern : the maximum size of flowering stems and the time to first flowering (Fig. 1), but these variations were caused by differences in a small number of families only and thus were not general (Table S3).

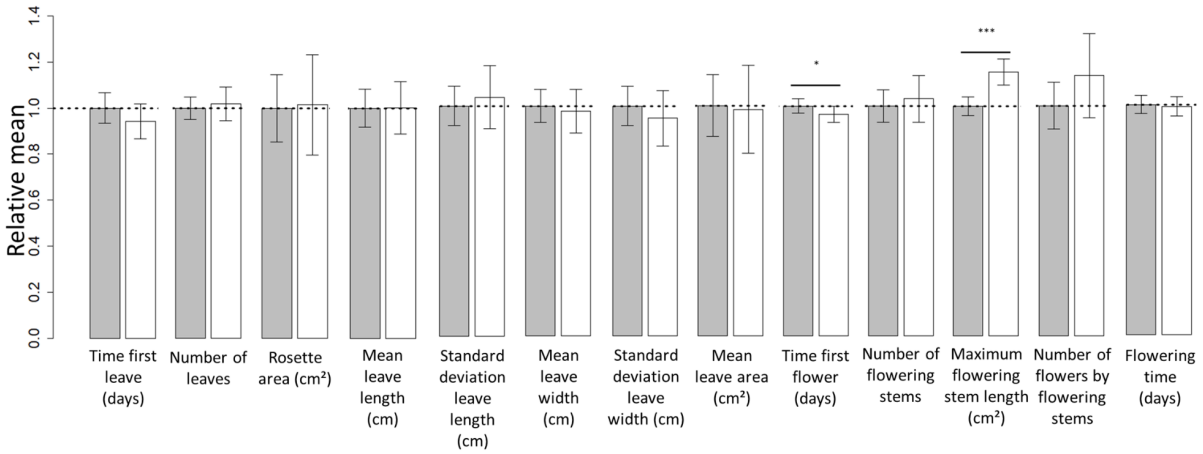


Figure 1 : Mean of the phenotypic traits in S-locus homozygotes (white bars) relative to heterozygotes (grey bars). *The distributions were compared by 10,000 random permutations. *** p-value <0.001, * p-value <0.05.*

A Generalised Linear Model (GLM, Table 2) confirmed that more dominant *S*-alleles did not have a more severe deleterious effect than recessive *S*-alleles when made homozygous. Overall, our phenotypic results did not confirm the conclusion of Llaurens et al. (2009a) that dominant *S*-alleles carried a more severe deleterious load than recessive *S*-alleles in the Nivelle population.

Table 2: Variation of phenotypic traits in homozygotes at the S-locus with dominance.

Trait (unit)	GLM model	Linear effect	P-value	Mean difference with heterozygotes
Time first leave (days)	~ Dom	-1.27	0.399	0.411
Leaves (counts)	~ Dom + (1 | oxy stress)	-0.258	0.405	-0.163
Rosette area (cm²)	~ Dom + (1 | oxy stress)	2.34	0.891	-1.918
Mean leave length (cm)	~ Dom + (1 | oxy stress)	-0.014	0.611	-0.092
St dev leave length (cm)	~ Dom	0.015	0.571	0.004
Mean leave width (cm)	~Dom + (1 | oxy stress)	-0.043	0.399	-0.099
St dev leave width (cm²)	~ Dom	0.004	0.855	-0.031
Mean leave area (cm²)	~ Dom + (1 | oxy stress)	-0.021	0.700	-0.318
Time first flower (days)	~ Dom + (1 | phytopathogen attack)	1.96	0.083	-0.602
Flowering stems (counts)	~ Dom	0.809	0.376	-0 .172
Longest flowering stem length (cm)	~ Dom	2.601	0.177	5.084
Flowers by stem (counts)	~ Dom	3.195	0.660	4.577
Flowering duration (days)	~ Dom + (1 | phytopathogen attack)	3.122	0.922	0.142

Notations: dominance at the S-locus (Dom); oxidative stress (oxy stress). Oxidative stress and phytopathogen attacks were implemented as random effects (1 | random effect).

S-alleles are associated with specific sets of linked mutations in a limited S-flanking regions size

The model of the sheltered load assumes that each class of *S*-alleles carries a specific set of linked deleterious mutations. In order to verify this prediction directly, we combined a parents-offspring trio phasing approach with sequencing of the *S*-locus flanking regions over 75kb on either side as described in Le Veve et al (Chapter I) to associate the mutations segregating in the flanking regions to their respective *S*-alleles. We analysed the *A. halleri* Nivelle population, a closely related *A. halleri* population (Mortagne) and three distant allogamous *A. lyrata* populations (named IND, PIN and TSS; Foxe et al., 2010). Overall, we were able to reconstruct 34 haplotypes linked to a total of 12 distinct *S*-alleles in Nivelle, 38 haplotypes linked to 11 distinct *S*-alleles in Mortagne and 16, 22 and 16

haplotypes associated with 6, 7 and 5 distinct *S*-alleles in populations IND, PIN and TSS, respectively (Table 3). Nine of the *S*-alleles were shared between the two *A. halleri* populations (Ah01, Ah03, Ah04, Ah05, Ah12, Ah20, Ah24, Ah25 and Ah59). In the populations of *A. lyrata*, four *S*-alleles were shared between PIN and TSS (Ah01*, Ah03*, Ah18* and Ah63*), five *S*-alleles were shared between PIN and IND (Ah01*, Ah03*, Ah46* and Ah63*), four *S*-alleles were shared between IND and TSS (Ah01*, Ah03*, Ah31* and Ah63*), and three were shared across all three (Ah01*, Ah03* and Ah63*). Note that for convenience, we used *A. halleri* notations (with the addition of a *) to refer to the trans-specifically shared *A. lyrata S*-alleles. Overall, we were able to obtain the phased sequence of 130 *S*-locus haplotypes, comprising a total of 4,854 variable sites. This enabled us to evaluate the conservation of association between the two *A. halleri* populations for the nine *S*-alleles listed above, and the comparison could be extended to *A. lyrata* for two of them (Ah01* and Ah03*).

Table 3 : Number of haplotypes phased for different S-alleles in different populations and species.

Species	Populations	Number of phased haplotypes	Number of *S*-alleles	Number of *S*-alleles with more than one copy	Mean number of gene copies per *S*-allele*
A. halleri	Nivelle	34	12	7	4.1
	Mortagne	38	11	9	4
A. lyrata	TSS	16	5	3	4.7
	IND	16	6	4	4.5
	PIN	22	7	3	5.3

**based on the S-alleles with more than one copy*

We first visualised the relationships among these extended *S*-locus haplotypes using phylogenetic trees. If the polymorphisms in the *S*-flanking regions were specific to each *S*-allele, we would expect a clustering by *S*-alleles rather than by population of origin. In the 25kb on both sides around the *S*-locus in *A. halleri,* the haplotypes linked to allele copies of Ah03, Ah05, Ah25 and Ah59 from Nivelle and Mortagne were completely clustered by allelic lineages (Fig. S1). With a few exceptions only, we also found global clustering by *S*-alleles for the haplotypes linked to allele copies of Ah04, Ah12, Ah20 and Ah54. Only the haplotypes linked to allele copies of Ah01 were not clustered into a single group and formed divergent clades for Nivelle and Mortagne, respectively. However, beyond the first 25kb regions, the haplotypes linked to allele copies of the nine *S*-alleles from Nivelle and Mortagne were not clustered by *S*-alleles any longer but by populations. We observed clearly two groups of haplotypes, the group of Nivelle and the group of Mortagne, with five exceptions, and within each population cluster some level of clustering by *S*-allele, but not strict (Fig. S2). In *A. lyrata*, the clustering of the flanking regions sequences by *S*-locus haplotype was less marked. Indeed, we had to reduce to only 5kb the flanking regions studied to observe substantial clustering by *S*-alleles for the haplotypes linked to allele copies of Ah03*, Ah18*, Ah29*, Ah31*, Ah46* and Ah63*, with a few exceptions (Fig. S3). Allele copies of Ah01* from population IND did cluster separately from Ah01*

copies from PIN and TSS populations. When considering flanking regions over 10kb, the associations with *S*-alleles become less pronounced than the clustering by populations (Fig. S4).

Following Charlesworth (2006), the polymorphisms in the *S*-locus flanking regions can be seen as segregating in a population subdivided at two different levels. A given linked mutation can be exchanged between *S*-alleles by recombination, and it can be exchanged between local populations by migration. The relative time scales of these two processes determine the distribution of the linked mutations and can be quantified by comparing the fixation index F_{ST} among local populations and among *S*-alleles. We analysed variation of the mean F_{ST} among populations and among *S*-alleles in the *A. lyrata* and *A. halleri* datasets, in non overlapping windows of 5kb around the *S*-locus (Fig. 2). Moreover, we compared these mean F_{ST} values with their distributions across one hundred 25kb control regions unlinked to the *S*-locus (see Le Veve et al,. Chapter I for more details).

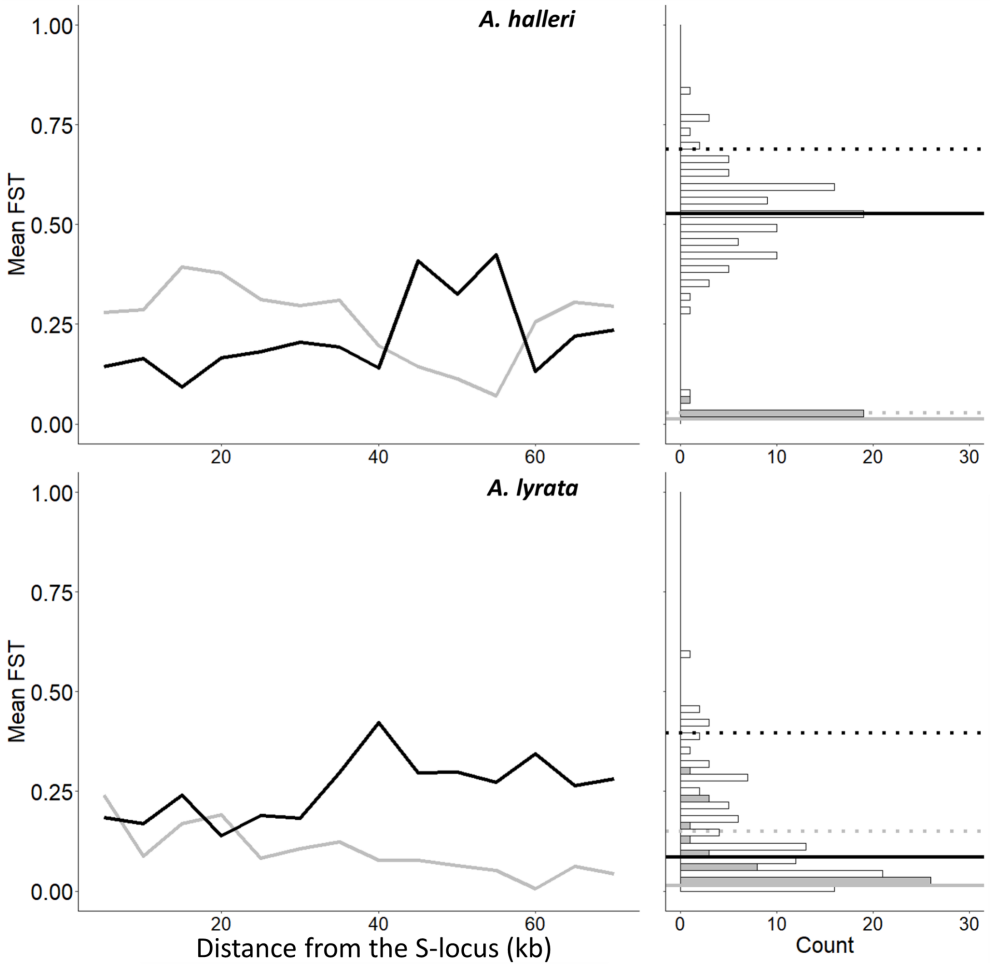


Figure 2 : Variation of the inter-allelic (grey lines) and inter-populations (black lines) F_{ST} for polymorphic sites linked to the S-locus (left) or in control regions (right) in A. halleri (top) and A. lyrata (bottom). *The distributions (count) of F_{ST} analysed by S-alleles (grey) or by populations (white) in the control regions were represented by histograms. The 95% percentile of the distributions are represented by dotted lines and the medians by solid lines for the estimations of the F_{ST} by S-alleles (grey) or by populations (black) .*

In both *A. halleri* and *A. lyrata*, F_{ST} values among *S*-alleles were high in regions close to the *S*-locus (0.28 and 0.24 respectively for the SNPs in the first 5kb) and quickly decreased to reach the background level (median values of 0.013 and 0.014, respectively; Fig. 2) as the distance from the *S*-locus increased. In parallel, the differentiation among populations followed roughly the opposite pattern, i.e. it was initially low in regions close to the *S*-locus and increased up to background level within the first few kilobases (median F_{ST}=0.52 and 0.09, respectively; Fig. 2). However, differentiation between populations started to exceed differentiation between *S*-alleles much closer to the *S*-locus in *A. lyrata* than in *A. halleri*. In addition, for *A. halleri*, the phased haplotypes became predominantly structured by *S*-alleles again after 55kb, and this pattern was associated with a concomitant decrease of the structure by populations. Overall, our results indicate that due to limited recombination, the *S*-alleles carry a specific set of polymorphic sites in the linked region. This association fades away for more distant sites, where population structure becomes predominant, as in the rest of the genome.

No evidence that dominant S-alleles accumulate more deleterious mutations in the S-flanking regions

In Llaurens et al (2009a), stochastic simulations predicted a positive correlation between the dominance class of *S*-alleles and their tendency to fix deleterious mutations. In Le Veve et al. (Chapter I), we found that the genetic sheltered load mainly accumulates in the first 25kb on either side of the *S*-locus. Thus, we investigated in these two regions the correlation between the level of dominance of the *S*-allele for each phased haplotype and either the total number of 0-fold mutations (S_{0f}) or the ratio of 0-fold to 4-fold mutations (S_{0f}/S_{4f}). We found a significant positive effect of *S*-allele dominance on S_{0f} and S_{0f}/S_{4f}, only in the *A. halleri* Nivelle population (p=0.001 and p=9.61e-4, respectively; Fig. 3A, Table 4). We note that the particular *S*-allele whose sheltered load was quantified in Llaurens et al. 2009a (Ah15) happens to be one of the *S*-alleles associated the highest number of 0-fold mutations among all *S*-alleles of the most dominant class (class IV). In contrast, no effect of the dominance class on the number of 0-fold mutations or on the ratio of 0-fold to 4-fold mutations was observed in the *A. halleri* Mortagne population (*p*=0.215 and *p*=0.362; Fig. 3B, Table 4), where the mean number of 0-fold mutations per haplotype was higher overall (Fig. 3B). For *A. lyrata*, we observed stark differences among populations in the accumulation of 0-fold mutations around the most recessive *S*-allele (Ah01*; Fig. 3C). Consequently, while there was no significant effect of *S*-allele dominance class on S_{0f} in the IND population, a positive effect was detected in the TSS population (linear effect=0.29, p value=3.7e-10, Fig. 3C), and a negative effect was detected in the PIN population (linear effect=-0.08, p value=0.008, Fig 3C). Thus, the correlation between *S*-allele dominance classes and the accumulation of putatively deleterious mutations appears to be highly dependent on the accumulation around the most recessive *S*-allele, and when considering the three *A. lyrata* populations jointly, we found again no overall effect (*p*-value=0.227, *p*-value=0.884; Table 4). Overall, we thus did not confirm that the total putative sheltered load increased with dominance class, but we noted a particularly high number of 0-fold mutations associated with allele Ah15.

Table 4: Effect of S-allele dominance on the number of 0-fold mutations (S_{0f}) or on the ratio of 0-fold to 4-fold mutations (S_{0f}/S_{4f}) accumulated in the two 25kb regions on either side of the S-locus.

Species	Populations	Statistic considered	Effect of the statistic	P Value
A. halleri	Nivelle	S_{0f}	0.074	0.001
		S_{0f}/S_{4f}	0.066	9.61e-4
A. halleri	Mortagne	S_{0f}	-0.031	0.215
		S_{0f}/S_{4f}	-0.019	0.362
A. lyrata	North America	S_{0f}	0.027	0.227
		S_{0f}/S_{4f}	-0.022	0.884

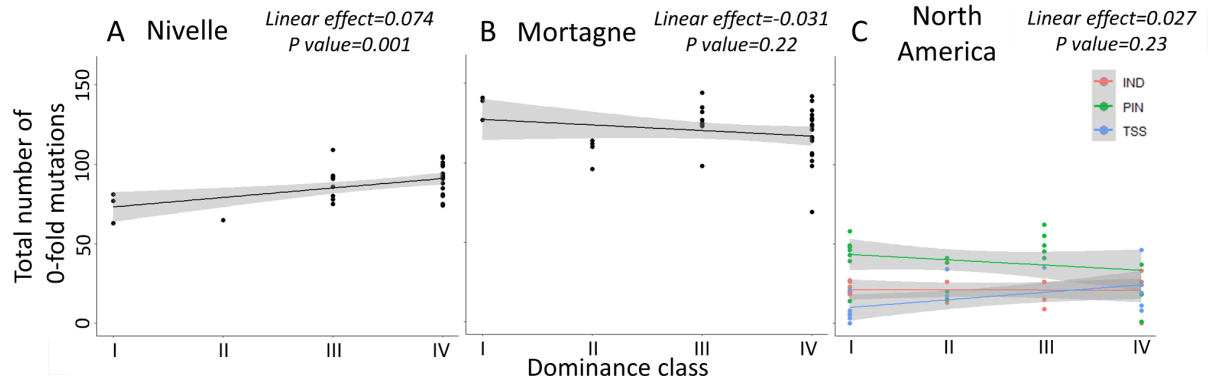


Figure 3 : Total number of 0-fold degenerate mutations (S_{0f}) in the regions of 25kb flanking S-alleles of different dominance classes in the A. halleri Nivelle (A) and Mortagne (B) populations, and in the A. lyrata populations (C). *Each point represents the value obtained for one S-haplotype. The correlations evaluated by a GLM model are represented by lines. The confidence intervals are represented in grey. The red arrow points to the copy of Ah15, corresponding to the S-allele whose sheltered load was quantified in Llaurens et al. 2009a.*

The structure of the genetic load differs between S-alleles dominance classes

To clarify the relationship between dominance and the sheltered load, we distinguished mutations contributing to the total sheltered load of a given haplotype that were fixed within allelic classes, i.e. that were shared by all haplotypes associated with a given *S*-allele. In line with the prediction of Llaurens et al. (2009a), we found within all populations a consistently positive relationship between the *S*-allele dominance class and the number of 0-fold degenerate mutations fixed within each class (Fig. 4; Table S4). These results suggest that the structure of the genetic load differs between dominant and recessive *S*-alleles: on the one hand dominant *S*-alleles tend to have more deleterious mutations that are fixed, but on the other hand recessive *S*-alleles compensate by having a larger number of segregating mutations, resulting in a similar number of deleterious mutations in total in most populations, except Nivelle in which the total number of 0-fold mutations remains higher overall in dominant *S*-alleles.

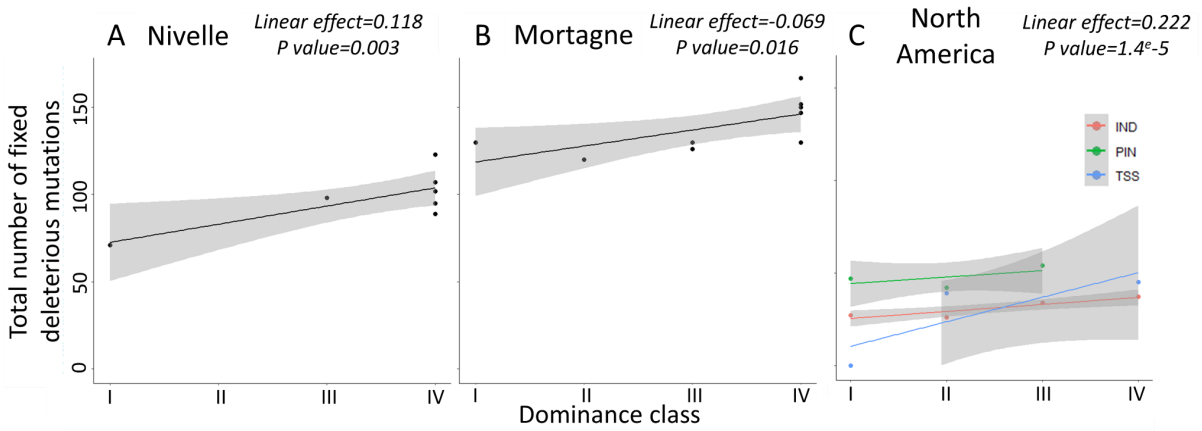


Figure 4 : Number of fixed 0-fold mutations in the 25kb regions flanking the S-locus in A. halleri Nivelle (A) and Mortagne (B) populations and in A. lyrata (C) as a function of the dominance class of the S-allele associated. *Each point represents the value obtained for one S-allele. The correlations by GLM are represented by lines. The confidence intervals are represented in grey. The A. lyrata subpopulations are represented by colours: red for IND, green for PIN, blue for TSS. For each population, we represent the p-value and the effect obtained.*

To investigate this intuition further, we modified the model proposed by Llaurens et al., (2009a) to examine the dynamics of accumulation of deleterious mutations linked to *S*-alleles in more details, focusing not only on deleterious mutations that are fixed but also on those that are segregating within allelic classes. These stochastic simulations confirmed that, at equilibrium, dominant *S*-alleles tend to accumulate a larger number of deleterious mutations that are fixed among allele copies within *S*-alleles (Fig. 5A, Fig S5). In contrast, the number of segregating mutations was higher linked to recessive than to dominant *S*-alleles. These two effects compensated each other, such that in the end the total number of deleterious mutations accumulated in linkage to each haplotype was not expected to change with dominance (Fig. 5B). These predictions are in line with our genomic analysis, and suggest a model where the dominance level of the *S*-alleles modifies the structure of the genetic sheltered load: dominant *S*-alleles accumulate more fixed deleterious mutations, but recessive *S*-alleles accumulate more segregating mutations, resulting in an equivalent total load.

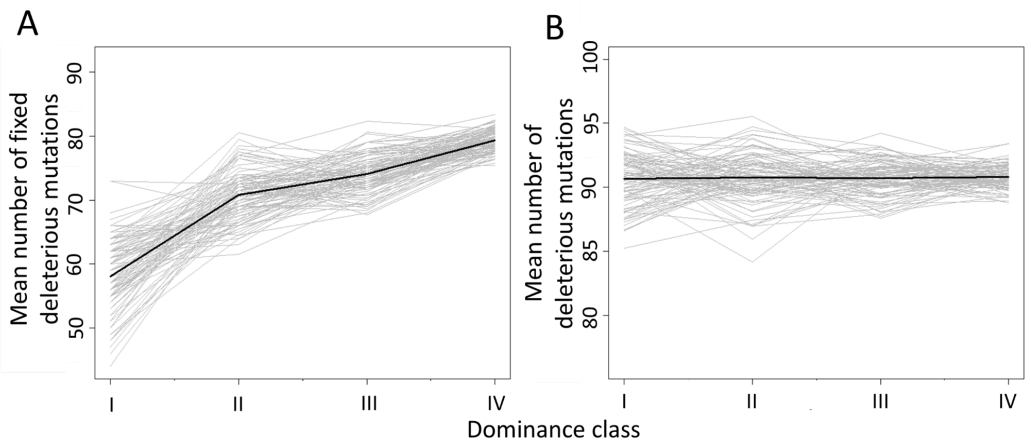


Figure 5: Predicted number of fixed (A) and total (B) deleterious mutations accumulated in the S-locus flanking regions according to the level of dominance after 100,000 generations of simulated populations. *The means were estimated by S-alleles dominance classes. The lines in bold represent the mean variation found in the 100 simulations.*

A major effect mutation linked to the Ah03 and Ah15 S-alleles?

In this study we found a transmission ratio distortion against Ah03 homozygotes, while in Llaurens et al. (2009a) the Ah15 homozygotes had smaller leaves with both lower width and length values, lower survival and lower number of seeds produced per cross. With our resequencing data we could examine the mutations linked to these two *S*-alleles. We found that these Ah03 and Ah15 *S*-alleles share a 1bp deletion in the fifth exon of *AT4G21323*, a gene involved in pollen tube growth. This mutation was not found in other alleles than the Ah03 and Ah15. In *A. thaliana*, a knock-out mutation in this gene promoted a segregation bias in the offspring against this mutation in cases where both partners presented the mutation (Qin et al., 2009). Interestingly, this mutation was also fixed in the three Ah03 copies from the Mortagne population.

Discussion

In this study, we expanded the phenotypic and theoretical analyses of the genetic load linked to self-incompatibility alleles in a SSI system, and we combined it with the first genomic assessment of the load within populations of *Arabidopsis halleri*. This allowed us to obtain a more nuanced view of the link between *S*-allele dominance and the sheltered load than was proposed by previous studies.

A first striking feature of our results is that the effect of the genetic load, measured for different traits in *A. halleri*, showed different patterns. For vegetative traits, we found no difference between homozygotes vs. heterozygotes at the *S*-locus, while for survival, we observed a significant decrease for Ah03/Ah03 homozygotes, compared with heterozygotes. This is in line with the observations of Stift et al. (2013) on *Arabidopsis lyrata* that find effect on survival of homozygotes at the *S*-locus, while Llaurens et al. (2009a) detected an effect on both vegetative traits and on survival in *A. halleri*. This heterogeneity for the fitness effect of the genetic load can be expected if we consider that each deleterious mutation associated with the different *S*-alleles measured can affect a different trait.

Another striking feature of our results is that we did not confirm the more severe deleterious load in dominant rather than recessive *S*-alleles observed by Llaurens et al. (2009a). There are notable experimental differences between our phenotypic study and that of Llaurens et al. (2009a). While Llaurens et al. (2009) used CO_2 treatment to by-pass the self-incompatibility system, we used the "natural" masking by dominant *S*-alleles to facilitate the obtention of homozygous genotypes. The main differences between the two approaches are : 1) our approach is experimentally simple and avoids possible contamination by offspring obtained by selfing, and which may contains combining effects of the sheltered load with those of genome-wide inbreeding depression (see Stift et al., 2013, for a discussion of this issue); 2) the *S*-locus homozygotes in Llaurens et al. (2009a) are true homozygotes (they possess two copies of the same parental *S*-haplotype), whereas in our study they possess two distinct copies of the same *S*-allele originating from the same population; given the low intra-allelic polymorphism expected for *S*-alleles (Castric et al. 2010), we reasoned that this difference should *a priori* be negligible; and 3) our approach is restricted to *S*-alleles that are recessive or intermediate along the dominance hierarchy, and so it is not applicable to the most dominant *S*-alleles. Overall, these technical differences are unlikely to account for the different patterns we observed, and we propose two possible explanations for why the results we obtained in our analysis differ from those of Llaurens et al. (2009a). First, a particular limitation in our approach as compared to Llaurens et al. (2009a) is that we did not include *S*-alleles from the highest dominance level for experimental reasons (as they were used for masking the more recessive *S*-alleles). It is therefore possible that the *S*-alleles we examined did not exhibit a sufficiently contrasted level of genetic load. Second, because quantifying the phenotypic effect of the genetic load is experimentally relatively demanding, both studies relied on the comparison of a limited number of *S*-alleles (three *S*-alleles in our study, two in Llaurens et al. 2009a). Hence, both studies had inherently low power for testing association between dominance and the level of sheltered load. However, our genomic analysis of the genetic load found in reconstituted haplotypes shows that one of the three particular *S*-alleles studied in Llaurens et al. (2009a), the most dominant Ah15, is indeed unusual in terms of the number of mutations it carries. We observed that it is one of the most "loaded" *S*-alleles among all the most dominant *S*-alleles present in this particular local population. This observation suggests that the choice of another *S*-allele belonging to the same dominance class could have led Llaurens et al. (2009a) to a different conclusion. As such, the fact that the mere comparison of the sheltered load between Ah01 and Ah15 fits with the theoretical expectation can be seen as a circumstantial coincidence. However the results of both studies confirm that some *S*-alleles have a substantial sheltered load. For the case of Ah15 the former result of Llaurens et al. (2009) is confirmed by our molecular analysis, even though the association with dominance is not obvious.

Beyond these experimental considerations, we confirmed the theoretical prediction by Llaurens et al. (2009a) that deleterious mutations can fix more readily in linkage with the most dominant *S*-alleles. However, we also demonstrated that *S*-alleles from the most recessive classes tend to maintain more linked standing variation for deleterious mutations, such that in the end *S*-allele dominance is not expected to influence the total amount of sheltered deleterious mutations of a given *S*-haplotype.

This observation involves that *S*-alleles, in particular the more recessive ones, can occur on a diversity of local haplotypes. Such "intra-allelic" polymorphism was documented by Miège et al. (2001) and Castric et al. (2010), but only for partial sequences of the *SRK* gene, hence for sites in complete (rather than partial) linkage within *S*-allele lineages. As expected because of the low effective population size within allelic lineages, these two studies observed very limited polymorphism overall, but Castric et al. (2010) confirmed that recessive *S*-alleles in *A. halleri* and *A. lyrata* tend to exhibit higher levels of nucleotide variation than dominant *S*-alleles.

The build-up of a sheltered load involves that *S*-allele*s* are associated with specific sets of linked mutations. We observed that the same suite of linked mutations were consistently associated with the different copies of a given *S*-allele when they were sampled from within the same population. As expected for outcrossing populations with short-scale linkage disequilibrium, this association was lost when examining sites at increasing distances from the *S*-locus along the chromosome (see also Le Veve et al. chapter 1). More importantly, the association with linked sites was lost when comparing gene copies of *S*-alleles sampled from different local populations. This suggests that crosses between individuals carrying identical *S*-alleles from distinct populations should not reveal as much load as observed within populations. The decrease of the intensity of population subdivision after 55kb observed in *A. halleri* can tentatively be explained by the presence of another gene under balancing selection. In Roux et al (2013), they found in *A. halleri*, an apparent exception of excess of polymorphism detected at the *At4g21480* gene, which is located 59 kb away from ARK3. This excess of polymorphism might be due to a distinct balancing selection process unrelated to SI that maintains both functional and nonfunctional alleles. Possibly, the signature of balancing selection at this gene is caused by long-term host-parasite interactions because this gene is implicated on the infection rate by male individuals of the nematode Heterodera schachtii (Hofmann et al. 2009). However, we can offer no explanation for why mutations in this distal region seemed to be specifically associated with the *S*-alleles. Uyenoyama (2003) showed that the existence of a sheltered load should influence the dynamics of apparition of new SI alleles. Specifically, antagonistic interactions are expected between ancestral and derived functional specificities if they initially share their linked deleterious mutations, slowing down the appearance of new SI alleles. Our observation that partially different sets of linked mutations are associated with *S*-alleles from the different populations raises the question of whether the (short) time scale at which recombination decouples *S*-alleles from their sets of linked mutation is sufficiently slow to allow such antagonistic interactions to take place. In other words, this effect should be important only in the case where the diversification dynamics of new *S*-alleles takes place within local populations, rather than involving a metapopulation-scale process (see Stetsenko et al. 2021).

The genetic load linked to *S*-alleles is an important factor that is also expected to impact the probability that new dominance relationships between *S*-alleles become established in a SSI system (Llaurens et al., 2009b). The question of whether the genetic load affects equally the evolution of dominance interactions between dominant *vs* recessive S-alleles remains open, and our observation that they have differently structured loads (in terms of the relative abundances of fixed vs. segregating mutations) represents an additional layer of complexity to this question. I explore this question in detail in chapter 3.

Material and methods

Source plant material

We worked on natural accessions from two closely related species, *A. halleri* and *A. lyrata*, each represented by two population samples named Mortagne and Nivelle for *A. halleri*, and three samples from three highly outcrossing populations from the North American Great Lakes, named IND (Indiana Dunes National Lakeshore in Michigan), PIN (Pinery Provincial Park in Ontario) and TSS (Tobermory Provincial Park in Ontario) (Foxe et al., 2010) for *A. lyrata* (Fig. 7, Table S5).

For these samples, we collected individuals from natural populations and we developed a dedicated sequence capture approach to sequence genomic regions of interest. The North American sample is represented by 9 individuals of IND, 11 individuals of PIN and 8 individuals of TSS, kindly provided by Barbara Mable (University of Glasgow; Fig. 7; Table S6). These populations colonised North America from ancestral European populations about 20.000-30.000 years ago (Clauss and Mitchell-Olds, 2006 ; Ross-Ibarra et al., 2008). We collected 60 individuals in Nivelle (50°47'N, 3°47'E, France) and the closely related Mortagne population (50°47'N, 3°47'E, France). These peripheral populations colonised the north of France during the last century from ancestral German populations (Pauwels et al., 2005).

We performed 92 and 91 controlled crosses between randomly chosen individuals within the Nivelle and Mortagne populations, respectively, and 40, 43 and 21 controlled crosses between randomly chosen individuals within the IND, PIN and TSS populations, respectively. We obtained 60, 66, 21, 21 and 10 successful crosses, respectively. We wanted to minimise the number of copies of the recessive *S*-allele Ah01 with the objective to reconstruct the maximum number of flanking region haplotypes linked to different *S*-alleles. We screened the individuals carrying this allele by PCR with *S*-allele-specific primers (Llaurens et al., 2008). We then preferentially selected the offspring with at most one parent with allele Ah01. For the population of Nivelle, we selected 33 individuals in Nivelle and 30 individuals in Mortagne, based on their genotype at the *S*-locus (Fig. 7; Table S6). We also selected their respective offspring (Table S7) and the offspring of five other crosses of the Nivelle population for the phenotypic measurements (Fig. 7). For the other populations, we selected one offspring of each parent from IND, PIN and TSS (Table S7).

Library preparation, capture and sequencing

Purified DNA was quantified by Qubit and 50 ng of DNA was fragmented mechanically with a Bioruptor (Diagenode) to obtain fragments of around 300 pb that we verified on 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. The extremities of fragments 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 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 synthesised to target 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). 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). For five individuals from the capture datasets (Table S6), we completed the sequencing with genome-wide resequencing (WGS) in order to distinguish the homozygous and heterozygous genotypes at the *S*-locus based on read depth (Genete et al. 2020), which is not possible using data from the capture protocol. The libraries previously prepared were sequenced by Illumina NovaSeq (2x 150pb, paired-end) from the GenoScreen platform (Lille, France).

Determination of the S-locus genotypes and dominance of S-alleles

We used a dedicated pipeline for genotyping the *S*-locus based on short reads sequencing (Genete et al., 2020), in order to determine the *S*-alleles present in each parental individual and their offsprings (Table S6 and S7). This pipeline implements sequential mapping of individual reads against each previously known *SRK* sequence from the literature, and computes mapping statistics to determine the identity of the *S*-alleles carried by each individual. The level of dominance of *S*-alleles found in our study was determined based on either previous assessments of dominance in *A. lyrata* and *A. halleri* (Schierup et al., 2001; Mable et al., 2003; Bechsgaard et al., 2004; Llaurens et al. 2008; Durand et al. 2014) or indirectly inferred based on the observed association between the phylogeny of *S*-alleles and level of dominance (Prigoda et al,. 2005).

Read mapping and variant calling in *A. halleri* and *A. lyrata* populations

Raw reads were mapped on the complete *A. lyrata* reference genome (V1.0.23, Hu et al., 2011) using Bowtie2 v2.4.1 (Langmead and Salzberg, 2012), as described in Le Veve et al (Chapter I). 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 at https://github.com/leveveaudrey/analysis-of-polymorphism-S-locus.

We obtained an average of 620 million properly mapped paired-end 300bp reads per population sample. For consistency, we conserved only reads which mapped to the *S*-locus flanking or control regions, even for samples sequenced by WGS, using the targetintercept option of bedtool v2.25.0 (Quinlan and Hall, 2010). In Le Veve et al (Chapter I), we demonstrated that only the first 25kb around the *S*-locus present an excess of polymorphism. Hence, here we focused on the 75 kb after the first base of the gene *Ubox* in 3' and the last base of the gene *ARK3* in 5' at the *S*-locus. This region contains 20 annotated genes. In this study we excluded the genes inside the *S*-locus itself (*SCR*, *SRK* and microRNAs). 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). Finally, we removed SNPs fixed in each population using the script 1_fix_pos_vcf.py (https://github.com/leveveaudrey/dominance_and_sheltered_load) .

Quantifying the sheltered load of deleterious mutations

We examined the genetic load signatures based on the accumulation of mutations on 0-fold degenerate sites, the vast majority of which are considered deleterious. The 0-fold and 4-fold degenerate sites were identified and extracted from the reference genome and the gene annotation using the script NewAnnotateRef.py (Williamson et al., 2014). The 3-fold and 2-fold degenerate sites were not considered. The number of variable positions considered for each dataset is summarised in table S8.

Phasing S-haplotypes

Based on the sequencing of parents and offsprings of 9, 11, 5, 6 and 5 of compatible crosses in Nivelle, Mortagne, IND, PIN and TSS populations respectively (Fig. 7), we phased the haplotypes of 126 *S*-allele copies, using the script 3_phase_S_allele.py (https://github.com/leveveaudrey/dominance_and_sheltered_load). This pipeline compares the polymorphic sites present in parents and offspring and attributes the commun variant to commun *S*-alleles and the others variant to the other allèle found in parents. We assumed no recombination between mutations in the offspring and their respective parents. Moreover, because some parents were used in many crosses, we avoided duplications of phased haplotypes.

Study of the structure of S-haplotypes

To compare the phased haplotypes onto which the *S*-alleles are found, we used maximum likelihood phylogenies based on the Tamura-Nei model (Tamura and Nei, 1993), with 1000 replicates. The analyses were conducted in MEGA X (Kumar et al,. 2018). Moreover, we divided the phased haplotypes into non overlapping windows of 5kb and examined the variation of F_{ST} between populations within each species (Nivelle and Mortagne for *A. halleri* and IND, PIN and TSS for *A. lyrata*) along the flanking region. We also examined the variation of F_{ST} along the flanking region obtained by grouping haplotypes by their linked *S*-allele rather than by population of origin. Finally, we compared these F_{ST} values computed in the *S*-locus flanking regions with those computed for the 100 control regions. The F_{ST} values were estimated with the DNAsp 6 software (Rozas et al., 2017).

Estimation of the number of fixed and segregating deleterious mutations within S-allele lineages

For each variable position considered in the phased haplotypes, we used the script 3_phase_S_allele.py (https://github.com/leveveaudrey/dominance_and_sheltered_load), to estimate the number of mutations on 0-fold (S_{0f}) and 4-fold degenerate sites (S_{4f}) compared with the reference genome. We distinguished SNPs that were fixed from those that were segregating within each of the allelic lines. We then tested by GLM whether the number of fixed and/or segregating deleterious mutations were associated with the dominance level of the allelic line.

Estimation of the phenotypic impact of homozygosity at the S-locus for three S-alleles

To determine if the genetic sheltered load putatively linked to the *S*-locus has a detectable phenotypic impact, we performed 45 crosses between offspring of the Nivelle individuals that shared one *S*-allele and obtained as described above (Fig. 7). Based on the dominance hierarchy in pollen (Durand et al., 2014, Table S1), these crosses should correspond to compatible partners. The general principle of the experiment was to take advantage of the dominance hierarchy to mask recessive *S*-alleles and generate full sibs that were either homozygous (because they inherited the *S*-allele that was shared by their two parents) or heterozygous at the *S*-locus, and thus isolate the effect of homozygosity at the *S*-locus. Note that all offspring in our experiments were thus ”naturally” outcrossed, whereas Llaurens et al. (2009a) based their comparisons on outcrossed progenies obtained by enforced incompatible crosses and Stift et al. (2011) based their comparisons on enforced selfed progenies. These crosses generated 399 seeds overall, with homozygous genotypes expected for the *S*-allele Ah01 of classe I, Ah03 of class II, and Ah04 of class III and the following dominance relationships : Ah01<Ah03<Ah04.

Seedlings were grown in a greenhouse between 14.5 and 23.1°C and a photoperiod of 16 hr day/8 hr night. Offsprings from the four selected crosses were randomly placed on the greenhouse tables, and their position randomised every 3 days. After three months of growing, all the germinated plants were vernalised under a temperature between 6 and 8°C and a natural photoperiod for two months (January-February). Then, all surviving plants began reproduction in a greenhouse under temperature

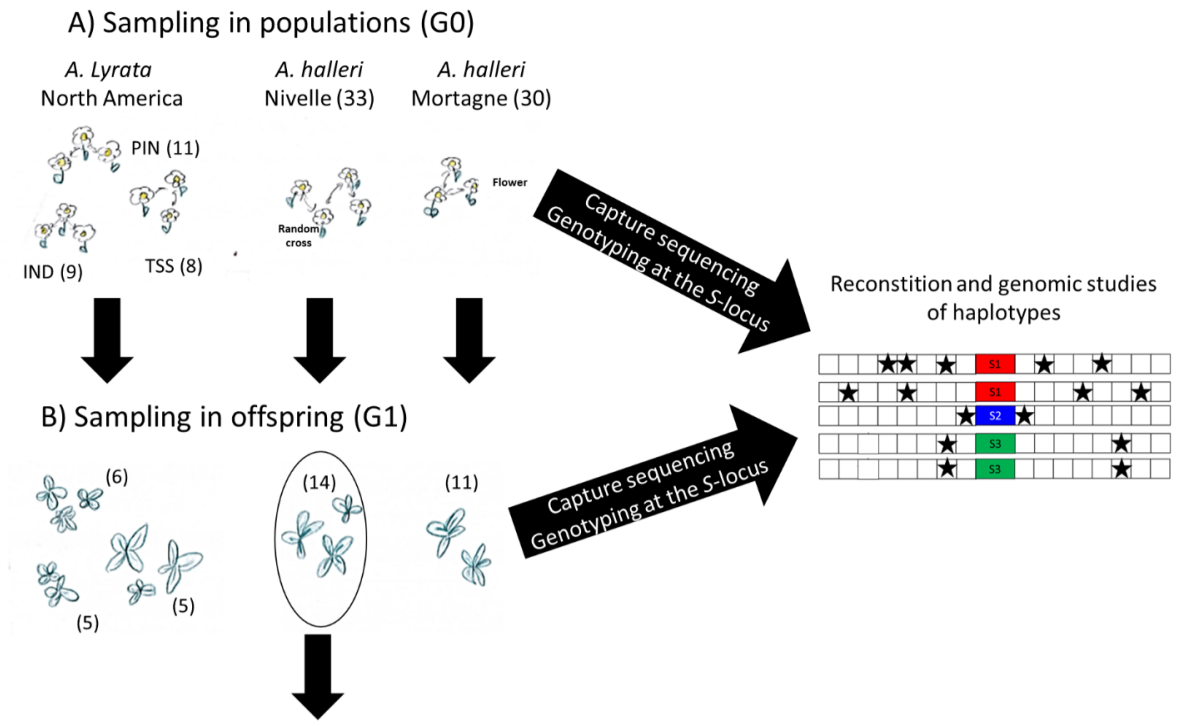
between 10.6 and 25.3°C and a natural photoperiod. The genotypes at the *S*-locus were determined in surviving plants by a PCR approach, using *S*-allele-specific primers for the pistil-expressed *SRK* gene. We assessed the reproductive success of offspring from the different crosses on the basis of a total of fourteen phenotypic traits (detailed below), and computed, within each family, the difference for the trait between homozygotes and heterozygotes. We also compared the observed proportions of each *S*-locus genotypic category in the family after the apparition of the first stem to their mendelian expectation. Departures from mendelian expectation were interpreted as reflecting differences in survival between homozygous and heterozygous *S*-locus genotypes. We used 10,000 replicate simulations of mendelian segregation based on the *S*-locus genotype of the parents. We expected the phenotypic impact of homozygosity at the *S*-locus to increase with dominance of the *S*-alleles and tested this expectation by GLM. The models used for GLM (poisson, gaussian…) depended on the type of trait analysed.

The fourteen phenotypic traits measured were : the time (days) for the first leaves measured by visual control every day during seven weeks after sowing the seeds, the number of leaves, the area of the rosette (cm^2), the mean length and width of leaves (cm), the standard deviation of length and width of leaves (cm) and the mean area of leaves (cm^2) measured by ImageJ (Schneider et al., 2012) based on photographs taken seven weeks (+/- five days) after the first leaf. At reproduction, we measured the time to the first flower bud (day), scored by visual control every three days during nine weeks, the number of flower buds per flower stem produced during four week after the appearance of the first bud, the number of flower stems, the length of the highest flower stem produced four weeks after the appearance of the first bud (cm), and finally the time of production of buds (days), scored by visual control every three days during eleven weeks after the appearance of the first bud. The last trait we measured was the proportion of homozygotes per family that survived until reproduction time. During the whole experiment, the presence of phytophages, pathogens and stress markers were scored as binary variables. The presence of phytophages and pathogen attacks were detected by the occurrence of gaps in leaves. The oxidative stress marker was defined qualitatively based on the occurrence of purple leafs. These effects were controlled by redistributing 1,000 times the values observed in groups of the same size observed for each effect (for example, presence or absence of pathogen attack) and comparing the difference for the trait observed with the distribution of the differences obtained in the permutations. We considered the impact of the effect on the trait if the observed difference between groups was higher than the 95% percentile of the distribution obtained randomly (Table S9, S10, S11 for phytopathogens, phytophages and oxidative stress respectively). When the test was significant , the effect was implemented as a random effect in the GLM. We used the same method to control for the family effect, which was included as a random effect in GLM if necessary (Table S12).

Simulations

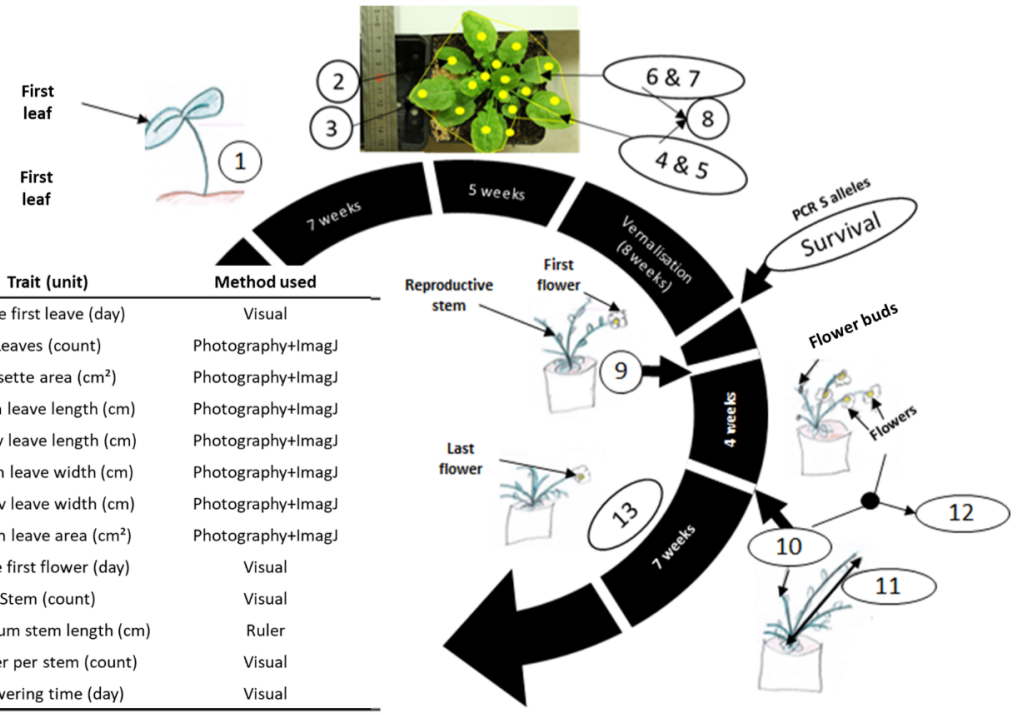
Finally, we refined the model of Llaurens et al., (2009a) to take into account the fact that a substantial proportion of linked mutations are segregating rather than fixed within allelic lineages. We modified the model of SSI with hierarchical dominance interactions in several ways. First, the size of the region (D) strictly linked to the *S*-locus was increased to one hundred potentially deleterious positions (only one position was used in Llaurens et al., 2009a)). Second, the population size was 10,000 diploid individuals (previously it was 1,000), so as to be large enough to avoid *S*-allele loss by drift during the simulations. Finally, four dominance classes (before only three were used), as observed in *A. halleri* (Durand et al., 2014), were implemented with fourteen *S*-alleles (eight alleles in the class IV, three in the class III, two in the class II and one allele in the class I). This distribution is more congruent with the observations in the natural populations studied (Table S6). The alleles in class IV were assumed to be codominant to each other, and dominant over all alleles of the other classes. The alleles in class III were codominant to each other and dominant over all alleles of classes II and I. The alleles in class II

were codominant to each other and dominant over the allele of class I. We also assumed that no new *S*-allele could appear by mutation during the simulations. We first ran simulations without deleterious mutations until a deterministic equilibrium for *S*-allele frequencies was reached, which was considered to be reached when the allelic frequencies changed between generation by less than 10^{-3}. Deleterious mutations were then allowed to accumulate at the D locus . Each simulation was performed with 100 independent replicates of 100,000 generations, and the frequency of the deleterious alleles was recorded every 1,000 generations. The coefficient of selection of mutations in the D locus was fixed at 0.01 and fitness was multiplicative. At the end of the simulation runs, we estimated the number of deleterious mutations found in each haplotype associated with each *S*-allele to determine the association between the accumulation of genetic sheltered load and dominance at the *S*-locus. Finally, we estimated the number of fixed deleterious mutations found in each allelic lineage to determine the association between the accumulation of the number of fixed deleterious mutations and the dominance level of the allelic lineage associated. The general experimental procedure is summarised in Fig. 7 and all data analyses were done in R ver. 3.1.2 (R Development Core Team 2014).



C) Compatible crosses of Nivelle G1 to obtain families G2

D) Traits measurements



N°	Trait (unit)	Method used
1	Time first leave (day)	Visual
2	Leaves (count)	Photography+ImagJ
3	Rosette area (cm²)	Photography+ImagJ
4	Mean leave length (cm)	Photography+ImagJ
5	St dev leave length (cm)	Photography+ImagJ
6	Mean leave width (cm)	Photography+ImagJ
7	St dev leave width (cm)	Photography+ImagJ
8	Mean leave area (cm²)	Photography+ImagJ
9	Time first flower (day)	Visual
10	Stem (count)	Visual
11	Maximum stem length (cm)	Ruler
12	Flower per stem (count)	Visual
13	Flowering time (day)	Visual

Figure 7: Experimental protocol. *A) We randomly crossed A. lyrata individuals from the PIN, TSS and IND populations in North America (left) and of A. halleri of Nivelle (middle) and Mortagne (right) populations. The individuals were selected after PCR against the recessive S-allele SRK01 to minimise the frequency of this allele in the dataset and constitute the G0 populations. The individuals selected were sequenced by a capture protocol. The numbers between parentheses represent the number of individuals per dataset. B) One offspring from each cross was sequenced along with its two parents for trio haplotyping. . The offspring from the Nivelle population (black circle) were conserved for the study of the impact of homozygosity at the S-locus. C) We used the dominance hierarchy between S-alleles expressed in pollen (Llaurens et al., 2008; Durand et al,. 2014) to cross the individuals of G1 of Nivelle populations and obtained six G2 families constituted of heterozygotes and homozygotes for the alleles Ah01 (class I) , Ah03 (class II) and Ah04 (class III). D) Description of the traits measured and the methods used to estimate the impact of homozygosity at the S-locus in homozygotes. Traits between the N° 1 to 8 are related to biomass and traits 9 to 13 are related to reproductive success. The times between each step are reported in the life cycle.*

Data Availability

All sequence data are available in the NCBI Short Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) with accession codes: PRJNA744343, PRJNA755829.

All scripts developed are available in Github (https://github.com/leveveaudrey/dominance_and_sheltered_load https://github.com/leveveaudrey/analysis-of-polymorphism-S-locus).

Acknowledgements

This work was funded by a grant from the France-Berkeley Fund to VC and Rasmus Nielsen, the European Research Council (NOVEL project, grant #648321), ANR TE-MoMa (grant ANR-18-CE02-0020-01). AL thanks the ERC and the University of Lille for funding her PhD project. We thank Barbara Mable for sharing seeds of *A. lyrata*. We thank also the greenhouse platform of the University of Lille for helps with the experiments.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. The authors thank the GenoScreen platform (Lille, France).

Bibliography

Bateman, A.J. (1952). Self-incompatibility systems in angiosperms: I. Theory. Heredity 6, 285–310.

Bechsgaard, J., Bataillon, T., and Schierup, M.H. (2004). Uneven segregation of sporophytic self-incompatibility alleles in Arabidopsis lyrata. Journal of Evolutionary Biology 17, 554–561.

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

Castric, V., and 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.S., Grenier, S., Noureddine, R., Schierup, M.H., and Vekemans, X. (2010). Molecular evolution within and between self-incompatibility specificities. Molecular Biology and Evolution 27, 11–20.

Charlesworth, B., and Charlesworth, D. (1999). The genetic basis of inbreeding depression. Genetics Research 74, 329–340.

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

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

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.

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.

Durand, E., Méheust, R., Soucaze, M., Goubet, P.M., Gallina, S., Poux, C., Fobis-Loisy, I., Guillon, E., Gaude, T., Sarazin, A., et al. (2014). Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. Science 346, 1200–1205.

Foxe, J.P., Stift, M., Tedder, A., Haudry, A., Wright, S.I., and 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.

Garrigan, D., and Hedrick, P.W. (2003). Perspective: Detecting Adaptive Molecular Polymorphism: Lessons from the Mhc. Evolution 57, 1707–1722.

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. Mol Biol Evol.

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.

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

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.

Kumar S, Stecher G, Li M, Knyaz C, and Tamura K (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution 35:1547-1549.

Kusaba, M., Dwyer, K., Hendershot, J., Vrebalov, J., Nasrallah, J.B., and 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.

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

Le Veve, A., Burghgraeve, N., Genete, M., Lepers-Blassiau, C., Rasmus, N., Takou, M., De Meaux, J., Mable, B., Durand, E., Vekemans, X., Castric, V. (Chapter I). Long-term balancing selection and the genetic load linked to the self-incompatibility locus in *Arabidopsis halleri* and *A. lyrata*.

Lenz, T.L., Spirin, V., Jordan, D.M., and 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., and 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., and 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., and Billiard, S. (2009a). 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., Billiard, S., Castric, V., and Vekemans, X. (2009b). Evolution of dominance in sporophytic self-incompatibility systems: I. Genetic load and coevolution of levels of dominance in pollen and pistil. Evolution 63, 2427–2437.

Lynch, M., Ackerman, M.S., Gout, J.-F., Long, H., Sung, W., Thomas, W.K., and Foster, P.L. (2016). Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17, 704–714.

Mable, B.K., Schierup, M.H., and Charlesworth, D. (2003). Estimating the number, frequency, and dominance of S -alleles in a natural population of Arabidopsis Lyrata (Brassicaceae) with sporophytic control of self-incompatibility. Heredity 90, 422–431.

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

Miege, C., Ruffio-Châble, V., Schierup, M.H., Cabrillac, D., Dumas, C., Gaude, T., and Cock, J.M. (2001). Intrahaplotype polymorphism at the Brassica S Locus. Genetics 159, 811–822.

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

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.

Pauwels, M., Saumitou-Laprade, P., Holl, A.C., Petit, D., and 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.

Prigoda, N.L., Nassuth, A., and Mable, B.K. (2005). Phenotypic and genotypic expression of self-incompatibility haplotypes in *Arabidopsis lyrata* suggests unique origin of alleles in different dominance classes. Molecular Biology and Evolution 22, 1609–1620.

Qin, Y., Leydon, A.R., Manziello, A., Pandey, R., Mount, D., Denic, S., Vasic, B., Johnson, M.A., and Palanivelu, R. (2009). Penetration of the Stigma and Style Elicits a Novel Transcriptome in Pollen Tubes, Pointing to Genes Critical for Growth in a Pistil. PLoS Genet 5, e1000621.

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

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., and 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.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., and Sánchez-Gracia, A. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol Biol Evol 34, 3299–3302.

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

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9, 671–675.

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

Stetsenko, R., Brom, T., Castric, V., and Billiard, S. (2021). Balancing selection and the crossing of fitness valleys in structured populations: diversification in the gametophytic self-incompatibility system. BioRxiv 2021.11.20.469375.

Stift, M., Hunter, B.D., Shaw, B., Adam, A., Hoebe, P.N., and 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.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10, 512–526.

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 Phytologist 165, 63–70.

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

Supplementary data

Table S1 : Crosses to obtain homozygotes for three S-alleles.

Genotype pollen donor	Genotype stigmate	Homozygote studied	Dominance level	Number of crosses	Number of seed
Ah20/Ah01	Ah12/Ah01	Ah01	1	5	48
Ah20/Ah01	Ah12/Ah01	Ah01	1	2	23
Ah02/Ah03	Ah03/Ah01	Ah03	2	6	69
Ah24/Ah03	Ah03/Ah01	Ah03	2	3	27
Ah20/Ah04	Ah04/Ah04	Ah04	3	21	167
Ah20/Ah04	Ah04/Ah04	Ah04	3	8	65

The S-alleles in bold represent the dominant S-allele expressed on pollen of each donor genotype.

Table S2 : Variation on trait on homozygous at the S-locus.

Trait (unit)	Mean heterozygotes	Effect of homozygosity	P-value
Time first leave (day)	15.15	0.93	0.09
Number of leaves	13.6	1	0.48
Rosette area (cm²)	37.15	0.95	0.35
Mean leaf length (cm)	1.93	0.96	0.29
St dev leaf length (cm)	0.29	1.03	0.39
Mean leaf width (cm)	1.54	0.94	0.16
St dev leaf width (cm)	0.27	0.93	0.16
Mean leaf area (cm²)	3.33	0.93	0.27
Time first flower (day)	**36.29**	**0.96**	**0.04**
Number of flowering stems	13.2	0.99	0.42
Maximum flowering stem length (cm)	**56.6**	**1.12**	**8.0e-4**
Number of flowers by flowering stem	58.86	1.17	0.07
Flowering time (day)	44.02	1.01	0.43

The P Values represent the proportions of the distribution equal to or less than the value observed obtained after 10,000 random resamples. The significant values are represented in bold. The effect of homozygosity represents the ratio between the mean value obtained in homozygotes on the mean value obtained in heterozygotes .

Table S3: Variation on trait on homozygous at the S-locus in each family.

Trait (unit)	Pollen donnor	Stigmate	Allele	Mean heterozygotes	Effect	P value
Time first leave (day)	d24.1	d29.1	Ah01	19.47	0.88	0.18
Number of leaves	d24.1	d29.1	Ah01	13.37	1.05	0.34
Rosette area (cm²)	d24.1	d29.1	Ah01	35.24	0.78	0.21
Mean leaf length (cm)	d24.1	d29.1	Ah01	1.83	0.9	0.25
St dev leaf length (cm)	d24.1	d29.1	Ah01	0.30	0.96	0.37
Mean leaf width (cm)	d24.1	d29.1	Ah01	1.49	0.93	0.32
St dev leaf width (cm)	d24.1	d29.1	Ah01	0.31	0.77	0.05
Mean leaf area (cm²)	d24.1	d29.1	Ah01	3.08	0.82	0.23
Time first flower (day)	d24.1	d29.1	Ah01	38.44	0.93	0.18
Number of flowering stems	d24.1	d29.1	Ah01	12.80	0.96	0.41
Maximum flowering stem length (cm)	d24.1	d29.1	Ah01	47.11	0.92	0.20
Number of flowers by flowering stem	d24.1	d29.1	Ah01	43.71	0.73	0.15
Flowering time (day)	d24.1	d29.1	Ah01	40.27	1.11	0.10
Time first leave (day)	d12.1	d29.1	Ah01	13.14	1.03	0.52
Number of leaves	d12.1	d29.1	Ah01	16	0.91	0.31
Rosette area (cm²)	d12.1	d29.1	Ah01	44.73	0.65	0.16
Mean leaf length (cm)	d12.1	d29.1	Ah01	2.08	0.86	0.26
St dev leaf length (cm)	d12.1	d29.1	Ah01	0.31	0.77	0.23
Mean leaf width (cm)	d12.1	d29.1	Ah01	1.83	0.76	0.13
St dev leaf width (cm)	d12.1	d29.1	Ah01	0.29	0.66	0.18
Mean leaf area (cm²)	d12.1	d29.1	Ah01	3.95	0.73	0.22
Time first flower (day)	**d12.1**	**d29.1**	**Ah01**	**35.71**	**0.87**	**0.02**
Number of flowering stems	d12.1	d29.1	Ah01	12.29	0.9	0.37
Maximum flowering stem length (cm)	d12.1	d29.1	Ah01	51.03	1.01	0.46
Number of flowers by flowering stem	d12.1	d29.1	Ah01	43.79	2.37	0.06

Flowering time (day)	d12.1	d29.1	Ah01	47.57	1.09	0.23
Time first leave (day)	d208.1	d17.1	Ah03	19	0.84	0.6
Number of leaves	d208.1	d17.1	Ah03	13	1.15	0.4
Rosette area (cm²)	d208.1	d17.1	Ah03	15.46	2.27	0.21
Mean leaf length (cm)	d208.1	d17.1	Ah03	1.37	1.82	0.21
St dev leaf length (cm)	d208.1	d17.1	Ah03	0.27	0.41	0.19
Mean leaf width (cm)	d208.1	d17.1	Ah03	1.15	1.58	0.21
St dev leaf width (cm)	d208.1	d17.1	Ah03	0.2	1.15	0.4
Mean leaf area (cm²)	d208.1	d17.1	Ah03	1.61	2.81	0.2
Time first flower (day)	d208.1	d17.1	Ah03	34.5	0.96	0.6
Number of flowering stems	d208.1	d17.1	Ah03	13.75	0.58	0.2
Maximum flowering stem length (cm)	d208.1	d17.1	Ah03	55.65	1.1	0.2
Number of flowers by flowering stem	d208.1	d17.1	Ah03	61.49	1.67	0.2
Flowering time (day)	d208.1	d17.1	Ah03	46.25	1.06	0.8
Time first leave (day)	d10.1	d17.1	Ah03	14.6	0.96	0.76
Number of leaves	**d10.1**	**d17.1**	**Ah03**	**12.3**	**1.63**	**0.04**
Rosette area (cm²)	d10.1	d17.1	Ah03	33.29	0.76	0.48
Mean leaf length (cm)	d10.1	d17.1	Ah03	1.79	0.88	0.52
St dev leaf length (cm)	d10.1	d17.1	Ah03	0.27	2.11	0.09
Mean leaf width (cm)	d10.1	d17.1	Ah03	1.49	0.87	0.43
St dev leaf width (cm)	**d10.1**	**d17.1**	**Ah03**	**0.24**	**2.5**	**0.05**
Mean leaf area (cm²)	d10.1	d17.1	Ah03	3.13	0.65	0.48
Time first flower (day)	d10.1	d17.1	Ah03	38.84	0.75	0.15
Number of flowering stems	d10.1	d17.1	Ah03	10.18	0.59	0.17
Maximum flowering stem length (cm)	d10.1	d17.1	Ah03	52.75	1.52	0.05
Number of flowers by flowering stem	d10.1	d17.1	Ah03	72.28	1.42	0.16
Flowering time (day)	d10.1	d17.1	Ah03	48.12	1.1	0.11
Time first leave (day)	d205.2	d191.2	Ah04	13.12	1	0.51

Number of leaves	d205.2	d191.2	Ah04	13.74	0.95	0.22
Rosette area (cm²)	d205.2	d191.2	Ah04	37.54	0.95	0.4
Mean leaf length (cm)	d205.2	d191.2	Ah04	1.99	0.94	0.27
St dev leaf length (cm)	d205.2	d191.2	Ah04	0.28	1.04	0.36
Mean leaf width (cm)	d205.2	d191.2	Ah04	1.62	0.91	0.14
St dev leaf width (cm)	d205.2	d191.2	Ah04	0.3	0.87	0.09
Mean leaf area (cm²)	d205.2	d191.2	Ah04	3.59	0.89	0.26
Time first flower (day)	d205.2	d191.2	Ah04	35.52	1	0.48
Number of flowering stems	d205.2	d191.2	Ah04	13.39	0.97	0.38
Maximum flowering stem length (cm)	**d205.2**	**d191.2**	**Ah04**	**62.1**	**1.09**	**0.007**
Number of flowers by flowering stem	d205.2	d191.2	Ah04	68.32	1.12	0.18
Flowering time (day)	d205.2	d191.2	Ah04	44.45	0.97	0.19
Time first leave (day)	d205.2	d50.2	Ah04	14.94	0.99	0.53
Number of leaves	d205.2	d50.2	Ah04	14.29	0.94	0.32
Rosette area (cm²)	d205.2	d50.2	Ah04	45.04	1.06	0.43
Mean leaf length (cm)	d205.2	d50.2	Ah04	2.16	0.95	0.4
St dev leaf length (cm)	d205.2	d50.2	Ah04	0.3	1.03	0.41
Mean leaf width (cm)	d205.2	d50.2	Ah04	1.45	0.97	0.42
St dev leaf width (cm)	d205.2	d50.2	Ah04	0.22	1.09	0.29
Mean leaf area (cm²)	d205.2	d50.2	Ah04	3.48	0.99	0.48
Time first flower (day)	d205.2	d50.2	Ah04	33.59	0.96	0.28
Number of flowering stems	d205.2	d50.2	Ah04	16.47	0.99	0.48
Maximum flowering stem length (cm)	d205.2	d50.2	Ah04	60.65	1.09	0.07
Number of flowers by flowering stem	d205.2	d50.2	Ah04	46.01	0.95	0.37
Flowering time (day)	d205.2	d50.2	Ah04	40.41	1.09	0.14

The P Values represent the proportions of the distribution equal to or less than the value observed obtained after 10,000 random resamples. The significant values are represented in bold. The effect represents the ratio between the mean value obtained in homozygotes on the mean value obtained in heterozygotes .

Table S4 : Test of the correlation between the number of fixed mutations between copies of a same S-allele in the regions of 25kb around the S-locus with dominance at the S-locus.

Species	***Population***	***Effect***	***P value***
A. halleri	Nivelle	0.097	1.95e-7
	Mortagne	0.066	1.27e-6
A. lyrata	North America	0.232	2.338e-16

P value obtained by GLM for correlation found with dominance at the S-locus.

Table S5 : Summary of the datasets used.

Species	Population (subpopulations)	Accession N°	Age estimated of the last bottleneck (reference)	Sample size (individuals)
A. halleri	Nivelle (1)	PRJNA744343	100 years (Pauwels et al., 2005)	28
	Mortagne (1)			26
A. lyrata	North America (3)	PRJNA755829	35000 years (Ross-Ibarra et al. 2008)	28

The individuals were mainly sequenced by capture approach. The genotypes of potential homozygotes were confirmed after a whole genome sequencing.

Table S6 : Genotypes at the S-locus of the individuals sequenced by capture.

Identity	Allele 1	Allele 2	Dominance 1	Dominance 2	Population	Species	SRA
Mor_19_13	Ah04	Ah25	III	III	Mortagne	*A. halleri*	SAMN20088356
Mor_19_14	Ah04	Ah01	III	I	Mortagne	*A. halleri*	SAMN20844087
Mor_19_19	Ah12	Ah36	IV	IV	Mortagne	*A. halleri*	SAMN20088358
Mor_19_2	Ah03	Ah24	II	IV	Mortagne	*A. halleri*	SAMN20088359
Mor_19_22	Ah03	Ah20	II	IV	Mortagne	*A. halleri*	SAMN20088354
Mor_19_23	Ah20	Ah25	IV	III	Mortagne	*A. halleri*	SAMN20844088
Mor_19_24	Ah12	Ah24	IV	IV	Mortagne	*A. halleri*	SAMN20088362
Mor_19_3	Ah03	Ah01	II	I	Mortagne	*A. halleri*	SAMN20844098
Mor_19_37	Ah59	Ah20	IV	IV	Mortagne	*A. halleri*	SAMN20088357
Mor_19_38	Ah36	Ah01	IV	I	Mortagne	*A. halleri*	SAMN20088371
Mor_19_4	Ah25	Ah12	III	IV	Mortagne	*A. halleri*	SAMN20088360
Mor_19_41	Ah25	Ah25	III	III	Mortagne	*A. halleri*	SAMN20088372
Mor_19_42	Ah12	Ah04	IV	III	Mortagne	*A. halleri*	SAMN20088373
Mor_19_45	Ah36	Ah02	IV	III	Mortagne	*A. halleri*	SAMN20088374
Mor_19_46	Ah03	Ah12	II	IV	Mortagne	*A. halleri*	SAMN20088375
Mor_19_51	Ah20	Ah05	IV	IV	Mortagne	*A. halleri*	SAMN20088361
Mor_19_53	Ah36	Ah12	IV	IV	Mortagne	*A. halleri*	SAMN20088365
Mor_19_54	Ah24	Ah12	IV	IV	Mortagne	*A. halleri*	SAMN20088366
Mor_19_55	Ah20	Ah05	IV	IV	Mortagne	*A. halleri*	SAMN20088377
Mor_19_56	Ah12	Ah03	IV	II	Mortagne	*A. halleri*	SAMN20088378
Niv_19_18	Ah15	Ah03	IV	II	Nivelle	*A. halleri*	SAMN20088349

Niv_19_19	Ah20	Ah24	IV	IV	Nivelle	*A. halleri*	SAMN20088350
Niv_19_22	Ah24	Ah01	IV	I	Nivelle	*A. halleri*	SAMN20088351
Niv_19_23	Ah20	Ah12	IV	IV	Nivelle	*A. halleri*	SAMN20088352
Niv_19_3	Ah04	Ah59	III	IV	Nivelle	*A. halleri*	SAMN20088336
Niv_19_31	Ah04	Ah24	III	IV	Nivelle	*A. halleri*	SAMN20088337
Niv_19_4	Ah01	Ah24	I	IV	Nivelle	*A. halleri*	SAMN20088382
Niv_19_42	Ah22	Ah12	IV	IV	Nivelle	*A. halleri*	SAMN20088341
Niv_19_45	Ah12	Ah04	IV	III	Nivelle	*A. halleri*	SAMN20088344
Niv_19_5	Ah25	Ah67	III	IV	Nivelle	*A. halleri*	SAMN20088340
Niv_19_52	Ah20	Ah22	IV	IV	Nivelle	*A. halleri*	SAMN20088331
Niv_19_53	Ah05	Ah20	IV	IV	Nivelle	*A. halleri*	SAMN20088339
Niv_19_54	Ah01	Ah12	I	IV	Nivelle	*A. halleri*	SAMN20844097
Niv_19_58	Ah02	Ah59	III	IV	Nivelle	*A. halleri*	SAMN20088343
Niv_19_59	Ah01	Ah67	I	IV	Nivelle	*A. halleri*	SAMN20844096
Niv_19_60	Ah04	Ah20	III	IV	Nivelle	*A. halleri*	SAMN20088332
Niv_19_7	Ah04	Ah67	III	IV	Nivelle	*A. halleri*	SAMN20088345
Niv_19_8	Ah04	Ah12	III	IV	Nivelle	*A. halleri*	SAMN20844090
Niv_19_9	Ah04	Ah67	III	IV	Nivelle	*A. halleri*	SAMN20088338
Pin_15_1	Ah03	Ah01	II	I	PIN	*A. lyrata*	SAMN20088324
Pin_16_1	Ah01	Ah46	I	IV	PIN	*A. lyrata*	SAMN20088318
Pin_4_24	Ah03	Ah29	II	III	PIN	*A. lyrata*	SAMN20088314
Pin_4_54	Ah03	Ah03	II	II	PIN	*A. lyrata*	SAMN20088319
Pin_5_1	Ah01	Ah01	I	I	PIN	*A. lyrata*	SAMN20844102
Pin_5_12	Ah29	Ah01	III	I	PIN	*A. lyrata*	SAMN20088325
Pin_5_2	Ah03	Ah01	II	I	PIN	*A. lyrata*	SAMN20088311
Pin_8_15	Ah01	Ah01	I	I	PIN	*A. lyrata*	SAMN20088328

Pin_8_2	Ah29	Ah63	III	III	PIN	*A. lyrata*	SAMN20088326
Pin_9_1	Ah42	Ah29	IV	III	PIN	*A. lyrata*	SAMN20088312
Tss_14_3	Ah01	Ah31	I	IV	TSS	*A. lyrata*	SAMN20088315
Tss_21_10	Ah01	Ah03	I	II	TSS	*A. lyrata*	SAMN20088329
Tss_22_24	Ah01	Ah31	I	IV	TSS	*A. lyrata*	SAMN20088306
Tss_22_7	Ah01	Ah31	I	IV	TSS	*A. lyrata*	SAMN20088317
Tss_3_10	Ah31	Ah01	IV	I	TSS	*A. lyrata*	SAMN20088330
Tss_3_23	Ah18	Ah01	IV	I	TSS	*A. lyrata*	SAMN20088307
Tss_5_1	Ah01	Ah31	I	IV	TSS	*A. lyrata*	SAMN20088316
Ind_1_1	Ah01	Ah01	I	I	IND	*A. lyrata*	SAMN20088320
Ind_10_3	Ah31	Ah24	IV	IV	IND	*A. lyrata*	SAMN20088327
Ind_15_1	Ah03	Ah24	II	IV	IND	*A. lyrata*	SAMN20088321
Ind_15_2	Ah63	Ah03	III	II	IND	*A. lyrata*	SAMN20088309
Ind_15_3	Ah18	Ah24	IV	IV	IND	*A. lyrata*	SAMN20088322
Ind_18_1	Ah01	Ah01	I	I	IND	*A. lyrata*	SAMN20844103
Ind_6_1	Ah01	Ah63	I	III	IND	*A. lyrata*	SAMN20088308
Ind_8_1	Ah03	Ah63	II	III	IND	*A. lyrata*	SAMN20088305

The individuals were mainly sequenced by capture approach. The genotypes of homozygotes were confirmed after a whole genome sequencing.

Table S7 : Genotypes at the S-locus of the offspring selected for the reconstitution of haplotypes and the crosses for the study of phenotypic traits (grey).

Identity	Allele 1	Allele 2	Pollen donor	Stigmate	SRA
d32	Ah03	Ah12	Mor_19_2	Mor_19_19	SAMN20844104
d33	Ah01	Ah12	Mor_19_3	Mor_19_4	SAMN20844105
d38	Ah04	Ah04	Mor_19_13	Mor_19_14	SAMN20844106
d42	Ah20	Ah12	Mor_19_23	*Mor_19_24*	SAMN20844107
d72	Ah20	Ah36	*Mor_19_37*	Mor_19_38	SAMN20844108
d74	Ah25	Ah12	Mor_19_41	Mor_19_42	SAMN20844109
d76	Ah03	Ah36	Mor_19_46	Mor_19_45	SAMN20844110
d89	Ah03	Ah05	Mor_19_56	Mor_19_55	SAMN20844111
d217	Ah12	Ah59	Mor_19_19	Mor_19_37	SAMN20844112
d250	Ah12	Ah20	Mor_19_24	Mor_19_51	SAMN20844113
d265	Ah36	Ah24	Mor_19_53	Mor_19_54	SAMN20844114
d3.1	Ah24	Ah25	Niv_19_4	Niv_19_5	SAMN20844115
d10.1	Ah03	Ah24	Niv_19_18	Niv_19_19	SAMN20844116
d12.1	Ah01	Ah20	Niv_19_22	Niv_19_23	SAMN20844117
d30.1	Ah02	Ah01	Niv_19_58	Niv_19_59	SAMN20844118
d48.1	Ah20	Ah25	Niv_19_52	*Niv_19_5*	SAMN20844119
d50.1	Ah12	Ah04	Niv_19_8	Niv_19_60	SAMN20844120
d51.1	Ah12	Ah04	Niv_19_54	*Niv_19_7*	SAMN20844121
d191.1	Ah04	Ah04	Niv_19_3	Niv_19_31	SAMN20844122
d206.1	Ah22	Ah25	Niv_19_42	*Niv_19_5*	SAMN20844123
d205.1	Ah20	Ah04	Niv_19_53	Niv_19_9	SAMN20844124
d208.1	**Ah03**	**Ah02**	***Niv_19_58***	***Niv_19_18***	**SAMN20844125**
d17.1	**Ah03**	**Ah01**	**Niv_19_33**	**Niv_19_32***	**SAMN20844126**

d24.1	**Ah20**	**Ah01**	**Niv_19_47***	**Niv_19_46***	**SAMN20844127**
d29.1	**Ah12**	**Ah01**	**Niv_19_57***	**Niv_19_56***	**SAMN20844128**
d50.2	**Ah04**	**Ah04**	***Niv_19_8***	***Niv_19_60***	**SAMN20844129**
d266.1	Ah04	Ah12	Niv_19_7	Niv_19_45	SAMN20844130
d122	Ah01	Ah63	Pin_16_1	Pin_8_2	SAMN20844131
d127	Ah01	Ah03	Pin_5_1	Pin_4_54	SAMN20844132
d173	Ah03	Ah18	Pin_5_2	Pin_16_3	SAMN20844133
d176	Ah03	Ah42	Pin_15_1	Pin_9_1	SAMN20844134
d177	Ah01	Ah29	Pin_16_1	Pin_5_12	SAMN20844135
d239	Ah29	Ah01	Pin_4_24	Pin_8_15	SAMN20844136
d113	Ah18	Ah31	Tss_3_23	Tss_22_7	SAMN20844137
d273	Ah03	Ah01	Tss_23_2	Tss_22_24	SAMN20844138
d275	Ah63	Ah31	*Tss_23_2*	Tss_14_3	SAMN20844139
d284	Ah01	Ah31	Tss_21_10	Tss_5_1	SAMN20844140
d285	Ah01	Ah31	*Tss_21_10*	Tss_3_10	SAMN20844141
d118	Ah03	Ah01	Ind_9_3	Ind_8_1	SAMN20844142
d163	Ah03	Ah01	*Ind_8_1*	Ind_1_1	SAMN20844143
d166	Ah03	Ah01	Ind_6_1	Ind_15_1	SAMN20844144
d170	Ah01	Ah63	*Ind_9_3*	Ind_15_2	SAMN20844145
d232	Ah01	Ah31	Ind_18_1	Ind_10_3	SAMN20844146

*The bold offspring were not used to reconstitute haplotype of parents. * parents are not sequenced by capture. The haplotypes suppressed from the dataset because already phased with another offspring are represented in italic. The homozygous genome of d38 was confirmed by genome wide sequencing. The homozygous genomes of d191 and d50.1 were confirmed by the genotypes of their offsprings found by PCR after cross.*

Table S8 : Summary of the number of variable positions in each dataset in the S-flanking regions of 25kb.

Species	Population	Variable positions
A. halleri	Nivelle	2441
	Mortagne	2435
A. lyrata	North America	2360

The variable positions represented the number of positions with one or more individuals with one or more variants compared with the reference genome of A. lyrata (Hu et al., 2011).

Table S9 : Effect of phytopathogen on the phenotypic traits.

Trait	Effect observed	Median difference expected (2.5-97.5%)
Time first leave (day)	-0.02	-0.02 (-1.6-1.65)
Number of leaves	-0.43	0 (-1.09-1.05)
Rosette area (cm^2)	-5.15	0.04 (-8.48-8.6)
Mean leave length (cm)	-0.13	0 (-0.26-0.25)
St dev leave length (cm)	0.01	0 (-0.04-0.04)
Mean leave width (cm)	-0.08	0 (-0.17-0.17)
St dev leave width (cm)	-0.01	0 (-0.04-0.04)
Mean leave area (cm^2)	-0.31	0 (-0.7-0.71)
Time first flower (day)	**2.68**	**0.01 (-1.67-1.75)**
Number of flowering stems	-1.32	-0.02 (-1.52-1.53)
Maximum flowering stem length (cm)	-2.49	0 (-3.93-3.93)
Number of flowers by flowering stem	-6.91	-0.13 (-10.66-11.07)
Flowering time (day)	**-3.15**	**0.01 (-2.6-2.61)**

The effect observed represents the mean difference observed between individuals with traces of attacks or not. The median difference expected represents the median value of difference observed between individuals after 10000 permutations tests. The values in parentheses represent these same values at 2.5 and 97.5% of the distribution. The traits in bold represent the trait that presented a mean difference observed between individuals with traces of attacks or not significantly different that the difference expected by random sampled.

Table S10 : Effect of phytophagous attacks on the phenotypic traits.

Trait	Effect observed	Median difference expected (2.5-97.5%)
Time first leave (day)	-1.07	-0.05 (-1.92-2.26)
Number of leaves	-0.85	0.04 (-1.47-1.37)
Rosette area (cm^2)	-3.63	-0.06 (-10.95-11.54)
Mean leave length (cm)	-0.21	0 (-0.32-0.33)
St dev leave length (cm)	0.01	0 (-0.05-0.05)
Mean leave width (cm)	-0.22	0 (-0.22-0.22)
St dev leave width (cm)	-0.05	0 (-0.05-0.05)
Mean leave area (cm^2)	-0.66	-0.01 (-0.88-0.94)
Time first flower (day)	2.1	-0.01 (-2.17-2.48)
Number of flowering stems	-0.47	-0.06 (-2.01-2.05)
Maximum flowering stem length (cm)	4.27	0.05 (-5.39-5.26)
Number of flowers by flowering stem	-7.06	-0.65 (-13.35-16.6)
Flowering time (day)	0.07	0.02 (-3.43-3.41)

The effect observed represents the mean difference observed between individuals with traces of attacks or not. The median difference expected represents the median value of difference observed between individuals after 10000 permutations tests. The values in parentheses represent these same values at 2.5 and 97.5% of the distribution.

Table S11: Effect of oxydative stress on the phenotypic traits.

Trait	Effect observed	Median difference expected (2.5-97.5%)
Time first leave (day)	-0.54	-0.08 (-1.91-2.25)
Number of leaves	**-3.89**	**0.01 (-1.46-1.43)**
Rosette area (cm²)	**-27.85**	**-0.22 (-10.99-11.53)**
Mean leave length (cm)	**-0.97**	**0 (-0.32-0.34)**
St dev leave length (cm)	-0.04	0 (-0.05-0.06)
Mean leave width (cm)	**-0.71**	**0 (-0.22-0.23)**
St dev leave width (cm)	-0.04	0 (-0.05-0.05)
Mean leave area (cm²)	**-2.48**	**-0.01 (-0.9-0.97)**
Time first flower (day)	2.03	-0.04 (-2.27-2.51)
Number of flowering stems	-0.22	0 (-2.11-2.11)
Maximum flowering stem length (cm)	3.53	-0.01 (-5.5-5.34)
Number of flowers by flowering stem	5.99	-0.69 (-13.77-16.88)
Flowering time (day)	2.57	0.02 (-3.53-3.63)

The effect observed represents the mean difference observed between individuals with traces of attacks or not. The median difference expected represents the median value of difference observed between individuals after 10000 permutations tests. The values in parentheses represent these same values at 2.5 and 97.5% of the distribution. The traits in bold represent the trait that presented a mean difference observed between individuals with traces of attacks or not significantly different that the difference expected by random sampled.

Table S12: Effect of family for each S-allele on the phenotypic traits.

Allele	Trait	Effect observed	Median difference expected (2.5-97.5%)
Ah01	Time first leave (day)	2.04	-0.11 (-3.34-4.19)
	Number of leaves	-1.55	-0.07 (-2.49-2.35)
	Rosette area (cm²)	-7.09	0.01 (-16.18-16.71)
	Mean leave length (cm)	-0.22	0 (-0.51-0.53)
	St dev leave length (cm)	-0.07	0 (-0.08-0.07)
	Mean leave width (cm)	-0.04	0.01 (-0.37-0.37)
	St dev leave width (cm)	-0.04	0 (-0.09-0.08)
	Mean leave area (cm²)	-0.36	0.01 (-1.32-1.36)
	Time first flower (day)	-1.65	-0.14 (-4.18-4.41)
	Number of flowering stems	-2.64	-0.08 (-3.15-3.16)
	Maximum flowering stem length (cm)	4.88	-0.1 (-7.02-7.36)
	Number of flowers by flowering stem	4.91	-0.35 (-16.91-18.75)
	Flowering time (day)	**8.76**	**-0.17 (-7.27-7.91)**
Ah03	Time first leave (day)	-3.43	0 (-6-4.71)
	Number of leaves	0.33	-0.1 (-3.74-3.76)
	Rosette area (cm²)	15.21	0.37 (-22.85-19.04)
	Mean leave length (cm)	0.26	0 (-0.67-0.64)
	St dev leave length (cm)	0.06	0.01 (-0.14-0.11)
	Mean leave width (cm)	0.25	0 (-0.52-0.5)
	St dev leave width (cm)	0.03	0 (-0.11-0.11)
	Mean leave area (cm²)	1.07	0.05 (-1.92-1.78)
	Time first flower (day)	5.02	0.47 (-7.12-5.88)
	Number of flowering stems	-1.72	0.06 (-3.5-3.39)
	Maximum flowering stem length (cm)	-2.99	0.08 (-11.06-11.08)

	Number of flowers by flowering stem	-1.1	2.22 (-41.97-31.86)
	Flowering time (day)	2.89	-0.22 (-5.33-6.89)
Ah04	Time first leave (day)	1.46	-0.06 (-1.57-1.73)
	Number of leaves	0.59	-0.01 (-1.52-1.4)
	Rosette area (cm²)	9.57	-0.14 (-13.08-13.6)
	Mean leave length (cm)	0.21	0 (-0.38-0.37)
	St dev leave length (cm)	0.01	0 (-0.06-0.06)
	Mean leave width (cm)	-0.1	0 (-0.24-0.25)
	St dev leave width (cm)	-0.05	0 (-0.05-0.06)
	Mean leave area (cm²)	0.11	-0.01 (-1.06-1.08)
	Time first flower (day)	**-2.4**	**-0.07 (-2.02-2.11)**
	Number of flowering stems	**3.15**	**-0.01 (-2.08-2.12)**
	Maximum flowering stem length (cm)	-2.38	0.01 (-4.38-4.26)
	Number of flowers by flowering stem	**-26.91**	**-0.92 (-12.11-15.57)**
	Flowering time (day)	-2.11	0.02 (-3.36-3.18)

The effect observed represents the mean difference observed between individuals of the two families with homozygotes of the S-allele specified. The median difference expected represents the median value of difference observed between individuals after 10000 permutations tests. The values in parentheses represent these same values at 2.5 and 97.5% of the distribution. The traits in bold represent the trait that presented a mean difference observed between individuals of each family significantly different that the difference expected by random sampled.

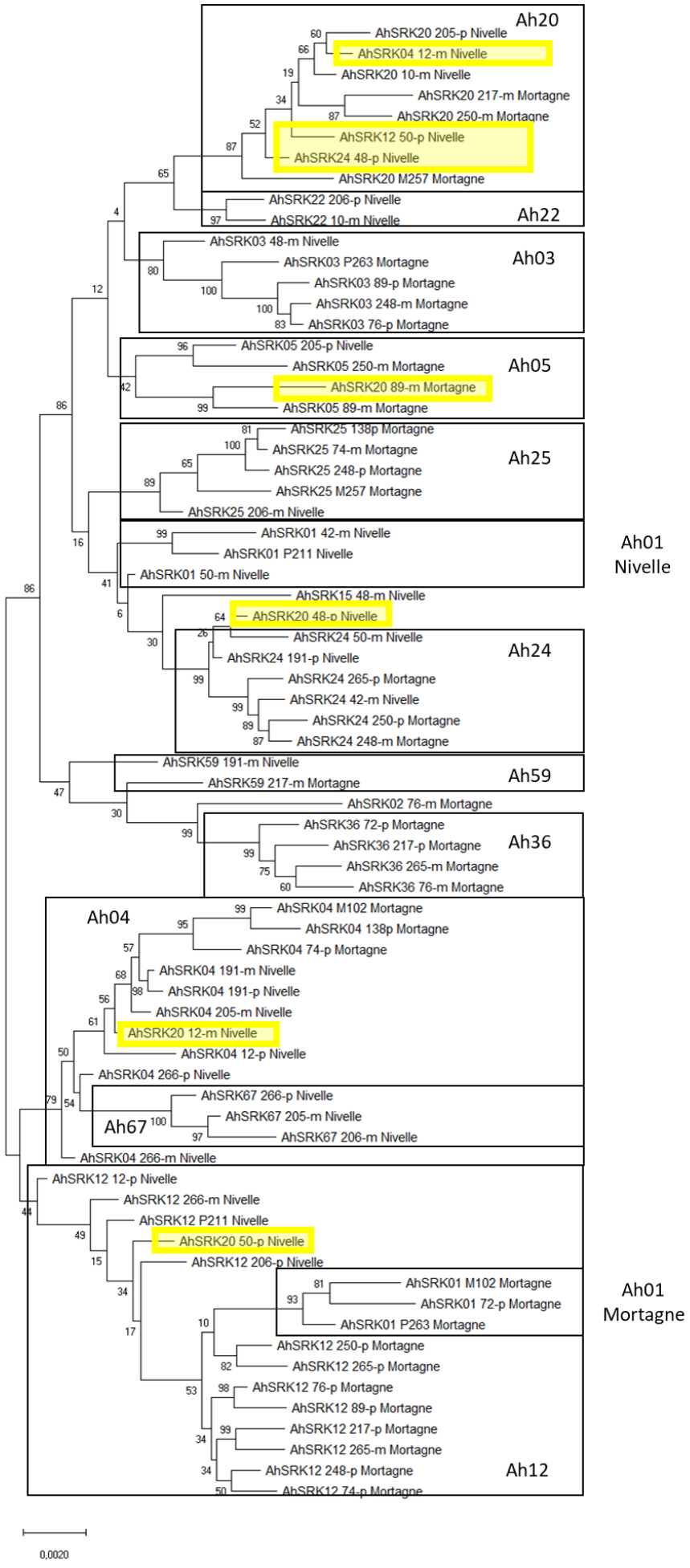


Figure S1 : Phylogenetic tree obtained by Maximum Likelihood for haplotypes of A. halleri of the Nivelle and Mortagne populations across the first 25kb flanking the S-locus. *The Tamura-Nei model was used and the percentage of trees in which the associated haplotypes clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The black boxes show the distribution of haplotypes by populations while the yellow boxes show the exceptions.*

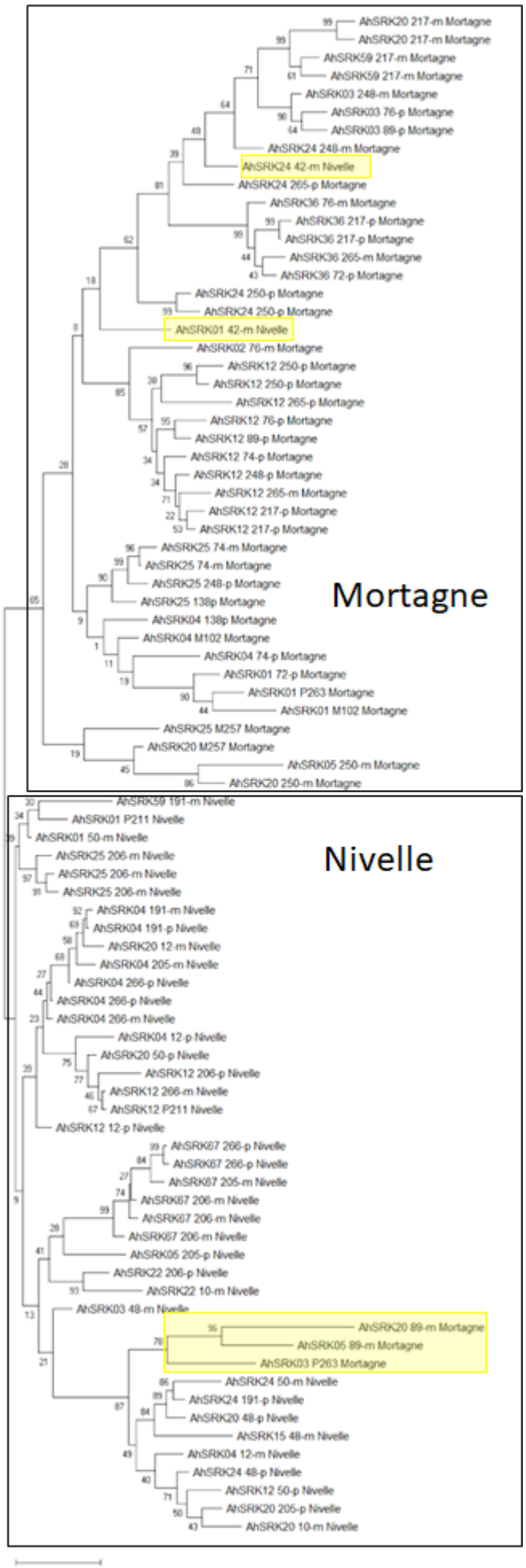


Figure S2 : Phylogenetic tree obtained by Maximum Likelihood for haplotypes of A. halleri of the Nivelle and Mortagne populations between 25 and 50kb flanking the S-locus. *The Tamura-Nei model was used and the percentage of trees in which the associated haplotypes clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The black boxes show the distribution of haplotypes by populations while the yellow boxes show the exceptions.*

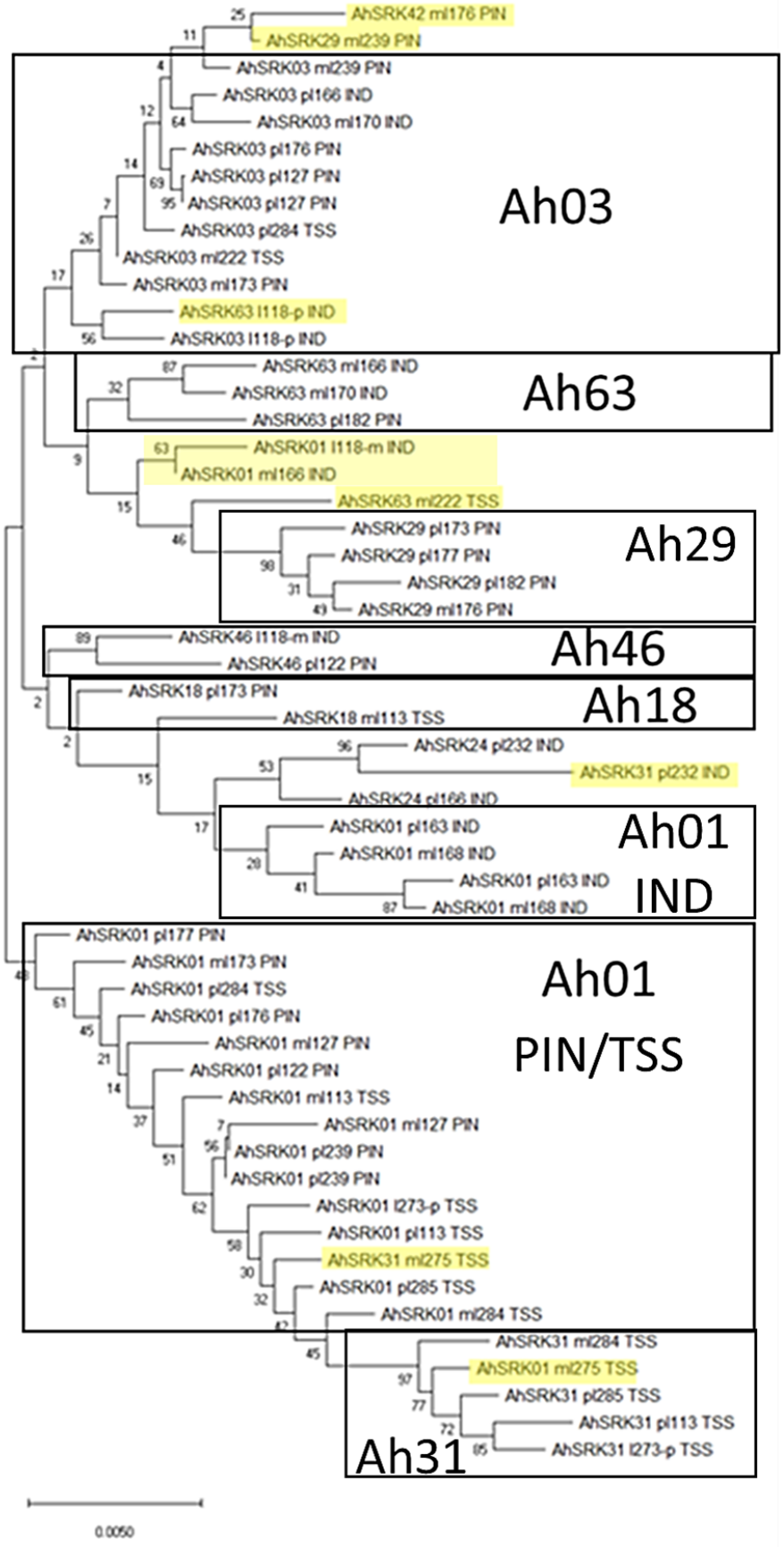


Figure S3 : Phylogenetic tree obtained by Maximum Likelihood for haplotypes of A. lyrata of North America across the first 5kb flanking the S-locus .*The Tamura-Nei model was used and the percentage of trees in which the associated haplotypes clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The black boxes show the distribution of haplotypes by populations while the yellow boxes show the exceptions*

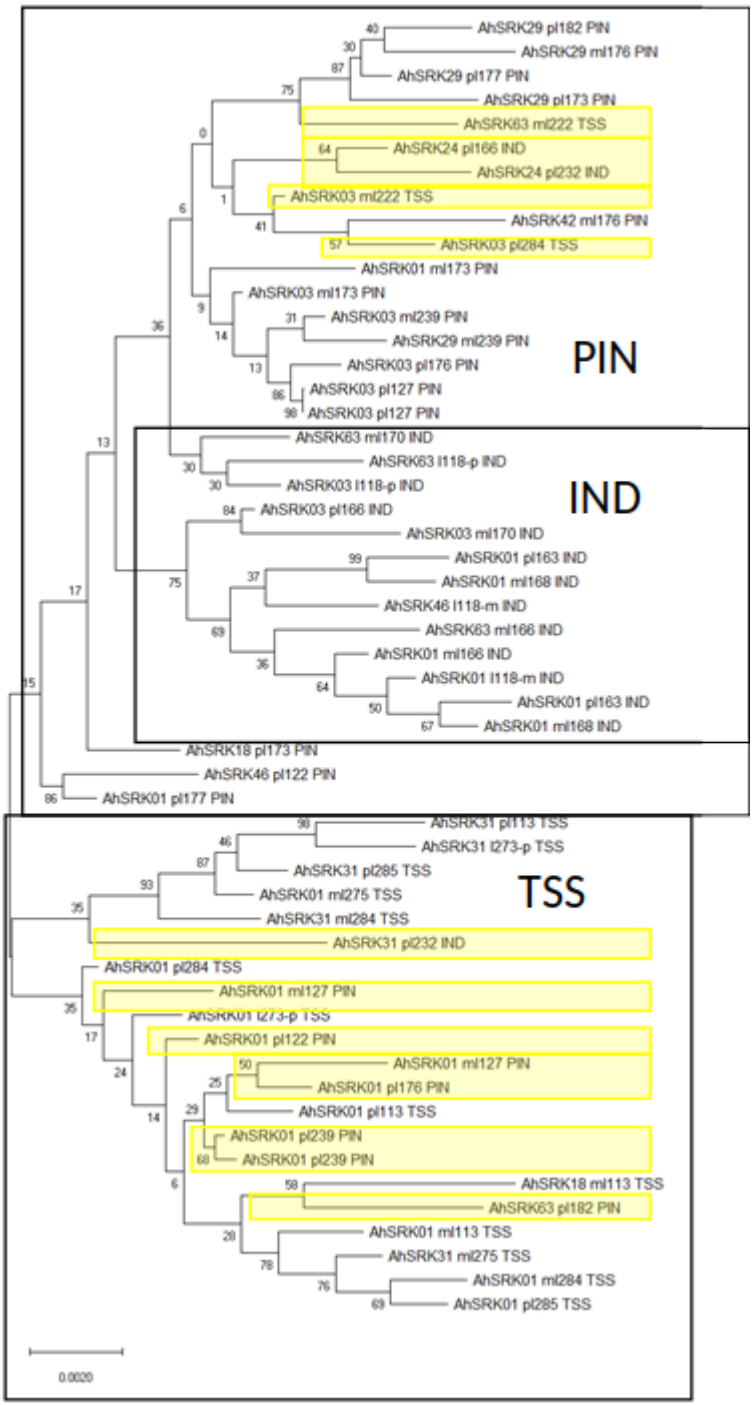
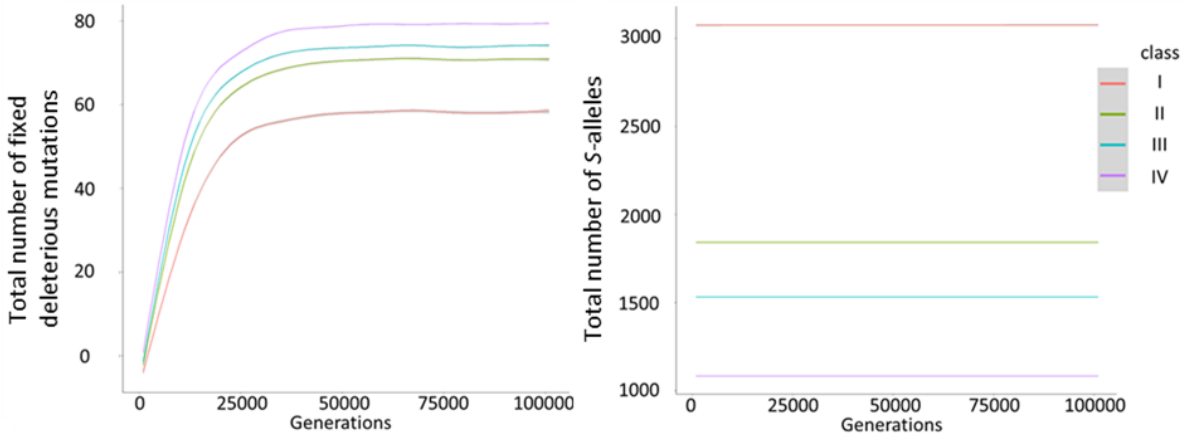


Figure S4 : Phylogenetic tree obtained by Maximum Likelihood for haplotypes of A. lyrata of North America between 5 and 10kb flanking the S-locus. *The Tamura-Nei model was used and the percentage of trees in which the associated haplotypes clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The black boxes show the distribution of haplotypes by populations while the yellow boxes show the exceptions*

Figure S5 : Evolution of the number of fixed deleterious mutations (left) and evolution of the number of S-alleles (right) with time in simulated S-flanking regions of 100 positions. *Each line represents the value obtained for each class of S-allele: red = class I, green = class II, blue = class III, purple= class. The confidence intervals are represented in grey.*



Chapter III

Scientific questions:

The existence of complex dominance relationships among alleles at the *S*-locus of Brassicaceae is a key feature that influences allele frequencies and patterns of mate availability within natural populations. However, the overall architecture of such a network of dominance interactions, its genetic/molecular determination, and the processes leading to its evolution at the *S*-locus remain largely under-documented.

In relation to these issues, the final focus of the thesis is to address the following questions:

- What is the overall shape of the network of dominance interactions in *Arabidopsis halleri*, and how is it determined by a limited set of sRNA precursor genes and sRNA targets?
- What are the observed impacts of the dominance interactions allowing dominance between *S*-alleles on the conservation of sRNAs and? Target sequences?
- What are the implications of a networked system on the expected evolution of dominance between *S*-alleles?
- Is the evolution of new dominance interactions equally likely at different levels of the network?
- Is dominance evolution driven mostly by mutations in targets or sRNAs?
- Is dominance evolution irreversible?
- What are the expected impacts of the genetic load linked to *S*-locus on the expected evolution of dominance between *S*-alleles?

Contributions:

In order to address these questions, I analysed phenotypic data on dominance between additional *S*-alleles that were collected by controlled crosses over several years by Chloé Ponitzki and Eléonore Durand, and I analysed additional full S-locus sequences produced in the team (from BAC clones) in order to obtain new *SRK*, *SCR* and sRNA sequences, and estimate the levels of sequence conservation. Note that the PCRs allowing identification of the S-locus genotypes of each crossed individual were performed by Anne-Catherine Holl.

I also developed a stochastic model, written in python, simulating the evolution of a mutation allowing to create or destroy an interaction between two *S*-alleles. The results of these models were compared and illustrated by examples of molecular interactions found for different *S*-alleles of *A. halleri*.

This chapter is written in English, as a draft article.

Empirical and theoretical investigations of the structure and dynamics of the network of dominance interactions among self-incompatibility alleles in *Arabidopsis halleri.*

Le Veve Audrey[1], Holl Anne-Catherine[1], Ponitzki Chloé[1], Durand Eleonore[1], Castric Vincent[1], Vekemans Xavier[1]

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

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

Abstract

Various forms of balancing selection can promote the evolution of dominance modifiers, especially when they are strictly linked to the genes they control. Previous theoretical studies demonstrated that evolution of dominance in a sporophytic self-incompatibility system by modifiers should lead to a process of ascending dominance because recessivity induces the expression of deleterious mutations linked to the *S*-locus in homozygous individuals. However, these models assumed a simplistic genetic basis for the dominance modifiers considered as single independent genetic entities linked to *S*-alleles. The first biological evidence that *bona fide* dominance modifiers indeed exist was provided by the identification of small non-coding RNAs (sRNA) that are linked to the sporophytic self incompatibility locus in *Brassicaceae,* where they regulate in heterozygous individuals the relative transcript level of the two alleles of the gene determining the pollen SI recognition phenotype (*SCR*). These dominance modifiers function as molecular interactions between a small RNA produced by the dominant allele and its target sequence on the recessive allele, rather than as independent genetic entities that would determine on their own whether an allele is dominant or recessive. How this peculiar genetic architecture constraints the evolution of these dominance modifiers has not been studied. Here, we combined phenotypic, genomic and theoretical approaches to address this question. First, we extended the phenotypic characterization of dominance in pollen to a nearly comprehensive network of dominance among 11 *S*-alleles *of A. halleri.* We then explored the molecular processes by which the molecular network has evolved through modifications of the sRNA regulators and/or their target sites. Finally, we performed stochastic simulations to compare the strength of natural selection on mutations creating new regulatory interactions according to 1) their level of pleiotropy, measured as the number of alleles involved in the new interaction, 2) their molecular nature (on the sRNA or on the target), 3) the initial level of dominance of the *S*-allele onto which they occur (initially dominant vs. initially recessive *S*-alleles). Finally, we evaluated the impact of the genetic sheltered load linked to the *S*-locus on the fate of these mutations. Overall, our results show that details of two-component genetic architecture of the dominance modifiers has important consequences for how they evolve.

Keywords: genetic dominance, sRNA, network, sheltered load, S-locus.

Introduction

Genetic dominance is a basic property of genetic systems in diploid species, and describes the phenotypic expression of only one the two alleles at a diploid locus in heterozygous individuals. The causes of genetic dominance have been debated in the evolutionary genetics community since the 1920's, as exemplified by the heated debate between Ronald A. Fisher et Sewall Wright (reviewed in Bagheri. H.C, 2006). According to Fisher (1928), dominance interactions between alleles can result from the intervention of genetic elements, which he called « dominance modifiers ». Without them, heterozygotes at a given diallelic locus exhibit an intermediate phenotype between the two homozygotes. In the presence of a dominance modifier conferring dominance to the wild type allele over a deleterious allele, heterozygotes demonstrate an equivalent fitness as the wild type (Fisher, 1928, 1941). Wright pointed out that dominance modifiers are active in heterozygotes only, which generally have low population frequencies for deleterious mutations. Hence, he claimed that the intensity of natural selection should generally not be sufficient to select for « dominance modifiers ». Haldane (1930) and Wright (1934) proposed instead that dominance may be explained by simple biochemical properties such as enzymatic activity. However, Wright and Haldane conceded to Fisher that special cases such as loci under balancing selection, which cause high levels of heterozygosity in natural populations, could still allow natural selection to promote the evolution of dominance modifiers. Balancing selection is defined as selective processes that maintain allelic diversity at a locus by reducing the rate of fixation of any allele, and it promotes long-term polymorphisms and the maintenance of high levels of heterozygosity. More recently, Otto et Bourguet (1999) demonstrated that selection of dominance modifiers is possible in case of overdominance, a well known form of balancing selection, mostly if the dominance modifier is highly genetically linked to the selected locus. However, these putative dominance modifiers postulated by Fisher in 1928 have mostly remained elusive. Recently, the first example of such dominance modifiers have been documented, and are involved in the control of dominance relationships between self-incompatibility alleles in pollen of the Brassicaceae (Tarutani et al. 2010, Billiard & Castric, 2011).

Self-incompatibility systems in plants are common genetic mechanisms based on the recognition and rejection of self-pollen which prevent self-fertilisation and promote outcrossing and heterozygosity in hermaphrodite plant species (Nettancourt, 2001). In Brassicaceae, the self-incompatibility system is called sporophytic because the incompatibility phenotype of the pollen is determined by the diploid genotype of the pollen parent. The sporophytic self-incompatibility system (SSI) of Brassicaceae is controlled by a single genomic region, the *S*-locus (Boggs et al., 2009), composed of two linked genes: *SCR* (*S*-locus cysteine-rich, named *SP11* in Brassica) and *SRK* (*S*-locus receptor kinase), encoding pollen and pistil proteins, respectively (Schopfer et al., 1999; Goubet et al., 2012). Pollination between partners carrying the same haplotype at the *S*-locus (e.g. in case of self-pollination) promotes the rejection of pollen (Ma et al., 2016). This system promotes heterozygosity at the *S*-locus and maintains high allelic diversity in the population because it advantages rare alleles (Wright 1939, Castric & Vekemans, 2004). Complex dominance relationships between *S*-alleles were demonstrated by controlled crosses in different Brassicaceae species (Bateman, 1952 ; Kowyama et al., 1994; Schoen & Busch, 2009), and more recently in *A. halleri* (Llaurens et al., 2008 ; Durand et al., 2014).

Two theoretical studies developed determinist evolutionary models and demonstrated that the strength of selection in SSI systems is sufficient to promote the evolution of « dominance modifiers » in strong linkage disequilibrium with the *S*-locus (Llaurens et al., 2009b ; Schoen & Bush 2009). Mutations that are either increasing or decreasing dominance of an *S*-allele relative to other

codominant alleles, without changing the allelic specificity (i.e., acting as dominance modifiers *sensu* Fisher 1928), will have the effect of increasing the number of compatible mates of heterozygous individuals (and thus their reproductive fitness, Vekemans et al., 1998), a phenomenon that Llaurens et al. (2009b) called the "hiding effect". These studies demonstrated that in SSI systems, mutations either increasing or decreasing dominance of an *S*-allele relative to other extant alleles are generally favoured in populations with codominant alleles, whereas mutations leading to codominance are disfavored when the population contains dominant or recessive alleles (Llaurens et al., 2009b ; Schoen and Busch., 2009). The main consequence of the occurrence of such dominance relationships at the *S*-locus is that the equilibrium frequencies differ substantially between recessive and dominant *S*-alleles, with the former reaching higher frequencies, a phenomenon known as the "recessive effect" (Bateman, 1952; Sampson, 1974). Schoen & Busch (2009) also showed theoretically that the hiding effect selecting for dominance is expected to be stronger for the pollen SI phenotype than for the pistil phenotype, and found that this was in line with the observation that dominance relationships are more frequently reported in pollen than in pistils.

The existence of a genetic sheltered load linked to *S*-alleles is expected to modify the perspective of evolution of dominance relationship between *S*-allele, essentially because *S*-alleles that become recessive can form homozygous combinations and hence express their sheltered load, whereas *S*-alleles that become dominant are prevented to express their sheltered load (Llaurens et al., 2009b). We previously demonstrated that the *S*-locus promotes a detectable accumulation of deleterious mutations in the flanking regions (Le Veve et al., chapter I). Llaurens et al (2009b) predicted that the genetic sheltered load promoted an evolution of dominance in an "ascending scale". However, Le Veve et al (Chapter II) revealed that recessive *S*-alleles could quantitatively accumulate a more important genetic sheltered load linked to *S*-locus than the dominant *S*-alleles. The predictions of models developed in Llaurens et al (2009b) have to be revised to estimate the effect of the genetic load linked to *S*-locus on the evolution of the dominance network, considering different structures of genetic load in *S*-alleles in the different dominance classes.

Recent studies showed that the molecular bases of dominance among *S*-alleles for pollen specificities are caused by different families of sRNAs acting as dominance modifiers. In *B. rapa*, *S*-alleles are divided in two dominance classes that determine the pollen phenotype : alleles of class I are all dominant over alleles of class II, the alleles in class I are codominant to each other, and the alleles in class II form a linear dominance hierarchy (Hatakeyama et al., 1998). However, all alleles are codominant for the stigma incompatibility phenotype. Different functional studies, conducted in *Brassica rapa*, have established that the dominance phenotype for pollen specificity is explained by two loci producing a RNA precursor with a hairpin structure, which is then processed into sRNAs (named *Smi1* and *Smi2*, for *SP11 methylation inducer 1 and 2*). Their sRNAs were genetically linked to class I and class II *S*-alleles respectively (Hatakeyama et al., 1998, Kakizaki et al., 2003; Tarutani et al., 2010, Yasuda et al., 2016). These sRNAs target the promoter of the recessive *SCR* allele and repress its transcription through the methylation of the promoter (Kusaba et al., 2002; Fujimoto et al., 2006; Shiba et al., 2006). In *A. halleri*, the dominance network in pollen is mostly linear with a dominance hierarchy among six *SCR* alleles and only one observed case of codominance (Durand et al., 2014). The regulatory network of this dominance hierarchy involves at least 8 families of sRNAs producing loci linked to different *S*-alleles and 21 sRNA-*SCR* target interactions identified (Durand et al., 2014). The molecular sRNA/*SCR* interaction involves a certain level of sequence complementarity between the sRNA and the target. Burghgraeve et al. (2020) showed that the sRNA/*SCR* interaction in *A. halleri* follows a threshold model, whereby sequence complementarity beyond a certain level (about three

mismatches over the 21 to 24 nucleotides of the sRNA molecule and its target) leads to an effective transcriptional silencing of recessive *SCR* alleles, whereas a complementarity below this threshold does not. This suggests that even point mutations on the sRNA, or in the target, can modify (either create or disrupt) the regulatory interaction by changing the sequence complementarity relative to the threshold.

The sRNA families involved in the complex dominance hierarchy observed among six *A. halleri SCR* alleles seem to have appeared successively and independently over the course of evolution (Durand et al., 2014). Some sRNA families are shared by many phylogenetically distant *S*-alleles (e.g. *mirS3* distributed between *S*-alleles in the four dominance classes), suggesting an ancient origin. *mirS3* is involved in the regulation of several recessive *S*-alleles and can be considered a "generalist" sRNA. The sequence of the *mirS3* precursor and its target sites varies slightly among the different *S*-alleles that carry them, but the consequences of these variations have never been investigated in detail. The interaction between a sRNA and its targets in plants usually causes selective constraints that create particular conservation profiles of both actors and controls the maintenance of particular sRNA-target interactions (Fahlgren et al., 2010; Ma et al., 2010), but these profiles have never been investigated in the case of sRNAs controlling the dominance networks in the SSI system. In contrast, other sRNAs families, like *mirS4* or *mir867*, are shared by only a few closely related *S*-allele lineages, suggesting a recent origin. These observations raise questions about the molecular mechanisms generating new sRNA families, or new *SCR* targets. The main hypothesis for the appearance of new miRNA genes, to which these sRNA precursors resemble, is the partial reverse duplication of the (future) target gene (Nozawa et al., 2012). It is expected that the recent reverse duplication of the target gene leaves a recognizable genomic signature because it will extend beyond the precise site of interaction with the sRNA (Allen et al., 2004; Zhang et al., 2011). The fine-scale study of ancient and recent sRNAs such as *mirS3* and *mir867* could provide insights on how the interaction network acquires new regulatory elements and how the interactions they establish evolve over time once they are in place.

Previous models for the evolution of dominance in the SSI system assumed that *S*-alleles could move in the dominance hierarchy independently of the other *S*-alleles (Llaurens et al., 2009b ; Schoen & Bush 2009), but molecular deciphering of the molecular nature of the dominance modifiers now makes it clear that the position of an *S*-allele in the dominance hierarchy rather involves a whole network of interactions with the other *S*-alleles. To properly understand how the position of an *S*-allele in the dominance hierarchy can evolve, it is thus necessary to take into account explicitly this network of molecular interactions. For example, because they can have different consequences on which interactions are disrupted or created, it is possible that the fate of mutations on the sRNA differs from that of mutations on its target site(s). Moreover, because the intensity of negative frequency-dependent selection differs between dominant and recessive *S*-alleles, it is also possible that the probability of fixation of mutations creating or disrupting dominance interactions differs according to whether they occur on *S*-alleles that are initially already high or low in the dominance hierarchy. This potential implication of the initial dominance state of *S*-alleles on the evolution of the dominance network was not considered in the previous theoretical studies (Llaurens et al., 2009a; Schoen & Bush 2009). Finally, previous evolutionary models demonstrated also that the sheltered load linked to *S*-alleles was an important determinant of the evolutionary dynamics of dominant modifiers. Because the sheltered load can be more readily expressed in recessive S-alleles that can form homozygous combinations, we hypothesise that the load may favour the fixation of mutations creating new sRNA regulators (making the *S*-allele become dominant towards more alleles) as compared to mutations creating new targets (making the *S*-allele recessive towards more alleles).

In this study, we performed theoretical and empirical investigations on the evolution of the dominance network of an SSI system. For the empirical part, we first extended the phenotypic data on the architecture of the dominance network to a nearly full array of 11 *S*-alleles in*Arabidopsis halleri*, and distributed among the four dominance classes. Then we used sequence data over the whole *S*-locus region of these 11 haplotypes to investigate the relationship between the occurrence of different sRNA precursor families and their potential targets sites in the *SCR* region, with the position of the different haplotypes in the phenotypic dominance network. To evaluate the consequences of the functional diversification on nucleotide sequences associated with the sRNA/*SCR* target interactions, we studied the patterns of nucleotide sequence conservation along the deeply conserved *mirS3* sRNA precursors and their putative targets on the *SCR* gene. Finally, we investigated the molecular scenarios of origin of the recently evolved *mir867* and *mirS4* sRNA precursor families and their putative target sites. Finally, we used stochastic simulations to investigate the evolution of a simplified linear dominance network under the two-components genetic architecture for the dominance modifiers. We studied the impact of (1) the initial dominance relationships between *S*-alleles, (2) the position of the mutations that creates the new interaction (on the modifier vs on the target), and (3) the existence of a sheltered load linked to *S*-alleles on the probability of fixation of each new type of interaction.

Results

Architecture of the dominance network in A. halleri

Previous studies characterised the full array of dominance relationships in pollen between six *S*-alleles in *A. halleri* (Llaurens et al,. 2008; Durand et al,. 2014). Durand et al (2014) also investigated the interaction network of sRNAs and *SCR* targets that is involved in these patterns of dominance. They identified eigth families of sRNA precursors distributed among these six *S*-alleles and 21 interactions with *SCR* targets that collectively explained 93% of the dominance relationships determined phenotypically. More recently, the full array of dominance relationships has been extended to a total of 11 *S*-alleles (Fig. 1A) by crossing each of the 55 heterozygous genotypic combinations to both of its respective "tester" lines (i.e., lines that express a single *S*-allele), and the additional *S*-alleles have been sequenced for searching potential *SCR* targets (Durand & Castric, personal communication).

Overall, the phenotypic analysis of dominance demonstrated dominance relationships in 51 pairwise combinations of *S*-alleles (92.7%), and codominance in only four cases (7.27%, Fig. 1A, 1B). We observed one special case of apparent self-compatibility (homozygous plants for the *S*-allele Ah20 produced siliques spontaneously). The observed phenotype could be explained by a particular dominance interaction but also by a loss of functional SSI. Overall, the resulting network was fully transitive, as noted by Durand et al (2014), and takes the form of a linear hierarchy with the single class I allele as the most recessive, then with increasing dominance appeared successively the two class II alleles, then the four class III alleles, and three codominant alleles of class IV at the top of the hierarchy (Ah13, Ah15 and Ah20) (Fig. 1B). One can note an exception in the linear hierarchy with the combination Ah15 and Ah29 that were codominant.

With the sequences of *SCR* of the eleven *S*-alleles and sRNA-seq data for eight of those, we aimed to identify changes in the molecular interactions that could putatively be involved in the evolution of the phenotypic dominance network. In total, with stringent alignment criteria (Smith and Waterman, 1981) and a minimum alignment threshold of 18 (Burghgraeve et al., 2020), we identified putative regulatory interactions for 13 new pairwise sRNA/*SCR* target pairs (Fig. 1C). Cumulated with the interactions predicted in Durand et al (2014), we could now predict 30 of the 53 interactions observed by phenotypic approaches (56.6%). For the remaining 23 interactions we have either no sRNA seq data yet (26.4%) or the data we have provided no candidate regulatory interaction (17%). Overall, the 30 putative interactions involved six different target sites within *SCR* (two in the first exon, two in the intron and two in the upstream region, Fig. 1C) and five different precursor families of sRNAs (*mirS2*, *mirS3*, *mirS4*, *mir1887* and m*ir867*, Fig. 1C). The predicted molecular interactions between Ah03 and Ah13 *mirS3* and one putative target site on *SCR* exon 1 (*E1a* on Fig. 1C), and between *mir4239* and the exon 2 of *SCR* (*E2* on Fig. 1C) were not congruent with the dominance phenotypically observed in pollen (Fig. 1A). They probably are false positives and we excluded them in the next target conservation analysis.

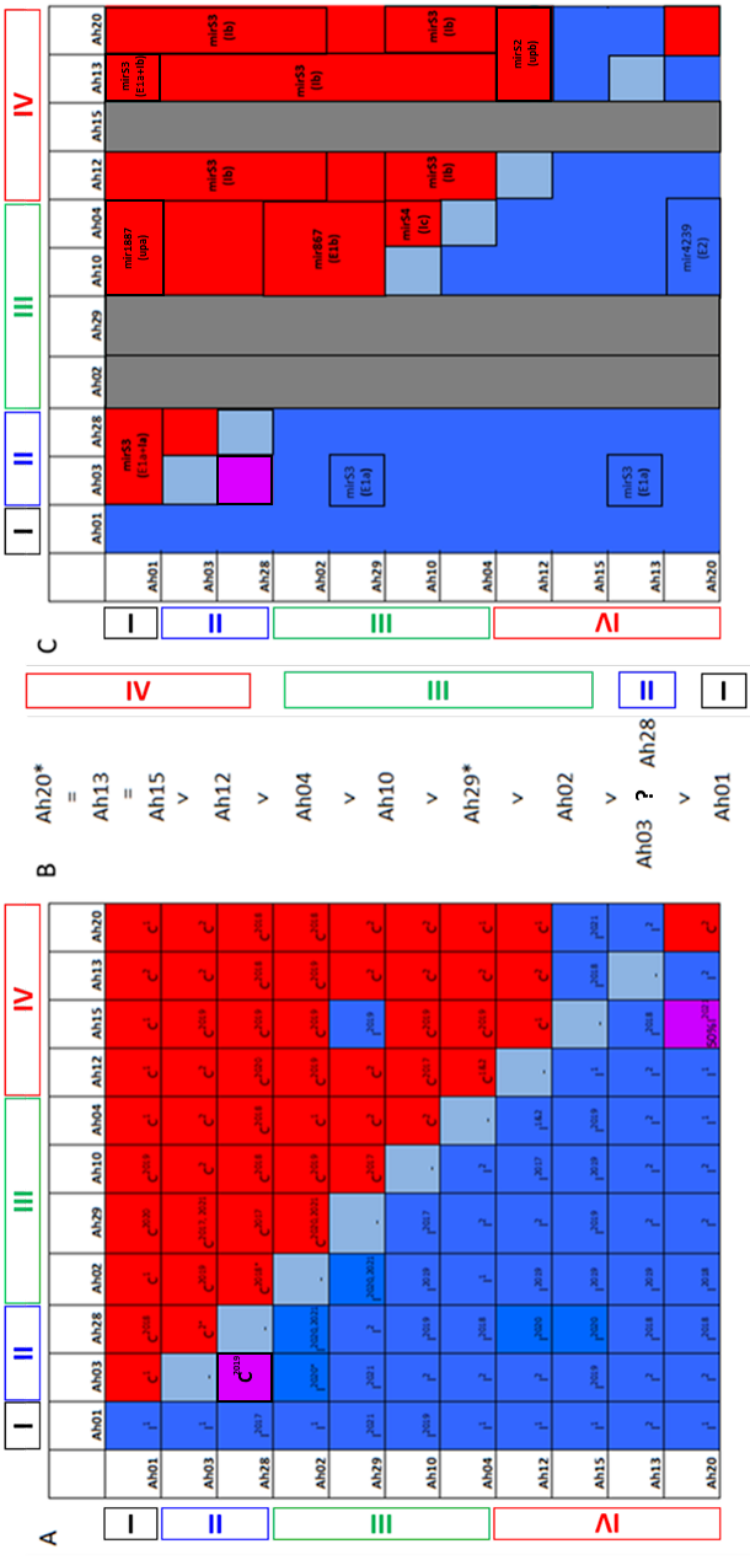


Figure 1 : Network of dominance-recessivity interactions between alleles. *(A) Dominance network obtained by controlled crosses and comparison with the molecular model. Superscripts indicate the source or the year of crosses of the data: 1, Durand et al,. 2014; 2, Llaurens et al,. 2008. We tested each heterozygous combination in the pistils of female lines with the S-allele in rows. Compatible crosses are represented in red, incompatible in blue and undetermined in purple. (B) Controlled crosses can be represented as a linear dominance hierarchy among the eleven S-alleles. ?, the relation of dominance is undetermined. *, complicated dominance hierarchy with Ah15. (C) Interactions predicted between the sRNAs produced by the S-alleles in columns with the SCR sequence of the S-allele in rows. The S-alleles without sRNA-seq data were presented in grey. Each interaction predicted is signified by the name of the precursor that produced the sRNAs followed by the name of the target localisation on SCR in parenthesis. E2=exon 2 first target, E1a=exon 1 first target, E1b=exon 1 second target, Ia=intron first target, Ib=intron second target, Ic=intron last target,upa= first target in upstream SCR, upb= second target in upstream SCR. interaction in bold represents the interactions congruent with dominance interaction observed by phenotypic approach. The dominance classes of each allele were reported (I, II, III, and IV) (Prigoda et al,. 2015).*

By combining phenotypically-determined dominance relationships with the occurrence of putative sRNA/*SCR* target interactions and the phylogenetic relationships among *S*-alleles, we investigated scenarios of evolution of the network of dominance among *S*-alleles (Fig. 2A). We illustrated the potential times of appearance of the different sRNA families and their *SCR* targets, based on the reconstruction of Durand et al. (2014). In this reconstruction, *mirS3* would be ancestral to all *S*-alleles, but a specific interaction would have evolved between the *mirS3* sequence from the ancestor of dominant *S*-alleles of class IV (Ah12, Ah13 and Ah20) and a novel target site in the intron of the ancestor of recessive alleles of class I to III (region Ib, except Ah29; Fig. 2B). The emergence of this interaction would have created two levels of dominance: one containing the ancestors of the *S*-alleles of class IV in a dominant group, and the other containing the ancestors of all other *S*-alleles in a recessive group. Within this recessive group, interactions between *S*-alleles of the intermediate class II (here represented by alleles Ah03 and Ah28) and the more recessive allele Ah01 involved the same *mirS3* family produced by Ah03 and Ah28 but a separate region of the intron of the most recessive *S*-allele Ah01 (region Ia). Hence, a single sRNA family, *mirS3*, determines dominance of class IV alleles over classes I to III alleles, and dominance of class II alleles over class I alleles. Surprisingly, dominance of *S*-alleles of class III (Ah04 and Ah10) over the most recessive *S*-allele Ah01 (class I) involves a distinct sRNA family, *mir1887*, targeting a third specific region of *SCR,* this time about upstream of the first exon rather than in the intron (region "up_a", Fig. 2B). In contrast, within allelic classes, the linear hierarchy among alleles seems to be associated with more specialised interactions: within class III, the dominance of Ah04 and Ah10 over Ah02 and Ah29, and the dominance of Ah04 over Ah10, involves class III-specific sRNA families, *mir867* and *mirS4*, respectively, and their specific *SCR* targets, E1b and Ic, respectively; and within class IV, the dominance of Ah13 and Ah20 over Ah12 involves the class IV-specific *mirS2* family (Fig. 2).

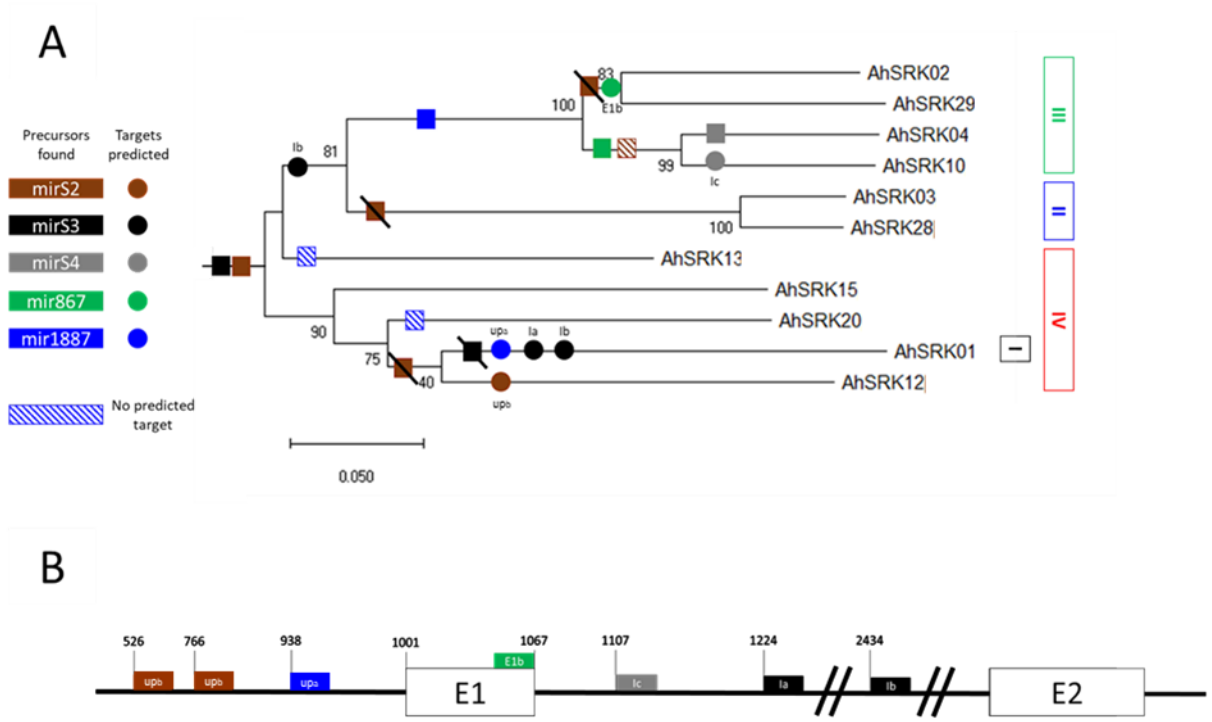


Figure 2 : Repertoire of sRNA genes and their targets involved in the six interactions predicted along the phylogeny of S-alleles congruent with phenotypic dominance observed. *A) The evolutionary history was inferred by using the Maximum Likelihood method and Tamura-Nei model (Tamura and Nei, 1993). The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 11 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. There were a total of 1625 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (Kumar et al,. 2018). Phylogenetic classes are reported (I, II, III, and IV) (Prigoda et al,. 2015). The scenario of gain and loss for sRNA precursors and their targets is represented on the branch where each event is inferred in Durand et al (2014). Circles indicate predicted targets. Ia represents the first target on the intron of SCR, Ib represented the second target on the intron of SCR, E1b represents the second target on the exon, E1c represents the last target on the exon, up_a represents the first target on the region of upstream of SCR, up_b represents the second targets on the region of upstream of SCR. Squares indicate precursors that are inferred by Durand et al,. (2014) to have been present in the ancestral repertoire . The hatched squares indicate precursors with no target detected on the eleven SCR alleles studied. B) Representation of the relative position of the different targets along a schematic SCR sequence (not drawn to scale).*

Patterns of sequence conservation of an ancient sRNA precursor and its putative SCR targets

The sRNAs produced by the *mirS3* precursor family explain the dominance of class IV alleles, by targeting region Ib of *SCR* of all the other more recessive alleles (classes I to III, except Ah29; Fig. 1C, Fig.2A). This suggests that the interactions established by *mirS3* represent substantial functional constraints. We analysed patterns of sequence conservation in the *mirS3* precursor gene and its *SCR*

target in region Ib (Fig. 3a). Our analysis included sequences of *mirS3* from the four class IV alleles described above (Ah12, Ah13, Ah15, Ah20), as well as three additional alleles of class IV (Ah32, Ah36, Ah43) for which the entire sequence of the *S*-locus region was available but the dominance phenotype has not been verified experimentally. We observed the highest levels of conservation either in the parts of the hairpin precursor that produce the sRNAs that are predicted to target the recessive alleles in region Ib of *SCR* (black box in Fig. 3b), or in their corresponding complementary sequence on the other side of the foldback (referred to as mir* for canonical miRNAs, red box in Fig. 3b). In comparison, nucleotide identity between the *mirS3* sequences of the class IV alleles and the *mirS3* sequences of the other classes was much lower in the portion encoding the sRNAs with predicted targets. Between two class IV *S*-alleles, the sequence identity in this region varies from 73 to 96% (median=77%), whereas identity in the same region varies from 4 to 84% when comparing sequences of *mirS3* from a class IV and a more recessive *S*-allele (median=62%). The difference between these median values (=15%) was highly significant (permutation test with 10,000 iterations, $p=3^{e-4}$).

For *SCR* target sequences, we analysed sequence identity among the classes I, II and III alleles described above, as well as an additional class III allele (Ah25) for which the full *SCR* sequence was also available (including the two exons and the intron). The average sequence identity over the entire region was low (49%), but we observed four relatively more conserved portions among the recessive *S*-alleles : a region in the first exon , two regions in the intron and a region in the second exon (Fig. 4). The region predicted to host the *mirS3* target (Ib) for recessive *S*-alleles was contained in the first conserved region in the intron (Fig. S1b). Hence, the *mirS3* target is one of the few conserved motifs along the otherwise highly diverged sequences of the *SCR* gene. The particular allele Ah29 of class III, was not predicted to be targeted by *mirS3* linked to *S*-allele of class IV and presents a long insertion in this particular region (Fig. S1B).

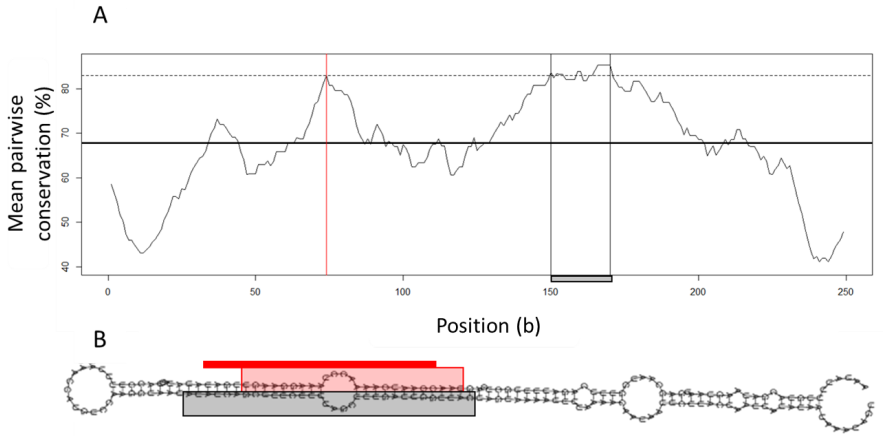


Figure 3 : Sequence conservation among mirS3 precursors of dominant S-alleles of class IV. *A) Mean pairwise sequence identity between windows of precursors of mirS3 linked to seven dominant class IV S-alleles. The window size (26nt) was chosen as the size of the sRNA region predicted to target the recessive S-alleles in region Ib of SCR. The median value across all windows is indicated by a solid horizontal line and the 95% percentile by a dashed horizontal line. The red vertical line indicates a first region of 26nt with high conservation. The black vertical lines indicate the second region with a high conservation. B) Structure of the precursor of mirS3 linked to the dominant S-allele Ah13 predicted by RNAfold (Minor Free Energy model). The red and black scares represent the conserved regions found in A. The red line represents the region that produces sRNA targeting the SCR of the other S-alleles.*

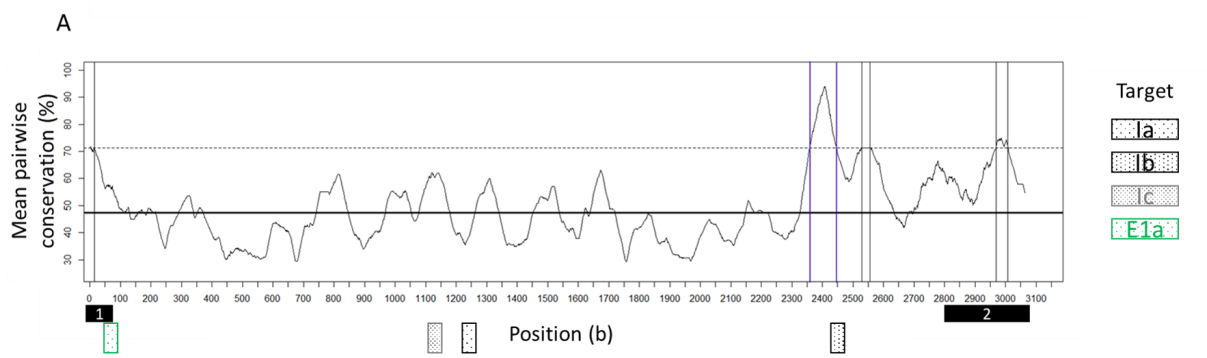


Figure 4 : Sequence conservation among recessive SCR alleles (classes I to III). *The window size (26nt) was chosen as the size of the sRNA region predicted to target the recessive S-alleles in region Ib of SCR. The median value across all windows is represented by the solid horizontal line and the 95% percentile by the dashed horizontal line. The blue vertical lines indicate the conserved portion that overlaps with the region predicted to be targeted in recessive S-alleles (target Ib, cf Fig. 2B). The black vertical lines indicate the three other conserved regions. A schematic representation of the SCR gene and its predicted targets are shown below the Figure, except the targets on upstream regions.*

Evolution of a new sRNA precursor by inverted duplication of its target: an example with mir867

Within allelic classes, the linear dominance hierarchy among *S*-alleles seems to be associated with more specialised interactions involving the recruitment of new sRNA precursors in the network. We took advantage of the occurrence of two recent sRNA families, *mir867* and *mirS4* (Fig. 2a), to identify signatures of the evolutionary scenarios by which they emerged. The *mir867* precursor is carried by alleles Ah04 and Ah10, and its predicted targets are located within the first exon of *SCR* of the other class III alleles (Ah02, Ah25 and Ah29). We compared the sequences of the extended *SCR* region of all class III *S*-alleles beyond the target site with sequences of the precursors of *mir867* of alleles Ah04 and Ah10 (also extended by 100bp). We observed strong sequence similarity between *mir867* from Ah10 and *SCR* from Ah29. Specifically, the nucleotide sequence just upstream and downstream from the *mir867* precursor of Ah10 is highly similar with a 126bp region of *SCR29* encompassing the first exon as well as 15bp of the promoter and 45bp of the intron (Fig. 5a and b). We found similar results between the *SCR* of Ah02 and Ah25 with the *mir867* precursor linked to Ah10 or between the *SCR* of Ah02, Ah25 and Ah29 with the *mir867* precursor linked to Ah04. Overall, these observations suggest that *mir867* arose on the ancestor of Ah10 and Ah04 as the result of a duplication of a portion of the *SCR* gene of the ancestor of Ah02, Ah25 and Ah29, followed by an inversion giving rise to a hairpin structure with strong sequence similarity to these *SCR* alleles. This relatively simple series of molecular events provide an elegant mutational mechanism by which new dominance interactions can emerge.

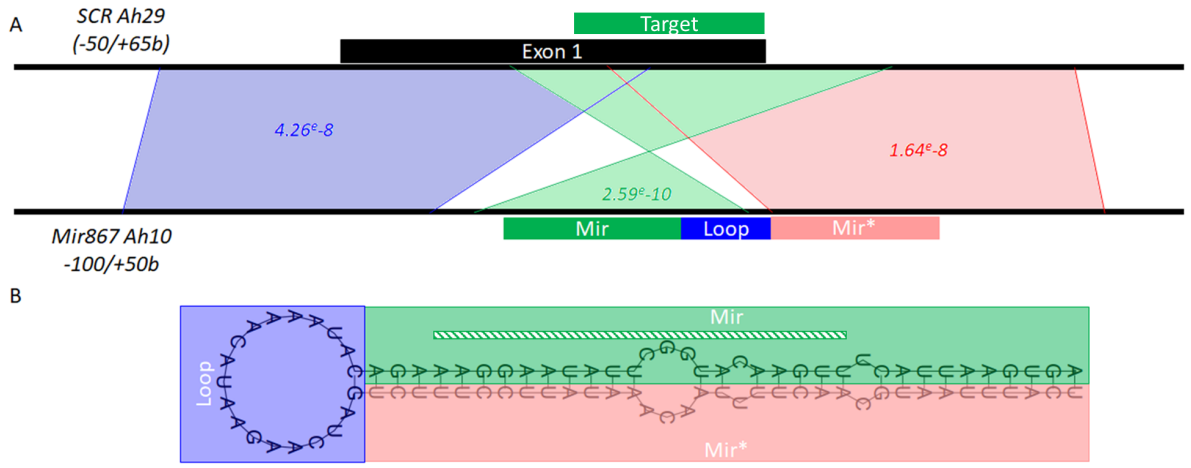


Figure 5 : Similarity between the SCR gene of the Ah29 S-allele (-50/+65b) and the mir867 precursor linked to Ah10 (-100/+50b). *A) The SCR gene is represented on top and the mir867 precursor on the bottom. The E-value for each colour-coded segment is shown (YASS alignment; Noé and Kucherov, 2005). B) Structure of the Ah10mir867 precursor predicted by RNAfold. The 21nt-long mir produced by the hairpin precursor with targets on SCR29 is represented by a hatched horizontal bar.*

Numerical simulations to investigate the fate of mutations conferring generalist versus specialist dominance interactions

Empirical analyses of the architecture of the dominance network had shown that some interaction implicated sRNAs and targets distributed among different *S*-alleles and could be considered as generalist interaction (e.g. *mirS3* from class IV alleles that target class I to III alleles, Fig.1). However, some other interactions implicated apparently a limited number of *S*-alleles, and were thus specialist interactions (e.g. *mirS4* of Ah04 that targets Ah10 only). Thus, the dominance network observed in *A. halleri* can be explained by the presence of both generalist and specialist interactions, although Durand et al. (2014) showed that generalist interactions are the most important in explaining this network. What distinguishes a generalist interaction from a specialist is ultimately only the number of alleles involved. However, previous models did not take into account that alleles evolve in ways that are not independent of each other. Thus, the selection for a mutation allowing the establishment of an interaction as a function of the number of alleles involved has never been estimated. Moreover, previous models do not predict expectations on the evolution of alleles involved, or not, in an interaction. Hence, we investigated the fate of mutations conferring generalist versus specialist dominance interactions, and also compared those occurring within sRNA sequences versus within *SCR* targets.

We then used stochastic simulations to study the consequences of the various aspects of the two-component genetic architecture for the dominance modifiers on the evolution of a simplified linear dominance network. We simulated a population with five co-occurring *S*-alleles that were initially co-dominant in both pistil and pollen and we compared the fate of various types of mutations creating or modifying dominance interactions relative to that of neutral mutations as a control. First, we observed that regardless of their properties, the probability of fixation of mutations generating dominance interactions between *S*-alleles was always higher than that of neutral mutations (Fig. 6), confirming the efficient selection of the dominance modifiers that we introduced. Second, our empirical analysis above showed that some sRNAs are highly specialist (regulating just a single or a

few *S*-alleles such as e.g. *mir867* and *mirS4*) whereas others are more generalist (e.g. *mirS3* from class IV alleles regulating all other extent *S*-alleles). Thus, we introduced mutations generating novel dominance interactions in pollen, but that differed with respect to the numbers of different *S*-alleles that they regulate (from one for specialist mutations to four for generalist mutations. We observed that the probability of fixation increased strongly with the level of generality of the mutations (Fig. 6). Hence, more pleiotropic mutations (creating regulatory interactions with a larger number of *S*-alleles) were favoured more strongly by natural selection.

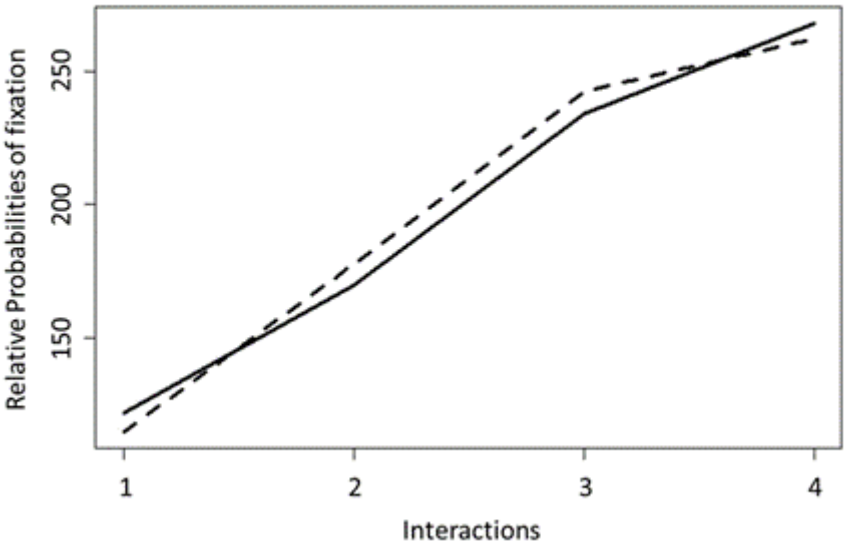


Figure 6 : Fixation probability of a new mutation as a function of the number of predicted regulatory interactions. *Relative probabilities of fixation of one mutation on modifier (solid line) or on target (dashed line) compared with the probability of fixation of neutral mutation linked to one S-allele with the number of S-allele implicated in the interaction.*

Third, we observed that in the simple model with no sheltered load, the probability of fixation of mutations occurring on sRNA modifiers or on their *SCR* targets were similar (Fig. 6). However, when introducing a sheltered load, the fate of mutations on the sRNA differed strongly from that of mutations on the targets (Fig 7). Indeed, mutations causing an *S*-allele to become more dominant (e.g. by creating a new sRNA) were favoured more strongly than mutations causing an *S*-allele to become more recessive (e.g. by creating a new target), possibly as a result of the possibility for recessive *S*-alleles to form homozygous combinations and thus express their sheltered load. Hence, the existence of a sheltered load is expected to create a bias towards the recruitment of new sRNA regulators rather than new targets.

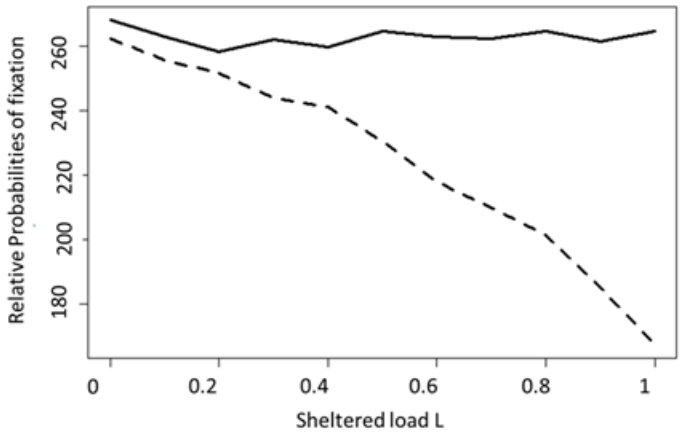


Figure 7 : Variation of the relative probabilities of fixation of mutations that change dominance hierarchy with genetic load linked to the S-locus (L) compared with neutral mutations. *Each plot represents the relative probabilities of fixation of each mutation with L. The probabilities to fix a mutation on a modifier are represented by solid lines. The probabilities to fix a mutation on target are represented by dashed lines.*

Our empirical analysis above also showed that the rare observed co-dominance interactions are restricted to pairs of *S*-alleles that are highly dominant overall. Thus, we asked whether the intensity of selection on elements of the dominance modifiers (either the sRNAs or the targets) was identical when introduced in linkage to *S*-alleles at different levels of the dominance hierarchy. We observed a higher probability of fixation for mutations creating novel regulatory interactions between *S*-alleles in the intermediate dominance class than in the highly dominant class of *S*-alleles (Fig. 7). Hence, natural selection to resolve pairwise co-dominance interactions is more intense between *S*-alleles that are low in the dominance hierarchy than between *S*-alleles that are already high in the dominance hierarchy. If we introduced a genetic load linked to *S*-alleles, we observed similar results (Table S2).

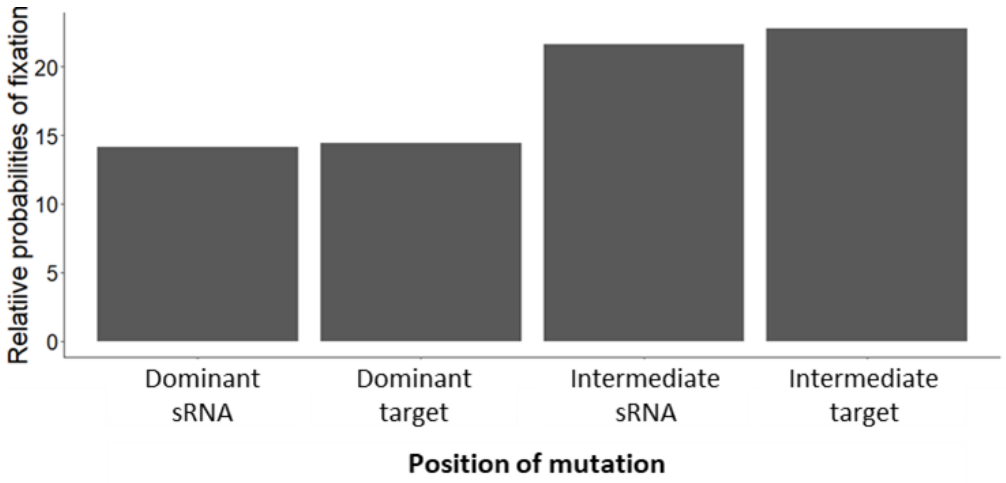


Figure 9 : Variation of the relative probabilities of fixation of mutations that change dominance hierarchy compared with neutral mutation.

Finally, we investigate the fate of mutations suppressing dominance interactions, i.e. suppressing an existing interaction. Such mutations are supposed to lead to codominance, and as such are supposedly disfavored in a sporophytic SI (Laurens et al., 2009b), but this has not been tested in models with dominance associated to two interactors (a modifier and a target). Our results confirmed that the fixation probability of all such mutations was lower than 0.01%. We supposed that the loss of an interaction could still be advantageous in case codominance could reduce the expression of the genetic load of the initially recessive *S*-allele. Hence, we reiterated the simulations with a strong genetic load linked to all alleles at the *S*-locus (lethal mutations, L=1). However, even so, the probabilities of fixation of loss of dominance relationship remained lower than 0.01%.

Discussion

The SSI system of the Brassicaceae offers the first example of dominance modifiers, and they take the form of molecular interactions between sRNAs and their target sites (Tarutani et al., 2010). In comparison with Brassica, where the network of dominance interactions presents a limited number of levels in the hierarchy and involves only two sRNAs (Yasuda et al., 2016), the number of dominance levels in *A. halleri* is very high and involves up to 8 families of sRNAs (Llaurens et al., 2008; Durand et al., 2014). Hence, the latter system offers an excellent model to investigate the evolution of a complex network of dominance interactions. We addressed this question through a combination of numerical simulations, experimental determination of the architecture of the dominance network, and an empirical study of the occurrence of genetic modifiers and their putative targets in a set of *S*-alleles.

Previous studies demonstrated that in SSI, the mutations either increasing or decreasing the dominance of an *S*-allele relative to other extant alleles are generally favoured in populations with codominant alleles, whereas mutations leading to codominance are disfavored when the population contains already dominant alleles (Llaurens et al., 2009b ; Schoen and Busch., 2009). However, the discovery that sRNAs/target interactions control the dominance network between *S*-alleles demonstrated that each interaction involves at least two alleles: the (dominant) allele associated with the sRNA, and the (recessive) allele carrying the *SCR* sequence target (Durand et al., 2014 ; Tarutani et al., 2010). Therefore, to understand the evolution of the dominance network in the SSI system, we should move from the idea that the evolution of one allele can occur independently from the other alleles, as was assumed in the previous models, and integrate the actual genetic architecture of the modifiers in models of evolution of dominance in the SSI system. Here, we first explored the molecular underpinning of changes in the dominance network of *A. halleri* and detailed the molecular events by which the network has acquired a new regulatory element (*mir867*), and how an existing element (*mirS3*) has been modified in the course of evolution. Then, we compared these phenotypic and molecular observations with the predictions from a new stochastic model.

Confirmation of a strong hierarchical structure of the network of dominance interactions among S-alleles in A. halleri

Full experimental determination of the architecture of the dominance network in pollen in *A. halleri* among 11 *S*-alleles revealed an almost strictly hierarchical structure of the interactions, i.e. along a linear ladder, except for the three most dominant *S*-alleles which showed codominance among them. Hence, the results of Llaurens et al. (2008) and of Durand et al. (2014) are confirmed and extended, showing a strongly hierarchical structure with at least nine dominance levels (among 11 *S*-alleles), in contrast to that described in Brassica with only five levels (Yasuda et al. 2016). Those results also confirm the suggestion by Prigoda et al. (2004) that dominance levels of *S*-alleles in *Arabidopsis lyrata* (which is closely related and shares most of its *S*-alleles with *A. halleri*, Castric et al. 2008) are associated with their phylogenetic relationships, as the position of alleles in the dominance hierarchy in *A. halleri* was strictly associated with their class grouping defined based on the phylogeny (classes I to IV, with increasing dominance, Fig. 1 & Fig. 2). Such remarkable observation is also in agreement with the theoretical studies of Llaurens et al. (2009b) and of Schoen et al. (2009) predicting that codominance relationships in the pollen of species with SSI would be reproductively disadvantageous as compared to dominance.

A complex dominance network achieved through the combination of generalist and specialist interactions involving sRNAs and targets within SCR S-alleles

In Brassica, dominance of class I over class II alleles is due to a generalist interaction between a sRNA, *SMI*, shared by all class I alleles, and a target present in the promoter region of all class II alleles (Tarutani et al., 2010). Our analysis suggests that the patterns of dominance among the four dominance classes occurring in *A. halleri* are achieved through similar molecular mechanisms. Hence we could explain the overall dominance of class IV alleles by a generalist interaction involving *mirS3* and a target region within the *SCR* intron shared among all alleles from classes I to III. We also found that, as observed by Fahlgren et al (2010) in the general case of miRNAs across the genome, this interaction appeared to promote sequence conservation among alleles of class IV of the region of *mirS3* involved in this interaction, and sequence conservation of the targeted region within the intron of *SCR* from class I to III alleles. This evolutionary constraint was also repercuted on the mir*. Similarly, dominance of alleles of class II on the single allele of class I involved a generalist interaction between *mirS3* from class II alleles and a different target in the first exon of *SCR* of class I. Dominance of class III alleles on the single allele of class I was mediated by a generalist interaction involving a distinct mir family, i.e. *mir1887*, and a target in the region upstream of the allele of class I. Up to now, the mechanism of dominance of alleles of class III over class II alleles remains unresolved. Overall, the evolution of additional classes of dominance seems to evolve through either the production of new mir families with new target sites on *SCR*, or through the divergence of mir sequences among classes (i.e. between *mirS3* of class IV and class II) and evolution of additional corresponding target sites on *SCR* of the most recessive class. Our numerical simulations have shown that evolution of new generalist interactions are more strongly favoured than new specialist interactions, or said differently, if they create a new dominance class. Thus we expect to observe more generalist than specialist interactions in the dominance network of the SSI system. This is in agreement with the observations of Durand et al (2014). In addition, the correspondence between phylogenetic relationships and dominance (Prigoda et al. 2005) suggests that these generalist interactions are transmitted vertically during the process of allelic diversification, where the new *S*-alleles in a given class inherit the sRNA precursors and target sites from their parental allele, i.e. inheriting their position within the dominance network. Then purifying selection acting on sequences of the mir region and of the target sites seems to maintain the functional interactions, despite the fact that target sites within *SCR* are located on non-coding regions of the gene. Because of the long timeframe of diversification of *S*-alleles (e.g. class I and class IV alleles are known to be shared between the tribes *Camelinae* and *Cardaminae* that diverged about 20 MY years ago; Gan et al., 2016), conservation of sRNA/target interaction should indeed involve very strong forces of purifying selection. An alternative scenario would involve evolution within the *S*-locus of one allele of a new sRNA family targeting a pre-existing target site shared by different alleles, and horizontal transfer of the sRNA precursor towards the other alleles of the same class. Moreover, the presence of the *mir1887* in distant phylogenetic alleles also suggested an enrichment of the network by horizontal transfer, like described in Durand et al (2014). However, it would then be difficult to explain the general association of dominance with phylogenetic relationships.

Although the generalist interactions described above explain reasonably well the observed patterns of dominance among allelic classes, they fail to explain the almost strictly hierarchical structure of the dominance network observed in *A. halleri*, i.e. they cannot account for dominance interactions within allele classes. Within the intermediate class III, we found two specific interactions between *S*-alleles, mediated by additional sRNA families (*mir867* and *mirS4*), and additional specific *SCR* targets. A

similar interaction involving a specific sRNA (*mirS2*) and a specific target was observed among some alleles within class IV, however codominance seemed also to be rather abundant in this class. Evolution of these specific interactions within allele classes could follow the same processes as between classes, i.e. be associated with recruitment of a new sRNA family, and a new target site, and vertical transmission of these among phylogenetically close alleles. Different molecular scenarios have been proposed in the literature to explain the generation of a new sRNA and its target (Carthew and Sontheimer., 2009), including the inverted duplication of a future target, the duplication of another sRNA, the spontaneous accumulation of mutations that create a stem-loop structure or the recruitment from transposable element-derived sequences. In this study, in the case of the most recent sRNA family identified at the *S*-locus, i.e. *mir867*, our analysis highlighted a strong genetic signature of a possible reverse duplication at the target *SCR* sequence, suggesting a scenario involving the inverted duplication of a future target.

Altogether, our analysis suggests that the overall structure of the dominance network is explained by a combination of generalist and specialist interactions between several sRNA families and a number of targets within the *SCR* gene. Our understanding of the molecular processes governing the overall dominance network is however still fragmentary, as only about half of the phenotypically determined dominance relationships could be formally explained by such molecular interactions. This is in part due to the fact that sRNA-seq data for some *S*-alleles were not yet available. In particular, the absence of sRNA data for allele Ah29 prevented us from explaining the observed dominance relationship between this *S*-allele and Ah02, but the apparent absence of precursors of *mir867* and *mirS4* from the full *S*-locus sequence of Ah29 suggested the implication of other sRNA families in interactions within this class. Similarly, we supposed that the dominance relationship between the class III alleles Ah02 and Ah29 and the recessive allele Ah01 was associated with the same sRNA/target interaction as that documented in the same as the class III alleles Ah04 and Ah10, but sRNA-seq data will be required to firmly establish this interpretation. In the same way, the molecular basis of the dominance relationship between the class IV alleles Ah15 and Ah12 should be checked for consistency with those described between the *mirS2* linked to Ah13 or Ah20 and the target on Ah12. Moreover the molecular processes that allow these interactions were only partially known. The diversity of targeted regions (upstream, CDS and intron) suggest that a diversity of molecular processes may actually regulate the expression of *SCR* beyond that proposed in *Brassica rapa* (Shiba et al,. 2006; Tarutani et al,. 2010). Indeed, while targets in the upstream region of *SCR* seem consistent with promoter methylation of the recessive allele in the heterozygote, this is not the case for targeted regions in the intron or CDS. For example, one may suspect that inhibition of some of the recessive *SCRs* may also occur by degradation of the targeted mRNA, which is the more classical mode of action of miRNAs. Similarly, the fact that the observed dominance relationships between class II and III alleles could not be associated to any targeting of class II *SCRs* by class III sRNAs may suggest that additional elusive molecular processes other than sRNAs may participate in the negative regulation.

A central question for the future will be to properly understand the relative importance of generalist vs specialist sRNA regulators. While the more generalist regulators tend to be favoured in the long term, it is hard to think of mutational mechanisms that would create them from scratch. Rather, it is likely that newly emerged regulators are initially regulating only pairs of *S*-alleles (such as *mirS4* and *mir867*), and that they later gain generalism as the *S*-alleles within the recessive class diversify. The fact that the intermediate dominance classes in *A. halleri* (II and III) have a relatively more recent history of allelic diversification than the most dominant class IV argues in favour of a more rapid rate

of emergence of new *S*-alleles in the recessive than in the more dominant classes. Explicitly testing this idea would require developing models that jointly consider these two processes (evolution of dominance and allelic diversification) that have so far been considered separately in the literature. Because of the complexity of both processes, this may be challenging to set up. Whether these models are able to explain why the most recessive class (I) has apparently not been able to diversify similarly will be an interesting test.

Occurrence of codominant relationships in the most dominant allelic class

In Brassica, codominance among *S*-alleles in the pollen is assumed to be the rule in the most dominant allelic class (Class II, Kakizaki et al., 2003). Our phenotypic analysis in *A. halleri* has shown that among alleles of the most dominant class (class IV), the three most dominant alleles are codominant to each other, while such codominant interactions are absent elsewhere in the dominance network (with one possible exception, see next section). These observations are in line with the theoretical prediction that the equilibrium frequencies of *S*-alleles are lower in the dominant than in the intermediate or recessive classes (the "recessive effect", Schierup et al., 1997; Billiard et al. 2006). This probably explains why in our simulations the probability of fixation of a new interaction was higher within the intermediate class of dominance than in the most dominant class. Indeed, heterozygotes carrying two most dominant *S*-alleles were expected to occur at lower frequencies than heterozygotes between alleles of the intermediate dominance class, so that the « hiding effect » due to the new dominance interaction (Llaurens et al., 2009b) would be weaker in the most dominant class than in the intermediate class. In short, this frequency effect creates a selective advantage for new dominance interactions evolving within intermediate rather than within the most dominant class, in line with the initial argument made by Wright (1929) in the different context of deleterious (rather than balanced) variants. These expectations are also congruent with the observation of a strictly linear hierarchy between *S*-alleles in the more recessive classes in *B. Rapa* (Kakizaki et al., 2003) and *A. halleri* (Durand et al., 2014 , and results of this study).

The sheltered genetic load promotes the evolution of the dominance network by mutations on the modifiers rather than on the sRNA targets

When a *S*-locus-linked genetic load was added to the model, it disfavored the mutations on the target because the genetic load was expressed in homozygous genotypes only, and mutations that create dominance relationships were eliminated as a by-product of the purge of newly recessive *S*-alleles. In contrast, mutations in the modifier were favoured. This suggested that, with genetic load linked to the *S*-locus, the mutations that create the interaction in the dominance network should appear on the modifier (Durand et al., 2014). A mutation on modifiers that create a new interaction in SSI system corresponds molecularly to a formation of new sRNA by inverted-duplication of an ancestral *SCR* sequence, by horizontal transfer or by a mutation on ancestral sRNA which increases the spectrum of *SCR* sequences recognized. However, a new interaction is always more likely to be established in recessive than in dominant classes, even in the presence of a high genetic load linked to *S*-alleles.

A few exceptions to the strictly linear hierarchical structure of the dominance network

A notable exception to the strongly hierarchical structure of the dominance network is found for allele Ah29 belonging to class III, which shows codominance with a single allele, i.e. allele Ah15 from class IV. Dominance of class IV alleles over alleles from class III is generally achieved through interaction of *mirS3* from class IV alleles with target Ib on the intron of *SCR* of class III alleles. At the

molecular level, we observed a large specific insertion within the intron of *SCR* of Ah29, which may explain why Ah29 was not phenotypically repressed by Ah15. Our numerical simulations showed that such mutations leading to a loss of a particular dominance interaction would be very unlikely to go to fixation within a population. Thus it remains to be investigated whether this specific insertion is fixed within all copies of allele Ah29 in *A. halleri*. We also ignore how other *S*-alleles of class IV maintained their dominance interaction with allele Ah29, because we did not predict interaction between sRNAs linked to dominant *S*-alleles and the *SCR* of Ah29.

Materials and Methods

Plant material and collection of S-genotypes

We used *Arabidopsis halleri* seeds collected in four French and Italian populations (Durand et al,. 2014) to constitute a set of individuals carrying 11 *S*-alleles selected to span the sequence diversity of the *S*-locus (Schopfer et al,. 1999) including class I (Ah01), class II (Ah03/Ah28), class III (Ah04/Ah10/Ah29/Ah02) and class IV (Ah12/ Ah13/Ah15/Ah20) alleles (dominance classes referred to, respectively, as A1, B, A3 and A4 by Prigoda et al., 2005). We grew the seeds in the greenhouse and DNA was extracted, treated and purified from 15 mg of dried leaves of each sample with Chemagic beads (PerkinElmer) following Holtz et al (2016) with an additional Agencourt AMPure beads (Beckman) purification step using the manufacturer's instructions. *S*-alleles were identified using PCR primers specific for the *SRK* alleles (Llaurens et al,.2008). Controlled pollinations were performed as described in Llaurens et al (2008), by manually depositing pollen from one individual chosen as the male parent on the pistil of the chosen female partner, within one day of flower opening. Plants were separated by at least 60 cm to avoid pollen contamination.

Controlled crosses and inference of the phenotypic dominance network

We determined the dominance relationships between each pair of *S*-alleles in pollen, say Sx and Sy, by using pollen from a heterozygous individual Sx/Sy to pollinate pistils of tester lines expressing either the Sx or Sy incompatibility type. When Sy is dominant, pollen from Sx/Sy should be compatible with the Sx tester line, but incompatible with the Sy tester line, whereas rejection by both tester lines indicates that Sx and Sy are codominant. Following Llaurens et al (2008), compatibility was scored using stigmate elongation 7 days after pollination. The tester lines were checked for proper rejection of pollen expressing their corresponding *SCR* allele. As negative controls (to assess fruit elongation for incompatible crosses), we self-pollinated each tester line. As positive controls, we crossed each tester line with an individual sharing no *S*-alleles to estimate fruit elongation values for compatible crosses. Based on these measures, we defined pollinations as successful if the fruit was at least 0.55 cm long 7 days after pollination. Using 4 to 15 replicates for each cross, we defined a cross as compatible when more than 50% of the pollination replicates were successful. Combining dominance phenotypes obtained in our experiment with those from Llaurens et al (2008) and Durand et al (2014) allowed us to define the dominance phenotype for all possible heterozygous combinations of the 11 *S*-alleles studied here. This experiment was developed over four years (between 2018 and 2021).

BAC sequencing

Two BAC clones containing Ah02 and Ah25 were newly obtained, using the protocol of Goubet et al (2012). Briefly, high molecular weight DNA was prepared from young leaves of *A. halleri* individuals carrying either *S*-alleles and used to construct separate BAC libraries. Libraries were screened based on the two flanking genes, and positive clones were sequenced using a PACBIO technology (www.pacb.com). For Ah01, we used a BAC containing the orthologous sequence in the closely related *A. lyrata* (Al01, Goubet et al,. 2012). The summaries of data used are resumed in table S5.

S-alleles phylogeny based on SRK sequences

The phylogeny reconstruction of fiveteen *SRK* (exon 1 minimum) amino acid sequences studied from *Arabidopsis halleri* was performed using maximum likelihood method and Tamura-Nei model (Tamura and Nei, 1993), with 1000 replicates. The *SRK S*-domain sequences are taken from a

reference database (X. Vekemans, personal communication) compiling Genbank sequences as well as de novo assemblies obtained from raw sequence data using the NGS genotype pipeline (Genete et al. 2021). The analyses were conducted in MEGA X (Kumar et al,. 2018). Phylogenetic classes were as defined by Prigoda et al. (2005).

Identification of sRNA precursor genes

Additional precursor motifs within the *S*-locus region of all available *S*-alleles were then searched for in the two BAC clones using a similarity search based on the YASS program (Noé and Kucherov, 2005) version 1.14 starting from the 55 initially identified sRNA motifs plus Smi (Tarutani et al,. 2010) and Smi2 (Yasuda et al,. 2017) and using an e-value threshold of 10-4 . Based on this first set of hits, we then iterated the procedure to identify further motifs that might have remained undetected because of their divergence. We aligned each candidate mir sequences with the initially identified sRNA motifs with MUSCLE (Edgar, 2004) implemented in MEGA X. We conserved candidates that exhibit at least 60% sequence identity and covering at least 75% of the length of both the query and subject sequences with one or more identified sRNA motifs, like described in Durand et al, 2014.

Target site predictions

Small RNA targets were predicted in *SCR* alleles including 1kb of flanking sequence ("*SCR*+/-1kb") using psRNATarget 2017 servor (Dai et al,. 2017), with the following scoring matrix: matches = +1; mismatches = -1; gaps = -2; G:U wobbles = -0.5, like described in Durand et al, 2014 . We conserved only predictions of interaction with a score >=18 (Burghgraeve et al., 2020).

Sequence alignment, conservation and structure of sRNA precursor genes

Firstly, we aligned each *mirS3* sequence with MUSCLE implemented in MEGA X. The alignments were visualised with Jalview V2.10.5. Then, we studied the pairwise conservation of overlapping windows of 26nt between all *mirS3* copies identified. This size corresponds to the size of the region predicted to target the seven recessive *S*-alleles (Al01, Ah02, Ah03, Ah04, Ah10, Ah28, Ah25). The window overlap was of 1nt. We estimated the mean pairwise conservation between the precursors linked to the seven *S*-alleles of class IV (Ah12, Ah13, Ah15, Ah20, Ah32, Ah36, Ah43). Moreover, we identified the windows with a mean pairwise conservation greater than or equal to 95% of the distribution obtained for all the windows obtained for the seven *S*-alleles of class IV. Finally, we compared the pairwise conservation of the region of 26nt on the *mirS3* precursor predicted to target the recessive *S*-alleles obtained after comparisons between *mirS3* linked to two dominant *S*-alleles or linked to one dominant and one recessive *S*-alleles. We tested if the difference of median value obtained is significantly different from a difference obtained randomly by a permutation test with 10000 reiterations.

Secondly, we aligned each *SCR* sequence with MUSCLE implemented in MEGA X. Then, we studied the pairwise conservation of overlapping windows of 81nt between all the *SCR* genes. This size corresponds to the size of the region of the seven recessive *S*-alleles predicted to be targeted by the sRNA of *mirS3* of dominant *S*-alleles. The window overlap was of 1nt. We estimated the mean pairwise conservation between the *SCR* of the seven recessive *S*-alleles. Moreover, we identified the windows with a mean pairwise conservation greater than or equal to 95% of the distribution obtained for all the windows obtained for the recessive *S*-alleles. Finally, we compared the pairwise conservation of the region of 81nt on the *SCR* predicted to be targeted after comparisons between *SCR* of two recessive *S*-alleles or of one dominant and one recessive *S*-allele. We tested if the

difference of median value obtained is significantly different from a difference obtained randomly by a permutation test with 10000 reiterations.

Finally, we predicted the structure of the sRNA precursor genes linked to the different *S*-alleles by a Minor Free Energy model with the web server RNAfold (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi).

Research of inverted duplication of SCR genes in the sRNA precursors of the mir867

We supposed that the evolution of the recent *mir867* is explained by an inverted duplication of the first exon of *SCR* targeted in the *S*-alleles of -class III (Ah02, Ah04, Ah10, Ah25, Ah29). To test this hypothesis, we compared by Yass, the *SCR*+/-100b sequences of all *S*-alleles and the sequences of the precursors of *mir867*+/-100b linked to the *S*-alleles Ah04 and Ah10. We expected to find similarities with the *SCR* more extended than the sequences that produced sRNA. We considered only *SCR* sequences with a similarity found in forward and in reverse with an e value <0.0001 and found for the two sequences of *mir867* precursors.

Numerical simulations

We modelled evolution of dominance relationships among *S*-alleles in a sporophytic SSI system using individual centred stochastic simulations. The model assumed that dominance relationships follow the DOMCOD model of Schierup et al. (1997), i.e. *S*-alleles followed hierarchical dominance in pollen, with co-dominance allowed at any level within the hierarchy, but only codominance relationships occur in pistil. The choice of this model was based on the observation, within *Brassicaceae*, that dominance among *S*-alleles tended to be more frequent in pollen than in pistils (Schoen & Busch, 2009), and on demonstration of the implications of sRNA dominance modifiers only in pollen (Tarutani et al., 2010; Durand et al., 2014).

a) Individuals definition

In the models, each individual was defined by its genotype at the *S*-locus SiSj, the mutations that created the dominance relationship between *S*-allele expression on pollen and a L region totally linked to the *S*-locus. The specificities on pollen depended on the interactions created by new mutations and by its genotype at the *S*-locus. An individual with genotype SiSj produced specificity i on pollen if Si was dominant, j if Si was recessive, or i and j if Si and Sj were codominant. In pistil, all alleles were codominant. We assumed mutations that impact dominance in pollen were completely linked to the *S*-allele but the number of specificities on stigma remained unchanged. Hence, individuals expressing an *S*-allele Si', can mate with Si only if Si' is masked in the genotype of male by a dominant *S*-allele. The genotype SiSj with i = j allowed in the models for all *S*-alleles present in the population except the most dominant (i.e. homozygotes are excluded for the most dominant *S*-alleles).

b) Initial state of models

We simulated a panmictic population of 1000 diploid individuals with nonoverlapping generations. The population size was chosen high enough so that no alleles were lost by genetic drift, hence the number of *S*-alleles remained identical throughout the simulations. At the initial state, we used a simple model of SSI with 5 *S*-alleles. First, we considered a system in which all alleles are codominant (Table 1). Then, we studied a system with hierarchical dominance interactions similar to those observed in *A. halleri* (Llaurens et al., 2008). In this system, the *S*-alleles are distributed in three dominance classes : two alleles in the most dominant class, two alleles in the intermediate

dominance classes, and one allele in the most recessive class. The alleles in the most dominant class were codominant relative to each other and dominant over all alleles of the intermediate and the recessive classes. The alleles in the intermediate dominance class were also codominant relative to each other and dominant over the allele in the most recessive class (Table 1). Si occurring at their expected equilibrium frequencies determined according to their dominance levels following the general equations given in Billiard et al. (2006). We also assumed that no new *S*-allele could appear by mutation during simulations.

c) Characteristic of the mutant

Mutations could occur in a region strictly linked to the *S*-allele determinants, altering its dominance relationship on the pollen side with another *S*-allele. Indeed, at some levels of dominance, an Si allele could be associated with the presence (noted Si') or absence (noted Si) of a given mutation altering its dominance relationship with another allele, Sj. Initially, a mutant Si' was introduced as a single copy in the population in the first generation and its fate at the end of each simulation run was used to compute the proportion of replicate simulations with the mutation either fixed (Si' replaces Si) or eliminated (Si' goes extinct) from the population. On the appearance of a mutant *S*-allele in the population, the number of *S*-alleles increases to 6, but the number of specificities on stigma remains unchanged.

First, we introduced in one individual one mutant Si' allele whose dominance relationships with other *S*-alleles does not differ from those of ancestral allele Si in one individual. Here, we simulated the evolution of neutral mutation linked to a *S*-allele in intermediate or dominant class. These simulations were used with the two initial states previously mentioned.

Then, we reiterate the simulations for one mutant Si' allele whose dominance relationships with other *S*-alleles differ from those of ancestral allele Si. In this individual, the modification of specificities on pollen depends on a new interaction created by mutation. Hence, individuals expressing an *S*-allele Si', can mate with Si only if Si' is masked in the pollen by a dominant *S*-allele. The mutant allele Si' was assumed to share the same deleterious allele as the ancestral allele Si so that heterozygotes Si' Si expressed the same decrease in fitness due to the sheltered load as homozygotes Si Si and Si' Si'.

We investigated whether *S*-alleles could evolve along a dominance hierarchy in function of position of mutant that create the new interaction (on target or on modifier). We simulated the evolution of mutants Si' if Si' became codominant, recessive or dominant in heterozygous Si'Sj.

We then investigated the evolution of mutants if Si, Si' and Sj are initially in the intermediate or in dominant class to evaluate the impact of the initial dominance class on the evolution of a new relationship of dominance in pollen.

All mutations tested and their consequences on dominance hierarchy are resumed on Table 1. Because the interactions don't permite self-cross, the probability of selfing is null.

The frequency of the mutant *S*-allele was recorded every 10 generations and simulation stopped when the mutant was lost or fixed. We estimated the frequency spectrum of the *S*-alleles in the whole population by generations.

Initial dominance hierarchy	New interaction created	Position of mutation	Allele with mutation
	S5>(S4=S3=S2=S1)		
	S5>(S3=S2=S1)		
	S5>(S2=S1)		
	S5>(S1)	modifier	S5'
	(S5=S4=S3=S2)<S1		
	(S4=S3=S2)<S1		
	(S3=S2)<S1		
S5=S4=S3=S2=S1	(S2)<S1	target	S1'
		target	S3'
	S2>S3	modifier	S2'
		target	S5'
	S4>S5	modifier	S4'
	S2/S3=S1	target	S1'
	S2=S1	modifier	S2'
	S4/S5=S1	target	S1'
	S4=S1	modifier	S4'
	S4/S5=S3	target	S3'
(S5=S4)>(S3=S2)>S1	S5=S3	modifier	S5'

Table 1 : Summary of the mutations studied that create a new dominance relationship. *Initial dominance hierarchy : 3 levels of dominance before mutations and in ancestral alleles, []= group of codominant alleles, >= left group dominate right group. New interaction created : modification in mutant in dominance hierarchy compared with ancestral allele. Position of mutation : position of the mutation that close the interaction, target=SCR allele mutation (SNP), modifier= mutation on sRNA (SNP on previous sRNA or creation of sRNA by inversal-duplication). Allele with mutation : identity of allele with mutation.*

d) Genetic sheltered load associated with the different S-alleles

Here, we assumed a sheltered genetic load due to recessive deleterious alleles at the L region fully linked to the *S*-locus. We studied the influence of sheltered genetic load accumulated on the evolution of dominance of *S*-alleles. We further assumed that each *S*-allele can have a specific sheltered load, in other words each *S*-allele can be in linkage with specific deleterious mutations. This sheltered load was assumed to be expressed in homozygotes at the *S*-locus only and was equivalent for all *S*-alleles. We thus assumed that the sheltered load decreases the survival of homozygotes by a factor $(1 - L)$ with $0 \leq L \leq 1$.

e) Life cycle

The life cycle in our simulations had three steps:

i. Gametogenesis: 1000 adult individuals produced infinity of ovules and pollen.

ii. Syngamy: Two individuals are randomly chosen in the population. Crosses between ovules and pollen were compatible when the specificities expressed in pollen and stigmas were different. For each compatible cross, we randomly sample one *S*-allele of each parent to form a new zygote.

iii. Viability selection and regulation: We assumed that the survival probability p of a zygote depends on its genotype at the *S*-locus : for homozygous individuals p= (1-d) with $0 \leq d \leq 1$. To form the next generation, we computed p for this zygote and randomly determined whether this individual survived.

We repeated these steps until 1000 surviving individuals were obtained. Each simulation was performed with 10000 independent replicates.

Data Availability

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.17025419.

The script developed for the models is available in Github (https://github.com/leveveaudrey/evolution-of-dominance-network.git).

Acknowledgements

This work was funded by the European Research Council (NOVEL project, grant #648321), the Agence Nationale de la Recherche (TE-MoMa project, grant #ANR-18-CE02-0020-01). AL thanks the ERC and the University of Lille for funding her PhD project. We thank Sylvain Billiard for helpful discussions.

Bibliography

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.-H., Spatafora, J.W., and Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nature Genetics 36, 1282–1290.

Bagheri, H.C. (2006). Unresolved boundaries of evolutionary theory and the question of how inheritance systems evolve: 75 years of debate on the evolution of dominance. J Exp Zool B Mol Dev Evol 306, 329–359.

Bateman, A.J. (1952). Self-incompatibility systems in angiosperms: I. Theory. Heredity 6, 285–310.

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

Billiard, S., and Castric, V. (2011). Evidence for Fisher's dominance theory: how many 'special cases'? Trends in Genetics 27, 441–445.

Boggs, N.A., Dwyer, K.G., Shah, P., McCulloch, A.A., Bechsgaard, J., Schierup, M.H., Nasrallah, M.E., and Nasrallah, J.B. (2009). Expression of distinct self-incompatibility specificities in *Arabidopsis thaliana*. Genetics 182, 1313–1321.

Burghgraeve, N., Simon, S., Barral, S., Fobis-Loisy, I., Holl, A.-C., Ponitzki, C., Schmitt, E., Vekemans, X., and Castric, V. (2020). Base-pairing requirements for small rna-mediated gene silencing of recessive self-incompatibility alleles in *Arabidopsis halleri*. Genetics 215, 653–664.

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and mechanisms of miRNAs and siRNAs. Cell 136, 642–655.

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

Dai, X., Zhuang, Z., and Zhao, P.X. (2018). psRNATarget: a plant small RNA target analysis server (2017 release). Nucleic Acids Res 46, W49–W54.

Durand, E., Méheust, R., Soucaze, M., Goubet, P.M., Gallina, S., Poux, C., Fobis-Loisy, I., Guillon, E., Gaude, T., Sarazin, A., et al. (2014). Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. Science 346, 1200–1205.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792–1797.

Fahlgren, N., Jogdeo, S., Kasschau, K.D., Sullivan, C.M., Chapman, E.J., Laubinger, S., Smith, L.M., Dasenko, M., Givan, S.A., Weigel, D., et al. (2010). MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. The Plant Cell 22, 1074–1089.

Fisher, R.A. (1928). The possible modification of the response of the wild type to recurrent mutations. The American Naturalist 62, 115–126.

Fisher, R.A. (1941). Average excess and average effect of a gene substitution. Annals of Eugenics 11, 53–63.

Fujimoto, R., Okazaki, K., Fukai, E., Kusaba, M., and Nishio, T. (2006). Comparison of the genome structure of the self-incompatibility (S) Locus in interspecific pairs of S haplotypes. Genetics 173, 1157–1167.

Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R.D., Hofhuis, H., Pieper, B., Cartolano, M., Neumann, U., et al. (2016). The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. Nature Plants 2, 1–7.

Genete, M., Castric, V., and Vekemans, X. Genotyping and de novo discovery of allelic variants at the Brassicaceae self-incompatibility locus from short read sequencing data. Mol Biol Evol.

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.

Haldane, J.B.S. (1930). A note on Fisher's theory of the origin of dominance, and on a correlation between dominance and linkage. The American Naturalist 64, 87–90.

Hatakeyama, K., Watanabe, M., Takasaki, T., Ojima, K., and Hinata, K. (1998). Dominance relationships between S -alleles in self-incompatible Brassica campestris L. Heredity 80, 241–247.

Holtz, Y., Ardisson, M., Ranwez, V., Besnard, A., Leroy, P., Poux, G., Roumet, P., Viader, V., Santoni, S., and David, J. (2016). Genotyping by Sequencing Using Specific Allelic Capture to Build a High-Density Genetic Map of Durum Wheat. PLOS ONE *11*, e0154609.

Kakizaki, T., Takada, Y., Ito, A., Suzuki, G., Shiba, H., Takayama, S., Isogai, A., and Watanabe, M. (2003). Linear dominance relationship among four class-II S haplotypes in pollen is determined by the expression of SP11 in brassica self-incompatibility. Plant and Cell Physiology 44, 70–75.

Kowyama, Y., Takahasi, H., Muraoka, K., Tani, T., Hara, K., and Shiotani, I. (1994). Number, frequency & dominance relationships of S-alleles in *diploid Ipomoea trifida*. Heredity 73, 275–283.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol 35, 1547–1549.

Kusaba, M., Tung, C.-W., Nasrallah, M.E., and Nasrallah, J.B. (2002). Monoallelic expression and dominance interactions in anthers of self-incompatible *Arabidopsis lyrata*. PLANT PHYSIOLOGY 128, 17–20.

Le Veve, A., Burghgraeve, N., Genete, M., Lepers-Blassiau, C., Rasmus, N., Takou, M., De Meaux, J., Mable, B., Durand, E., Vekemans, X., Castric, V. (Chapter I). Long-term balancing selection and the genetic load linked to the self-incompatibility locus in *Arabidopsis halleri* and *A. lyrata*.

Le Veve, A., Genete, M., Lepers-Blassiau, C., Ponitzki, C., Durand, E., Castric, V., Vekemans, X. (Chapter II). The structure of the linked genetic load differs between dominant and recessive self-incompatibility alleles in*Arabidopsis halleri* and *A. lyrata*.

Llaurens, V., Billiard, S., Leducq, J.-B., Castric, V., Klein, E.K., and 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., and Billiard, S. (2009a). 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., Billiard, S., Castric, V., and Vekemans, X. (2009b). Evolution of dominance in sporophytic self-incompatibility systems: I. Genetic load and coevolution of levels of dominance in pollen and pistil. Evolution 63, 2427–2437.

Ma, Z., Coruh, C., and Axtell, M.J. (2010). Arabidopsis lyrata small RNAs: Transient MIRNA and small interfering RNA loci within the Arabidopsis Genus. The Plant Cell 22, 1090–1103.

Ma, R., Han, Z., Hu, Z., Lin, G., Gong, X., Zhang, H., Nasrallah, J.B., and Chai, J. (2016). Structural basis for specific self-incompatibility response in Brassica. Cell Research 26, 1320–1329.

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

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

Nozawa, M., Miura, S., and Nei, M. (2012). Origins and evolution of microRNA genes in plant species. Genome Biology and Evolution 4, 230–239.

Otto, S.P., and Bourguet, D. (1999). Balanced polymorphisms and the evolution of dominance. The American Naturalist, 153(6), 561-574.

Prigoda, N.L., Nassuth, A., and Mable, B.K. (2005). Phenotypic and genotypic expression of self-incompatibility haplotypes in *Arabidopsis lyrata* suggests unique origin of alleles in different dominance classes. Molecular Biology and Evolution 22, 1609–1620.

Sampson, D. R. (1974). Equilibrium frequencies of sporophytic self-incompatibility alleles. Canadian Journal of Genetics and Cytology, 16(3), 611-618.

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

Schoen, D.J., and Busch, J.W. (2009). The evolution of dominance in sporophytic self-incompatibility systems. II. Mate availability and recombination. Evolution 63, 2099–2113.

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

Shiba, H., Kakizaki, T., Iwano, M., Tarutani, Y., Watanabe, M., Isogai, A., and Takayama, S. (2006). Dominance relationships between self-incompatibility alleles controlled by DNA methylation. Nature Genetics 38, 297–299.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10, 512–526.

Tarutani, Y., Shiba, H., Iwano, M., Kakizaki, T., Suzuki, G., Watanabe, M., Isogai, A., and Takayama, S. (2010). Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. Nature 466, 983–986.

Vekemans, X., Schierup, M.H., and Christiansen, F.B. (1998). Mate availability and fecundity selection in multi-allelic self-incompatibility systems in plants. Evolution 52, 19–29.

Wright, S. (1929). Fisher's Theory of dominance. The American Naturalist 63, 274–279.

Wright, S. (1934). Physiological and evolutionary theories of dominance. The American Naturalist 68, 24–53.

Wright, S. (1939). The Distribution of Self-Sterility Alleles in Populations. Genetics 24, 538–552.

Yasuda, S., Wada, Y., Kakizaki, T., Tarutani, Y., Miura-Uno, E., Murase, K., Fujii, S., Hioki, T., Shimoda, T., Takada, Y., et al. (2017). A complex dominance hierarchy is controlled by polymorphism of small RNAs and their targets. Nature Plants 3.

Zhang, Y., Jiang, W., and Gao, L. (2011). Evolution of microRNA genes in Oryza sativa and Arabidopsis thaliana : An update of the inverted duplication model. PLoS ONE 6, e28073.

Supplementary data

Mutation implicated	L	Freq_S1	Freq_S2	Freq_S3	Freq_S4	Freq_S5	Freq_S'
intermed_target	0	0.38	0.17	0	0.12	0.12	0.21
intermed_modifier	0	0.38	0	0.21	0.12	0.12	0.17
dom_target	0	0.41	0.16	0.16	0.13	0	0.15
dom_modifier	0	0.4	0.16	0.16	0	0.15	0.13
intermed_target	0.1	0.36	0.17	0	0.13	0.13	0.21
intermed_modifier	0.1	0.36	0	0.21	0.13	0.13	0.17
dom_target	0.1	0.38	0.16	0.16	0.14	0	0.15
dom_modifier	0.1	0.38	0.16	0.16	0	0.16	0.14
intermed_target	0.2	0.33	0.22	0	0.13	0.14	0.18
intermed_modifier	0.2	0.33	0	0.18	0.13	0.13	0.22
dom_target	0.2	0.36	0.17	0.17	0.15	0	0.16
dom_modifier	0.2	0.36	0.17	0.17	0	0.16	0.15
intermed_target	0.3	0.31	0.18	0	0.14	0.14	0.22
intermed_modifier	0.3	0.31	0	0.22	0.14	0.14	0.18
dom_target	0.3	0.33	0.17	0.17	0.15	0	0.17
dom_modifier	0.3	0.33	0.17	0.17	0	0.17	0.15
intermed_target	0.4	0.29	0.19	0	0.15	0.15	0.22
intermed_modifier	0.4	0.29	0	0.22	0.15	0.15	0.19
dom_target	0.4	0.31	0.18	0.18	0.16	0	0.17
dom_modifier	0.4	0.31	0.18	0.18	0	0.17	0.16
intermed_target	0.5	0.28	0.19	0	0.15	0.15	0.22
intermed_modifier	0.5	0.28	0	0.22	0.15	0.15	0.19
dom_target	0.5	0.29	0.18	0.18	0.17	0	0.18
dom_modifier	0.5	0.29	0.18	0.18	0	0.18	0.17
intermed_target	0.6	0.26	0.2	0	0.16	0.16	0.22
intermed_modifier	0.6	0.26	0	0.22	0.16	0.16	0.2
dom_target	0.6	0.27	0.18	0.18	0.17	0	0.19
dom_modifier	0.6	0.27	0.18	0.18	0	0.19	0.17
intermed_target	0.7	0.24	0.2	0	0.17	0.17	0.22
intermed_modifier	0.7	0.24	0	0.22	0.17	0.17	0.2
dom_target	0.7	0.26	0.18	0.18	0.18	0	0.19
dom_modifier	0.7	0.26	0.18	0.18	0	0.19	0.18
intermed_target	0.8	0.23	0.2	0	0.17	0.17	0.21
intermed_modifier	0.8	0.23	0	0.22	0.18	0.17	0.2
dom_target	0.8	0.24	0.19	0.19	0.19	0	0.2
dom_modifier	0.8	0.24	0.19	0.19	0	0.2	0.19
intermed_target	0.9	0.22	0.21	0	0.18	0.18	0.21
intermed_modifier	0.9	0.22	0	0.21	0.18	0.18	0.21
dom_target	0.9	0.23	0.19	0.19	0.2	0	0.2
dom_modifier	0.9	0.23	0.19	0.19	0	0.2	0.2
intermed_target	1.0	0.21	0.21	0	0.19	0.19	0.21
intermed_modifier	1.0	0.21	0	0.21	0.19	0.19	0.21
dom_target	1.0	0.21	0.19	0.19	0.21	0	0.21
dom_modifier	1.0	0.21	0.19	0.19	0	0.21	0.21
intermed_target	LLAURENS_2009	0.37	0.17	0	0.14	0.14	0.18
intermed_modifier	LLAURENS_2009	0.37	0	0.18	0.14	0.14	0.17
dom_target	LLAURENS_2009	0.38	0.16	0.16	0.15	0	0.16
dom_modifier	LLAURENS_2009	0.38	0.16	0.16	0	0.16	0.15

Table S1: Frequencies of S-alleles in function of type of mutations implicated and equal genetic load linked to the S-locus (L). *Mutation implicated : intermed_target=mutation on target in intermediate class, intermed_modifier=mutation on modifier in intermediate class, dom_target=mutation on target in dominant class, dom_modifier=mutation on modifier in dominant class. d : values between 0 and 1 for equal genetic load. Freq_S1= mean frequency at equilibrium for recessive S-allele, Freq_S2 and Freq_S3= mean frequency at equilibrium for ancestral intermediate S-alleles, Freq_S4 and Freq_S5= mean frequency at equilibrium for ancestral dominant S-alleles, Freq_S'= mean frequency at equilibrium for mutant S-allele.*

Mutation implicated	L	P_fix	MIN_time	MEAN_time	MED_time	MAX_time
intermed_target	0	6.61	120	339.91	300	1070
intermed_modifier	0	6.27	100	304.85	280	1010
dom_target	0	4.58	130	340.02	310	910
dom_modifier	0	4.38	100	338.17	320	1020
intermed_target	0.1	7.09	100	345.49	300	1270
intermed_modifier	0.1	6.98	100	306.52	270	1060
dom_target	0.1	5.11	110	331.27	310	950
dom_modifier	0.1	5.22	120	314.04	290	790
intermed_target	0.2	7.1	80	283.27	250	960
intermed_modifier	0.2	7.04	110	345.04	310	1250
dom_target	0.2	5.91	110	322.84	290	900
dom_modifier	0.2	5.83	110	305.61	280	860
intermed_target	0.3	6.82	120	325.07	290	1000
intermed_modifier	0.3	8.47	90	285.74	250	1030
dom_target	0.3	5.81	120	320.02	300	1090
dom_modifier	0.3	6.34	100	296.51	280	900
intermed_target	0.4	6.91	80	326.18	280	1330
intermed_modifier	0.4	8.74	80	280.43	240	990
dom_target	0.4	6.16	100	316.23	290	960
dom_modifier	0.4	7.17	90	282.38	260	950
intermed_target	0.5	7.78	110	310.53	270	1020
intermed_modifier	0.5	9.01	90	270.93	240	810
dom_target	0.5	6.03	120	311.46	290	950
dom_modifier	0.5	7.66	90	269.91	250	1100
intermed_target	0.6	7.59	120	318.54	280	1150
intermed_modifier	0.6	9.52	90	256.76	230	1040
dom_target	0.6	7.15	110	306.39	280	900
dom_modifier	0.6	8.93	100	259.46	240	910
intermed_target	0.7	6.92	90	313.96	280	980
intermed_modifier	0.7	10.3	80	246.46	220	870
dom_target	0.7	6.93	100	307.86	280	940
dom_modifier	0.7	10.17	80	254.75	230	840
intermed_target	0.8	6.98	100	319.66	290	890
intermed_modifier	0.8	10.91	80	244.01	220	890
dom_target	0.8	7.45	110	300.62	270	1230
dom_modifier	0.8	11.2	80	245.61	220	910
intermed_target	0.9	6.52	120	307.73	280	780
intermed_modifier	0.9	11.45	80	230.73	200	1120
dom_target	0.9	7.04	120	299.15	270	820
dom_modifier	0.9	12.36	80	238.06	210	890
intermed_target	1.0	5.96	130	309.55	290	910
intermed_modifier	1.0	11.55	70	225.55	200	890
dom_target	1.0	7.6	110	293.54	270	910
dom_modifier	1.0	12.99	80	224.4	200	1160
intermed_target	LLAURENS_2009	4.42	130	367.96	340	1070
intermed_modifier	LLAURENS_2009	6.55	100	291.59	270	920
dom_target	LLAURENS_2009	3.81	90	391.31	370	1360
dom_modifier	LLAURENS_2009	6.4	100	312	280	1050

Table S2 : Probability of fixation of mutations in function of type of mutations implicated and equal genetic load linked to the S-locus (L). *Mutation implicated : intermed_target=mutation on target in intermediate class, intermed_modifier=mutation on modifier in intermediate class, dom_target=mutation on target in dominant class, dom_modifier=mutation on modifier in dominant class. d : values between 0 and 1 for equal genetic load. P_fix= Probability of fixation of mutation (%). MIN_time= minimum generation to fix mutation. MEAN_time= mean generation to fix mutation. MED_time= median generation to fix mutation. MAX_time= maximum generation to fix mutation.*

Mutation implicated	L	P_fix
dominant	0	0.31
intermediate	0	0.295
dominant	0.1	0.35
intermediate	0.1	0.28
dominant	0.2	0.33
intermediate	0.2	0.19
dominant	0.3	0.31
intermediate	0.3	0.23
dominant	0.4	0.29
intermediate	0.4	0.19
dominant	0.5	0.3
intermediate	0.5	0.25
dominant	0.6	0.32
intermediate	0.6	0.23
dominant	0.7	0.32
intermediate	0.7	0.28
dominant	0.8	0.24
intermediate	0.8	0.19
dominant	0.9	0.32
intermediate	0.9	0.25
dominant	1	0.31
intermediate	1	0.3

Table S3 : Probability of fixation of neutral mutations in function of class of S-allele linked and equal genetic load linked to the S-locus (L). *Mutation implicated : mutation on modifier of S-allele in dominant or intermediate class. d : values between 0 and 1 for equal genetic load. P_fix= Probability of fixation of mutation (%).*

class	L (dom)	L (inter)	P_fix_interaction (%)	P_fix_neutral (%)	MIN_time	MEAN_time	MED_time	MAX_time
intermed	0.8	0.8	6.54	0.19	100	303.26	270	1050
dom	0.8	0.8	6.31	0.24	110	308.91	290	820
intermed	0.8	0.7	6.33	0.3	110	300.88	270	1340
dom	0.8	0.7	6.31	0.37	110	315.78	290	840
intermed	***0.8***	***0.6***	***6.44***	***0.34***	***110***	***302.11***	***270***	***990***
dom	***0.8***	***0.6***	***6.04***	***0.26***	***100***	***316.59***	***300***	***790***
intermed	0.8	0.5	6.81	0.31	100	300.63	270	910
dom	0.8	0.5	5.91	0.34	130	305.08	280	980
intermed	0.8	0.4	7.06	0.23	90	297.83	270	910
dom	0.8	0.4	6.08	0.36	120	309.57	295	1030
intermed	0.8	0.3	7.25	0.18	100	295.32	270	880
dom	0.8	0.3	5.92	0.18	110	304.51	280	1040
intermed	0.8	0.2	7.05	0.19	80	280.77	250	970
dom	0.8	0.2	5.09	0.38	110	316.17	300	810
intermed	0.7	0.7	6.35	0.14	110	291.09	260	1020
dom	0.7	0.7	5.97	0.25	110	309.28	290	850
intermed	0.7	0.6	6.33	0.2	110	299.46	270	930
dom	0.7	0.6	6.16	0.41	100	306.36	280	850
intermed	0.7	0.5	6.38	0.23	100	298.62	260	1210
dom	0.7	0.5	5.65	0.37	100	311.43	290	830
intermed	0.7	0.4	7.03	0.18	100	298.71	260	1000
dom	0.7	0.4	6.04	0.31	110	307.67	290	840
intermed	0.7	0.3	7.33	0.24	100	286.62	260	1050
dom	0.7	0.3	5.55	0.43	100	312.88	280	1020
intermed	0.7	0.2	7.66	0.26	100	295.95	260	1220
dom	0.7	0.2	5.77	0.36	100	313.6	280	1100
intermed	0.6	0.6	6.67	0.19	90	296.99	270	950
dom	0.6	0.6	6.41	0.30	120	308.95	280	1020
intermed	0.6	0.5	6.53	0.36	100	296.06	270	1340
dom	0.6	0.5	6.31	0.3	100	307.77	280	800
intermed	0.6	0.4	7.28	0.31	110	300.66	260	840
dom	0.6	0.4	5.86	0.33	100	316.42	290	870
intermed	0.6	0.3	7.28	0.15	100	294.45	260	1050
dom	0.6	0.3	6.06	0.31	110	303.09	280	1180
intermed	0.6	0.2	8.06	0.23	80	296.35	260	1210
dom	0.6	0.2	5.89	0.43	100	312.92	280	1260
intermed	0.5	0.5	6.49	0.23	110	303.14	270	990
dom	0.5	0.5	6.19	0.33	100	308.9	280	880
intermed	0.5	0.4	6.59	0.34	100	296.54	270	920
dom	0.5	0.4	6.24	0.35	110	316.04	290	960
intermed	0.5	0.3	6.81	0.27	100	295.99	260	1000
dom	0.5	0.3	5.81	0.23	100	314.27	290	1110
intermed	0.5	0.2	7.48	0.29	100	295.84	260	1430
dom	0.5	0.2	5.89	0.3	110	310.25	280	880
intermed	0.4	0.4	7.08	0.3	110	306.57	270	1080
dom	0.4	0.4	6.39	0.49	110	298.73	270	950
intermed	0.4	0.3	6.84	0.31	100	292.38	265	1410
dom	0.4	0.3	5.63	0.3	100	299.36	270	920
intermed	0.4	0.2	7.52	0.29	90	292.82	260	1230
dom	0.4	0.2	5.73	0.31	100	306.13	280	900

Table S4 : Probability of fixation of mutations on modifiers in function of the class of allele associated and genetic load linked to the S-locus difference in alleles in dominant (L(dom)) and intermediate class (L(inter)). *Class : class of allele with mutation on modifier, intermed =mutation intermediate class, dom =mutation in dominant class. P_fix_interaction= Probability of fixation of mutation (%). P_fix_neutral= Probability of fixation of equivalent neutral mutation (%). MIN_time= minimum generation to fix mutation. MEAN_time= mean generation to fix mutation. MED_time= median generation to fix mutation. MAX_time= maximum generation to fix mutation. Italic : models with values of genetic load presented in Llaurens et al., 2009b.*

Name	Accession EMBL database	Reference	Dominance class	Phenotyping	Small RNA sequencing
Ah02	**Aha-B-MtgEte-61O5**	**this study**	**III**	X	
Ah03	KJ772378-KJ772385	Goubet et al,. 2012	II	X	X
Ah04	KJ461484	Durand et al,. 2014	III	X	X
Ah10	KM592810-KM592817	Durand et al,. 2014	III	X	X
Ah12	KJ772373-KJ772377	Durand et al,. 2014	IV	X	X
Ah13	KJ461479-KJ461483	Goubet et al,. 2012	IV	X	X
Ah15	KJ772386-KJ772395	Goubet et al,. 2012	IV	**X**	
Ah20	KJ772396-KJ772400	Goubet et al,. 2012	IV	X	X
Ah25	**Aha-B-MtgEte-57C6**	**this study**	**III**		
Ah28	KJ461475-KJ461478	Goubet et al,. 2012	II	**X**	X
Ah29	KM592798-KM592803	Durand et al,. 2014	III	**X**	
Ah32	KJ461470-KJ461474	Goubet et al,. 2012	IV		
Ah36	KM592804-KM592809	Durand et al,. 2014	IV		
Ah43	KJ461485-KJ461492	Goubet et al,. 2012	IV		
Al1	KJ772401-KJ772404	Goubet et al,. 2012	I	X	X

Table S5: S-locus sequences. *The fasta files for the BAC clones used in this study are available in the figshare database at 10.6084/m9.figshare.17025419. The others in the EMBL database. Phenotyping: S-alleles with dominance hierarchy tested by phenotyping approach. Small RNA sequencing: libraries of sRNAs produced by different S-alleles extracted by Durand et al,. 2014.*

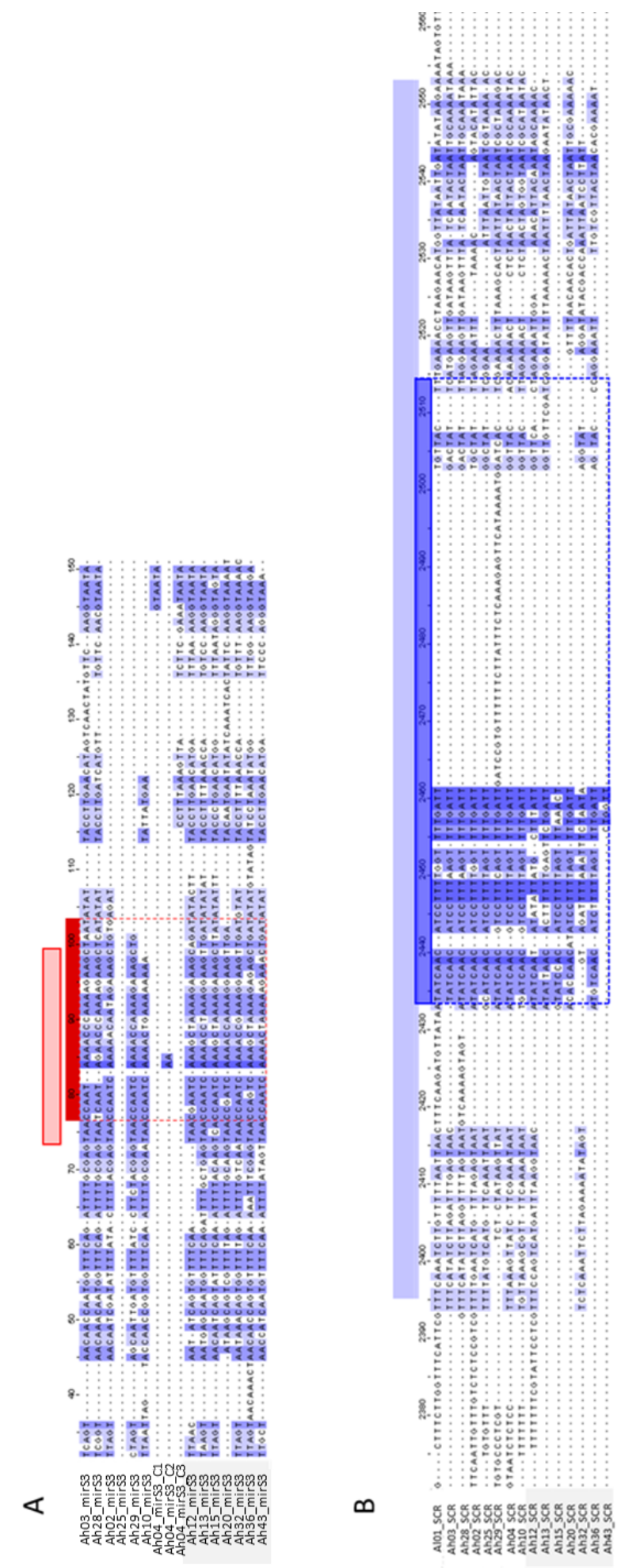


Figure S1 : Alignment of the different precursors of mirS3 (A) and SCR genes (B) of the recessive (white) and dominant (grey) S-alleles*. Visualisation by Jalview V.2.10.15. The intensity of blue on position increases with the conservation score.(A) We focused on the region predicted to target recessive S for the sRNA produced by the dominant S-alleles Ah12, Ah13, Ah20 (red scare). Light red scare= first region of 26nt with higher mean pairwise conservation. B) We focused on the region predicted to be targeted on recessive S-allele by the sRNA produced by the dominant S-alleles Ah12, Ah13, Ah20 (blue scare). Light blue scare= Region with higher mean pairwise conservation that overlapped the region predicted to be targeted in recessive S-alleles.*

Discussion and perspectives

In this thesis, I first focused on the interaction between balancing selection at the *S*-locus and the accumulation of polymorphism in its flanking genomic regions, focusing in particular on the accumulation of potentially deleterious mutations (the linked genetic load). Secondly, I combined theoretical, genomic and empirical approaches to study the interaction between dominance at the *S*-locus and the accumulation of the genetic load in the regions flanking the *S*-locus. Finally, I have furthered our theoretical knowledge on the interaction between balancing selection and the evolution of the dominance network involving sRNAs, and between the linked load and the dominance network. The main results, summarised in Figure 1, have clarified some of the effects of this complex interaction between the partners of this ménage à trois : balancing selection, the genetic load and dominance, in the special context of the SSI system. However, some questions remained only partially resolved, and some new questions arose. I will now describe these new problems and outline potential avenues of research for each question.

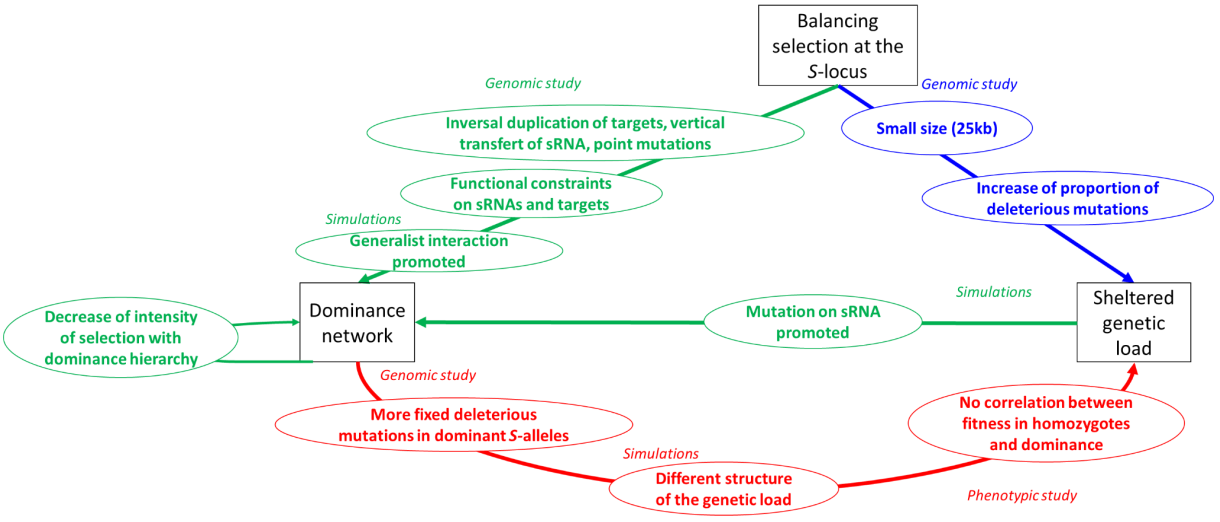


Figure 1: Schematic representation of the main results presented in the thesis. *The lines symbolise the relations between the partners of the ménage à trois (black square). The blue, red and green lines represent the relations studied in the first, second and third chapter, respectively. The main results obtained relative to each relation are outlined in the ovals and correspond to the questions in Figure 15 of the general introduction. The methods used to tackle each question are noted in italic.*

1) Defining the sheltered genetic load remains a major challenge

The architecture of the genetic load linked to the *S*-locus that we characterised (chapter I) is sharply distinct from the one observed in the other well-studied example of balancing selection, the *HLA* system (Lenz et al., 2016). In fact, contrary to genes linked to *HLA*, the *S*-flanking regions present an increase in the proportion of deleterious mutations but not an

increase in allelic frequency, compared with control genes. Understanding how and why different processes of balancing selection result in such contrasted patterns of genome diversity would clearly merit further theoretical investigation. Indeed, it would be interesting to model the evolution of genomes linked to loci under different models of balancing selection beyond the general processes of balancing selection that have been considered so far. This has been started for the *MHC* locus (Lenz et al., 2016), but it would clearly be interesting to incorporate more cases of balancing selection, including cases of frequency-dependent selection such as a gametophytic SI, but also sporophytic SI with or without dominance interactions for example.

The effect of dominance on the accumulation of deleterious mutations around the *S*-locus (chapter II) are more complex than predicted in Llaurens et al (2009a), and lead to a modification of the structure of haplotypes linked to the *S*-locus along the dominance hierarchy. In fact, the proportion of segregating deleterious mutations was higher in recessive than in dominant *S*-alleles. However, the effect on fitness of homozygotes at the *S*-locus remains unclear at the end of this project. This relation must be studied by further developing models but also by improving phenotypic approaches. While Llaurens et al. (2009a) observed a positive relationship between dominance and the linked load, Stift et al. (2013) and my own results (Chapter II) could not confirm this observation. A first difficulty to compare the results of these three studies is that they used different types of crosses to obtain homozygotes at the *S*-locus. Llaurens et al (2009a) uses forced incompatible crosses under CO_2. This method allows the analysis of a large sample size (872 individuals), but is not able to distinguish between homozygotes derived from outcrossing and homozygotes derived from self-fertilisation. Hence, the *S*-alleles that are made homozygous by the method can be carried by chromosomes that are either the same or different allelic copies from the populations where the parents of the cross were collected. Stift et al (2013) used forced self-fertilisation to avoid this problem, but in their case, the sample size analysed was smaller (112 individuals). The approach I used is based on compatible crosses and takes advantage of the “natural” masking of recessive *S*-alleles by dominant *S*-alleles to generate homozygotes at the *S*-locus. This method is less fastidious than the two previously mentioned as it does not require by-passing the SI system. As in Llaurens et al. (2009a), homozygotes at the *S*-locus are formed from two different copies of the same allele. However, this method requires a lot of different genotypes at the *S*-locus to perform compatible crosses. For example, during the thesis project, I tried to make compatible crosses to obtain homozygotes for an allele of class IV, Ah12. However, the number of possible partners for these crosses was too small to obtain the expected offspring. Moreover, this method does not allow the formation of homozygotes for the most dominant *S*-alleles, such as the Ah13, Ah15 or Ah20. Given the contradictory results between the results of the three studies, one can consider that the question of whether the genetic load increases with dominance is not yet settled. Obtaining experimental evidence to test this hypothesis conclusively will require a broader, more powerful analysis. Such an empirical study will certainly require reverting to forced crosses under CO_2 so as to be able to include the most dominant *S*-alleles again. The second limitation is, for the three papers, the number of *S*-alleles analysed (three in Llaurens et al

(2009a), four in Stift et al (2013) and three in chapter II of this thesis). To understand the relation between dominance and the genetic load, we will require more alleles to be analysed. In addition, the choice of the phenotypes to be measured will need to be carefully considered. In Llaurens et al (2009a), Stift et al (2013) and our own study, survival to the reproductive stage is the only trait that was consistently impacted by homozygosity at the *S*-locus. This trait is easily measured compared to the thirteen others and requires less work, both in terms of experimental manipulation and in plant maintenance. It could therefore be measured easily on a very large number of individuals, which would allow the analysis of the effect of homozygosity at the *S*-locus for a large number of different alleles. Finally, the genomic analyses I performed suggest that the accumulation of potentially deleterious mutations varies strongly across populations. Yet, the empirical data accumulated to date to estimate the phenotypic effect of homozygosity at the *S*-locus are based on just one *A. halleri* population (Nivelle in Llaurens et al (2009a) and the present study) and one *A. lyrata* population (Stift et al. 2013). In Nivelle, I observed a positive relationship between the total genetic load accumulated and dominance (in terms of the number of 0-fold degenerate mutations), whereas in PIN I observed the reverse relationship. It would therefore be essential to repeat these empirical analyses on other populations to determine the generality of the eventual effect of dominance and its direction.

An important contribution of this project, which I achieved by combining sequence capture with parents-offspring trio phasing, was to reconstruct haplotypes for a large number of *S*-alleles from different populations. Previous methods were based on the construction of BAC libraries, that were time and resource intensive (Goubet et al. 2012) and were not able to reconstitute as many haplotypes over such a long distance. Sequence capture had the advantage of specifically amplifying our region of interest, even for *A. lyrata*, whereas the probes were designated from the *A. halleri* genome. This approach has its drawbacks, however. For instance, as compared to BAC clones this technique cannot reliably represent paralogous or repeat-rich sequences. Despite the filtration to remove sites with exceptionally high sequencing coverage, we cannot be sure that we correctly considered sites as polymorphic due to the alignment of amplified reads of paralogs, for example. Furthermore, as this technique is based on a limited number of probes that did not contain all of the *SRK* or *SCR* alleles present in *A. halleri* or *A. lyrata*, it is not highly reliable to comprehensively identify *S*-alleles themselves because of their very high level of sequence divergence. The genotype of homozygotes in Chapter II thus had to be confirmed by genome-wide sequencing to distinguish true homozygotes from individuals carrying two *S*-alleles, one of which was not included in the probes. Another limitation of the capture approach is that it does not allow the analysis of structural mutations such as transposable elements. For these types of mutations, the use of BAC clones can be considered, but this method is tedious and does not ensure that the entire region of interest can be analysed. The use of new technology, such as Nanopore or PACBIO HiFi sequencing, seems more reliable. For the time being though, the capture still has the advantage of being cost-effective because it offers the possibility of analysing a large number of variants for our region of interest on a large number of individuals. Another important difficulty in the course of this project was to

determine how to define deleterious mutations. We chose to stick to the strict definition that mutations on 0-fold degenerate sites were deleterious. However, defining a deleterious mutation on the basis of genomic data remains a real challenge. An increasing number of tools are aiming to be "efficient" in detecting deleterious mutations. Some of these tools, such as SNPeff (Cingolani et al,. 2012), have been developed for variant annotation on a functional basis. But generally these tools don't present the annotation for sites that are not polymorphic, even though this is required for the analysis of polymorphism. Other tools, such as SIFT4G (Vaser et al,. 2016), make the assumption that a mutation at a highly conserved position across multiple species is likely to be deleterious. However, these tools are performed for a limited choice of reference genome and a limited proportion of annotated position. For SIFT4G, we had to align the reads to the *A. lyrata* reference genome and, for this reference genome, on average, only 12% of the variable positions in each dataset were annotated. These tools appeared to be inefficient in our study. An alternative to these tools is to manually calculate the phylogenetic conservation of a reference of our choice after alignment with homologous sequences present in other species. This is a possible extension to the work I presented. Moreover, in all the methods mentioned, only the mutation on genes can be considered. The potential deleterious mutations in intergenic regions, like the transposable element previously mentioned, cannot be studied with these methods.

2) Dominance as a regulatory network rather than as the property of an allele

The main theoretical advance of our study of dominance at the *S*-locus in Arabidopsis is that it is considered for the first time dominance as a network of interactions. Our theoretical models, supported by molecular observation, offer new predictions on the evolution of dominance. For example, the results predicted that evolution of dominance is promoted in recessive classes. Moreover, this concept of dominance network offers new perspectives of research. For example, modifiers such as sRNAs offer new possibilities for potential interactions that have not been considered before : can an allele of a recessive class become capable of inhibiting the expression of a more dominant allele in heterozygotes? If so, under what circumstances and what would be the consequences on the initial network? Alternatively, one can envision that an allele acquires a modifier capable of targeting itself, thus generating a self-compatible allele. The circumstances allowing such an interaction and its consequences remain to be studied, which would allow, in the long run, to consider a new way to lose the incompatibility system, or to explain complex incompatibility systems. Nevertheless, one can wonder if the predictions of our model are not restricted to the SSI system observed in some Brassicas. Testing these predictions will remain challenging. The first challenge lies with the very high level of multi allelism observed at the *S*-locus. An ongoing analysis in the team shows that very close to 65 *S*-alleles segregate species-wide in *A. halleri*. Yet, dominance interactions were studied only between eleven *S*-alleles. Extending the analysis to this more complete catalogue of *S*-alleles would be necessary to obtain a more comprehensive view of the properties of the dominance network. Unfortunately,

determining phenotypic relationships remains a long and difficult task. Another possibility would be to extend the analysis to other plant species with an SSI system, but again the amount of experimental work to document these dominance networks can be daunting. Finally, to even further generalise the properties of dominance networks, it would be very interesting to ask whether other balanced polymorphisms with dominance interactions (beyond the SSI system) also show the general properties that we predicted. Given the diversity of balancing selection processes, it would probably be necessary to tailor the models to the exact form of balancing selection at play.

From a molecular point of view, our understanding of the molecular control of dominance interactions is quite good. However, some of the dominance relationships observed at the pollen level are still not understood, like the dominance relationship between Ah29 and Ah02. Further molecular studies will be needed to identify the mechanism. Moreover, the molecular processes involved (i.e. methylation of promoter, degradation of mRNA…) stay unclear. With the diversity of the targeted regions on *SCR* alleles (Exon, intron, upstream) and the diversity of the size of the sRNAs (essentially between 21 and 24nt), we can suppose that different molecular processes are involved. Finally, the particular dominance phenotype that we observed between Ah29 and the *S*-alleles Ah15 is associated with an insertion in the intron of Ah29 at the miRNA target site (Chapter III). It will be necessary to reiterate crosses with Ah15 to exclude the possible experimental artefact. Moreover, the recessive phenotype observed for Ah29 in heterozygotes with the other *S*-alleles of class IV (Ah12, Ah13 and Ah20 is not explained by molecular interaction between *SCR* of Ah29 and sRNA linked to *S*-alleles of class IV. We supposed that Ah29 presents an intra-allelic polymorphism. Exactly, we supposed that some copies of Ah29 present the insertion (found in our BAC clone) and they are not targeted by *S*-alleles of class IV and some others copies without the insertion are targeted by *S*-alleles of class IV. To verify this hypothesis, we must sequence more copie of Ah29. We can use BAC clones to verify this assumption, but the ongoing whole-genome Nanopore assemblies being produced in the lab may help resolve this question.

3) The effect of dominance and the genetic load on the maintenance and diversification of the balanced polymorphism

During this phD project, I studied most the effect of balancing selection on dominance and on genetic load. However, I did not directly consider the reverse interactions. Yet, we have expectations about the effect of dominance and the genetic load on the maintenance and on the diversification of balanced polymorphisms.

The architecture of the sheltered genetic load linked to a locus or gene in balancing selection could have an influence on the maintenance of balancing selection for two main reasons: first, the absence of genetic load linked to the *S*-locus is expected to lose the SI system. In the GSI system, the maintenance of self incompatibility with a genetic load was theoretically studied: purging of deleterious mutations may greatly enhance the spread of self compatible mutants when deleterious alleles have strong fitness effects (Gervais et al., 2014). Congruent

with these expectations, the absence of genetic load linked to the *S*-locus was observed in the self compatible populations of *A. lyrata* of North America (Carleil et al., 2017). However, the theoretical model for the SSI system is not developed yet. Second, the genetic load associated with some alleles at a locus under balancing selection can cause an overdominance process to emerge, which accumulates at the negative frequency dependent selection process. Indeed, if homozygotes at the *S*-locus express a high associated genetic load, this genetic load may force the *S*-alleles to remain in the heterozygous state to benefit from improved fitness (Llaurens et al., 2017 for review). This further reinforces the persistence of balanced polymorphism in natural populations.

The architecture of the sheltered load linked to the *S*-locus or gene in balancing selection is also expected to prevent the diversification of *S*-allele if the new *S*-allele *Si*, favoured by negative frequency dependant selection, expressed the genetic load associated with the ancestral *S*-allele *Sj* in heterozygous *SiSj* (Uyenoyama, 2003). However, our study demonstrates that the number of deleterious mutations in segregation between copies of a same *S*-allele could be high, particularly when the dominance level of the *S*-allele is low. This interaction between dominance and genetic load must be considered to correctly predict the perspectives of diversification of *S*-allele in future models. Another effect of dominance on diversification must be considered: dominance relationships between *S*-alleles modified their frequencies in population (Llaurens et al., 2009b). We supposed that an increase in frequency in population promotes the diversification of the *S*-alleles. However, this assumption deserves a theoretical study, by stochastic or deterministic models for example.

Bibliography

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.-H., Spatafora, J.W., and Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana.* Nature Genetics 36, 1282–1290.

Avise, J.C. (1989). A role for molecular genetics in the recognition and conservation of endangered species. Trends in Ecology & Evolution 4, 279–281.

Barton, N.H. (1995). Linkage and the limits to natural selection. Genetics 140, 821–841.

Bateman, A.J. (1952). Self-incompatibility systems in angiosperms : I. Theory. Heredity 6, 285–310.

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

Billiard, S., and Castric, V. (2011). Evidence for Fisher's dominance theory : how many 'special cases'? Trends in Genetics 27, 441–445.

Burghgraeve, N., Simon, S., Barral, S., Fobis-Loisy, I., Holl, A.-C., Ponitzki, C., Schmitt, E., Vekemans, X., and Castric, V. (2020). Base-pairing requirements for small rna-mediated gene silencing of recessive self-incompatibility alleles in *Arabidopsis halleri*. Genetics 215, 653–664.

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell 136, 642–655.

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

Castric, V., Bechsgaard, J.S., Grenier, S., Noureddine, R., Schierup, M.H., and Vekemans, X. (2010). Molecular evolution within and between Self-Incompatibility Specificities. Molecular Biology and Evolution 27, 11–20.

Charlesworth, D., and Charlesworth, B. (1975). Theoretical genetics of *batesian mimicry* III. Evolution of dominance. Journal of Theoretical Biology 55, 325–337.

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

Charlesworth, B., and Charlesworth, D. (1999). The genetic basis of inbreeding depression. Genetics Research 74, 329–340.

Charlesworth, B., Charlesworth, D., and Barton, N.H. (2003). The effects of genetic and geographic structure on neutral variation. Annual Review of Ecology, Evolution, and Systematics 34, 99–125.

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

Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. Nature Reviews Genetics 10, 783–796.

Charlesworth, B., and Charlesworth, D. (2010). Elements of evolutionary genetics (Roberts and Company).

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

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92.

Cope, F.W. (1962). The effects of incompatibility and compatibility on genotype proportions in populations of *Theobroma cacao L*. Heredity 17, 183–195.

Das, G., Patra, J.K., and Baek, K.-H. (2017). Insight into MAS: A Molecular Tool for Development of Stress Resistant and Quality of Rice through Gene Stacking. Frontiers in Plant Science 8.

Decaestecker, E., Gaba, S., Raeymaekers, J.A.M., Stoks, R., Van Kerckhoven, L., Ebert, D., and De Meester, L. (2007). Host–parasite 'Red Queen' dynamics archived in pond sediment. Nature 450, 870–873.

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

Durand, E., Méheust, R., Soucaze, M., Goubet, P.M., Gallina, S., Poux, C., Fobis-Loisy, I., Guillon, E., Gaude, T., Sarazin, A., et al. (2014). Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. Science 346, 1200–1205.

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

Finnegan, E.J., Liang, D., and Wang, M.-B. (2011). Self-incompatibility : *Smi* silences through a novel sRNA pathway. Trends in Plant Science 16, 238–241.

Fisher, R.A. (1928). The possible modification of the response of the wild type to recurrent mutations. The American Naturalist 62, 115–126.

Frankham, R. (2005). Genetics and extinction. Biological Conservation 126, 131–140.

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. Mol Biol Evol.

Gervais, C., Awad, D.A., Roze, D., Castric, V., and Billiard, S. (2014). Genetic architecture of inbreeding depression and the maintenance of gametophytic self-incompatibility. Evolution 68, 3317–3324.

Gilbert, W., and Maxam, A. (1973). The nucleotide sequence of the lac operator. PNAS 70, 3581–3584.

Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2011). Unlocking Mendelian disease using exome sequencing. Genome Biol 12, 228.

Glémin, S., Bataillon, T., Ronfort, J., Mignot, A., and Olivieri, I. (2001). Inbreeding depression in small populations of self-incompatible plants. Genetics 159, 1217–1229.

Gloag, R., Ding, G., Christie, J.R., Buchmann, G., Beekman, M., and Oldroyd, B.P. (2016). An invasive social insect overcomes genetic load at the sex locus. Nat Ecol Evol 1, 1–6.

Gloss, B.S., and Dinger, M.E. (2018). Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med 50, 1–8.

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., and Weigel, D. (2011). Evolution of the S-Locus region in Arabidopsis relatives. Plant Physiology 157, 937–946.

Guttmacher, A.E., and Collins, F.S. (2002). Genomic medicine — A primer. New England Journal of Medicine 347, 1512–1520.

Haldane, J.B.S. (1924). A mathematical theory of natural and artificial selection. Part II the influence of partial self-fertilisation, inbreeding, assortative mating, and selective fertilisation on the composition of mendelian populations, and on natural selection. Biological Reviews 1, 158–163.

Haldane, J.B.S. (1930). A note on Fisher's theory of the origin of dominance, and on a correlation between dominance and linkage. The American Naturalist 64, 87–90.

Hamblin, M.T., Buckler, E.S., and Jannink, J.-L. (2011). Population genetics of genomics-based crop improvement methods. Trends in Genetics 27, 98–106.

Hedrick', P.W., and Thomson, G. (1983). Evidence for balancing selection at *HLA*. Genetics 104, 449–456.

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(1), 153-159.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36, 338–345.

Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., and Joron, M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. Nat Genet 53, 288–293.

Kacser, H., and Burns, J.A. (1981). The molecular basis of dominance. Genetics 97, 639–666.

Kakizaki, T., Takada, Y., Ito, A., Suzuki, G., Shiba, H., Takayama, S., Isogai, A., and Watanabe, M. (2003). Linear dominance relationship among four class-II S haplotypes in pollen is determined by the expression of *SP11* in brassica self-incompatibility. Plant and Cell Physiology 44, 70–75.

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

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature 217, 624–626.

Klein, J., Sato, A., and Nikolaidis, N. (2007). *MHC*, *TSP*, and the origin of species: from immunogenetics to evolutionary genetics. Annual Review of Genetics 41, 281–304.

Kowyama, Y., Takahasi, H., Muraoka, K., Tani, T., Hara, K., and Shiotani, I. (1994). Number, frequency & dominance relationships of S-alleles in diploid *Ipomoea trifida*. Heredity 73, 275–283.

Krier, J.B., Kalia, S.S., and Green, R.C. (2016). Genomic sequencing in clinical practice: applications, challenges, and opportunities. Dialogues in Clinical Neuroscience 18, 299.

Kulski, J. (2016). Next Generation Sequencing : advances, applications and challenges (BoD – Books on Demand).

Kusaba, M., Tung, C.-W., Nasrallah, M.E., and Nasrallah, J.B. (2002). Monoallelic expression and dominance interactions in anthers of self-incompatible *Arabidopsis lyrata*. PLANT PHYSIOLOGY 128, 17–20.

Lamoril, J., Ameziane, N., Deybach, J.-C., Bouizegarène, P., and Bogard, M. (2008). Les techniques de séquençage de l'ADN : une révolution en marche. Première partie. immuno-analyse & biologie spécialisée 23, 260–279.

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

Llaurens, V., Billiard, S., Leducq, J.-B., Castric, V., Klein, E.K., and 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., and Billiard, S. (2009a). 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., Billiard, S., Castric, V., and Vekemans, X. (2009b). Evolution of dominance in sporophytic self-incompatibility systems: I. Genetic load and coevolution of levels of dominance in pollen and pistil. Evolution 63, 2427–2437.

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

Loewe, L., and Hill, W.G. (2010). The population genetics of mutations : good, bad and indifferent. Philosophical Transactions of the Royal Society B : Biological Sciences 365, 1153–1167.

Lynch, M., Conery, J., and Bürger, R. (1995). Mutational meltdowns in sexual populations. Evolution 49, 1067–1080.

Ma, R., Han, Z., Hu, Z., Lin, G., Gong, X., Zhang, H., Nasrallah, J.B., and Chai, J. (2016). Structural basis for specific self-incompatibility response in Brassica. Cell Research 26, 1320–1329.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. Trends in Genetics 24, 133–141.

Masel, J. (2011). Genetic drift. Current Biology 21, R837–R838.

Mendel, G. (1913). Experiments in plant hybridisation. Trad. C. T. Druery. Bateson, W. Mendel's principles of heredity. 2nd ed. Cambridge: Cambridge University Press, 335–379.

Mills, M.C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. Commun Biol 2, 1–11.

Navarro, A., and Barton, N.H. (2002). The effects of multilocus balancing selection on neutral variability. Genetics 161, 849–863.

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

Nguyen, K.L., Grondin, A., Courtois, B., and Gantet, P. (2019). Next-Generation Sequencing Accelerates Crop Gene Discovery. Trends in Plant Science 24, 263–274.

Nozawa, M., Miura, S., and Nei, M. (2012). Origins and evolution of microRNA genes in plant species. Genome Biology and Evolution 4, 230–239.

Nuismer, S.L., and Otto, S.P. (2005). Host–Parasite interactions and the evolution of gene expression. PLOS Biology 3, e203.

O'Brien, S.J. (1994). Genetic and phylogenetic analyses of endangered species. Annual Review of Genetics 28, 467–489.

Oleksyk, T.K., Smith, M.W., and O'Brien, S.J. (2010). Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B : Biological Sciences 365, 185–205.

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.

Orr, H.A. (1991). A test of Fisher's theory of dominance. PNAS 88, 11413–11415.

Otto, S.P., and Bourguet, D. (1999). Balanced polymorphisms and the evolution of dominance. The American Naturalist, 153(6), 561-574.

Peischl, S., and Schneider, K.A. (2010). Evolution of dominance under frequency-dependent intraspecific competition in an assortatively mating population. Evolution 64, 561–582.

Robinson, J.A., Räikkönen, J., Vucetich, L.M., Vucetich, J.A., Peterson, R.O., Lohmueller, K.E., and Wayne, R.K. (2019). Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. Science Advances 5, eaau0757.

Roux, C., Pauwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V., and 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., and 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.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. PNAS 74, 5463–5467.

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., and Charlesworth, D. (2000). The effect of subdivision on variation at multi-allelic loci under balancing selection. Genetics Research 76, 51–62.

Schoen, D.J., and Busch, J.W. (2009). The evolution of dominance in sporophytic self-incompatibility systems. II. Mate availability and recombination. Evolution 63, 2099–2113.

Shiba, H., Kakizaki, T., Iwano, M., Tarutani, Y., Watanabe, M., Isogai, A., and Takayama, S. (2006). Dominance relationships between self-incompatibility alleles controlled by DNA methylation. Nat Genet 38, 297–299.

Slotte, T. (2014). The impact of linked selection on plant genomic variation. Briefings in Functional Genomics 13, 268–275.

Smith, J.M., and Haigh, J. (2007). The hitch-hiking effect of a favorable gene. Genet Res 89, 391–403.

Stift, M., Hunter, B.D., Shaw, B., Adam, A., Hoebe, P.N., and 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., and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124, 967–978.

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

Tarutani, Y., Shiba, H., Iwano, M., Kakizaki, T., Suzuki, G., Watanabe, M., Isogai, A., and Takayama, S. (2010). Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. Nature 466, 983–986.

Thompson, K.F., and Taylor, J.P. (1966). Non-linear dominance relationships between S alleles. Heredity 21, 345–362.

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.

Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nat Protoc 11, 1–9.

Veitia, R.A. (2006). The biology of genetic dominance (CRC Press).

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

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

Wright, S. (1929). Fisher's theory of dominance. The American Naturalist 63, 274–279.

Wright, S. (1934). Physiological and evolutionary theories of dominance. The American Naturalist 68, 24–53.

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

Yasuda, S., Wada, Y., Kakizaki, T., Tarutani, Y., Miura-Uno, E., Murase, K., Fujii, S., Hioki, T., Shimoda, T., Takada, Y., et al. (2017). A complex dominance hierarchy is controlled by polymorphism of small RNAs and their targets. Nature Plants 3.

Zhang, Y., Jiang, W., and Gao, L. (2011). Evolution of microRNA genes in *Oryza sativa* and *Arabidopsis thaliana* : An update of the inverted duplication model. PLoS ONE 6, e28073.

Abstracts

Sporophytic self-incompatibility is a genetic system preventing self-fertilisation by self-recognition. In many species, this system is controlled by a single locus, the *S*-locus, composed of two linked genes coding for the pistil and pollen recognition proteins. The self-incompatibility locus is a classical case of a particular form of balancing selection called negative frequency dependent selection. This form of selection is predicted to cause an accumulation of polymorphism in the flanking regions of the *S*-locus, including sheltered deleterious mutations. In the Brassicaceae, this system exhibits a linear dominance hierarchy between *S*-alleles. This dominance network is controlled by interactions between sRNAs linked to dominant *S*-alleles and their target sequences on recessive alleles of the gene controlling the pollen specificities *SCR*. The dominance level is predicted to have an effect on the accumulation of polymorphisms in regions immediately linked to the *S*-locus, with a higher accumulation of the genetic load sheltered by dominant *S*-alleles than by recessive *S*-alleles.

In my PhD project, I first studied the effect of balancing selection at the *S*-locus on polymorphism in the flanking regions in order to determine the magnitude of the peak of polymorphism and to characterise its molecular properties. I used whole genome resequencing data from several populations of *A. halleri* and *A. lyrata* to specifically determine the chromosomal distance up to which the effect of the *S*-locus can still be observed. I observed an increase of polymorphism in the first 25kb around the *S*-locus, mainly explained by an increase of the proportion of polymorphic sites.

I then tested if dominance of the *S*-alleles influences the genetic load they accumulate. I combined a genomic approach using parent-offspring trios to phase haplotypes and compare the number of deleterious mutations linked to dominant vs. recessive *S*-alleles, with a phenotypic approach to experimentally measure the severity of the load. I demonstrated that dominance promotes contrasted profiles of the genetic load between the recessive and the dominant *S*-alleles.

Finally, I used a modelling approach based on stochastic simulations to predict the evolution of the dominance network between *S*-alleles, taking interactions between sRNAs and their *SCR* targets explicitly into account. My results show that mutations have different fixation probabilities according to whether they occur on dominant vs. recessive *S*-alleles, and also whether they hit the sRNAs producing locus or its target sites. The distribution of the sheltered genetic load between dominant and recessive *S*-alleles is also an important determinant of the evolution of the dominance network.

French version

L'auto-incompatibilité sporophytique est un système génétique empêchant l'autofécondation par reconnaissance du soi. Chez de nombreuses espèces, ce système est contrôlé par un locus unique, le locus S, composé de deux gènes liés codant pour les protéines de reconnaissance du pistil et du pollen. Le locus d'auto-incompatibilité est un cas classique d'une forme particulière de sélection équilibrante appelée sélection fréquence dépendante négative. Cette forme de sélection est censée provoquer une accumulation de polymorphisme dans les régions flanquantes du locus S, y compris des mutations délétères. Chez les Brassicacées, ce système présente une hiérarchie de dominance linéaire entre les allèles S. Ce réseau de dominance est contrôlé par des interactions entre les allèles S. Ce réseau de dominance est contrôlé par des interactions entre les petits ARN liés aux allèles S dominants et leurs séquences cibles sur les allèles récessifs du gène contrôlant les spécificités polliniques *SCR*. Il est prédit que le niveau de dominance a un effet sur l'accumulation de polymorphismes dans les régions immédiatement liées au locus S, avec une accumulation plus importante de la charge génétique par les allèles S dominants que par les allèles S récessifs.

Dans mon projet de thèse, j'ai d'abord étudié l'effet de la sélection équilibrante au locus S sur le polymorphisme dans les régions flanquantes afin de déterminer l'ampleur du pic de polymorphisme et de caractériser ses propriétés moléculaires. J'ai utilisé les données de séquençage du génome entier de plusieurs populations d'*A. halleri* et d'*A. lyrata* pour déterminer spécifiquement la distance chromosomique jusqu'à laquelle l'effet du locus S peut encore être observé. J'ai observé une augmentation du polymorphisme dans les premiers 25kb autour du locus S, principalement expliquée par une augmentation de la proportion de sites polymorphes.

J'ai ensuite testé si la dominance des allèles S influence la charge génétique qu'ils accumulent. J'ai combiné une approche génomique utilisant des trios parents-descendant pour mettre en phase les haplotypes liés au locus S et comparer le nombre de mutations délétères liées aux allèles S dominants ou récessifs, avec une approche phénotypique pour mesurer expérimentalement l'impact de cette charge génétique protégée. J'ai démontré que la dominance favorise des profils contrastés de la charge génétique entre les allèles S récessifs et dominants.

Enfin, j'ai utilisé une approche de modélisation basée sur des simulations stochastiques pour prédire l'évolution du réseau de dominance entre les allèles S, en prenant explicitement en compte les interactions entre les petits ARN et leurs cibles *SCR*. Mes résultats montrent que les mutations ont des probabilités de fixation différentes selon qu'elles se produisent sur des allèles S dominants ou récessifs, et aussi selon qu'elles touchent le locus producteur de petits ARN ou ses sites cibles. La distribution de la charge génétique abritée entre les allèles S dominants et récessifs est également un déterminant important de l'évolution du réseau de dominance.

Remerciements

Je remercie tout d'abord les membres du jury pour la lecture de ce manuscrit, en espérant qu'il a été aussi plaisant de le lire qu'il a été pour moi de l'écrire.

Ce travail a été financé par une subvention du Fonds France-Berkeley à VC et RN, le Conseil Européen de la Recherche (projet NOVEL, subvention #648321), ANR TE-MoMa (subvention ANR-18-CE02-0020-01). Je remercie l'ERC et l'Université de Lille pour le financement de mon projet de thèse.

Je remercie mes encadrants pour m'avoir fait confiance, suivie et soutenue durant ces courtes années. Je n'oublierais pas ces discussions du midi avec Vincent Castric, qui m'a aiguillé sans contrainte tout le long de cette thèse et qui a toujours fait de son mieux pour me soutenir et m'apprendre le maximum dans le minimum de temps permis par son agenda. Tu vas aussi beaucoup me manquer. Je suis très reconnaissante envers Xavier Vekemans pour s'être autant investi et pour son enthousiasme malgré un emploi du temps parfois très chargé. Enfin, je souhaite témoigner de ma gratitude envers Eléonore Durand qui m'a soutenue et encouragée même lors des moments de doutes.

Je remercie évidemment l'ensemble des collaborateurs à ce projet, sans qui rien n'aurait été possible : merci à Mathieu Genete pour tous ces moments passés à développer et/ou debugger les pipelines et les softwares utilisés (ou pas) lors de cette thèse, mais aussi pour ces moments plus conviviaux passés ensemble. Tu resteras pour moi le superman du labo. Merci aussi à ma bonne fée, Christelle Lepers-Blassiau pour avoir résolu tous mes problèmes de biologie moléculaires et mes doutes d'un coup de pipette magique. Chloé Ponitzki, Eric Schmitt et Nathalie Faure pour s'être occupés aussi bien de mes bébés, même des plus rebelles, durant ces trois années. Merci à Sylvain Billiard pour m'avoir aidé avec son éternelle franchise rafraîchissante pour la conception intellectuelle et factuelle des modèles développés pour cette thèse. Je remercie Camille Roux pour son aide et ses conseils, toujours prodigués dans la bonne humeur. Je remercie enfin mes autres collaborateurs: S. Wright, R. Nielsen, M. Takou, J. De Meaux et B. Mable.

Je remercie aussi mes collègues thésard(e)s/postdoc rencontrés au cours de cette thèse: les anciennes, Mathilde et Béné, pour leur accueil chaleureux qui a rendu l'arrivée ici plus facile, Estelle et Thomas pour les sorties et les plus importants, Rita, Flavia, Emilie, Agathe, Zoé, et François pour le dynamisme et la joie de vivre qu'ils m'ont apportées. Je ne doute pas que les plus jeunes soient bien entourés avec vous, et je vous souhaite le meilleur pour les années à venir. Profitez un max, elles passent trop vite ces années.

Je remercie, bien entendu, ma famille, ma belle-famille et mes amis pour leurs soutiens et pour s'être autant intéressée à mes travaux. Merci Bernard et Béa pour ce super pointeur de pro!

Je remercie aussi cette personne de l'ombre, Christopher Sauvage, qui m'a aidée à croire en mes compétences pour devenir chercheuse et m'a orientée sur ce sujet de thèse il y a quatre ans. Comme un jour tu m'as dit qu'il n'est pas toujours facile de trouver de la reconnaissance alors je te le dis haut et fort: MERCI!!!

Je remercie enfin Nicolas Pothin pour tout l'amour qu'il m'apporte, pour m'avoir suivie dans cette aventure dans le Nord, pour m'avoir supportée les jours plus sombres et pour tous les moments encore à venir. Ma plus belle réussite restera toujours de t'avoir trouvé.

Je souhaite dédier ce manuscrit à Catherine Le Veve Salabert, disparue trop tôt, à l'aube de cette aventure. Personne d'autre que toi n'aurait autant voulu être là aujourd'hui et chaque jour, une pensée pour toi s'est envolée dans l'espoir de t'atteindre. Malgré le temps qui passe, j'ai toujours autant besoin de toi.