Revisiting the number of self-incompatibility alleles in finite populations: from old models to new results

Peter Czuppon[1], Sylvain Billiard[2]

[1] Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany

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

Abstract. Under gametophytic self-incompatibility (GSI), plants are heterozygous at the self-incompatibility locus (S-locus) and can only be fertilized by pollen with a different allele at that locus. The last century has seen a heated debate about the correct way of modeling the allele diversity in a GSI population that was never formally resolved. Starting from an individual-based model, we derive the deterministic dynamics as proposed by Fisher (1958), and compute the stationary S-allele frequency distribution. We find that the stationary distribution proposed by Wright (1964) is close to our theoretical prediction, in line with earlier numerical confirmation. Additionally, we approximate the invasion probability of a new S-allele, which scales inversely with the number of resident S-alleles. Lastly, we use the stationary allele frequency distribution to estimate the population size of a plant population from an empirically obtained allele frequency spectrum, which complements the existing estimator of the number of S-alleles. Our expression of the stationary distribution resolves the long-standing debate about the correct approximation of the number of S-alleles and paves the way to new statistical developments for the estimation of the plant population size based on S-allele frequencies.

Keywords: gametophytic self-incompatibility; allele frequency distribution; parameter estimation; invasion probability; diffusion approximation; stochastic model

1 Introduction

Gametophytic self-incompatibility (GSI) is a genetically controlled mating system, common in flowering plants. GSI prevents crossing between individuals sharing identical alleles at the self-incompatibility locus (S-locus), especially self-fertilization (Durand et al., 2020). It was early observed that S-allele diversity in natural populations was surprisingly large: in a population of *Oenothera organensis* at least 45 different S-alleles had been identified despite the population comprising at most one thousand individuals (Emerson, 1938, 1939, 1949). The large S-allele diversity is now known to be prevalent in GSI populations across various species and families (Castric and Vekemans, 2004). In the 1930's, explaining the origin, dynamics and maintenance of this diversity was a real challenge for the young fields of population and evolutionary genetics, and to some extent still is (*e.g.* Durand et al., 2020).

Wright (1939) was the first to propose an approximation of the stationary distribution of S-allele frequencies in a finite population, one of the first applications of theoretical population genetics as introduced by Fisher (1930) and Wright (1937). Wright (1939) used his approximation of the stationary distribution to estimate the expected number of S-alleles in a given population of finite size with recurrent mutation. His prediction, applied to *O. organensis* populations, was much lower than the observed number of S-alleles (Emerson, 1938) under a reasonable mutation rate and population size. Wright (1939) then suggested that the discrepancy between theoretical prediction and empirical observation could be due to population subdivision of the *O. organensis* population.

The failure of Wright's attempt to apply a population genetic model to an empirical case generated a fierce and long debate about the good formulation of stochastic models of population genetic dynamics. Fisher (1958) and Moran (1962) were the first to criticize Wright's initial model, mostly for a lack of mathematical rigor. Wright further refined his GSI model (Wright, 1960, 1964), and provided, with others, computer simulations (Kimura and Crow, 1964; Ewens and Ewens, 1966; Mayo, 1966) that confirmed that his approximation of the stationary distribution of S-allele frequencies, as well as his prediction of the expected number of S-alleles, were correct.

Yet, to the best of our knowledge, there is still no theoretical derivation and justification of the allele frequency dynamics and stationary state in a GSI population, even though, more than 50 years ago, Moran (1962, p.163) already suggested that the probabilistic model underlying GSI should be properly specified. Wright (1964) tentatively answered Moran's criticism by framing it as "a basic difference in viewpoint". Ewens and Ewens (1966) then argued that "[Wright's] approximations are far better than could reasonably be expected". An explicit derivation of the (macroscopic) properties of a GSI population from the individual (microscopic) level is necessary to determine the assumptions under which the approximation remains valid, and therefore can be used to interpret empirical observations and to conduct parameter inference. Indeed, paraphrasing Ewens and Ewens (1966): the validity of Wright's approximation is "possibly fortuitous", as suggested by some of the results by Mayo (1966) where the approximation can be incorrect (see Ewens (1964) and Ewens and Ewens (1966) for technical arguments).

In the end, one question remains unsolved: Wright's approximation of the GSI stationary distribution seems to work well, but we do not know why. Here, we address this theoretical gap. We start by defining an individual-based model of gametophytic self-incompatibility. Taking the infinite population size limit, we first obtain the deterministic dynamics of gametophytic self-incompatibility, an equation that has already been derived and studied in great detail (Fisher, 1958; Nagylaki, 1975; Boucher, 1993). Second, we derive a diffusion approximation and compute the stationary distribution based on the central limit theorem for density-dependent Markov processes (Kurtz, 1971). We obtain an explicit expression for the stationary distribution that only depends on the population size and the number of S-alleles in the population, which formally confirms the validity of Wright's approximation under general conditions. Finally, we also provide new results: 1) the invasion probability of a novel S-allele in a resident population, and 2) a method to estimate the population size from S-allele frequency data, based on our explicit result of the stationary distribution of S-allele frequencies.

2 Stochastic model of gametophytic self-incompatibility

We consider a diploid plant population with fixed size N with overlapping generations (Moran-type model). The analogous model definition of the self-incompatibility dynamics with non-overlapping generations (Wright-Fisher-type model) is given in the Supplementary Material (SM), Section C. We assume a fixed number of S-alleles in the population and denote it by M. The number of plants with genotype $\{ij\}$ is denoted A_{ij} with no allele order (*i.e.* $A_{ij} = A_{ji}$). Because of GSI, a $\{ij\}$ individual can only be fertilized by pollen of type $k \neq i, j$ and then produces an offspring of type $\{ik\}$ or $\{jk\}$ with equal probability, consequently $A_{ii} = 0$ for all i. The dynamics of the number of $\{ij\}$ individuals is described in terms of coupled birth and death events to maintain the fixed population size N, as is common in Moran-type models. Then, the number of $\{ij\}$ individuals, A_{ij}, increases by one if

(b-i) a type i pollen fertilizes a $\{j\bullet\}$ individual that transmits allele j, or

(b-ii) a type j pollen fertilizes a $\{i\bullet\}$ individual that transmits allele i,

and a non-$\{ij\}$ individual dies and is replaced.

The probability of fertilization of a $\{jk\}$ individual with an i pollen is given by $p_i/(1-p_j-p_k)$, where p_i is the frequency of i-type pollen in the population ($p_i = \sum_{k\neq i} \frac{A_{ik}}{2N}$). The form emerges from conditioning on successful fertilization, which reflects the assumption that the number of pollen is essentially infinite, or at least not limiting the (female) reproductive success of a plant, *i.e.* we assume no pollen limitation.

The rate at which the number of $\{ij\}$ individuals increases by 1, the *birth rate*, is then given by:

$$T^+_{ij} = \left(\underbrace{\frac{1}{2}\sum_{k\neq i,j} A_{jk}\frac{p_i}{1-p_j-p_k}}_{i \text{ pollen fertilization}} + \underbrace{\frac{1}{2}\sum_{k\neq i,j} A_{ik}\frac{p_j}{1-p_i-p_k}}_{j \text{ pollen fertilization}} \right) \underbrace{\frac{N-A_{ij}}{N}}_{\text{non-}\{ij\}\text{ replacement}} . \tag{1}$$

The reproduction terms between brackets were first derived by Fisher (1958).

Using the same arguments, the number of $\{ij\}$ individuals decreases by one if

(d-i) a non-$\{ij\}$ plant is fertilized (by any pollen), or

(d-ii) a $\{i\bullet\}$ plant is fertilized but the offspring is not $\{ij\}$, or

(d-iii) a $\{j\bullet\}$ plant is fertilized but the offspring is not $\{ij\}$,

and a $\{ij\}$ individual dies and is replaced.

The rate at which the number of $\{ij\}$ individuals decreases by one, the *death rate*, is then given by (detailed derivation in the SM, Section A):

$$T^-_{ij} = \underbrace{\left(N - \frac{1}{2}\sum_{k\neq i,j} A_{ik}\frac{p_j}{1-p_i-p_k} - \frac{1}{2}\sum_{k\neq j,i} A_{jk}\frac{p_i}{1-p_j-p_k} \right)}_{\text{non-}\{ij\}\text{ reproduction}} \underbrace{\frac{A_{ij}}{N}}_{\{ij\}\text{ replacement}} . \tag{2}$$

Since all individuals are equally likely to reproduce (at rate 1), *i.e.* the overall rate of reproduction is N, the reproduction part of the death rate is given by N minus the reproduction rate of an $\{ij\}$ individual (Eq. (1)), which explains the term between brackets in Eq. (2).

We now proceed with the theoretical analysis of this stochastic model.

3 Results

3.1 Approximation of the dynamics in a finite population

First, we derive a diffusion approximation of the stochastic dynamics of genotypes and S-alleles in a large but finite GSI plant population. To this end, we apply standard mathematical arguments (reviewed in the context of evolutionary applications in Czuppon and Traulsen (2021)). Denoting by $a_{ij}(t) = A_{ij}/N$ the density of $\{ij\}$ individuals in the population at time t, we find (ignoring terms related to covariances between the different genotypes)

$$da_{ij}(t) = \mu_{ij}(t)dt + \sqrt{\frac{\sigma^2_{ij}(t)}{N}}dW^{ij}_t, \tag{3}$$

where $\mu_{ij}(t)$ is the deterministic dynamics of the genotype frequencies (see below), $\sigma^2_{ij}(t)$ is the infinitesimal variance of genotype frequencies (explicitly given in SM, Section A, Eq. (A.4)), and $(W^{ij}_t)_{t\geq 0}$ are standard Brownian motions related to the stochastic fluctuations of genotype frequencies, which model the inherent randomness of births and deaths in a finite population. Notably the stochastic fluctuations scale with $1/\sqrt{N}$ and vanish in the limit of infinite populations ($N \to \infty$), where only the deterministic dynamics μ_{ij} remain. The deterministic change of the genotype frequencies is given by

$$\mu_{ij} = \frac{1}{2}\left(\sum_{k\neq i,j} a_{jk}\frac{p_i}{1-p_j-p_k} + \sum_{k\neq i,j} a_{ik}\frac{p_j}{1-p_i-p_k}\right) - a_{ij}. \tag{4}$$

Solving $\mu_{ij} = 0$ in Eq. (4) gives the steady state $p^*_i = 1/M$ and $a^*_{ij} = 2/(M(M-1))$ where all genotypes on the one hand, and all S-alleles on the other hand, have identical frequencies (Nagylaki, 1975; Boucher, 1993).

The diffusion approximation of the dynamics given in Eq. (3) shows that trajectories randomly fluctuate with order $1/\sqrt{N}$ around the deterministic trajectory (we refer to Ethier and Kurtz (1986) for a rigorous treatment). The precise scaling of these fluctuations is obtained by explicitly computing the infinitesimal variance $\sigma^2_{ij}(t)$ (SM, Section A).

The derivation of the S-allele dynamics is obtained by similar arguments and using the relation $p_i = \sum_{j\neq i} a_{ij}/2$ (details in SM, Section A). The deterministic change in S-allele frequency, *i.e.* the infinitesimal mean, is then given by

$$\frac{dp_i}{dt} = \mu_i = \frac{1}{2}\sum_{j\neq i}\mu_{ij} = \frac{p_i}{2}\left(\frac{1}{2}\sum_{j\neq i}\sum_{k\neq i,j} a_{jk}\frac{1}{1-p_j-p_k} - 1\right). \tag{5}$$

Eq. (5) shows that, since we assume no pollen limitation, the deterministic dynamics of allele i is driven by pollination of non-i plants (first term in the brackets), rather than pollination of i-plants by pollen of a different type (Fisher, 1958; Nagylaki, 1975; Boucher, 1993).

The diffusion approximation of S-allele frequencies is

$$dp_i(t) = \mu_i(t)dt + \sqrt{\frac{\rho^2_i(t)}{N}}dW^i_t, \tag{6}$$

where $(W^i_t)_{t\geq 0}$ is a standard Brownian motion that reflects the stochastic fluctuations of the i-th S-allele frequency around the deterministic trajectory μ_i. The exact expression of the infinitesimal variance of the S-allele frequency change (ρ^2_i) is computed in SM, Section A, Eq. (A.9). Note that the infinitesimal variance of S-alleles depends on the covariance between genotype frequencies.

3.2 Approximation of the stationary distribution

Since the deterministic fixed point is globally stable (Boucher, 1993), the stochastic trajectories of S-allele frequencies will fluctuate around the value $1/M$, at least as long as the probability of extinction of a S-allele is negligibly small. This will typically be the case if $M \ll \sqrt{N}$. Using arguments motivated by the central limit theorem of density-dependent Markov processes (Ethier and Kurtz, 1986; van Kampen, 2007), we can then approximate the fluctuations around the deterministic steady state by a normal distribution (details in SM, Section B). We find that the stationary distribution of S-allele frequencies, denoted by ψ, is distributed as ($\sim$ sign)

$$\psi \sim \mathcal{N}\left(\frac{1}{M}, \frac{(M-2)^3}{NM^2(2M-3)}\right), \tag{7}$$

where $\mathcal{N}(m, s^2)$ denotes a normal distribution with mean m and variance s^2.

The same derivation is possible for a model with non-overlapping generations. In this case, we find that the stationary distribution differs only by a factor 2 in the denominator of the variance (SM, Section C, Eq. (C.9)), in line with the generally found difference between the variances of the Moran and Wright-Fisher model (Czuppon and Traulsen, 2021).

Fig. 1 shows that the approximation of the stationary distribution in Eq. (7) predicts well the stationary distributions obtained by stochastic simulations, especially if the number of S-alleles (M) is small. The approximation of the stationary distribution is slightly worse for large numbers of S-alleles because the probability that an allele is lost by chance increases with the number of S-alleles in the population, as shown by the bar located at zero for $M = 15$ in Fig. 1. As a consequence of an allele extinction, the mean allele frequency is displaced from $1/M$ to $1/(M-1)$ in some simulations, or even lower values of M if multiple extinction events occurred. This explains the slightly worse fit of the approximated stationary distribution for $M = 15$, compared to lower numbers of S-alleles.

3.3 The number of S-alleles maintained in a finite population

To determine the number of S-alleles stably maintained in a population of finite size, we follow the partially heuristic reasoning by Wright (1939) and fill in the analytical gaps using the stationary distribution and the individual-based model. The basic idea is as follows (Wright, 1939): First, given an approximation of the stationary distribution of the allele frequencies for a given number of S-alleles M in a finite population of size N (Eq. (7)), the *loss rate* of a focal S-allele is computed by multiplying the probability for this S-allele to be at frequency $1/2N$ with its death rate T^- at frequency $1/2N$ (computation in SM, Section D). Second, the loss rate is balanced with the *gain rate* of new S-alleles which is proportional to the mutation rate, denoted u. Assuming an infinitely many alleles model, the rate at which new S-alleles arise is simply given by the mutation rate u (because the overall rescaled reproduction rate in the population is 1 due to the conditioning on successful reproduction). This yields the following equality

$$u = \underbrace{\frac{M}{N}}_{T^- \times M} \times \underbrace{\int_{-\infty}^{1/2N} \psi(x)\,dx}_{S\text{-allele at freq. } 1/2N}, \tag{8}$$

which can be solved numerically for M. Fig. 2 shows a comparison between the prediction obtained by solving Eq. (8) for different values of M and stochastic simulations. The number of S-alleles maintained in a finite population is reasonably well predicted by our model, even though it is overestimated as population size increases.

The overall good fit confirms the previous conclusion from numerical simulations (Ewens and Ewens, 1966; Mayo, 1966) that Wright's (1939) methodology indeed correctly predicts the number of

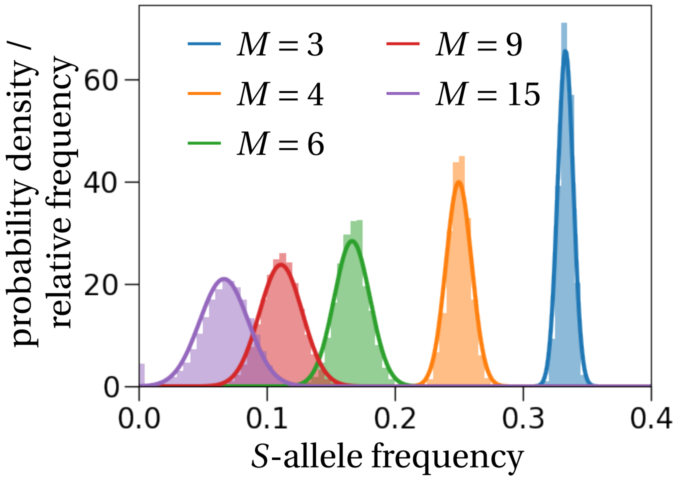

Figure 1: **Stationary distribution of S-alleles.** As predicted (solid lines, Eq. (7)), the stationary distribution of the allele frequency is well approximated by a normal distribution centered at $1/M$. The histograms are calculated from 100 stochastic simulations that ran for 1,200 generations in a population of size $N = 1,000$, where we recorded the frequency of the same S-allele every generation. To ensure no dependence on the initial state we started recording frequencies after 200 generations. If the number of S-alleles becomes too large, *e.g.* $M = 15$, alleles might get lost, which is shown by the purple bar at the frequency equal to zero.

S-alleles in a population. Note that we adapted Wright's (1939) methodology for predicting the number of S-alleles in a finite population with two small adjustments: we used our new approximation of the stationary distribution (Eq. (7)) that is theoretically justified by the central limit theorem (yet neglecting allele covariances), and we used the explicit expression of the death rate of a single individual with the focal S-allele derived from the individual-based description of the model (instead of different intuitively proposed rates by Wright (1939, 1964)).

3.4 Invasion probability of a new S-allele

We now derive new results on the invasion behavior of a novel S-allele. The invasion probability can be computed from the diffusion approximation (Eq. (6)) by applying results from stochastic diffusion theory (mathematical details in SM, Section E). We find that a new S-allele, appearing at frequency $1/2N$ by mutation (or immigration), establishes in a population with M resident S-alleles with probability

$$\varphi = \frac{2}{2M-3} \quad (M \geq 3). \tag{9}$$

This approximation agrees well with simulation results (Fig. 3). Surprisingly though, it fits the simulation results better than the non-approximated formula from which Eq. (9) is derived; more details in SM, Section E.

The invasion probability decreases with M because the rare allele advantage is smaller when the number of resident S-alleles M increases. This is explained by an increase in the number of plants that can be fertilized by a single resident S-allele for larger numbers of resident alleles. The invading allele is initially present in a single copy, which means that the number of individuals it can fertilize is always $N-1$. All the resident S-alleles, assuming that genotype frequencies are in equilibrium at

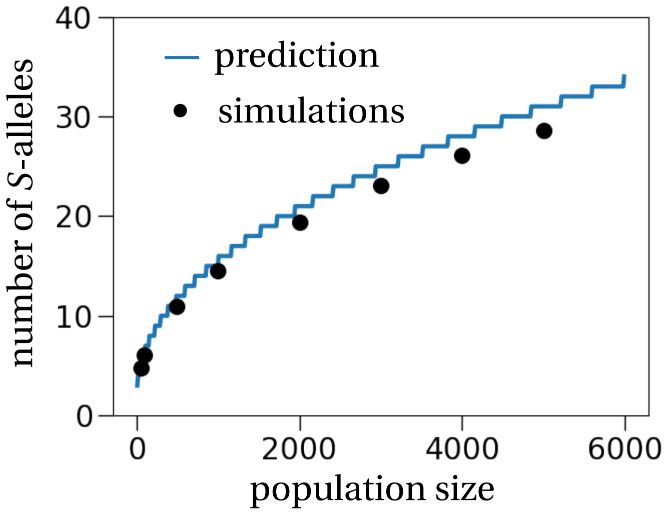

Figure 2: **The number of S-alleles in a finite population.** The general methodology proposed by Wright (1939), made explicit by our theoretical approach (solid line computed by Eq. (8)), predicts well the number of S-alleles in a finite population, obtained from stochastic simulations (dots). Each simulation was initiated with 30 different S-alleles and the number of S-alleles was recorded every 100 generations for 10,000 generations after a 10,000 generation burn-in period. The mutation rate was set to $u = 1/(100N)$, *i.e.*, one mutation every 100 generations on average. This procedure was repeated 50 times for every population size.

$2/M(M-1)$ (except for one genotype that is at frequency $2/M(M-1) - 1/N$ because of the invading allele), can fertilize

$$\underbrace{\frac{(M-1)(M-2)}{2}}_{\text{no. of non-}S\text{-allele genotypes}} \times \underbrace{\frac{2}{M(M-1)}N}_{\text{no. of individuals with same genotype}} = \frac{M-2}{M}N, \quad (M \geq 3) \tag{10}$$

individuals. This shows that as M increases, the difference in the rate of successful pollen fertilization between the invading S-allele, *i.e.* $N-1$, and the resident alleles, *i.e.* $N(M-2)/M$, decreases. In contrast, since we assume no pollen limitation, the probability for a plant to be fertilized only depends on its genotype frequency in the population.

The invasion probability of a new S-allele under GSI, denoted by φ (Eq. (9)), can also be compared to the case of haploid self-incompatibility (HSI), a mating system found for example in some fungi and ciliates. Czuppon and Rogers (2019) showed that a good approximation for the invasion probability of an S-allele in an HSI population with M resident S-alleles is $1/M$. Here, for GSI, we find $\varphi > 1/M$, *i.e.* the invasion probability of a new S-allele is always larger for GSI than for HSI. This can be understood by comparing the ratio between the number of individuals that are compatible with the invading and with resident S-alleles in both situations. Applying the same reasoning that led to Eq. (10), we find

$$\begin{aligned} (N-1) \times \frac{M}{N(M-2)} &\approx \frac{M}{M-2}, \text{ for GSI,} \\ (N-1) \times \frac{M}{N(M-1)} &\approx \frac{M}{M-1}, \text{ for HSI.} \end{aligned} \tag{11}$$

This shows that an invading (or rare) S-allele always has a larger fertilization advantage relative to a resident S-allele in GSI than in HSI. As a consequence, for similar population sizes, we would a higher

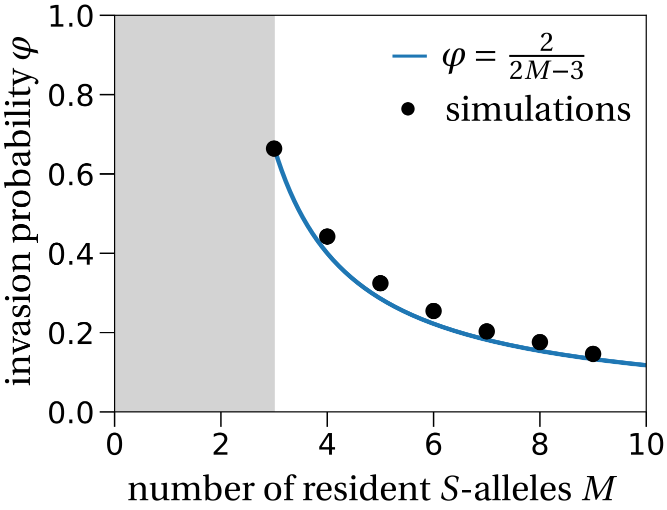

Figure 3: **Invasion probability of a new S-allele.** The invasion probability for different numbers of resident S-alleles, M, fits the result obtained from the diffusion approximation in Eq. (9) (solid line). The population size is set to $N = 10,000$ and the number of independent stochastic simulations is 10,000 per number of resident S-alleles.

S-allele diversity under GSI than under HSI dynamics because the rare allele advantage is stronger in the GSI mating system. The equations also show that as M increases, the difference between GSI and HSI vanishes as new S-alleles are getting closer to being neutral. Hence, as M increases, GSI and HSI tend to have similar invasion dynamics.

3.5 Estimation of the number of S- alleles and the population size from data

Our goal in this section is to provide a joint estimator of both the population size and the number of S-alleles in a sampled plant population genotyped at the S-locus. The main idea is to use the mean and the variance of the stationary distribution (Eq. (7)) to estimate the two population parameters of interest, M and N, given the sample size n and the number of S-alleles observed in the sample.

We first extended Paxman's (1963) urn model (details are given in SM, Section F). Adopting the same idea, we derived a likelihood estimator for the number of S-alleles and the population size from the mean and variance of the number of occurrences Y of a given S-allele observed in the sample (SM, Section G). We show that the mean $\mathbf{E}[Y]$ and the variance $\mathbf{V}[Y]$ can be used to estimate the number of S-alleles and the population size. However, the likelihood maximization procedure showed convergence instability: even though the number of S-alleles was correctly estimated, the population size estimator performed rather poorly (SM, Section G, Fig. C).

Second, we defined a simpler estimate by directly fitting the normal distribution from the approximated stationary distribution to the empirical S-allele frequency distribution. From the estimated mean and variance of the fitted normal distribution, denoted by $\widehat{\mu}$ and $\widehat{\rho}^2$, we can directly compute the estimates for the number of S-alleles, $\widehat{M}$, and the population size, $\widehat{N}$, using Eq. (7):

$$\widehat{M} = \frac{1}{\widehat{\mu}} \quad \text{and} \quad \widehat{N} = \frac{(\widehat{M}-2)^3}{\widehat{\rho}^2 \widehat{M}^2 (2\widehat{M}-3)}. \tag{12}$$

Note that by definition, the estimated number of S-alleles, $\widehat{M}$, is exactly the observed number of S-

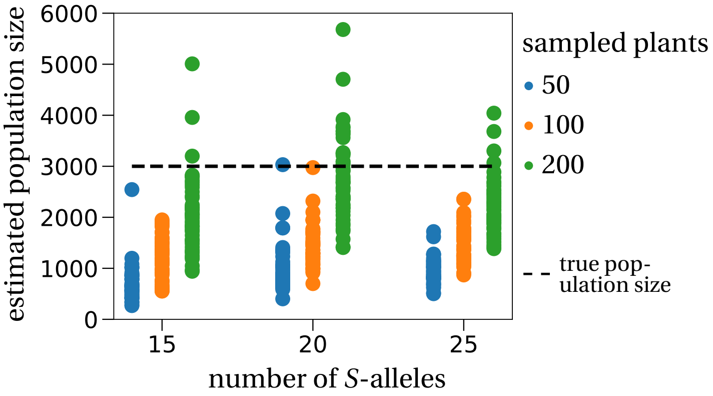

Figure 4: **Estimated population sizes from simulated data.** Overall, we observe a tendency to underestimate the true population size (3000 plants, dashed line). This effect is most pronounced for small sample sizes (50 sampled plants, blue dots) and becomes less prominent for intermediate (100 sampled plants, orange dots) and large sample sizes (200 sampled plants, green dots). The single dots are population size estimates for different random samples taken from the same data set, implemented by a hypergeometric distribution (sampling without replacement). The simulated population of size 3000 with 15, 20 or 25 different S-alleles was started close to its steady state. The random sample is taken after 500 generations.

alleles in the sample (the expectation of the allele frequency spectrum equals the number of observed S-alleles). This estimator therefore underestimates the true number of S-alleles in the population.

We applied this second estimate to a simulated plant population of size 3000 with different numbers of S-alleles. We find that the estimated population sizes mostly underestimate the true population size that is depicted as a dashed line in Fig. 4. This effect is most pronounced for small sample sizes (blue dots in Fig. 4). This is explained by smaller samples showing a larger variance in the allele frequency distribution, which directly translates to smaller values of the estimated population size $\widehat{N}$ (Eq. (12)). The larger the sample size, the better is the population size approximated, yet there is still considerable variation in the estimates.

Lastly, we applied our estimator to an empirical dataset from Stoeckel et al. (2012), where a population of wild cherry (*Prunus avium*) was quasi-exhaustively sampled ($n = 249$). We estimated $\widehat{M} = 18$ and $\widehat{N} = 120$, which is of the same order as, though underestimates, the census population size n. From the 95% confidence interval (CI) of the sample variance, we computed the 95% CI for the population size, which spans from $\widehat{N} = 53$ to $\widehat{N} = 213$, still not covering the sample size. The discrepancy between the estimation and the census population size can be due to a bias of our estimator. Another possible reason for this underestimation is violation of our modeling assumptions, *e.g.* spatial structure of the population or fecundity differences between individuals due to their size (Stoeckel et al., 2012). Overall, our results may serve as a starting point to develop new statistical estimators that can be used to infer ecological parameters from the diversity observed at a S-locus. In particular, as a perspective, it might be possible to combine Paxman's model with a fitted normal distribution to jointly estimate the number of S-alleles and the population size, similar to our first attempt in SM, Section G.

4 Discussion

Self-incompatibility as an historical illustration of probabilistic thinking

As emphasized by Wright (1937), one of the central goals of population genetics is to find the frequency distribution of genetic variants under various conditions, in particular, for convenience, under stationarity. GSI was one of the first genetic systems used to validate theoretical predictions from the young field of population genetics in the 1930-40's. Before detailing how GSI was used to bring together theory and empirical observations, we first briefly summarize the general theory as stated by Fisher and Wright in the 1930's.

The approach taken by Wright was to split the frequency changes into two parts: a deterministic part (describing mutation, migration or selection) and a stochastic part (describing random sampling errors). Wright generally first states how the frequency of a given allele p_i changes in expectation from one generation to the next (by a difference equation), and second derives the form of the stochastic fluctuations, here denoted ρ_i^2. To derive the stochastic fluctuations, he usually assumes that the gene frequency dynamics are a succession of binomial draws, where the probability for a specific allele to be transmitted is equal to its frequency in the gamete offspring pool, assumed to be infinite. These allele frequencies in the offspring pool could have changed, compared to the parental generation, because of the processes of selection, migration or mutation (*e.g.* Wright, 1937). In this model, the stochastic fluctuations from one generation to the next are given by the variance of a binomial distribution $\Delta\rho_i^2 = \frac{p_i(1-p_i)}{2N}$. The deterministic change of the allele frequency Δp_i should appear in the binomial distribution parameters, but Wright explicitly assumes that it is negligible and thus that stochastic fluctuations only depend on the population size and the actual allele frequency, generally referred to as "genetic drift".

Using a continuous approximation of the discrete stochastic process, Wright (1937, 1938) calculated the change of the mean and variance of the allele frequencies, giving a possible approximation of the distribution of allele frequencies for various cases regarding selection, migration and mutation, and its explicit stationary distribution. In his papers, Wright (1937, 1938) acknowledged that Fisher (1930) found similar results, despite Fisher's approach being different. Thanks to a letter from the mathematician Kolmogorov, Wright (1945) realized that the form he had found for the allele distribution was the solution of the more general Fokker-Planck equation used in physics, which describes a stochastic process combining deterministic directed change on the one hand, and random motion on the other hand. At the same time, Malécot considered the change of the allele frequencies as a Markov process (reviewed in Nagylaki, 1989). In particular, Malécot rigorously demonstrated that assuming weak selection and a large population size, the form conjectured by Wright for the distribution of the allele frequencies was correct, probably the first application of the diffusion approximation to population genetics (Malécot, 1945). After a little more than two decades, Wright, Fisher, Kolmogorov and Malécot founded the basics of the stochastic models of population genetics. It then remained to apply this general framework to a real scenario. GSI was an ideal problem since it combines selection, mutation and migration, with a parameter easily estimated from natural populations, the number of S-alleles.

Very rapidly, even before the early general theory of population genetics was definitely established, Wright (1939) aimed at using the theory to resolve an apparent paradox from observations in one population of the flowering plant *O. organensis*. In this population, 45 self-incompatibility alleles were detected by Emerson (1938, 1939, 1949) even though the population size was thought to be not larger than one thousand individuals. Wright (1939) followed the general approach developed before (Fisher, 1930; Wright, 1938): he first provided the deterministic change of the frequency of a focal S-allele given the frequency of all the other S-alleles in the population, summarized in the parameter R, which measures the relative success of fertilization of the focal S-allele compared to the other S-alleles: $\Delta p_i = \frac{p_i(1-R-p_i(3-R))}{2(R+p_i(1-R))}$. Wright remained elusive on the derivation and biological interpretation of R, but it can be seen as a selection coefficient, which depends on the genotype structure of the population.

Second, as in his previous works (Wright, 1937, 1938), Wright assumed that the stochastic fluctuations of the i-th S-allele frequency in a generation is given by the variance of a binomial distribution, from which he derived an approximation of the S-allele frequency distribution. As there was no explicit solution for the stationary distribution, Wright (1939) provided numerical results. He finally concluded that the large number of S-alleles observed in the *O. organensis* population would be possible if the mutation rate was as large as $10^{-3} - 10^{-4}$ per generation, or if the observed population was a small part of a much larger subdivided population with a low dispersal rate and a mutation rate of $10^{-5} - 10^{-6}$.

During almost twenty years, no important advances were made in the theory of stochastic processes applied to self-incompatibility systems. After experimental estimations of the mutation rate, found to be much lower than 10^{-6} in *O. organensis* by Emerson (1939) and Lewis (1948, 1949), there was a revival of interest for having the best approximation of the frequency distribution of S-alleles in a population, launched in particular by a reexamination of the theory by Fisher (1958). Fisher (1958) challenged Wright's (1939) model, called for developing the best stochastic approximation of the GSI system, and using it to identify the mechanisms underlying the pattern first highlighted by Emerson (1938). Fisher adopted the same approach as Wright (1939): first he derived the deterministic frequency change and then he assumed a form for the variance of the allelic change from a binomial distribution. In contrast to Wright however, Fisher (1958) motivated his deterministic change in allele frequency by an individual-based reasoning on the level of the plants. He was thus the first to give an explicit expression for the deterministic change of a given S-allele in the form given in Eq. (5) above. Moreover, he also proposed a different form, compared to Wright (1939), for the variance of the S-allele frequency change, $\Delta\rho_i^2 = \frac{p_i(1-2p_i)}{2N}$. Similar to Wright, Fisher did not provide an explicit derivation of the infinitesimal variance from a probabilistic model. Finally, after deriving an allele frequency distribution under some technical simplification, Fisher came back to the plausible mechanisms explaining the large number of S-alleles observed in the *O. organensis* population (Emerson, 1938). He concluded that the population had most likely suffered from a strong population bottleneck, resulting in the population not being at drift-mutation equilibrium. Fisher discarded Wright's (1939) hypothesis that the population could be strongly subdivided, with no clear reasons since Fisher did not consider any population structure or dispersal in his model. Fisher also took the opportunity to criticize that Wright (1939) only relied on numerical computations to analyze his model, and was not able to derive any explicit formula.

Wright (1960) answered Fisher's (1958) criticisms mostly by comparing his 1939's and new 1960's models, and Fisher's model through numerical simulations. Wright found no notable difference between the different models. Still he pointed out that Fisher's approximation gave the worst estimation of S-allele frequency change among the three models. Overall, Wright (1960) and Fisher (1958), compared to Wright (1939), brought no important conceptual progress to the theory of GSI. Yet, these two papers illustrate that finding the correct form of approximation of the genetic dynamics of a GSI system was a major challenge for population genetics that led to another dispute between Wright and Fisher, beside the famous controversy about genetic dominance (Bagheri, 2006; Billiard and Castric, 2011).

A few years after this dispute, many papers were at least partly devoted to address the question of the correct approximation of the stationary distribution of S-allele frequencies in a finite population (Moran, 1962; Wright, 1964; Kimura and Crow, 1964; Ewens, 1964; Mayo, 1966; Ewens and Ewens, 1966; Wright, 1966). Yet again, no theoretical advances were made; most results relied on computer simulations that confirmed that Wright's (1964) approximation was good enough. Later, Yokoyama and Nei (1979) and Yokoyama and Hetherington (1982) obtained new results by novel diffusion theory methods: instead of computing the allele frequency of a single allele, they computed the entire allele frequency distribution to directly infer the number of stably maintained *S*-alleles in a population. Yet again, similar to Wright's different models (1939, 1960, 1964), they relied on a numerical approximation of the stationary distribution. Finally, and more recently, Muirhead and Wakeley (2009) derived recursion formulae to approximate the *S*-allele frequency distribution. This latest approach does not take a continuum limit (infinite population size limit) but approaches the problem in the discrete state space. While giving good results when compared to simulation results, this approach makes the formulae intractable and complicated to evaluate.

It is remarkable that the population genetics of self-incompatibility systems generated a lot of theoretical progress motivated by the aim to have theoretical predictions compatible with empirical observations. There was clearly a challenge for theoretical population geneticists, as if the stochastic theory of population genetics would be valuable only if it could help to explain the genetic diversity observed in natural populations, or at least if it could give correct predictions about parameter values like the mutation rate. In short, the main question that theoreticians tried to address was: what is the correct way to model the stochastic dynamics of allele frequencies in finite populations?

Two points of view opposed: the need for rigorous *vs.* practical derivations. Moran (1962) particularly criticized the lack of rigour of Wright's approach, while Wright (1964) claimed that he only aimed for a *good* rather than a *justified* approximation. Such an opposition raised two difficulties: Is it possible to evaluate the validity of a model independently of data obtained from natural populations? How can the robustness and precision of the model's approximation be evaluated? As emphasized by Moran (1962) and Malécot (Nagylaki, 1989), these goals can be achieved by using a probabilistic way of thinking, *i.e.* by clearly justifying a stochastic model, and by properly deriving its approximation with probabilistic mathematical tools. Here, we applied this theoretical framework to GSI.

Reexamination of the number of *S*-alleles in a finite population

One aim of our study was to analytically revisit the results on the number of *S*-alleles under GSI, as derived by Wright (1939, 1964). Even though Wright's results were numerically confirmed (see previous section), a theoretical confirmation was still lacking (to the best of our knowledge). Our results on the stationary distribution fit the numerically derived normal distribution obtained by Wright (1964) almost perfectly, *e.g.* Fig. 5, thus confirming Wright's results also from a theoretical perspective.

Following Wright's (1939) strategy, we computed the number of *S*-alleles that can be maintained in a finite population of fixed size and for a fixed mutation rate (Fig. 2). Our analytical prediction slightly overestimates the simulated values. This is explained by the idealized assumption that all *S*-allele frequencies are centered around the steady state. Therefore, deviations from this stationary distribution are underestimated, which results in the overestimation of the simulated number of *S*-alleles. If we were to compute the number of *S*-alleles without the normal approximation of the stationary distribution, we would instead slightly underestimate the number of *S*-alleles as found in Czuppon and Rogers (2019) in the related HSI system.

Our results were derived assuming neither fitness differences between different *S*-alleles, nor phenomena such as partial self-compatibility. Accounting for these mechanisms would alter the prediction of the number of *S*-alleles. For example, it was shown that variation in zygote viability reduces the diversity at the *S*-locus (Uyenoyama, 2003), which is similar to results found in HSI systems (Krumbeck et al., 2020). In addition, we can speculate that the effect of self-fertilization on *S*-allele diversity maintained under GSI would be similar to that under HSI. We would therefore expect that the number of *S*-alleles is reduced for increasing rates of self-fertilization (Constable and Kokko,

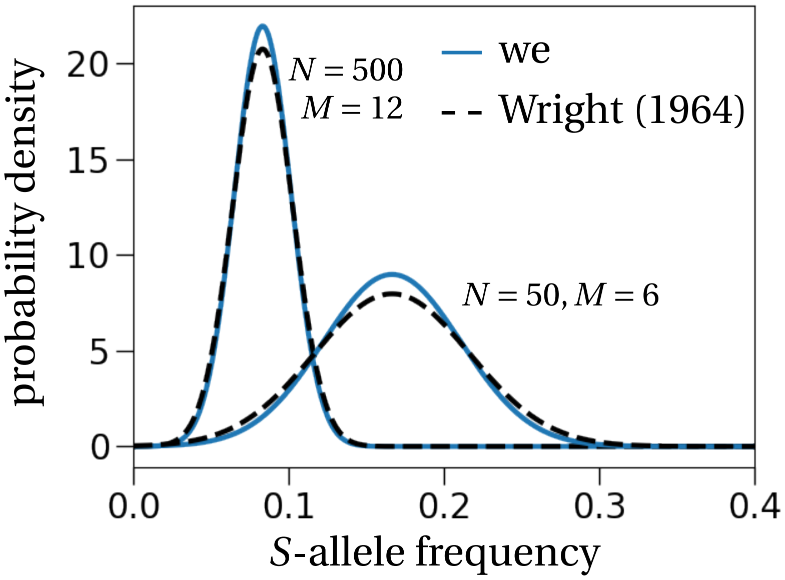

Figure 5: **Stationary allele frequency distribution – comparison between Wright (1964) and our result.** Our prediction is the result from Eq. (7) adapted to non-overlapping generations as considered by Wright (1964). The exact expression is stated in Eq. (C.9) in the SM. The black dashed lines correspond to the expressions computed by Wright (1964), for $M = 12$ S-alleles in a population of size $N = 500$ (left) and $M = 6, N = 50$ (right). We stress that Wright remains vague on the evaluation of the variance of this distribution and computes it numerically, *i.e.* he does not derive it analytically as we do (we refer to his discussion in section "Some examples" in Wright (1964)).

2018; Czuppon and Constable, 2019; Berríos-Caro et al., 2021).

Connecting theoretical results to data

Returning to the question about the observed diversity at the S-locus in *O. organensis*, Wright (1960) concluded that the empirically estimated low mutation rate would not be possible to maintain the 45 sampled S-alleles in an estimated population of 500 individuals (Emerson, 1949). Instead, he proposed that population subdivision could be an explanation. However, as shown by simulations, population subdivision reduces the number of S-alleles in the population (Schierup, 1998), which contrasts the idea advocated by Wright (1939). Another hypothesis for the large number of observed S-alleles is a recent decrease in population size (Fisher, 1958) or that the population size has been underestimated (Levin et al., 1979). To our knowledge, the question about the large number of S-alleles in *O. organensis* has never been resolved, at least partly because appropriate theoretical and statistical tools were lacking. Dispersal kernels have been estimated in other GSI populations (*e.g.* Stoeckel et al., 2012), but estimating the effective population size is still necessary to identify and disentangle the processes underlying the observed diversity at the S-locus.

We used our approximation of the stationary distribution to derive a joint estimator for the number of S-alleles and the plant population size from a sample. The procedure is rather simple: we fit a normal distribution to the sampled allele frequency spectrum. This procedure will always estimate that there are as many S-alleles in the population as there are observed S-alleles in the sample. In this respect, our estimation is less accurate than Paxman's (1963) model that accounts for sampling error (details in SM, Section F). Yet, when estimating the population size from the allele frequency spectrum (by Eq. (12)), this results in much better estimates than an extension of the method proposed by Paxman (1963) (compare Fig. 4 with Fig. C in SM, Section G). We therefore suggest to use the classical Paxman-estimator to estimate the number of S-alleles in the population and then to proceed with our estimator for the population size as defined in Eq. (12). Yet, we recommend to take the estimator of population size with care because it has a tendency to underestimate the true population size (Fig. 4). One possible way to check whether the obtained estimate is reasonable, is to compare the estimated population size with a prediction of the number of maintained S-alleles in the population (Fig. 2). If the population size is too small compared to the number of observed S-alleles, the estimate is very likely an underestimate of the true population size. Alternatively, it is of course possible that other assumptions of our neutral model are violated, *e.g.* non-stationarity of the S-allele frequency distribution, which could suggest a recent population bottleneck. A better understanding of the limitations of this new estimator is therefore needed to reliably apply it to data sets.

Conclusion

To summarize, we have revisited the question of the possible number of S-alleles maintained in a finite population from a theoretical point of view. Our explicit expression of the stationary allele frequency distribution confirms the numerically obtained results by Wright (1964). Additionally, we provide an approximation for the establishment probability of a new S-allele and find that it is similar to the corresponding expression in populations with haploid self-incompatibility. Lastly, we define an estimator for the population size of a GSI population, which provides a starting point to assess the state of the sampled population. For example, a larger estimated diversity of S-alleles than predicted by the estimated population size in Fig. 2 might be indicative for a recent population decline; information that can be important from a conservation biology perspective.

Data availability

The C++ codes, data files and Python scripts used to generate the figures are available at `https://gitlab.com/pczuppon/stoch_gsi`.

Acknowledgments

We warmly thank Solenn Stoeckel and Stéphanie Mariette who kindly shared their *Prunus avium* dataset. We also thank Vincent Castric and Xavier Vekemans for fruitful discussions.

References

Bagheri, H. Unresolved boundaries of evolutionary theory and the question of how inheritance systems evolve: 75 years of debate on the evolution of dominance. *Journal of Experimental Zoology Part B*, 306:329–359, 2006. doi: 10.1002/jez.b.21069.

Berríos-Caro, E., Galla, T., and Constable, G. W. Switching environments, synchronous sex, and the evolution of mating types. *Theoretical Population Biology*, 138:28–42, 2021. doi: 10.1016/j.tpb.2021.02.001.

Billiard, S. and Castric, V. Evidence for Fisher's dominance theory: how many 'special cases'? *Trends in Genetics*, 27:441–445, 2011. doi: 10.1016/j.tig.2011.06.005.

Boucher, W. A deterministic analysis of self-incompatibility alleles. *Journal of Mathematical Biology*, 31(2), 1993. doi: 10.1007/bf00171223.

Castric, V. and Vekemans, X. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. *Molecular Ecology*, 13:2873–2889, 2004. doi: 10.1111/j.1365-294X.2004.02267.x.

Constable, G. W. A. and Kokko, H. The rate of facultative sex governs the number of expected mating types in isogamous species. *Nature Ecology & Evolution*, 2(7):1168–1175, 2018. doi: 10.1038/s41559-018-0580-9.

Czuppon, P. and Constable, G. W. A. Invasion and extinction dynamics of mating types under facultative sexual reproduction. *Genetics*, 213(2):567–580, 2019. doi: 10.1534/genetics.119.302306.

Czuppon, P. and Rogers, D. W. Evolution of mating types in finite populations: The precarious advantage of being rare. *Journal of Evolutionary Biology*, 32(11):1290–1299, 2019. doi: 10.1111/jeb.13528.

Czuppon, P. and Traulsen, A. Understanding evolutionary and ecological dynamics using a continuum limit. *Ecology and Evolution*, 11(11):5857–5873, 2021. doi: 10.1002/ece3.7205.

Durand, E., Chantreau, M., Le Veve, A., Stetsenko, R., Dubin, M., Genete, M., Llaurens, V., Poux, C, R., C, S., Billiard, Vekemans, X., and Castric, V. Evolution of self-incompatibility in the Brassicaceae: Lessons from a textbook example of natural selection. *Evolutionary Applications*, 13:1279–1297, 2020. doi: 10.1111/eva.12933.

Emerson, S. The genetics of self incompatibility in *Oenothera organensis* . *Genetics*, 23(2):190–202, 1938.

Emerson, S. A preliminary survey of the *Oenothera organensis* population. *Genetics*, 24(4):524–537, 1939.

Emerson, S. Incompatibility in flowering plants. *Biological Reviews*, 24:472–496, 1949.

Ethier, S. and Kurtz, T. *Markov processes: characterization and convergence.* Wiley series in probability and mathematical statistics. J. Wiley & Sons, New York, Chichester, 1986. doi: 10.1002/9780470316658.

Ewens, W. J. and Ewens, P. M. The maintenance of alleles by mutation — Monte Carlo results for normal and self-sterility populations. *Heredity*, 21(3):371–378, 1966. doi: 10.1038/hdy.1966.38.

Ewens, W. On the problem of self-sterility alleles. *Genetics*, 50(6):1433–1438, 1964.

Fisher, R. *The genetical theory of natural selection.* Oxford University Press, New York, Chichester, 1930.

Fisher, R. *The genetical theory of natural selection - A complete Variorum edition.* Oxford University Press, 1958.

Kimura, M. and Crow, J. Number of alleles that can be maintained in a finite population. *Genetics*, 49 (4):725–738, 1964.

Krumbeck, Y., Constable, G. W. A., and Rogers, T. Fitness differences suppress the number of mating types in evolving isogamous species. *Royal Society Open Science*, 7(2):192126, 2020. doi: 10.1098/rsos.192126.

Kurtz, T. G. Limit theorems for sequences of jump markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356, 1971. doi: 10.2307/3211904.

Levin, D. A., Ritter, K., and Ellstrand, N. C. Protein polymorphism in the narrow endemic *Oenothera organensis*. *Evolution*, 33(2):534–542, 1979. doi: 10.1111/j.1558-5646.1979.tb04708.x.

Lewis, D. Structure of the incompatibility gene. I Spontaneous mutation rate. *Heredity*, 2(2):219–236, 1948. doi: 10.1038/hdy.1948.12.

Lewis, D. Structure of the incompatibility gene. II Induced mutation rate. *Heredity*, 3(3):339–355, 1949. doi: 10.1038/hdy.1949.25.

Malécot, G. La diffusion des gènes dans une population mendélienne. *Comptes Rendus de l'Académie des Sciences*, 221:340–342, 1945.

Mayo, O. On the problem of self-incompatibility alleles. *Biometrics*, 22(1):111, 1966. doi: 10.2307/2528218.

Moran, P. *The statistical process of evolutionary theory*. Oxford University Press, Amen House, London, 1962.

Muirhead, C. and Wakeley, J. Modeling multiallelic selection using a Moran model. *Genetics*, 3182: 1141–1157, 2009. doi: 10.1534/genetics.108.089474.

Nagylaki, T. Gustave Malécot and the transition from classical to modern population genetics. *Genetics*, 122:253–268, 1989.

Nagylaki, T. The deterministic behavior of self-incompatibility alleles. *Genetics*, 79(3):545–550, 1975.

Paxman, G. J. The maximum likelihood estimation of the number of self-sterility alleles in a population. *Genetics*, 48(8):1029–1032, 1963.

Schierup, M. H. The number of self-incompatibility alleles in a finite, subdivided population. *Genetics*, 149(2):1153–1162, 1998.

Stoeckel, S., Klein, E., Oddou-Muratorio, S., Musch, B., and Mariette, S. Microevolution of S-allele frequencies in wild cherry populations: respective impacts of negative frequency dependent selection and genetic drift. *Evolution*, 66:486–504, 2012. doi: 10.1111/j.1558-5646.2011.01457.x.

Uyenoyama, M. K. Genealogy-dependent variation in viability among self-incompatibility genotypes. *Theoretical Population Biology*, 63(4):281–293, 2003. doi: 10.1016/s0040-5809(03)00020-0.

van Kampen, N. *Stochastic processes in physics and chemistry*. North Holland, 2007. doi: 10.1016/B978-0-444-52965-7.X5000-4.

Wright, S. The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences of the United States of America*, 23:307–320, 1937. doi: 10.1073/pnas.23.6.307.

Wright, S. The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 24:253–259, 1938. doi: 10.1073/pnas.24.7.253.

Wright, S. The distribution of self-sterility alleles in populations. *Genetics*, 24(4):538–552, 1939.

Wright, S. The differential equation of the distribution of gene frequencies. *Proceedings of the National Academy of Sciences of the United States of America*, 31:382–389, 1945. doi: 10.1073/pnas.31.12.382.

Wright, S. On the number of self-incompatibility alleles maintained in equilibrium by a given mutation rate in a population of given size – A reexamination. *Biometrics*, 16(1):61–85, 1960. doi: 10.2307/2527956.

Wright, S. The distribution of self-incompatibility alleles in populations. *Evolution*, 18(4):609–619, 1964. doi: 10.1111/j.1558-5646.1964.tb01675.x.

Wright, S. Polyallelic random drift in relation to evolution. *Proceedings of the National Academy of Sciences*, 55(5):1074–1081, 1966. doi: 10.1073/pnas.55.5.1074.

Yokoyama, S. and Nei, M. Population-dynamics of a sex-determining alleles in honey bees and self-incompatibility alleles in plants. *Genetics*, 91(3):609–626, 1979. doi: 10.1093/genetics/91.3.609.

Yokoyama, S. and Hetherington, L. E. The expected number of self-incompatibility alleles in finite plant populations. *Heredity*, 48(2):299–303, 1982. doi: 10.1038/hdy.1982.35.