Reconciling conflicting selection pressures in the plant collaborative non-self recognition self-incompatibility system

Amit Jangid[1], Keren Erez[1], Ohad Noy Feldheim[2], Tamar Friedlander[1]

[1]The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, Faculty of Agriculture, food and environment, The Hebrew University of Jerusalem, Rehovot, Israel.

[2]Einstein Institute for Mathematics, The Hebrew University of Jerusalem, Jerusalem, Israel.

November 8, 2024

Abstract

How do complex biological systems reconcile conflicting selection pressures acting on their components? We study this question in the context of the collaborative non-self recognition self-incompatibility, a system that inhibits self-fertilization in plants. Self-incompatibility relies on specific molecular recognition between proteins, expressed in female and male reproductive organs. These proteins co-evolved to enable fertilization between individuals and disable self-fertilization simultaneously. We study the impact of these opposing selection pressures on the amino acid frequencies in these proteins' recognition domain. We construct a theoretical framework enabling promiscuous recognition between proteins and multiple partners each, as found empirically, and employ stochastic simulations. We find asymmetric responses to selection affecting mostly the female, but not the male protein composition. Using large deviations theory, we well-approximate the simulated frequencies and find agreement with genomic data. We expect our methodology to be broadly applicable to analyze the balance between conflicting pressures in additional biological systems.

Introduction

Hermaphroditic flowering plants are at high risk of self-fertilization, which would produce less fit offspring (known as "inbreeding depression" [1]). Hence, more than 100 flowering plant families have developed various mechanisms to avoid self-fertilization, generally called "self-incompatibility" [2, 3, 4, 5, 6, 7, 8, 9, 10]. Under these mechanisms, the species is sub-divided into multiple 'types' or 'classes', such that a pollen grain cannot fertilize a maternal plant of its type. The type is encoded by a single highly diverse genomic region called S-haplotype that encodes both male and female type-specifying genes.

Here we focus on a particular self-incompatibility mechanism found in several large plant families, including some common crops: the Solanaceae (tomato, potato, tobacco, Petunia), Rosaceae (apple, pear, loquat), and Plantaginaceae (snapdragon). This mechanism relies on non-self-recognition (NSR), reminiscent of an immune system. Namely, fertilization is by default disabled, and only if the incoming pollen is positively identified as having a non-self type fertilization is unlocked. Thus, non-self-recognition requires each individual pollen to be equipped with multiple identifiers, to allow it to be recognized by multiple non-self female types (Fig. 1a).

The female protein in this mechanism is a cytotoxic RNase [11], and the male protein is an

F-box protein termed SLF [12, 13]. To enable fertilization, the RNase molecules in the pollen-receiving plant must recognize at least one of the incoming pollen's SLFs, which later leads to the degradation of the RNase. In this case, we say that the paternal S-haplotype (pollen) is compatible with the maternal S-haplotype (pollen recipient). Otherwise, the RNases arrest the pollen tube growth, and fertilization is inhibited, in which case we say that the pollen is incompatible with the recipient.

Multiple SLF genes are encoded in the S-haplotype to be collaboratively recognized by various non-self RNases and allow for a sufficient number of mating partners [14]. To avoid self-fertilization, an S-haplotype must not contain an SLF gene encoding for a protein that is recognized by its own (self) RNase (Fig. 1b). A plant carrying an S-haplotype that includes such an SLF is capable of self-fertilization and considered 'self-compatible' (Fig. 1c). To avoid an accidental insertion of an SLF gene that could cause self-compatibility, and preserve an SLF combination capable of fertilizing various female types, the S-haplotype genes are tightly linked, i.e. recombination is nearly suppressed in this genomic region. This mechanism was termed the 'collaborative non-self recognition' (CNSR) self-incompatibility [14].

This system requires molecular recognition that could distinguish between different molecular types, but how exactly the specificity of different RNase and SLF proteins is encoded remains elusive [15]. Detailed mapping of the molecular recognition between various protein pairs uncovered that an SLF protein could detoxify a variable number of distinct RNases [14, 16]. A combination of empirical and computational methods [17, 18, 19, 20, 21, 22] revealed that very few residues determine these protein specificities. It was demonstrated in one particular case that replacing only three residues in an RNase was sufficient to modify its compatibility phenotype [18] and computational studies estimated that only 10-20 residues determine the RNase [19, 23] and SLF specificities [22, 24, 25]. Given that these protein specificities are governed by very few amino acids, it remains unclear how they reconcile the multiple and possibly conflicting selection pressures, to avoid self-compatibility and enhance cross-compatibility. Addressing this question requires an explicit model of the molecular recognition implemented by these proteins. Yet, previous studies [26, 27] employed simplified allelic descriptions that were insufficient for this purpose.

We have recently proposed a novel theoretical framework that incorporates the molecular recognition between male and female type-specifying proteins into the evolutionary model to study the emergence and maintenance of mating specificities in this self-incompatibility system [28] (see Fig. 1). Notably, our approach does not require the definition of arbitrary notions of fitness, but rather derives fitness from the compatibility phenotype which directly depends on interaction energies between male and female type-specifying proteins. This modeling framework crucially relies on interaction promiscuity and facilitates many-to-many interactions between the female and male type-specifying proteins, in agreement with empirical evidence [14, 16]. Using this framework, we found that most S-haplotypes in our model spontaneously self-organize into 'compatibility classes', such that members of each class are incompatible with each other but compatible with all members of all other classes, as found empirically. We demonstrated a dynamic balance between class birth and death in this model, reaching a dynamically stable equilibrium in the class number and researched the mutational trajectories by which class birth and death occur. These desired properties of the model and, in particular, the explicit inclusion of molecular recognition render it highly suitable to study how the selection pressures shape these proteins'

specificities.

Here we employ this model to research the balance between the multiple selection pressures acting on the male and female type-specifying proteins in this system, as reflected by the amino acid frequencies in their specificity-determining residues. We analyze the amino acid frequencies of single amino acids in the RNase and SLF separately and the joint frequencies of corresponding amino acid pairs in both. Using Sanov's Theorem, we derive the amino acid frequencies as the maximizers of the entropy under a self incompatibility constraint. We demonstrate that a reduced model with only selection against self-compatibility closely approximates the frequencies in the full model. This result highlights the dominance of selection against self-compatibility in the evolution of this system. To disentangle the impacts of attractive vs. repulsive selection pressures we distinguish between same-S-haplotype and cross-S-haplotype gene pairs. We recover two strategies these sequences employ in response to selection pressures: the 'global' strategy reflected by biases in single amino acid frequencies, and the 'local' strategy that entails specific geometric adjustment to particular partner sequences. We show that avoidance of self-compatibility is universally obtained by the global strategy, as each RNase needs to dodge multiple SLFs. In contrast, compatibility with partners is often obtained via a local strategy by matching the locations of positively and negatively charged amino acids, especially when the interaction is non-promiscuous.

Results

The model

We assume a population of N S-haplotypes, each consisting of a single RNase and multiple SLF genes (Fig. 1a). Each of the proteins encoded by these genes is represented by a sequence of L amino acids, that could belong to either of four biochemical classes: hydrophobic H, neutral polar P, positively charged +, or negatively charged −. An RNase-SLF pair is considered matching if the interaction energy between their interfaces is below a threshold value E_{th}. This interaction energy between an RNase R_i and SLF F_j is defined as the sum of the pairwise interaction energies between the corresponding amino acids (see Table 1) of these proteins,

$$E^L_{\text{tot}}(R_i, F_j) = \sum_{\ell=1}^{L} E[R_i(\ell), F_j(\ell)], \tag{1}$$

where $R_i(\ell), F_j(\ell) \in \{\text{H}, \text{P}, +, -\}$. If $E^L_{\text{tot}}(R_i, F_j) < E_{\text{th}}$, these proteins are considered matching, namely fertilization of a maternal S-haplotype carrying the RNase R_i by a paternal S-haplotype carrying the SLF F_j is enabled (Fig. 1d). Hence, the S-haplotype composition determines its possible mating partners. In particular, an S-haplotype is termed 'self-incompatible' if it does not contain any SLF that matches its own RNase (Fig. 1b) and 'self-compatible' if it does (Fig. 1c). Self-compatible S-haplotypes can be fertilized by either self or non-self pollen, whereas self-incompatible ones can only be fertilized by non-self pollen.

Each individual produces both male and female gametes, but the male gametes typically outnumber the female ones. This natural asymmetry in abundance is captured in our model by the number of fertilization attempts $k > 1$ per female. We assume a well-mixed population with no spatial structure, such that every female is equally likely to encounter pollen from any other

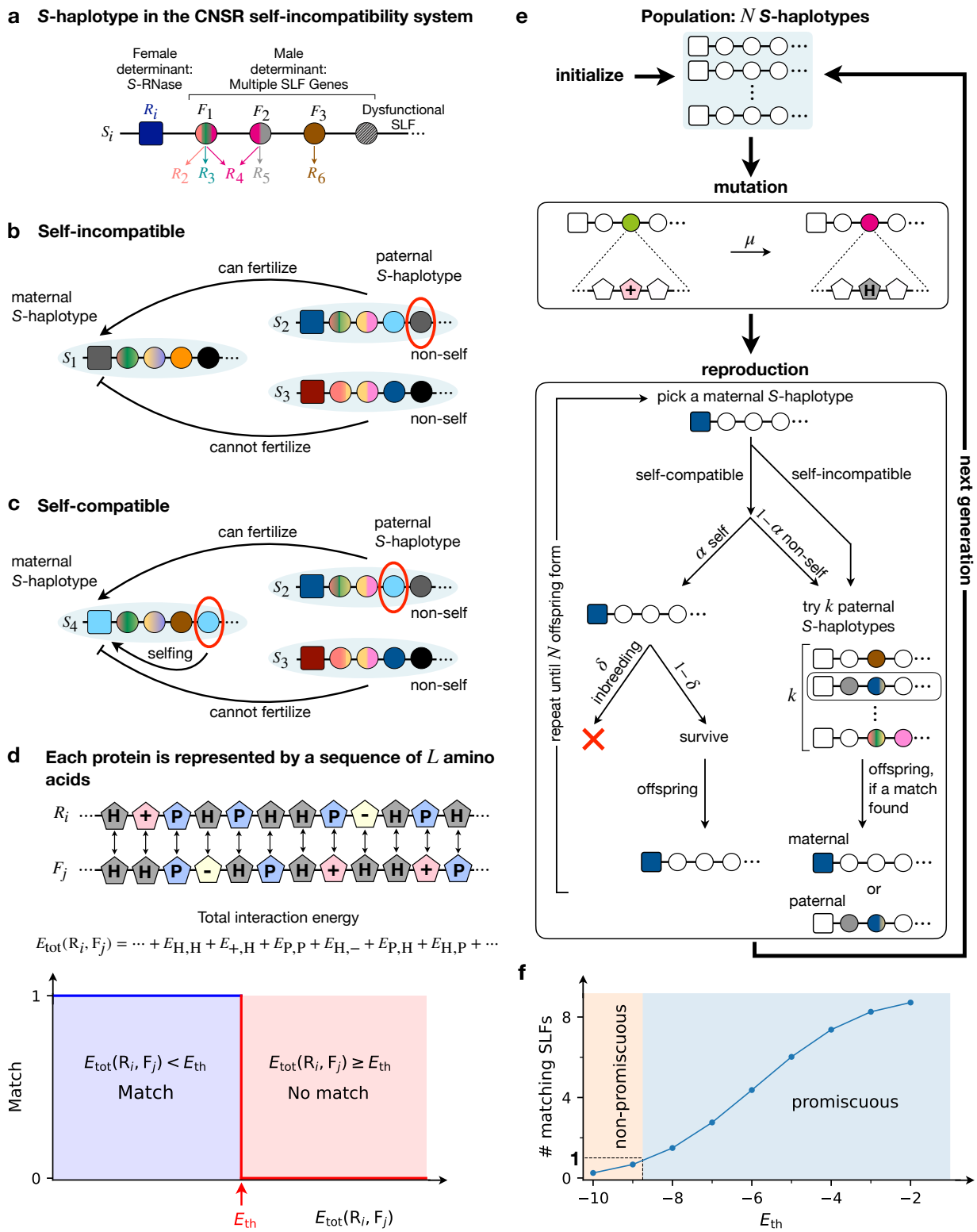


Figure 1

population member. Every female gamete receives both self-pollen and non-self-pollen originating uniformly from any other population member. The results below refer to the regime of pollen abundance $k \gg 1$ in which a female gamete is essentially guaranteed to be fertilized, and only the male gametes compete for matching female ones.

The population life-cycle is illustrated in Fig. 1e. We initialize a population of N S-haplotypes, constructed as shown in Fig. 1a,d. We assume non-overlapping generations. Every generation,

Figure 1 *(previous page)*: **Model description**. **(a)** Each S-haplotype consists of a single female type-specifying gene (RNase R_i, square) and multiple male-type-specifying genes (SLFs $F_1, F_2, ...$, circles). Each SLF could match one or several different RNases or none. **(b)** For a paternal S-haplotype to successfully fertilize a maternal S-haplotype it must be equipped with an SLF matching the maternal RNase. A maternal S-haplotype can be fertilized by either self (genetically identical) or non-self pollen. An S-haplotype is considered 'self-incompatible' if it does not contain an SLF matching its own RNase, rendering fertilization by self-pollen is impossible, and 'self-compatible' **(c)** if it does contain such an SLF facilitating fertilization by self-pollen. **(d)** Each protein encoded in the S-haplotype is represented by a sequence of L amino acids of four biochemical classes. The total interaction energy E^L_{tot} between an RNase R_i and each SLF F_j is defined as the sum of the pairwise interaction energies between their corresponding amino acids. Two such proteins are considered matching, if E^L_{tot} is smaller than an energy threshold E_{th}, and otherwise they are non-matching. **(e)** The population life cycle implemented in our stochastic simulation. The population consists of a fixed number N of S-haplotypes, as shown in (a). Every generation, each of the amino acids in each of the proteins can mutate with probability μ. We assume non-overlapping generations. Every generation, maternal S-haplotypes are picked at random for reproduction. If the maternal S-haplotype chosen is self-compatible it can be fertilized by either non-self-pollen (with probability $1 - \alpha$) or by self-pollen (with probability α), where self-fertilization offspring survive with probability $1 - \delta$ relative to offspring formed by non-self fertilization. Alternatively, if the maternal plant is self-incompatible only non-self pollination is possible. Under non-self pollination, the maternal plant is given up to k attempts to find a compatible paternal S-haplotype, by randomly picking S-haplotypes from the entire population. If a match is found, an offspring inheriting either the paternal or the maternal S-haplotype is generated. **(f)** We consider 'promiscuity' as the state that on average more than one SLF per S-haplotypes matches an arbitrary RNase. Here we plot the mean number of matching SLFs (assuming 9 SLFs per S-haplotype, our default value), against E_{th}, where all sequences are randomly drawn from the prior distribution. We find that the system is promiscuous for $E_{\text{th}} \geq -9$ (see also Table 3).

each of the amino acids in each of the proteins can be mutated with some probability. The new values of amino acids chosen to be mutated are randomly drawn from the prior amino acid distribution (Methods). In each generation, the offspring population is produced one by one from the parental one by repeating the following procedure. An S-haplotype is selected uniformly at random from the parental population to serve as a maternal parent. If the maternal S-haplotype is self-compatible, with probability α it is self-fertilized (here α corresponds to the proportion of self-pollen out of the total pollen received). If self-fertilization did not occur (whether or not the S-haplotype is self-compatible), the maternal S-haplotype is given k opportunities for fertilization by cross-pollination. In each of these, a random S-haplotype of the parental population is selected and its compatibility as male with the maternal parent is tested. If it is found compatible, it serves as a paternal parent, and an offspring S-haplotype is formed. This offspring is an identical copy of either of its two parents, with equal probability. To represent the deficiency of offspring produced via self-fertilization, if an offspring was produced via self-fertilization it is eliminated with probability δ, a parameter representing the inbreeding depression intensity. Hence every such round can either result in a single offspring, or none (due to failure of all fertilization attempts, or elimination of an offspring produced via self-fertilization). This procedure is repeated until a population of N offspring is formed. The offspring population then replaces the parental one, completing one generation.

Following this life-cycle, the population evolves under two types of selection pressures: avoid-

ance of self-compatibility exerted by the inbreeding depression penalty on offspring produced via self-fertilization and selection for reproductive success where S-haplotypes compete on being compatible with as many (non-self) population members as possible. This model is a haploid version of the diploid model presented in [28]. Although simplified, the haploid model behaves qualitatively very similarly to the more realistic diploid model (see detailed comparison in supplementary text), and we opt for it due to its analytical and numerical tractability.

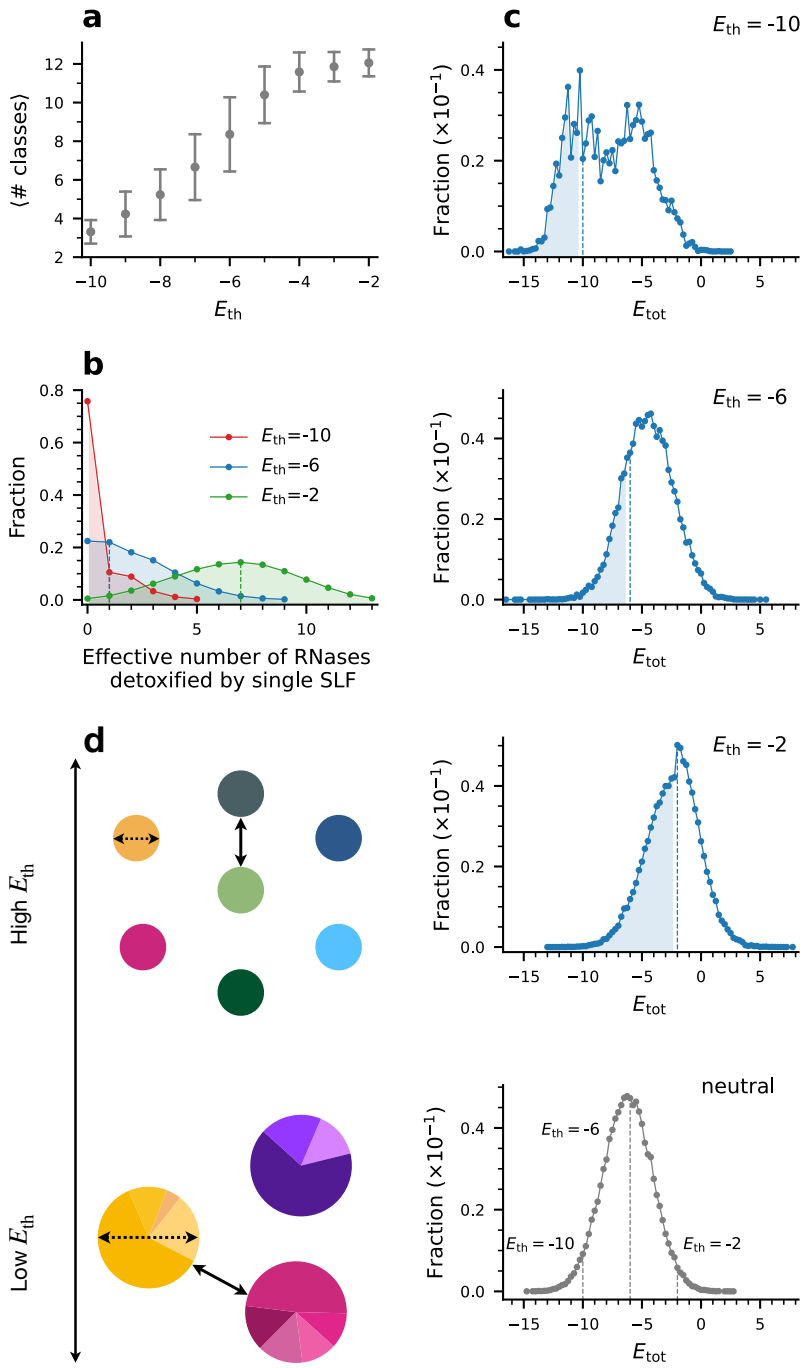


Figure 2

Figure 2 *(previous page)*: **Single SLFs can potentially interact with multiple distinct partners, and the number of partners increases with the interaction energy threshold E_{th}.** This figure shows steady-state characteristics of the population following evolution via our model. **(a)** The mean number of classes increases with E_{th}. Classes are defined such that every class member is compatible (as male and as female) with all members of all other classes and incompatible with its own class members. **(b)** Distributions of the effective number of distinct RNases in the population that a single SLF could detoxify (and hence enable fertilization) for different E_{th} values. These results refer to RNases and SLFs co-evolving in the same population. We observe that the distribution mean increases with E_{th}. SLFs with no partners are considered dysfunctional, and their fraction increases as E_{th} decreases. **(c)** Distributions of RNase-SLF interaction energies at steady state amongst populations evolved under three different E_{th} values: -10, -6, -2. The shaded regions mark the population proportion of matching sequences. For reference, we also show the distribution of RNase-SLF interaction energies for neutral sequences, namely drawn at random from the prior distribution (bottom). This figure demonstrates the shift in energies following evolution. **(d)** The higher the promiscuity (E_{th}), the larger the number of classes and the between-class distances, while classes become smaller and less diverse – a schematic figure (see also Fig. S1). Parameter values used are detailed in Table 2.

The population spontaneously self-organizes into compatibility classes; class number increases with E_{th}.

A crucial property of our model is the spontaneous self-organization of almost all S-haplotypes into compatibility classes, such that all individuals in the same class are bidirectionally incompatible with each other and bidirectionally compatible with all individuals of all other classes [28] (See Fig. S13, and S14). This behavior underlies most of the model outcomes and is tightly linked to the promiscuity of interaction, embodied by the energy threshold E_{th}. E_{th} is a fundamental parameter of the model capturing the probability that randomly drawn proteins match (Fig. 1f, Table 3). In the following we test the model behavior under a broad range of E_{th} values, spanning match probabilities between 0.03 to 0.97. Intuitively, we consider the population state as 'promiscuous', if on average a neutral S-haplotype, i.e. drawn at random from the prior distribution, has more than a single SLF matching an arbitrary RNase. Naturally, this depends on E_{th} and the number of SLFs per S-haplotype. Considering 9 SLFs per S-haplotype (our default value) promiscuity is obtained for $E_{\text{th}} \gtrapprox -9$ (see Fig. 1f and Table 3). An additional consequence of promiscuity is the multi-partner property, which was absent in previous models. In Fig. 2a we show the number of classes obtained in evolving populations at steady-state for different values of E_{th}, representing various levels of promiscuity. We find that classes form under a broad range of E_{th} values, but for low promiscuity, these classes are fewer and larger, whereas for high promiscuity they are more numerous and smaller (Fig. 2a,d). Related to that, we find that the multi-partner property too persists under a broad range of E_{th} values. In Fig. 2b we show the distribution of the number of RNase partners per SLF in the same populations as in Fig. 2a for different E_{th} values. We find a shift in the partner number distribution, such that for high promiscuity the number of partners is larger and for low promiscuity, it is smaller, in agreement with the total class number dependence on E_{th}. Interestingly, for intermediate-to-low promiscuity, ($E_{\text{th}} \leq -6$) a significant proportion of SLFs have no partner at all. Such dysfunctional SLFs likely survive by hitchhiking to linked functional ones.

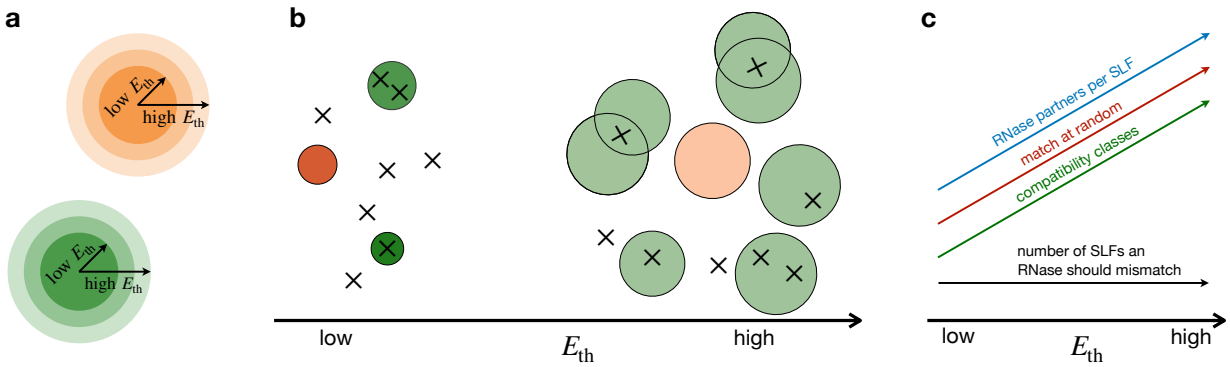


Figure 3: **Schematic description of the selection pressures acting in our model**. We use an abstract illustration of the high dimensional sequence space of the S-haplotype genes. **(a)** A colored circle represents here the set of SLF sequences matching a particular RNase. The size of this circle increases with E_{th}. The orange circle represents the self-RNase that should be avoided and the green circle represents non-self RNase that should be matched. **(b)** The suite of S-haplotype SLFs, each marked graphically by X, should all avoid binding their self-RNase (avoid the orange circle) and collectively bind all the non-self RNases (have at least one representative in each green circle). The size of the all these circles increases with E_{th}, hence overlaps (an SLF that matches multiple RNases) become more common. Some SLFs may match none of the RNases (are in none of the circles) and are considered dysfunctional. **(c)** A schematic summary of the dependence of difference factors in our model on E_{th}.

Avoidance of self-compatibility governs the amino acid composition of the RNases, while the SLF composition remains nearly neutral.

RNase and SLF genes evolve under a combination of two types of selection pressures, acting simultaneously to avoid within-S-haplotype but promote between-S-haplotype compatibility among RNase and SLF pairs - see Fig. 3 for schematic description. In Fig. 2c we illustrate the distributions of interaction energies in populations evolved under different E_{th} values, compared to neutral sequences. We observe, that the distribution shifts to higher energies (lower match) under higher E_{th}, suggesting that selection affects the sequence composition. As the likelihood of a coincidental match increases with E_{th}, the pressure to repel should increase and conversely the pressure to attract decrease with E_{th}.

In the following section, we investigate the molecular underpinnings of the shift in interaction energies propelled by selection. We begin by assessing the RNase-SLF match probability and the amino acid frequencies of either one and quantify their deviation from the neutral case using an information theoretic measure of distance between distributions. We find that in most of the E_{th} range selection lowers the RNase-SLF match probability relative to neutrality and biases the RNase AA frequencies. We then demonstrate that a reduced model with only selection against self-compatibility provides a good approximation to the AA frequencies in the full model.

Since the S-locus genes are tightly linked, the entire S-haplotype is the heritable evolutionary unit. As such, it should reconcile the evolutionary pressures acting on its constituent genes. For example, for an S-haplotype to match a particular RNase, it is sufficient that only one of its SLFs does so, thus forming a synergy between the same-S-haplotype SLFs. In particular, dysfunctional SLFs as observed in Fig. 2b can survive by 'hitchhiking' their neighboring functional ones.

To assess the net pressure, we begin by examining the match probabilities between evolved RNase-SLF pairs, for different E_{th} values, compared to neutral sequences. The evolved sequences were obtained in simulation following the life cycle of Fig. 1 and were extracted after an initial

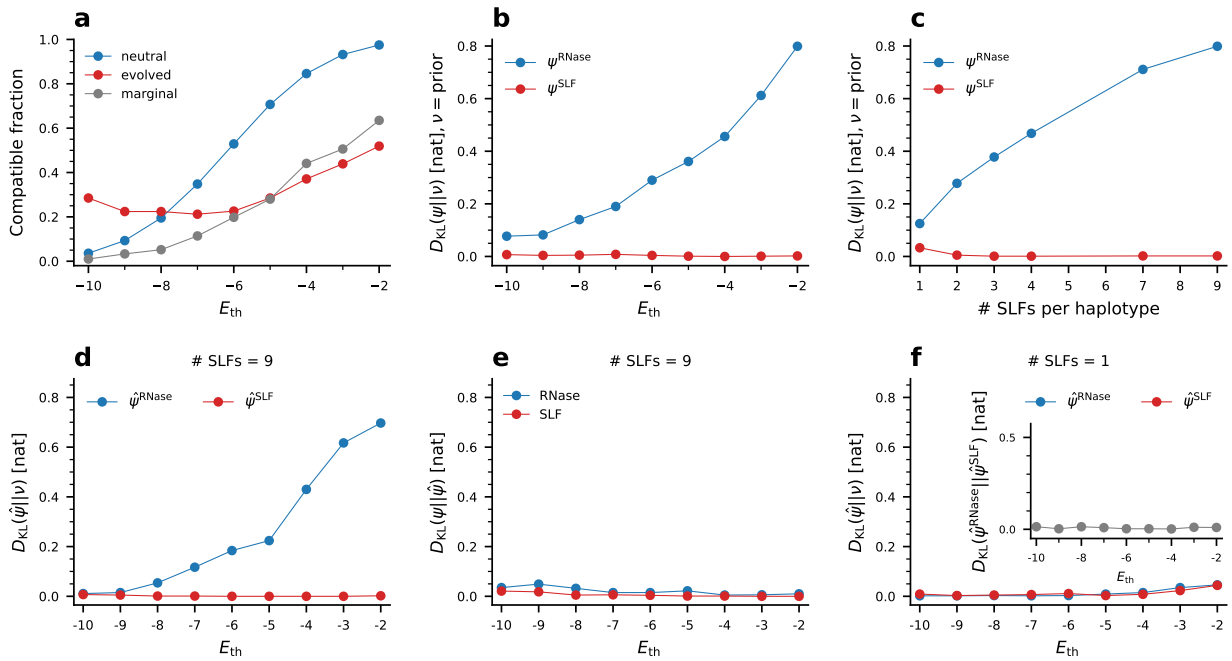


Figure 4: **Avoidance of self-compatibility, biases the amino acid composition for RNases, but not for SLFs. (a)** Compatible fraction amongst RNase-SLF pairs against E_{th} under three scenarios: neutral sequences drawn from the prior AA frequencies (blue), functional RNase and SLF sequences evolved in simulation (red), and random sequences drawn from the marginal AA frequencies of the evolved RNases and functional SLFs (gray). To calculate the neutral and marginal frequency cases, we drew a total of 50 RNases, and for each, we drew 1500 SLFs. **(b-c)** The Kullback-Leibler divergence between the marginal AA distribution of evolved sequences and the prior AA distribution for RNases (blue) and functional SLFs (red) as a function of E_{th} (b) or as a function of the number of SLFs per S-haplotype with fixed E_{th} =-2 (c). The calculations are based on 25 independent runs, using 600 data points from each run with 25 generation intervals between consecutive time points (Methods). **(d-f)** Simulation results of the reduced model including only selection against self-compatibility, with no reproductive cycle. **(d)** The divergence between the marginal distributions of RNases (blue circles) and SLFs (red circles) evolved under the reduced model relative to the prior, shows a similar pattern to (b) - where the RNase deviates from the prior, but the SLF does not. **(e)** The divergence between the marginal distributions evolved under the full (ψ) and reduced ($\hat{\psi}$) models, is close to zero, demonstrating that the reduced model is an excellent approximation in the entire range of E_{th}. **(f)** Similar to (d) but with only one SLF per S-haplotype, we obtain full symmetry between the RNase and SLF marginal distributions. The inset shows the divergence between the RNase and SLF marginal distributions for the reduced model, again demonstrating the full symmetry in their behaviors. Parameter values are as is Table 2 except for variable numbers of SLFs per S-haplotype in (c) and (f).

transient time was discarded from the analysis, such that the population state is essentially independent of its initial condition and its statistical measures have already stabilized (see Fig. S2). We refer below to the distributions obtained as being at steady-state (see Methods). We find (Fig. 4a) that the evolved sequences (red) behave very differently from the neutral ones (blue), and for most of the E_{th} range are less likely to match relative to the neutral case. We conclude that for a great part of E_{th} values, a match at random is highly expected and the net selection pressure operates to reduce it.

We denote the prior distribution from which neutral sequences are drawn by $\nu(i)$, $i \in \{+, -, H, P\}$, the marginal distributions of single amino acid frequencies in the evolved se-

quences by $\psi^{\mathrm{RNase}}(i)$, $\psi^{\mathrm{SLF}}(i)$ respectively, and the joint distribution of amino acid pairs in the same sequence position in RNase and SLF located on the same-S-haplotype by $\varphi(i,k), \quad i,k \in \{+,-,H,P\}$.

Our mutation mechanism replaces amino acids at random with those drawn from the prior distribution. Hence, if mutations are repeatedly applied without any selection pressure, the marginal distributions $\psi^{\mathrm{RNase}}(i)$, $\psi^{\mathrm{SLF}}(i)$ should follow the prior distribution $\nu(i)$. Since mutations occur independently in the RNase and SLF, their joint AA distribution φ would then be the product of these two marginals, namely $\varphi(i,k) = \nu(i) \times \nu(k)$. Thus, deviations of either ψ or φ from the prior distribution ν are a signature of selection pressure operating beside the mutations.

To quantify the strength of this evolutionary pressure, and test how it is distributed between the RNase and SLF sequences, we employ the Kullback–Leibler divergence D_{KL} (a.k.a relative entropy). D_{KL} is an information-theoretic measure of divergence between two distributions $P(x)$ and $Q(x)$, defined as [29]

$$D_{\mathrm{KL}}(P||Q) := \sum_x P(x) \log \frac{P(x)}{Q(x)},$$

and measured in units of nats (natural bits). We calculated $D_{\mathrm{KL}}(\psi^j(i)||\nu(i)), j \in \{\mathrm{RNase}, \mathrm{SLF}\}$ the divergence of each marginal distribution from the prior $\nu(i)$ for different levels of promiscuity E_{th} (Fig. 4b). Since a match at random between RNase and SLF is more likely under higher promiscuity, we expect the selection pressure for match avoidance to intensify with E_{th}. Indeed, we observe a larger deviation from the prior as E_{th} increases for the RNase distribution ψ^{RNase} but not for the SLF distribution. The latter remains close to the prior for the entire E_{th} range, namely $D_{\mathrm{KL}}(\psi^{\mathrm{SLF}}(i)||\nu(i)) \approx 0$.

Fig. 4a suggests that in most of the range ($E_{\mathrm{th}} > -8$) the dominant selection pressure affecting AA frequencies is self-compatibility avoidance. To further validate this hypothesis, we ran an additional simulation of a reduced model in which we only applied selection pressure to avoid self-compatibility, where individuals were regularly mutated and every self-compatible mutant was discarded, but no reproductive cycle was applied. We denote the marginal AA distributions in the reduced model by $\hat{\psi}^j(i), j \in \{\mathrm{RNase}, \mathrm{SLF}\}$, respectively. Similar to the full model (Fig. 4b), we found that in the reduced model too, only the RNase AA frequencies $\hat{\psi}^{\mathrm{RNase}}$ were biased, but the SLF AA frequencies closely followed the prior (Fig. 4d). To directly compare the two models we compute the divergence of the full simulation with respect to the reduced model (Fig. 4e) and observe that it is less than 0.05 nat per AA base, so that the two models are almost in perfect agreement. This similarity between the behaviors of the two models again confirms that the net selection pressure here acts to avoid self-compatibility whereas the selection to match non-self is only secondary.

If the selection pressure to avoid within-S-haplotype matches dominates, we also expect it to increase with the number of SLF genes per S-haplotype. Indeed, in Fig. 4c we illustrate that $D_{\mathrm{KL}}(\psi^{\mathrm{RNase}}||\nu)$ between the evolved and prior distributions also increases with the number of SLF genes per S-haplotype for fixed E_{th}. Again, we observe that the RNase bears the vast majority of the selection pressure with little impact on the SLF sequences. We also simulated selection to avoid self-compatibility alone in a symmetric setting of one RNase and only one SLF per S-haplotype (Fig. 4f). Here, for the first time, we observe a symmetric response to selection, where both RNase and SLF AA distributions $\hat{\psi}^{\mathrm{RNase}}$, $\hat{\psi}^{\mathrm{SLF}}$ are equally biased relative to the prior distribution $D_{\mathrm{KL}}(\hat{\psi}^{\mathrm{RNase}}||\nu) \approx D_{\mathrm{KL}}(\hat{\psi}^{\mathrm{SLF}}||\nu)$.

The balance between the selection pressures to attract and repel also depends on the pollen-to-ovule ratio, represented in our model by the number of fertilization attempts k per female. Under pollen abundance, RNase evolution is almost entirely governed by the pressure to avoid self-compatibility. Pollen limitation (low k), gives rise to an additional selection pressure on the RNase to match SLF partners, while the previous pressures on SLFs to match, and on both to avoid self-compatibility, persist. This tips the balance between the attraction and repulsion pressures. The results above were obtained under pollen abundance. We found however that only for very low values of $k = 1, 2$ the amino acid content differs, whereas for higher $k > 5$ it is insensitive to k (Fig. S9).

The amino acid frequencies evolve to minimize the divergence from the prior distribution under the constraint of no self-compatibility.

We found that the dominant selection pressure governing AA composition is avoidance of self-compatibility, but why would it act almost exclusively on the RNases with only little effect on the SLFs? At first glance, this seems non-intuitive, given that under pollen abundance as simulated here, selection to match acts mostly on the SLFs and much less on the RNases receiving multiple fertilization attempts each. Thus, one may naively expect that the SLFs should be the ones changing their AA content in response to selection, whereas the RNases will evolve almost neutrally. Below we introduce a theoretical argument using large deviations theory to establish that self-compatibility avoidance alone well-approximates the amino acid frequencies in the S-haplotype and explains the asymmetric response to selection in the RNase and SLF.

Recall, that in the absence of selection or any other constraint, when the process is solely governed by mutations drawn from the prior distribution $\nu(i)$, the stationary pairwise AA distribution $\varphi(i, k)$ follows $\varphi(i, k) = \nu(i) \times \nu(k)$. Here we analyze this system subject only to self-incompatibility elimination constraint.

Fixing E_{th}, we analyze the stationary pairwise AA distribution via energy-entropy considerations, applying large deviations theory. A balance between two forces determines the simulated AA distribution. The first is the entropy of the system, namely the number of sequences with a given AA distribution. The second is the long term survival probability of a sequence, namely the long term probability that mutations will not induce self-compatibility. Large deviations theory tells us that for such systems, when L, the sequence length, tends to infinity, the AA composition will be close, with high probability, to the distribution which maximizes the product of the number of configurations and the survival probability. This phenomenon is known as the energy-entropy maximization principle [30]. While this product is analytically intractable, we can well approximate the survival probability by a monotone function of the energy of the system, as higher energy configurations require a larger number of mutations, before they become self-compatible. In such a setting, the energy-entropy maximizing distribution could also be recovered as the distribution which maximizes only the entropy (i.e., the number of configurations), subject to a harsher effective self-compatibility energy threshold. A decoupling argument justifying this appears in the appendix. Given this effective energy threshold, such a maximizer could be analytically computed.

The main theoretical tool from large deviations theory which justifies this reasoning is Sanov's Theorem [30]. Denote by Γ the set of joint RNase-SLF AA distributions such that the total interaction energy between an RNase and each of the m SLFs drawn from this distribution,

$E^L_{\text{tot}}(R_i, F_j)$, is greater than E_{th}, the energy threshold, scaled to preserve energy per position. Formally, the theorem states that the probability $\mathbb{P}$ that S-haplotype sequences randomly drawn according to a pair distribution $\varphi(i,k)$ are not self-compatible with respect to E_{th}, converges as L tends to infinity to satisfy

$$\frac{1}{L}\log \mathbb{P}(\varphi \in \Gamma) \to \sup_{q\in\Gamma} \sum_{a,b_1,\ldots,b_m\in\{+,-,H,P\}} -q(a,b_1,\ldots,b_m)\log\frac{q(a,b_1,\ldots,b_m)}{\nu(a)\nu(b_1)\cdots\nu(b_m)}. \tag{2}$$

Moreover, the sampled distribution itself converges in probability to the maximizer of this expression. By convexity and symmetry arguments, we can express the right-hand side in terms of ψ and $\varphi^{(\text{RNase,SLF})}$, as minimizing

$$\sum_{a\in\{+,-,P,H\}} \psi^{\text{RNase}}(a)\log\frac{\psi^{\text{RNase}}(a)}{\nu(a)} + m\times \sum_{a,b\in\{+,-,P,H\}} \varphi^{(\text{RNase,SLF})}(a,b)\log\frac{\varphi^{(\text{RNase,SLF})}(a,b)}{\psi^{\text{RNase}}(a)\nu(b)} \tag{3}$$

under the constraints that $\varphi \in \Gamma$, i.e., the interaction energy induced by these AA pair frequencies is above E_{th}, and that ψ^{RNase} is the marginal of $\varphi^{(\text{RNase,SLF})}$. Further entropic considerations allow us to predict the pairwise AA frequencies in the simulated model when avoiding self-compatibility is the only selection pressure at play. These frequencies are the minimizer of Eq. (3) for suitably calibrated E_{th} which we denote by $\tilde{E}_{\text{th}}$. These predictions should be valid regardless of the inbreeding depression δ, which is only accounted for by altering the effective energy threshold $\tilde{E}_{\text{th}}$ (see appendix). This should persist as long as avoiding self-compatibility is the main evolutionary pressure at work, indicating that in this regime AA frequencies should depend smoothly on E_{th} and δ (see Fig. S7).

In Fig. 5 we compare the AA frequencies obtained in the simulations of the full and reduced models to the prediction obtained by numerical minimization of Eq. (3) for the entire range of E_{th} simulated. Except for the lowest promiscuity, the two models yield very similar AA frequencies, and Eq. (3) provides an excellent approximation to the reduced model. Both models predict a depletion of hydrophobic and an enrichment of charged AAs in the RNase, where these trends increase with E_{th}. We compared these theoretical predictions to the AA frequencies of RNase residues computationally estimated as being under positive selection [24] and likely serving as specificity-determinants (Fig. 5, shaded region in each panel). In this empirical data too we find a similar trend with under-representation of hydrophobic and over-representation of charged AAs, in agreement with our theoretical prediction.

To understand the asymmetric response to selection of the RNase and the SLFs, observe that in Eq. (3) every modification in the frequencies of the SLFs is m times as costly as its counterpart for the RNase. This is because a single RNase has to avoid compatibility with m SLFs. As a result, to minimize the entropic cost of avoiding self-compatibility, the RNase AA distribution should diverge from the prior distribution significantly more than the SLF AA distribution should.

The full simulation sees an additional pressure on the SLFs to match their RNase partners. While this has little effect on the behavior of the RNases, it could, in principle, tilt the AA frequencies of the SLFs. The fact that this additional pressure does not alter the joint distribution $\varphi^{(\text{RNase,SLF})}$ in the full model simulations is further explained below.

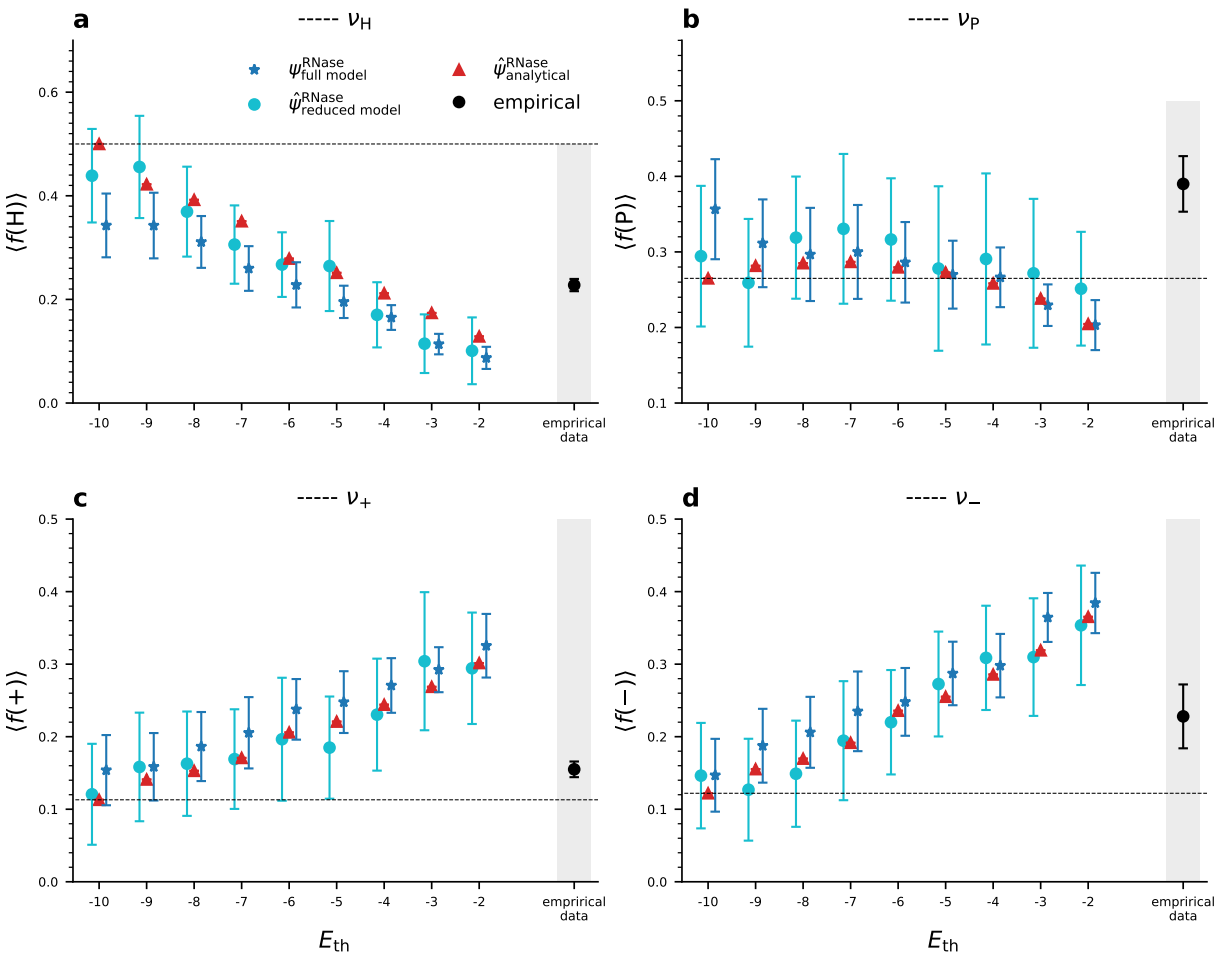
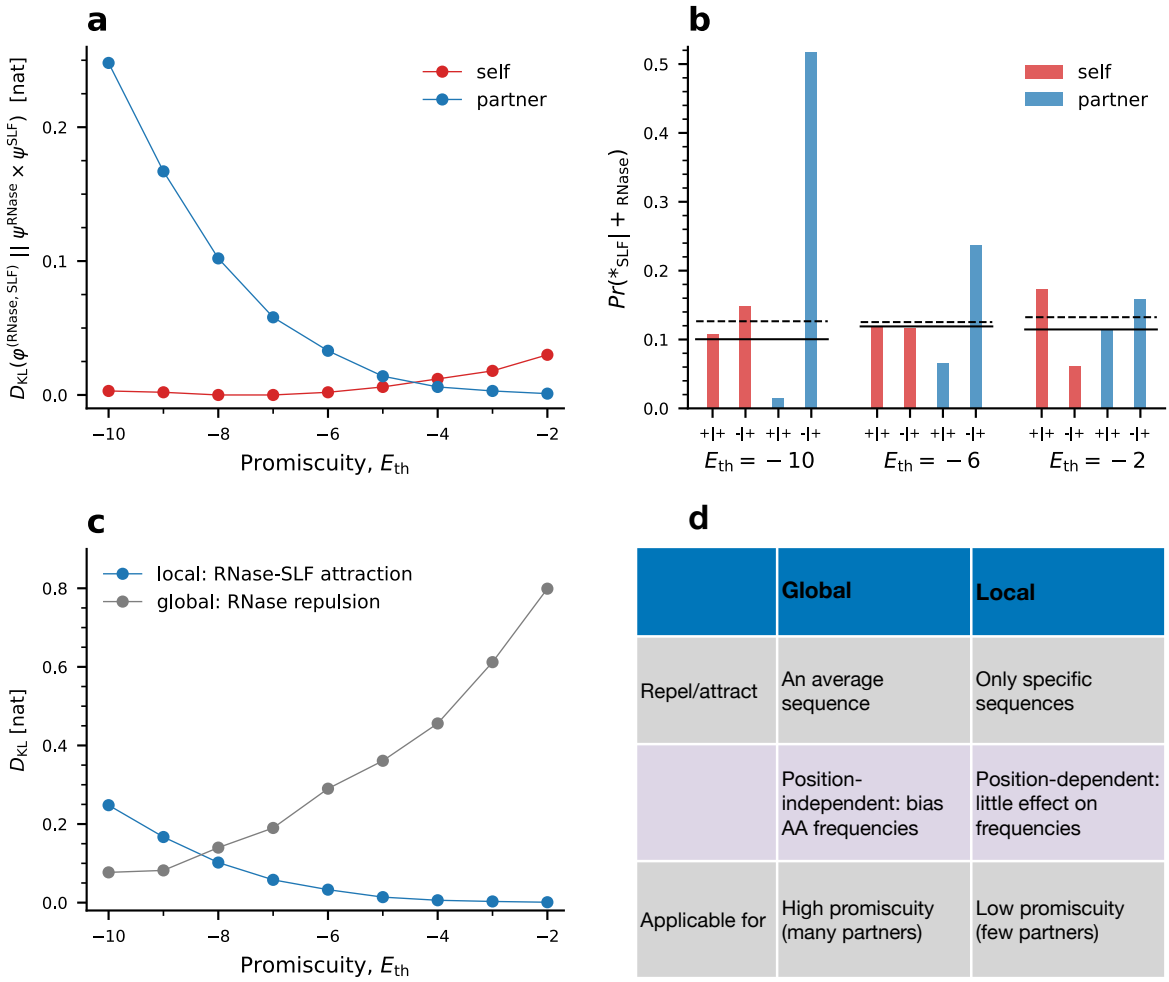


Figure 5: **The marginal frequencies of the various AA categories in RNases in the full model are well-approximated by the reduced model, and the analytical approximation is in excellent agreement with the reduced model simulations; Empirical data shows a similar trend to our theoretical predictions.** We plot the frequencies of the four AA categories (H, P, +, −) in the full model (blue star) and reduced model (cyan circle) simulations, and the analytical approximation (red triangle) for different $E_{\rm th}$ values with 9 SLFs per S-haplotype. The black circle (shaded background) in each panel shows the frequency of the corresponding AA category in empirical data of RNase residues computationally estimated to be under positive selection, [24] (Methods). The horizontal dashed lines mark the prior frequencies of the corresponding AA categories. Symbols denote averages over multiple simulation data points, and bars represent ±STD relative to the mean. The figure is based on 25 independent runs for each $E_{\rm th}$ value, 450 data points taken from each run, with a 25-generation interval between consecutive time points. All these data points were taken after the entire population descended from a single common ancestor and the amino acid frequencies had already reached their steady-state proportions. The total runtime for each $E_{\rm th}$ value is 1.5×10^5 generations. Parameter values are as in Table 2.

Matches between proteins can be implemented by either 'global' or 'local' strategies.

Evolving a sequence to match or mismatch other sequences can employ either of two basic strategies (or their combination), which we term 'global' and 'local'. The global strategy increases or decreases the sequence compatibility universally with respect to a large number of partner-sequences, by biasing its marginal AA frequencies in a position-independent manner. In contrast,



	Global	Local
Repel/attract	An average sequence	Only specific sequences
	Position-independent: bias AA frequencies	Position-dependent: little effect on frequencies
Applicable for	High promiscuity (many partners)	Low promiscuity (few partners)

Figure 6

the local strategy tailors the sequence to a particular partner in a position-dependent manner. The local strategy can generate a significant change in the interaction energy by controlling relatively few AAs and does not require a drastic change in their marginal frequencies. Hence, whenever (in)compatibility with respect to only one (or very few) partners is selected for, it is expected that a local strategy will dominate, while otherwise, a global strategy is preferable (see Fig. 6d for summary). In this section we distinguish between the contributions of the local and the global strategies to attraction and to repulsion between proteins, by analyzing separately interactions between same-S-haplotype pairs and interactions between partners from different S-haplotypes. To quantify the contribution of the local strategy we calculate the deviation of the joint RNase-SLF AA frequencies from the product of the marginal frequencies again applying the Kullback-Leibler divergence.

To quantify to which extent the gap between the compatible fractions among neutral and evolved sequences is purely due to biases in the AA frequencies ('global'), we calculated the compatible fraction among RNase-SLF pairs randomly drawn from the marginal distributions of evolved RNases and SLFs (Fig. 4a, gray curve, 'marginal'). We observe that for high promiscuity, the 'marginal' curve is close to the 'evolved'. Hence, most of the compatibility reduction of the evolved sequences relative to the neutral ones is achieved by position-independent biases in AA frequencies, namely the global strategy. Yet, for low-to-intermediate promiscuity the 'marginal' (gray curve) deviates significantly from the evolved (red curve), suggesting that the local strategy

Figure 6 *(previous page)*: **The local strategy is most useful for matches under low promiscuity when each protein has one or few partners (a)** The Kullback-Leibler divergence between the empirical joint distribution φ of RNase-SLF AA pairs evolved in simulation and the product of their marginal distributions $\psi^{\mathrm{RNase}} \times \psi^{\mathrm{SLF}}$ against the promiscuity E_{th}. This divergence quantifies the magnitude of the local strategy. We show this separately for the same S-haplotype RNase-SLF pairs (self, red) that should repel each other and for partners on distinct S-haplotypes (partner, blue). The number of partners per protein strongly depends on the promiscuity. Indeed the divergence for partners is greatest under lowest promiscuity, when partners are few, and diminishes to nearly zero under high promiscuity when partners are numerous. We observe an opposite trend with an increase of the divergence with promiscuity amongst same-S-haplotype proteins ('self'), but its relative contribution is much smaller compared to the partner case. **(b)** Frequencies of charged amino acids in SLFs conditioned on the RNase having a positively charged (+) residue in the corresponding position for RNase-SLF pairs either from the same S-haplotype (self, red) or for partners from different S-haplotypes (partner, blue) under three different E_{th} values. The horizontal lines show for reference the marginal SLF frequencies ((+) solid lines, and (−) dashed lines), where deviation from these marginals is a signature of the local strategy. Same-S-haplotype RNase and SLF are under selection pressure to repel each other. Indeed, we find there an enrichment of same-charge AA and an under-representation of opposite-charge AA pairs relative to the marginal frequencies. We find the opposite trend for RNase-SLF pairs that are partners and consequently attract each other. The relative role of attraction and repulsion pressures also depends on the value of promiscuity. In the extreme of high promiscuity E_{th} =-2 these over and under-representations exist but are mild, suggesting dominance of the global strategy with only little role of local adjustment to specific partners. In contrast, in the extreme of low promiscuity E_{th} =-10 we observe a significant over-representation of opposite-charge AA pairs and under-representation of same-charge AA pairs, compared to the corresponding marginal distributions. **(c)** Comparing the intensity of the global strategy, manifested by the RNase deviation from the prior $D_{\mathrm{KL}}(\psi^{\mathrm{RNase}}||\nu)$ (gray curve) and the local strategy, manifested by the deviation of RNase-SLF partner pairs from the product of their marginals $D_{\mathrm{KL}}(\varphi^{(\mathrm{RNase,SLF})}||\psi^{\mathrm{RNase}}\psi^{\mathrm{SLF}})$ (blue curve) shows that in most of the E_{th}, the global supersedes the local. **(d)** Qualitative comparison between the global and local strategies.

is at play there.

While the global strategy only biases the sequence AA frequencies independently of its partners, the local strategy induces dependencies between same-position residues amongst partner and anti-partner RNase-SLF pairs. Thus, to quantify the contribution of the local strategy, we calculated the deviation of the joint RNase-SLF distribution $\varphi^{(\mathrm{RNase,SLF})}$ relative to the product of its marginals, namely $\psi^{\mathrm{RNase}} \times \psi^{\mathrm{SLF}}$, using Kullback-Leibler divergence $D_{\mathrm{KL}}(\varphi^{(\mathrm{RNase,SLF})}||\psi^{\mathrm{RNase}} \times \psi^{\mathrm{SLF}})$, where any deviation of D_{KL} from zero is indicative of the local strategy contribution.

To disentangle the opposite effects of the two selection pressures operating simultaneously to attract (across S-haplotypes) and repel (within-S-haplotype), in Fig. 6a we calculated D_{KL} separately for pairs of same-S-haplotype RNases and SLFs, that are under pressure to repel each other (red curve), and pairs of partner RNase-SLF located on different S-haplotypes, that are under pressure to attract each other (blue).

Promiscuity affects these two selection pressures in two ways: firstly, as E_{th} increases, the pressure to repel (within the S-haplotype) rises, while the pressure to attract (across S-haplotypes) lessens. Secondly, the number of partners per protein increases with promiscuity (Fig. 2b) rendering the local strategy less useful, as it is difficult to match multiple distinct partners via local sequence arrangements. Only for the lowest promiscuity, most proteins have only one partner (or

none), making the local strategy feasible. In contrast, the number of anti-partners each protein should repel is determined by the number of genes per S-haplotype, a fixed parameter in our model, regardless of E_{th} (Fig. 3c). Indeed, we found that the local strategy was dominant for partner attraction (blue) under low promiscuity but its contribution decreased to zero for high promiscuity. For within-S-haplotype repulsion, we found an opposite trend, where the contribution of the local strategy increased with promiscuity, yet remained at much lower values even at its highest end (Fig. 6a).

As the Kullback-Leibler divergence is only indicative of a deviation from the marginal distribution but not of its sign, we further refined our analysis to distinguish between the opposite effects of repulsion and attraction. In Fig. 6b we present the conditional RNase-SLF AA frequencies of charged AA distinguishing between same-charge (repulsive) and opposite charge (attractive) AA pairs, compared to the marginal frequencies (horizontal lines). Again, we look separately at pairs of same-S-haplotype RNase and SLF evolving to repel each other (red) and at pairs of different S-haplotype RNase and SLF partners, evolving to attract each other (blue).

In line with the dominance of the local strategy under low E_{th} as demonstrated in Fig. 6a, we find among partner RNase-SLF pairs, an over-representation of +/- (attractive) and under-representation of +/+ (repulsive) AA pairs, but the magnitude of this effect decreases with E_{th} (Fig. 6b). This suggests a dominant role for the local strategy in partner attraction under low E_{th} (few partners). We find qualitatively similar trends for H-H and P-P AA pairs that also promote attraction (Fig. S8).

In contrast, for same-S-haplotype RNase-SLF pairs that evolve to repel each other (red bars) we find much smaller deviations from the marginals (Fig. 6b), showing only little contribution of the local strategy to avoidance of self-compatibility. This result too is in agreement with Fig. 6a demonstrating a lesser role for the local strategy among same-S-haplotype pairs.

Conclusions

Combining the results of Fig. 6 and Fig. 4, we conclude that under most of the parameter regime, there are many-to-many matches where proteins evolve by tuning their attractiveness/repulsion towards a generic partner and do so by biasing their amino acid frequencies to deviate from the prior ones (global strategy). This strategy allows to easily match newly formed proteins, facilitating the emergence of new class . Only in the extreme of low promiscuity, the proteins mostly have only a single partner (or none), hence the strategy of tailoring their sequence to match a specific partner becomes useful. This local strategy is mostly used for partner attraction and has only a minor role in anti-partner repulsion. See Fig. 6c for comparison between the magnitudes of the global and local strategies in the entire E_{th} range.

Roughly speaking, RNases develop non-specific locks with calibrated difficulty to guarantee self-incompatibility (global strategy, Fig. 4a). The SLFs, in turn, develop tailored keys to unlock particular RNases (local strategy), so that each paternal S-haplotype can fertilize maternal S-haplotypes with nearly all non-self RNases. As E_{th} decreases, the SLFs specialize in unlocking a decreasing set of RNase locks.

This also explains why the reduced model, for which the selection to match was not accounted for, provides a good approximation to the frequencies in the full model: when promiscuity is high, the selection pressure on the SLFs to match is minor, as they have a high probability of matching at random, and when promiscuity is lower, the selection pressure is expressed by a local matching

strategy, which has only a minor effect on the AA frequencies in a single S-haplotype.

Discussion

The collaborative non-self recognition self-incompatibility system in hermaphroditic plants is an example of a complex biological system comprised of multiple components that need to co-evolve while satisfying a set of potentially conflicting constraints. The evolutionary unit in this system is an S-haplotype that includes a female type-specifying gene (RNase) and multiple male type-specifying genes (SLFs) that are tightly linked and consequently affect each other's evolution. To elucidate the evolution of new mating specificities in this system, we recently proposed a theoretical framework that incorporates molecular recognition between the male and female type-specifying proteins into the evolutionary model [28]. Our model builds on the cornerstones of a degenerate genotype-to-phenotype mapping allowing multiple genotypes to share a common compatibility phenotype and on the promiscuity of molecular interactions. This model uniquely facilitates a general interaction pattern between the male and female proteins, rather than solely one-to-one interactions, as assumed in previous models [26, 27]. One-to-one interactions are obtained as a special case of our model in the extreme of low promiscuity. Using this model we study how its numerous microscopic degrees of freedom (in the single gene and in the individual levels) robustly give rise to a few population-level macroscopic phenotypes. Our model does not include an explicit definition of fitness as often employed in evolutionary models. Instead, an individual's reproductive success is determined by the interaction energies amongst its self-proteins and between its proteins and those of other population members, inducing dependencies of an individual fitness on the entire population composition. Matches and mismatches between individuals establishing the macroscopic system states can be achieved by a variety of amino acid (AA) sequences constituting its microscopic states. Hence, our model exhibits a degenerate genotype-to-phenotype mapping.

In plant populations, pollen is often more abundant than ovules. Hence, many pollen grains compete over fewer ovules resulting in asymmetry between the selection pressures acting on the hermaphroditic plant in its roles as male and female. Thus, the pressure for cross-compatibility in the collaborative non-self recognition system is mostly borne by the SLF (male type-specifying) and much less by the RNase (female type-specifying) genes. In contrast, the pressure to avoid self-compatibility is shared by both, but is much more expressed in the RNase which has to avoid match with multiple SLFs. If selection were only to match, one could have expected the RNases to employ the neutral AA frequencies, while the SLF frequencies were biased by the pressure to match. However, because of the additional and dominant pressure to avoid within-S-haplotype match, we observe a very different response, where the SLF frequencies are neutral and the RNases are pushed to avoid match and hence their frequencies deviate from neutrality. While the magnitude of both selection pressures varies with promiscuity, the pollen-to-ovule ratio and the inbreeding depression, the dominance of selection to avoid self-compatibility holds under a broad range of these parameter values (Fig. 4, Fig. S9, Fig. S12), and it is only reversed in the non-promiscuous region $E_{\mathrm{th}} \leq -9$ or under severe pollen limitation. In the latter case, the weight of the selection pressure to match acting on the RNases increases as pollen becomes scarce, while the pressure to avoid self-compatibility remains intact, thus tipping the balance between these two pressures acting on the RNases (Fig. S9).

The asymmetry in response to selection between the RNases and SLFs is explained by an entropic argument: due to the multiplicity of SLFs, biasing their AA frequencies bears a larger cost of entropy reduction compared to the cost required to bias the RNase, that has only a single copy per S-haplotype. We frame this argument using large deviations theory, whose accuracy improves as the sequence length L increases. While we use here $L = 18$, in accordance with the typical size of protein binding domains and typical numbers of residues estimated to be under positive selection, we obtain a very good fit between the theoretical prediction and the simulation results. The balance between multitude and fitness (reminiscent of the energy-entropy balance in statistical physics) can be used to infer the statistics of biological systems exhibiting a degenerate mapping between genotype to phenotype. Another example for a system in which similar considerations were applied to infer the stationary distribution is transcription factor binding sites [31, 32, 33]. Similar to the self-incompatibility system, there too the phenotype is determined by the binding energy alone and can be achieved by a multitude of distinct sequences.

We found that except for the non-promiscuous region, evolved sequences match each other much less than neutral ones do. This dominance of selection to avoid self-compatibility is in line with an experimental study by Zhao *et al.* [34] that demonstrated 85% compatibility between two Petunia RNases and several SLFs of different species or families, which obviously cannot fertilize Petunia. In contrast, only 19% of same-species SLFs exhibited such compatibility. In light of our model, this empirical result can be interpreted as compatibility being the default between RNase and SLF proteins in the absence of selection, as is the case for different species that cannot fertilize each other. Compatibility decreases only amongst same-species proteins due to selection against self-compatibility. This empirical result suggests though that local matches play a more dominant role in AA frequencies in the elimination of self-compatibility than global biases do, whereas our model predicts that elimination is almost exclusively achieved by a global change in AA frequencies. The default being high compatibility between proteins, unless selection operates to reduce it, is counter-intuitive. It is intriguing whether this is also the case in additional multi-component systems relying on molecular recognition.

A decrease in hydrophobicity as a means for a protein to be indiscriminately less attractive, as predicted by our model and found in empirical RNase data (Fig. 5), was previously reported as a transient state when proteins switch between different partners. Likewise, an increase in charged amino acid content was proposed to later follow as a means to specifically tune to attract or repel certain proteins [35]. Assessment of the amino acid content of hub proteins also showed a decrease in hydrophobic and an increase in electrostatic interactions compared to no-hub proteins [36], suggesting that these features are more broadly applicable.

A simplifying assumption in our model is that the proteins bind each other in one orientation only. It remains unknown whether SLFs having multiple RNase partners use a single shared binding domain for all partners or alternatively have a separate interface dedicated to each [36]. It would be interesting to extend our model to allow for multiple binding geometries and study how this affects the balance between the global and local match strategies and the resultant amino acid frequencies.

Our model exhibits several key microscopic properties that give rise to its macroscopic behavior: promiscuity, degenerate mapping between genotype-to-phenotype, and a co-evolution of multiple components under a combination of positive and negative design pressures. These properties are not unique to the self-incompatibility system and were observed in several other

systems. For example, a degenerate genotype to phenotype mapping is known in transcription factor-DNA binding [33] and RNA secondary structures [37]. Co-evolution of components that should match certain partners but avoid others is known in a variety of biological systems, such as gene regulatory networks [33], two-component signaling systems [38], bacterial toxin-antitoxin [39], host-pathogen [40], and venoms [41, 42], and are often accompanied by selection for diversification [43, 40, 39]. The superposition of positive and negative designs [44, 45] plays a key role in the evolution of the immune system whose aim is to distinguish between self and non-self peptides [46] and in gene regulatory networks where transcription factors should distinguish between cognate and non-cognate binding sites [47]. Promiscuity in molecular recognition was observed in signaling proteins [48], and in transcription factor-DNA binding [49] and proposed to be pivotal in the evolvability of gene regulatory networks [50, 51] and in the immune receptor functionality [52, 53]. Following these commonalities, we propose that the incorporation of molecular recognition into the evolutionary models of these systems may be more broadly applicable and further ramifications of our framework could be exploited to study additional systems. It would be interesting to see whether these systems also share some of the behaviors exhibited by the plant collaborative self-incompatibility system, as shown here.

Methods

Four biochemical categories of amino acids (AA)

We have classified the 20 AAs into four different biochemical categories: hydrophobic (H), neutrally polarized (P), positively charged (+), and negatively charged (–), see Table S1. The prior frequencies of the different categories are given by $\nu(i|i = \{\mathrm{H}, \mathrm{P}, +, -\}) = \{0.5, 0.265, 0.113, 0.122\}$, respectively [28]. Table 1 shows the interaction energies between pairs of amino acids of all possible categories. These values were adopted from [54].

Parameters and default values

The model parameters and their default values (unless specified otherwise) are summarized in Table 2. The parameter E_{th}, is the main parameters that vary from -10 to -2 with step size 1.

The stochastic simulation

Initialization of the S-haplotype population

We initialize the S-haplotype population using n distinct types, such that each type is self-incompatible and bidirectionally compatible with the remaining $n-1$ types. We initialize the population with equal frequencies of all types. We employ the following three steps to draw these initial types:

1. **Draw RNase sequences**: First, generate n distinct RNases $\mathrm{R} = \{\mathrm{R}_1 \ldots \mathrm{R}_n\}$ by randomly drawing sequences from the prior distribution ν.

2. **Draw SLF sequences**: Construct n pools where each pool $\mathrm{P}_i = \{\mathrm{S}_m\}$, $i \in (1 \ldots n)$, contains m SLF sequences, that can detoxify the i-th RNase, and potentially others as well. To construct the SLF pools, we generate SLFs, one at a time, test which RNase(s) it detoxifies, and add that SLF into the corresponding pool(s). Any SLF that does not

detoxify any RNase or that detoxifies all of them is discarded. Repeat this step until each pool has m SLF genes.

3. **Form N self-incompatible and complete S-haplotypes**: For each RNase $\mathrm{R}_i \in \mathrm{R}$, choose in total $n-1$ SLF genes, one from each pool P_j, $j \in (1 \ldots n)$, $j \neq i$ and make sure none of these SLFs detoxifies R_i. If it does, replace it with another SLF from the same pool until an SLF that does not detoxify R_i is found. After generating n distinct S-haplotypes, we construct an initial population of size N using exactly N/n copies of each S-haplotype, and then shuffle them.

The above procedure produces SLF that can potentially match multiple RNases each. Alternatively, to construct an initial population with only one-to-one RNase-SLF interactions, each pool P_i should have only those SLFs that detoxify only R_i.

The simulation time-course: mutation and reproduction

1. **Mutation**: Each amino acid in each gene can mutate with probability μ per generation. An amino acid chosen to mutate is replaced by a randomly chosen amino acid drawn from the prior distribution ν.

2. **Maternal S-haplotype picking**: Maternal SC S-haplotypes can be fertilized by either non-self paternal S-haplotype or by the self one. In the latter case, only a proportion $1-\delta$ of the offspring produced via self-fertilization survives. Maternal S-haplotypes chosen to reproduce can be either self-compatible (SC) with either self or non-self fertilization or self-incompatible (SI) allowing only non-self fertilization. Each generation we calculate the probabilities of either group forming offspring:

$$f^{\mathrm{s}}_{\mathrm{SC}}(t) = \frac{x(t)\ \alpha\ (1-\delta)}{(1-x(t)\ \alpha\ \delta)} \tag{4}$$

$$f^{\mathrm{ns}}_{\mathrm{SC}}(t) = \frac{x(t)\ (1-\alpha)}{(1-x(t)\ \alpha\ \delta)} \tag{5}$$

$$f_{\mathrm{SI}}(t) = \frac{1-x(t)}{(1-x(t)\ \alpha\ \delta)}, \tag{6}$$

where $x(t)$, and $1-x(t)$ are the population fraction of SC and SI S-haplotypes in the current generation, α is the proportion of self-pollen received and δ is the inbreeding depression, i.e. the proportion of self-fertilization offspring that do not survive. Full derivation of $f^{\mathrm{s}}_{\mathrm{SC}}$, $f^{\mathrm{ns}}_{\mathrm{SC}}$, and f_{SI} is provided in the supplementary information file.

3. **Offspring formation**: Pick a maternal S-haplotype H_i from either group with probabilities $\{f^{\mathrm{s}}_{\mathrm{SC}}, f^{\mathrm{ns}}_{\mathrm{SC}}, f_{\mathrm{SI}}\}$. If the maternal S-haplotype H_i belongs to $f^{\mathrm{s}}_{\mathrm{SC}}$, the offspring is identical to the parental one. If the maternal S-haplotype H_i belongs to either $f^{\mathrm{ns}}_{\mathrm{SC}}$ or f_{SI}, it is fertilized by non-self pollen. In this case, it is given k opportunities to match a randomly chosen non-self S-haplotype partner, in the paternal role. If such a match is found, the offspring would be one of the two parental S-haplotypes with equal probability. Repeat this step until N offspring are formed.

Steps 1-3 constitute a single generation. Repeat these steps until the desired number of generations is reached.

Total simulation run times for the default parameter values are listed in Table 4. The analyses are usually based on a total of 25 independent runs for each parameter set, where 1200 data points are taken from each run, with a 25-generation interval between consecutive time points.

Simulation run-time

We simulated the model for different numbers of generations as described in Table 4. The two criteria should be met to stop the simulation: the amino acid frequencies of RNases and SLFs should stabilize and the entire population should be the descendants of only one of the n initial S-haplotypes. For E_{th} = -2, it takes a larger number of generations to achieve the latter, compared to the former criterion.

Data analysis

Effective number of alleles and effective number of classes

We use the inverse participation ratio to calculate the effective number of alleles and the effective number of classes. It is calculated as follows,

$$\widetilde{n}_{\text{eff}} = \frac{1}{\sum_{i=1}^{M} f_i^2} \tag{7}$$

where f_i is the frequency of the i-th allele or class, and M is the total number of distinct allele types or classes.

Consider a set $T = \{T_l\}$ of l data points in each run over which we want to calculate the average allele or class numbers. The average of the effective number of alleles $\langle \widetilde{n}_{\text{allele}} \rangle$ and its standard deviation $\sigma_{\widetilde{n}_{\text{allele}}}$ are calculated as:

$$\langle \widetilde{n}_{\text{allele}} \rangle = \frac{1}{I\ n(T)} \sum_{j=1}^{I} \sum_{t \in T} \frac{1}{\sum_{k=1}^{M_j^t} f_{jtk}^2} \tag{8}$$

$$\sigma_{\widetilde{n}_{\text{allele}}} = \sqrt{\frac{1}{I\ n(T)} \left(\sum_{j=1}^{I} \sum_{t \in T} \left(\frac{1}{\sum_{k=1}^{M_j^t} f_{jtk}^2} - \langle \widetilde{n}_{\text{allele}} \rangle \right)^2 \right)}, \tag{9}$$

where $n(T)$ is the number of data points in T and I is the number of runs. f_{jtk}, and M_j^t are the frequency of k-th allele type and the number of distinct functional alleles in run j at generation t. We only consider functional RNases and functional SLFs in this calculation. The average effective number of classes is calculated similarly.

RNase partners per SLF gene

For a given run j, and at generation t, consider $R^{jt} = \{R_i^{jt}\}$ to be the set of all unique RNases, where $f_R^{jt} = \{f_{R_i}^{jt}\}$ is the set of their corresponding frequencies. Similarly, let $S^{jt} = \{S_k^{jt}\}$, and $f_S^{jt} = \{f_{S_k}^{jt}\}$ be the set of all unique SLFs and their corresponding frequencies. Further, consider $R^{jt}(S_k^{jt}) \subset R^{jt}$ to be the set containing all distinct RNases that are detoxified by a particular

SLF gene S_k^{jt}. The effective number of RNase partners of SLF gene S_k^{jt} is given by:

$$\mathrm{P}(S_k^{jt}) = \widetilde{n}_{\mathrm{eff}}^{jt}(\mathrm{RNase}) \times \sum_{x \in R^{jt}(S_k^{jt})} f_x \tag{10}$$

$$\widetilde{n}_{\mathrm{eff}}^{jt}(\mathrm{RNase}) = \frac{1}{\sum_{q \in R^{jt}} f_q^2}. \tag{11}$$

where f_x is the population frequency of unique RNase x.

Distribution of the number of partners per SLF

Let $X^{jt} = \bigcup\{\mathrm{P}(S_k^{jt})\}$ be the set of unique partner numbers of all unique SLF genes. For a fixed bin size h, and maximum number of RNases M over all independent runs and for all data points a fraction of SLF genes x^S detoxifying a number of RNases in the range $[i\ h, (i+1)\ h)$, $i \in (0, M)$ is calculated as,

$$x^{\mathrm{S}} = \sum_{j=1}^{I} \sum_{t=1}^{l} \sum_{\mathrm{P}(S_k^{jt}) \in [i*h,(i+1)*h)} f_{\mathrm{P}(S_k^{jt})} \tag{12}$$

where $f_{\mathrm{P}(S_k^{jt})}$ is the frequency of the k-th SLF gene that has a number of partners $\mathrm{P}(S_k^{jt}) \in [i*h, (i+1)*h)$. In Fig. 2a,b we took $h = 1.0$.

RNase partners per S-haplotype

For a given run j, and at generation t, consider $H^{jt} = \{H_k^{jt}\}$ to be the set of unique S-haplotypes and $f_H^{jt} = \{f_{H_k}^{jt}\}$ their corresponding frequencies. Further, consider $R^{jt}(H_k^{jt}) \subset R^{jt}$ to be the set containing all distinct RNases that can be detoxified by a particular S-haplotype H_k^{jt} as male. The effective number of RNase partners of S-haplotype H_k^{jt} is given by:

$$\mathrm{P}(H_k^{jt}) = \widetilde{n}_{\mathrm{eff}}^{jt}(\mathrm{RNase}) \times \sum_{x \in R^{jt}(H_k^{jt})} f_x \tag{13}$$

$$\widetilde{n}_{\mathrm{eff}}^{jt}(\mathrm{RNase}) = \frac{1}{\sum_{q \in R^{jt}} f_q^2} \tag{14}$$

where f_x is the population frequency of unique RNase x.

Compatibility classes

We classified the set of S-haplotypes in the population according to their compatibility phenotype. The classification should obey the following two requirements: an S-haplotype affiliated with a class should not be compatible neither as male nor as female with any of the other members of the same class, but should be compatible both as male and as female with all members of all other classes, it is not affiliated with. Following this definition, it is not guaranteed that all S-haplotypes are classified into any of the classes. Specifically, self-compatible S-haplotypes cannot be associated with any class, because they can fertilize their own class members. Moreover, the division of S-haplotypes into classes is not unique. To ensure that the most frequent S-haplotypes

are classified, we first ordered all the self-incompatible S-haplotypes in the population in descending order of their copy number. We then defined the most frequent S-haplotype to be the first class. Then, for every unclassified S-haplotype in its turn (following the descending copy number order), we tested its compatibility with the already existing classes. If it was bidirectionally incompatible with all members of exactly one class and bidirectionally compatible with all members of all the remaining classes, we associated it with the class it was incompatible with. If it was bidirectionally compatible with all members of all existing classes, we defined a new class and associated this S-haplotype with that class. Otherwise, this S-haplotype remained unclassified and we moved on to the next one. Importantly, our class definition refers only to the compatibility phenotype of S-haplotypes, rather than to their genotype. As our model allows for multiple partners per RNase/SLF, classes could be genetically heterogeneous. Classification was applied every 25 generations, always independently of previous classifications. The full classification algorithm is provided in the supplementary information.

Hamming distance

The Hamming distance between two sequences $x = (x_1, x_2 \dots x_L)$, and $y = (y_1, y_2 \dots y_L)$, $\forall$ $x_i, y_j \in \{\mathrm{H}, \mathrm{P}, +, -\}$ is defined as the number of positions in which these sequences differ from each other

$$d_{\mathrm{h}} \quad = \quad n\{1 | x_i \neq y_i, i = 1 \dots L\}, \quad n\{\} : \text{number of elements in the set.} \tag{15}$$

The effective energy threshold used in the analytical approximation

In the analytical calculation, we considered an effective energy threshold $\tilde{E}_{\mathrm{th}} = E_{\mathrm{th}}/C$, where the value of C was extracted from our simulations of the reduced model (Fig. S10), as the ratio of the E_{th} value used in simulation and the average of total interaction energy obtained in evolved populations, $C = E_{\mathrm{th}}/\langle E^L_{\mathrm{tot}} \rangle$, thus requiring a harsher energy threshold to avoid self-compatibility than used in the simulations (see appendix for more details). See supplementary Mathematica file for the numerical calculation.

Amino acid composition in RNase empirical data

The empirical data points in Fig. S3 are based on the positively selected positions in RNases of *Petunia inflata*, *Petunia hybrida*, and *Petunia axillaris* as inferred computationally by [24]. We classify the amino acids in these positions to either of the four biochemical classes based on the classification in Table S1.

Table 1: Interaction energies between pairs of amino acids.

	H	**P**	**+**	**-**
Hydrophobic (H)	-1	0	0	0
Neutrally polarized (P)	0	-0.75	-0.25	-0.25
Positively charged (+)	0	-0.25	1	-1.25
Negatively charged (-)	0	-0.25	-1.25	1

Table 2: Description of parameters and default values in the main model simulation

Parameters	Descriptions
$N = 2000$	Total number of S-haplotypes or population size
$n = 10$	Initial number of distinct S-haplotypes
$m = 9$	Number of SLF paralogs per S-haplotype
E_{th}	Interaction energy threshold
$\mu = 10^{-4}$ [1/generation]	Per-residue mutation rate
$L = 18$	Number of residues (amino acids) in each gene
$\alpha = 0.90$	Proportion of self-pollen received by haploid maternal plant
$\delta = 0.90$	Inbreeding depression - the proportion of self-fertilization inviable offspring
$k = N$	The number of non-self fertilization attempts for a female S-haplotype

Table 3: Compatible fraction and average number of SLFs per S-haplotype interacting with 1 RNase. The first 3 rows of the Table show the data of Fig. 4a.

E_{th}	**-10**	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**
Compatible fraction (neutral)	0.032	0.084	0.182	0.334	0.515	0.698	0.84	0.929	0.975
Compatible fraction (evolved)	0.285	0.224	0.224	0.212	0.226	0.285	0.371	0.439	0.519
Compatible fraction (marginal)	0.008	0.027	0.052	0.107	0.182	0.278	0.413	0.54	0.604
Average number of SLFs per S-haplotype interacting with 1 RNase (all randomly drawn from the prior distribution)	0.25	0.67	1.49	2.76	4.37	6.02	7.37	8.26	8.72

Table 4: Typical simulation times for different E_{th} values.

E_{th}	**-10**	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**
Generations ($\times 10^5$)	1.5	1.5	1.5	1.5	2.0	2.75	4.5	4.5	4.5

Data availability The authors declare that all data supporting the findings of this study are available within the article and its Supplementary Information file.

Code availability The code will be made available on a public repository upon acceptance for publication.

Appendix

Sanov derivation

To establish the fact that avoiding self-compatibility is the primary evolutionary force governing the evolution of the amino acid frequencies in the S-haplotype and explain the asymmetric response to the selection of the RNase and SLF, we estimate the empirical amino-acid distribution

in an S-haplotype, evolved strictly under pressure to avoid self-compatibility. We then show that this distribution is nearly identical to the distribution in our simulation, by comparing the relative AA frequencies.

To complement this empirical computation, we also consider a theoretical estimate derived in the following fashion. Firstly we consider an even simpler non-evolutionary model. Given an S-haplotype, denote by E_i its interaction energy between the RNase and the i-th SLF as per Eq. (1). In the simplified model we pick at random a neutral S-haplotype with a single RNase and m SLFs each of length L, drawn from the prior distribution, conditioned on $E_i > E_{\text{th}}$ for all i. Observe that this corresponds to a situation in which the population evolves neutrally and then suddenly every self-compatible S-haplotype is eliminated, and we are interested in the typical distribution of a surviving S-haplotype.

To understand the behavior of such a sample, assuming that L is sufficiently large, we may apply Sanov's Theorem, which tells us that, as the sequence length L tends to infinity, the sampled distribution conditioned to be in any non-empty set Γ will be close to the minimizer of the relative entropy with respect to the prior distribution.

Let us state this formally. Let q be a probability distribution on an alphabet Σ, and write J_n^X for the AA-frequencies of a sequence $X = (X_1, \ldots, X_n) \in \Sigma^n$ whose coordinates are drawn independently according to q (in our case the simulated S-haplotypes amino acids' distribution). Consider a restricted AA-frequencies space Γ (in our case those for which the interaction energy between the RNase and every SLF is lower than the threshold). Then

Theorem 1 (Sanov '57 [30][Theorem 2.1.10])

$$-\inf_{q' \in \Gamma^\circ} H(q' \mid q) \leq \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(J_n^X \in \Gamma) \leq -\inf_{q\in\Gamma} H(q'|q),$$

where $H(q'|q)$ is the relative entropy of q with respect to q' given by

$$H(q' \mid q) = \sum_{a\in\sigma} q'(a) \log \frac{q(a)}{q'(a)},$$

and Γ° is the interior of Γ with respect to the weak topology. Moreover, the empirical distribution is, with high probability, in any neighborhood of the minimizers of this relative entropy.

Applying this to our model, our alphabeth consists of $n = L(m+1)$ residues, each being one of $\{+, -, H, P\}$. We take Γ to be the set of such distribution whose total energy between every SLF and the RNase E_{tot}^L is greater than E_{th}. For large sequence length L we obtain that

$$\frac{1}{n} \log \mathbb{P}(\psi \in \Gamma) \to -\inf_{q\in\Gamma} \sum_{a,b_1,\ldots,b_k\in\{+,-,H,P\}} q(a, b_1, \ldots, b_k) \log \frac{q(a, b_1, \ldots, b_k)}{\nu(a)\nu(b_1)\cdots\nu(b_k)}.$$

We now apply convexity and symmetry arguments (namely that the optimiser must be obtained when conditioned on the RNase locus, all the SLF distributions for that locus must be identical), and express this in terms of ψ and φ, as minimizing

$$\sum_{a\in\{+,-,P,H\}} \psi^{\text{RNase}}(a) \log \frac{\psi^{\text{RNase}}(a)}{\nu(a)} + m \sum_{a,b\in\{+,-,P,H\}} \phi(a,b) \log \frac{\phi(a,b)}{\psi^{\text{RNase}}(a)\nu(b)}$$

under the constraint that ψ^{RNase} is the marginal of φ.

Next, let us explain how to use this analysis for a model closer to our simulated reduced evolutionary model, governed only by avoiding self-compatibility. In the model analysed each S-haplotype undergoes neutral evolution, but if at any point in the course of its evolution is crosses the energy threshold – it is eliminated.

Formally, this allows us to rewrite the survival probability as $\mathbb{P}(\psi \in \Gamma)$ as

$$\mathbb{P}(\psi \in \Gamma) \propto \sum_{\psi \in \Gamma} \mathbb{P}'(\psi) \mathbb{P}_{evo}(\psi),$$

where $\mathbb{P}'(\psi)$ is the neutral probability of obtaining ψ as in the simplified non-evolutionary model, and $P_{evo}(\psi)$ is the probability of surviving above the threshold throughout the evolution, conditioned on achieving final AA-frequencies ψ. In the special case in which $\mathbb{P}_{evo}(\psi) = f(E_{\mathrm{tot}}(\psi))$ for some monotone decreasing function f, we can rewrite this sum as:

$$\mathbb{P}(\psi \in \Gamma) \propto \sum_{E > E_{\mathrm{th}}} \sum_{\psi : E_{\mathrm{tot}}(\psi) = E} \mathbb{P}'(\psi) f(E) = \sum_{E > E_{\mathrm{th}}} f(E) \sum_{\psi : E_{\mathrm{tot}}(\psi) = E} \mathbb{P}'(\psi).$$

Writing $\mathbb{P}'(E) = \sum_{\psi : E_{\mathrm{tot}}(\psi) = E} \mathbb{P}'(\psi)$ we obtain

$$\mathbb{P}(\psi \in \Gamma) \propto \sum_{E > E_{\mathrm{th}}} f(E) \mathbb{P}'(E).$$

This sum has polynomially many terms, so that we may obtain a polynomially tight estimate of $\mathbb{P}(\psi \in \Gamma)$ as its maximizer, namely:

$$\mathbb{P}(\psi \in \Gamma) \approx C \max_{E > E_{\mathrm{th}}} f(E) \mathbb{P}'(E),$$

for some constant C. Denoting by $E_{\max}$ the energy maximizing this expression, we may further write

$$\mathbb{P}(\psi \in \Gamma) \approx f(E_{max}) \sum_{E \geq E_{max}} \mathbb{P}'(E).$$

We may now apply Sanov's theorem to the set $\Gamma = \{\psi : E_{\mathrm{tot}}(\psi) \geq E_{\max}\}$ to recover the asymptotic decay of this probability and the limiting empirical distribution, which is equivalent to an application of the standard Sanov's theorem with a harsher energy threshold $\tilde{E} = E_{max}$. This justifies our usage of Sanov's theorem with an effective energy threshold as per Figure S10. Observe that δ, the probability that an offspring formed by self-fertilization does not survive, affects this analysis only by altering the effective threshold $\tilde{E}$, as long as the assumption that survival is roughly a function of the energy holds.

It turns out that analyzing this simpler model in which the population members are statistically independent, and the assumption that survivability is energy dependent, are good enough to provide a good approximation of the frequencies obtained in the simulated evolutionary model. Indeed, we have formally derived the balance between the RNase and SLF amino acid frequencies and solved it numerically for the case of only selection to avoid self-compatibility. We then optimized it on the effective energy threshold and found excellent agreement with our simulation results (suppl. Mathematica file).

References

[1] Lande, R. & Schemske, D. W. The Evolution of Self-Fertilization and Inbreeding Depression in Plants. I. Genetic Models. *Evolution* **39**, 24–40 (1985). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.1985.tb04077.x/abstract.

[2] East, E. M. & Mangelsdorf, A. J. A New Interpretation of the Hereditary Behavior of Self-Sterile Plants. *Proceedings of the National Academy of Sciences* **11**, 166–171 (1925). URL https://www.pnas.org/doi/abs/10.1073/pnas.11.2.166. Publisher: Proceedings of the National Academy of Sciences.

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

[4] de Nettancourt, D. The Genetic Basis of Self-Incompatibility. In de Nettancourt, D. (ed.) *Incompatibility in Angiosperms*, Monographs on Theoretical and Applied Genetics, 28–57 (Springer, Berlin, Heidelberg, 1977). URL https://doi.org/10.1007/978-3-662-12051-4_2.

[5] Igic, B., Lande, R. & Kohn, J. Loss of Self-Incompatibility and Its Evolutionary Consequences. *International Journal of Plant Sciences* **169**, 93–104 (2008). URL http://www.journals.uchicago.edu/doi/abs/10.1086/523362.

[6] Takayama, S. & Isogai, A. Self-Incompatibility in Plants. *Annual Review of Plant Biology* **56**, 467–489 (2005). URL http://dx.doi.org/10.1146/annurev.arplant.56.032604.144249.

[7] Fujii, S., Kubo, K.-i. & Takayama, S. Non-self- and self-recognition models in plant self-incompatibility. *Nature Plants* **2**, 16130 (2016). URL http://www.nature.com/articles/nplants2016130.

[8] Shimizu, K. K. & Tsuchimatsu, T. Evolution of Selfing: Recurrent Patterns in Molecular Adaptation. *Annual Review of Ecology, Evolution, and Systematics* **46**, 593–622 (2015). URL https://doi.org/10.1146/annurev-ecolsys-112414-054249.

[9] Iwano, M. & Takayama, S. Self/non-self discrimination in angiosperm self-incompatibility. *Current Opinion in Plant Biology* **15**, 78–83 (2012). URL http://www.sciencedirect.com/science/article/pii/S1369526611001452.

[10] Gervais, C., Awad, D. A., Roze, D., Castric, V. & Billiard, S. Genetic Architecture of Inbreeding Depression and the Maintenance of Gametophytic Self-Incompatibility. *Evolution* **68**, 3317–3324 (2014). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/evo.12495.

[11] McClure, B. A. *et al.* Style self-incompatibility gene products of Nicotlana alata are ribonucleases. *Nature* **342**, 955–957 (1989). URL https://www.nature.com/articles/342955a0. Number: 6252 Publisher: Nature Publishing Group.

[12] Lai, Z. *et al.* An F-box gene linked to the self-incompatibility (S) locus of Antirrhinum is expressed specifically in pollen and tapetum. *Plant Molecular Biology* **50**, 29–41 (2002). URL https://link.springer.com/article/10.1023/A:1016050018779.

[13] Sijacic, P., Wang, X., Skirpan, A. L., Wang, Y. & al, e. Identification of the pollen determinant of S-RNase-mediated self-incompatibility. *Nature; London* **429**, 302–5 (2004). URL https://search.proquest.com/docview/204530198/abstract/9ABEA5AF71D048EAPQ/1.

[14] Kubo, K.-i. *et al.* Collaborative Non-Self Recognition System in S-RNase–Based Self-Incompatibility. *Science* **330**, 796–799 (2010). URL http://science.sciencemag.org/content/330/6005/796.

[15] Williams, J. S., Wu, L., Li, S., Sun, P. & Kao, T.-H. Insight into S-RNase-based self-incompatibility in Petunia: recent findings and future directions. *Frontiers in Plant Science* **6** (2015). URL https://www.frontiersin.org/articles/10.3389/fpls.2015.00041.

[16] Kubo, K.-i. *et al.* Gene duplication and genetic exchange drive the evolution of S-RNase-based self-incompatibility in *Petunia*. *Nature Plants* **1**, 14005 (2015). URL https://www.nature.com/articles/nplants20145/fig_tab.

[17] Matton, D. P. *et al.* Hypervariable Domains of Self-Incompatibility RNases Mediate Allele-Specific Pollen Recognition. *The Plant Cell Online* **9**, 1757–1766 (1997). URL http://www.plantcell.org/content/9/10/1757.

[18] Matton, D. P. *et al.* Production of an S RNase with Dual Specificity Suggests a Novel Hypothesis for the Generation of New S Alleles. *The Plant Cell* **11**, 2087–2097 (1999). URL http://www.plantcell.org/content/11/11/2087.

[19] Vieira, J., Morales-Hojas, R., Santos, R. A. M. & Vieira, C. P. Different Positively Selected Sites at the Gametophytic Self-Incompatibility Pistil S-RNase Gene in the Solanaceae and Rosaceae (Prunus, Pyrus, and Malus). *Journal of Molecular Evolution* **65**, 175–185 (2007). URL https://link.springer.com/article/10.1007/s00239-006-0285-6.

[20] Li, J. *et al.* Electrostatic potentials of the S-locus F-box proteins contribute to the pollen S specificity in self-incompatibility in Petunia hybrida. *The Plant Journal* **89**, 45–57 (2017). URL http://onlinelibrary.wiley.com/doi/10.1111/tpj.13318/abstract.

[21] Wu, L. *et al.* Use of Domain-Swapping to Identify Candidate Amino Acids Involved in Differential Interactions between Two Allelic Variants of Type-1 S-Locus F-Box Protein and S3-RNase in Petunia inflata. *Plant and Cell Physiology* **59**, 234–247 (2018). URL https://academic.oup.com/pcp/article/59/2/234/4632919.

[22] Pratas, M. I. *et al.* Inferences on specificity recognition at the Malus×domestica gametophytic self-incompatibility system. *Scientific Reports* **8**, 1717 (2018). URL .0.

[23] Vieira, J., Ferreira, P. G., Aguiar, B., Fonseca, N. A. & Vieira, C. P. Evolutionary patterns at the RNase based gametophytic self - incompatibility system in two divergent Rosaceae groups (Maloideae and Prunus). *BMC Evolutionary Biology* **10**, 200 (2010). URL https://doi.org/10.1186/1471-2148-10-200.

[24] Vieira, J. *et al.* Predicting Specificities Under the Non-self Gametophytic Self-Incompatibility Recognition Model. *Frontiers in Plant Science* **10** (2019). URL https://www.frontiersin.org/articles/10.3389/fpls.2019.00879/full.

[25] Aguiar, B. *et al.* Patterns of evolution at the gametophytic self-incompatibility Sorbus aucuparia (Pyrinae) S pollen genes support the non-self recognition by multiple factors model. *Journal of Experimental Botany* **64**, 2423–2434 (2013). URL https://academic.oup.com/jxb/article/64/8/2423/645638.

[26] Boďová, K., Priklopil, T., Field, D. L., Barton, N. H. & Pickup, M. Evolutionary Pathways for the Generation of New Self-Incompatibility Haplotypes in a Non-self Recognition System. *Genetics* genetics.300748.2018 (2018). URL http://www.genetics.org/content/early/2018/04/30/genetics.118.300748.

[27] Harkness, A., Goldberg, E. E. & Brandvain, Y. Diversification or Collapse of Self-Incompatibility Haplotypes as a Rescue Process. *The American Naturalist* **197**, E89–E109 (2020). URL https://www.journals.uchicago.edu/doi/full/10.1086/712424. Publisher: The University of Chicago Press.

[28] Erez, K., Jangid, A., Feldheim, O. N. & Friedlander, T. The role of promiscuous molecular recognition in the evolution of RNase-based self-incompatibility in plants. *Nature Communications* **15**, 4864 (2024). URL https://www.nature.com/articles/s41467-024-49163-7. Publisher: Nature Publishing Group.

[29] Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

[30] Dembo, A. & Zeitouni, O. *Large Deviations Techniques and Applications*, vol. 38 of *Stochastic Modelling and Applied Probability* (Springer, Berlin, Heidelberg, 2010). URL http://link.springer.com/10.1007/978-3-642-03311-7.

[31] Berg, J., Willmann, S. & Lässig, M. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology* **4**, 42 (2004). URL http://dx.doi.org/10.1186/1471-2148-4-42.

[32] Sella, G. & Hirsh, A. E. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9541–9546 (2005). URL http://www.pnas.org/content/102/27/9541.

[33] Friedlander, T., Prizak, R., Barton, N. H. & Tkačik, G. Evolution of new regulatory functions on biophysically realistic fitness landscapes. *Nature Communications* **8**, 216 (2017). URL https://www.nature.com/articles/s41467-017-00238-8.

[34] Zhao, H. *et al.* Origin, loss, and regain of self-incompatibility in angiosperms. *The Plant Cell* **34**, 579–596 (2022). URL https://doi.org/10.1093/plcell/koab266.

[35] Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J. & Shakhnovich, E. I. Structural Similarity Enhances Interaction Propensity of Proteins. *Journal of Molecular Biology* **365**, 1596–1606 (2007). URL https://www.sciencedirect.com/science/article/pii/S002228360601549X.

[36] Peleg, O., Choi, J.-M. & Shakhnovich, E. I. Evolution of Specificity in Protein-Protein Interactions. *Biophysical Journal* **107**, 1686–1696 (2014). URL https://www.sciencedirect.com/science/article/pii/S0006349514008418.

[37] Wagner, A. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences* **275**, 91–100 (2008). URL http://rspb.royalsocietypublishing.org/content/275/1630/91.short.

[38] Deeds, E. J. & Rowland, M. A. The Evolution of Crosstalk in Signaling Networks. *Biophysical Journal* **106**, 643a (2014). URL http://www.cell.com/biophysj/abstract/S0006-3495(13)04819-4.

[39] Aakre, C. *et al.* Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell* **163**, 594–606 (2015). URL http://www.sciencedirect.com/science/article/pii/S0092867415012726.

[40] Łuksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014). URL https://www.nature.com/articles/nature13087.

[41] Holding, M. L., Drabeck, D. H., Jansa, S. A. & Gibbs, H. L. Venom Resistance as a Model for Understanding the Molecular Basis of Complex Coevolutionary Adaptations. *Integrative and Comparative Biology* **56**, 1032–1043 (2016). URL https://academic.oup.com/icb/article/56/5/1032/2420622. Publisher: Oxford Academic.

[42] Arbuckle, K., Rodríguez de la Vega, R. C. & Casewell, N. R. Coevolution takes the sting out of it: Evolutionary biology and mechanisms of toxin resistance in animals. *Toxicon* **140**, 118–131 (2017). URL http://www.sciencedirect.com/science/article/pii/S0041010117303367.

[43] Nguyen, C. C. & Saier, M. H. Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS Letters* **377**, 98–102 (1995). URL http://onlinelibrary.wiley.com/doi/10.1016/0014-5793(95)01344-X/abstract.

[44] Sear, R. P. Highly specific protein–protein interactions, evolution and negative design. *Physical Biology* **1**, 166 (2004). URL http://stacks.iop.org/1478-3975/1/i=3/a=004.

[45] Sear, R. P. Specific protein–protein binding in many-component mixtures of proteins. *Physical Biology* **1**, 53 (2004). URL http://stacks.iop.org/1478-3975/1/i=2/a=001.

[46] George, J. T., Kessler, D. A. & Levine, H. Effects of thymic selection on T cell recognition of foreign and tumor antigenic peptides. *Proceedings of the National Academy of Sciences* **114**, E7875–E7881 (2017). URL http://www.pnas.org/content/114/38/E7875.

[47] Friedlander, T., Prizak, R., Guet, C. C., Barton, N. H. & Tkačik, G. Intrinsic limits to gene regulation by global crosstalk. *Nature Communications* **7**, 12307 (2016). URL http://www.nature.com/ncomms/2016/160804/ncomms12307/abs/ncomms12307.html.

[48] McClune, C. J. & Laub, M. T. Constraints on the expansion of paralogous protein families. *Current Biology* **30**, R460–R464 (2020). URL https://www.cell.com/current-biology/abstract/S0960-9822(20)30282-7. Publisher: Elsevier.

[49] Maerkl, S. J. & Quake, S. R. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science* **315**, 233–237 (2007). URL http://www.sciencemag.org/cgi/content/abstract/sci;315/5809/233.

[50] Pougach, K. *et al.* Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. *Nature Communications* **5**, 4868 (2014). URL http://www.nature.com/ncomms/2014/140910/ncomms5868/full/ncomms5868.html?WT.ec_id=NCOMMS-20140917.

[51] Shultzaberger, R. K., Maerkl, S. J., Kirsch, J. F. & Eisen, M. B. Probing the Informational and regulatory plasticity of a transcription factor DNA–binding domain. *PLoS genetics* **8** (2012). URL http://europepmc.org/articles/PMC3315485.

[52] Adams, R. M., Mora, T., Walczak, A. M. & Kinney, J. B. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife* **5**, e23156 (2016). URL https://elifesciences.org/content/5/e23156v3.

[53] Mayer, A., Balasubramanian, V., Mora, T. & Walczak, A. M. How a well-adapted immune system is organized. *Proceedings of the National Academy of Sciences* **112**, 5950–5955 (2015). URL https://www.pnas.org/doi/full/10.1073/pnas.1421827112. Publisher: Proceedings of the National Academy of Sciences.

[54] Noivirt-Brik, O., Unger, R. & Horovitz, A. Analysing the origin of long-range interactions in proteins using lattice models. *BMC Structural Biology* **9**, 4 (2009). URL https://doi.org/10.1186/1472-6807-9-4.

Acknowledgements

This research was supported by the Israel Science Foundation, grant number 1889/19 (T.F.).

Author contributions

T.F. conceptualized the project, acquired funding and supervised. A.J. wrote and ran the simulation, collected and analyzed results and prepared figures. K.E wrote an earlier version of the simulation. O.F. contributed to the mathematical analysis. T.F and O.F developed the methodology and wrote the paper.

Competing interests

The authors declare no competing interests.