Effects of fine-scale population structure on the distribution of heterozygosity in a long-term study of *Antirrhinum majus*

Parvathy Surendranadh*[1], Louise Arathoon*[1], Carina A. Baskett[1], David L. Field[2], Melinda Pickup[1,3], Nicholas H. Barton[1,4]

*Joint first authors
[1]IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria
[2]School of Science, Edith Cowan University, 270 Joondalup Drive, Joondalup WA 6027 Australia
[3]Greening Australia, 8 St Georges Terrace, Perth, WA 6000, Australia
[4]Corresponding author

Abstract

Many studies have quantified the distribution of heterozygosity and relatedness in natural populations, but few have examined the demographic processes driving these patterns. In this study, we take a novel approach by studying how population structure affects both pairwise identity and the distribution of heterozygosity in a natural population of the self-incompatible plant *Antirrhinum majus*. Excess variance in heterozygosity between individuals is due to identity disequilibrium (ID), which reflects the variance in inbreeding between individuals; it is measured by the statistic g_2. We calculated g_2 together with F_{ST} and pairwise relatedness (F_{ij}) using 91 SNPs in 22,353 individuals collected over 11 years. We find that pairwise F_{ij} declines rapidly over short spatial scales, and the excess variance in heterozygosity between individuals reflects significant variation in inbreeding. Additionally, we detect an excess of individuals with around half the average heterozygosity, indicating either selfing or matings between close relatives. We use two types of simulation to ask whether variation in heterozygosity is consistent with fine-scale spatial population structure. First, by simulating offspring using parents drawn from a range of spatial scales, we show that the known pollen dispersal kernel explains g_2. Second, we simulate a 1000-generation pedigree using the known dispersal and spatial distribution and find that the resulting g_2 is consistent with that observed from the field data. In contrast, a simulated population with uniform density underestimates g_2, indicating that heterogeneous density promotes identity disequilibrium. Our study shows that heterogeneous density and leptokurtic dispersal can together explain the distribution of heterozygosity.

Keywords

heterozygosity, identity disequilibrium, population structure, isolation by distance

Introduction

For most organisms, gene dispersal and therefore relatedness are spatially structured, such that individuals closer in space are more likely to mate, and be more closely related, than individuals further apart [1], [2]. Such spatial population structure causes decreasing genetic similarity with geographic distance (isolation-by-distance [3]); this reduces the mean heterozygosity of the whole population relative to a well-mixed population. Despite the ubiquity of these patterns in nature, the role of demography and gene dispersal in determining the spatial pattern of genetic variation has not been thoroughly explored. Commonly used spatial models typically assume discrete demes and/or a uniform population density. However, natural populations are typically patchy, with heterogeneity in both the distribution

and density of individuals. Patchy and heterogeneous spatial distributions within natural populations should result in spatial variation in inbreeding and, consequently, excess variance in heterozygosity. Despite this prediction, the effect of spatial heterogeneity on heterozygosity has rarely been examined in the population structure literature. Moreover, it is the interplay of heterogeneous density and dispersal that likely shapes the spatial structuring of genetic relatedness between individuals. This highlights the importance of understanding the factors (e.g., life history, demography, population structure) that contribute to shaping the full distribution of heterozygosity and relatedness in a spatially structured population.

Understanding the drivers of variation in inbreeding within populations is fundamental, given its importance to genetic diversity and to fitness. Quantifying variation in inbreeding and combining this with measures of fitness (or fitness proxies) makes it possible, in principle, to estimate inbreeding depression either through pedigrees [4], [5] or heterozygosity-fitness correlations (HFCs). For HFCs, inbreeding depression is estimated by comparing proxy measures of fitness against heterozygosity, with the expectation that offspring from related individuals will have lower heterozygosity. Variance in inbreeding is therefore essential for HFCs to be detected [6]. In addition, variance in inbreeding is interesting *per se* because it depends on both demographic history (e.g., [7]) and mating system (selfing, partial selfing or outcrossing) [8]. Outcrossing species, with generally low levels of inbreeding, provide an opportunity to examine factors other than mating system variation that may affect inbreeding variation, and thus, variance in heterozygosity.

If there is variation in inbreeding between individuals, heterozygosity at different loci will be correlated. The covariance between loci in heterozygous state is termed identity disequilibrium (ID), by analogy with linkage disequilibrium, which is the covariance in allelic state between loci. ID can be calculated across individuals and divided by the square of the mean heterozygosity to calculate the population statistic g_2, which is a measure of variance in identity by descent [6] amongst individuals. For an outcrossing organism with fine-scale population structure, spatial patterns of density and mating could have strong effects on the degree of mating with related individuals, and thus affect identity disequilibrium and g_2. Furthermore, as sessile organisms, mating and offspring dispersal in plants are mediated by external vectors (pollinators and seed dispersal mechanisms) [9]. Consequently, the shape of the distribution of dispersal of both pollen and seed will also have an impact on g_2. Additionally, as partial selfing will produce identity disequilibria across loci for selfed individuals, g_2 can be used to estimate the selfing rate of a population, with this estimator being robust to null alleles and biparental inbreeding [10], [11]. If the sources of variation in inbreeding are better understood, we may be able to combine g_2 with other statistics of population structure to improve inferences about demographic history [12], [13]. .

For over a decade, we have sampled a population of the self-incompatible plant *Antirrhinum majus*, the long-term aim being to build a pedigree that will allow us to estimate fitness and dispersal directly. Through that project, we have collected an exceptionally large sample of individuals with SNP genotypes that are spatially mapped. This dataset enables a powerful test of whether the observed density and dispersal in this population can account for both the decay of pairwise relatedness with distance, and for the distribution of heterozygosity across individuals. Here, we first verify that there is excess variance in heterozygosity, which reflects an underlying variance in inbreeding. Second, to understand the role of spatial patterns of dispersal in generating variance in heterozygosity, we compare the empirical distribution of heterozygosity with that of offspring from simulated matings where parents were drawn from different dispersal scales. Third, we ask whether heterogeneous population

density promotes variation in inbreeding, by comparing simulated pedigrees conditioned on uniform density versus on the observed locations of plants. Taken together, addressing these questions provides insight into the underlying drivers of the distribution of heterozygosity and relatedness, and provides novel ways to study the effects of mating patterns and demography in nature.

Methods

Study system

Antirrhinum majus is a self-incompatible, hermaphroditic, short-lived perennial herb native to the Iberian Peninsula. It has a seed bank with most individuals' parents recorded 3-4 years before they are sampled (D. Field, unpublished data). It grows in a variety of microhabitats with relatively bare soil or frequent disturbance, including rail embankments, rocky cliffs, and regularly mowed roadsides. Our study includes two "subspecies" that differ only in flower color: *A. majus pseudomajus* has magenta flowers and occurs in northern Spain and south-western France, including the Pyrenees. *A. majus striatum* has yellow flowers and a smaller range, encircled by *A. m. pseudomajus*. The subspecies are parapatric; narrow clines with intermediate color hybrids form wherever they meet, and there is no evidence for post-zygotic reproductive barriers [14]. We focus on such a hybrid zone in the Vall de Ribès, Spain [15], where we have collected demographic data annually since 2009. Across nearly all of the genome, there is little divergence within our study area between plants with different flower color, except for limited regions associated with floral pigmentation, which show steep clines [16]. Thus, the study area can be considered as a single population for studying neutral genetic variation.

Figure 1: Distribution of *A. majus* individuals (shown as white circles) in Vall de Ribès, Spain from the years 2009 to 2019.

Field sampling

Genetic samples were obtained annually from 2009-2019 from every accessible flowering individual in ~5 km stretches of two parallel roads that cross the Vall de Ribès, dubbed the "lower road" (GIV-4016; ~1150 m elevation) and "upper road" (N-260; ~1350 m) (Fig. 1). We also sampled along small side roads, railroad embankments, rivers, and hiking trails. The plants grow preferentially along exposed areas such as roads, therefore, density was very low away from these disturbed areas between the main sampling sites of the lower and upper

roads. In some years, we were limited to genotyping only in the core area, ~1 km along each road. The total genotyped sample summed over the eleven years is 22,353 plants, ranging from ~750 plants in the smallest year (2018), to ~5500 plants in the largest year (2014). Eighteen percent of individuals were sampled in more than one year. Sampling was conducted during peak flowering (early June to late-July). Each year there were fewer than 100 visible but inaccessible plants; consequently, we estimate that we found the majority of individuals in the sampled area.

For each plant, we collected leaf samples for genotyping, and recorded spatial locations with GeoXT handheld GPS units (Trimble, Sunnydale, CA, USA). These devices are accurate to within 3.7 m, determined by the mean distance between samples comparing samples that had been inadvertently recorded twice in the field (individuals with similar geographic location and near-identical genotypes, allowing for SNP errors). Leaf samples were refrigerated upon return to the field station, dried in silica gel and stored for several weeks.

SNP panel

Previously, a panel of 248 SNPs spread throughout the genome was designed for the focal population (see methods in [17]). We follow these methods but include an additional five years of data (2015-2019) and use a subset of 91 non-clinal SNPs; the mean sample size per SNP was 21,212, or ~95% of the total. (see Supplemental Material 1.1 (SM1.1) for SNP filtering methods).

Identity by descent vs identity in state

Throughout this paper, it will be important to distinguish between identity by descent (IBD) and identity in state. We denote the probability that two genes are identical by descent by *F*; this is defined relative to an ancestral reference population, and can in principle be calculated from the pedigree that descends from that population, independent of the actual allelic state. What we observe are biallelic SNP genotypes; the two homologous genes in a diploid individual will be identical in state if the genes are identical by descent, or if the ancestral genes carried the same allele. Thus, probabilities of identity by descent (*F*) can be estimated from observed identities in state. We denote the heterozygosity at locus i in a particular individual by h_i, with h_i=0 if the genes are identical in state, and h_i=1 otherwise. The mean heterozygosity of an individual is the average of h_i over n loci, denoted multilocus heterozygosity $H = \frac{1}{n}\sum_{i=1}^{n} h_i$.

Isolation by distance

The panel of 91 SNPs was used to calculate F_{ST} and isolation by distance, both of which relate to the mean heterozygosity. We imputed the ~5% missing genotypes for each SNP by randomly assigning genotypes according to the population-wide allele frequencies at each marker. F_{ST} is defined as the average identity by descent among individuals within a subpopulation, F_S, relative to the total population, F_T: $F_{ST} = \frac{F_S - F_T}{1 - F_T}$ [18]. These identities are estimated from SNP genotypes since we do not have the full pedigree. Two genes will have a different allelic state only if they are not identical by descent, and if they derive from different alleles in the ancestral population. Given overall ancestral allele frequencies p+q=1, the expected heterozygosity ($\overline{H}$) of offspring from parents whose genes have a probability of identity by descent *F* is $\overline{H} = (1 - F)\overline{2pq}$, where $\overline{2pq}$ is an average over loci. Thus, there is a

direct relation between F_{ST} and the mean heterozygosity: $F_{ST} = \frac{\overline{H_T} - \overline{H_S}}{\overline{H_T}}$. We use this relation to compute F_{ST} for this dataset [19]. (Note that here, H is the probability of non-identity in state, which depends on the SNP genotype. The subscripts S and T refer to the specified quantity within subpopulation and total population, respectively). Since we have a single continuous population, a subpopulation is defined as the set of pairs of individuals within a geographic separation of 20m and total population denotes all distinct pairs of individuals in the population. Note that 20m is an arbitrary choice of distance class used to define F_{ST}.

Isolation-by-distance – the decay of genetic similarity with geographic distance – can be observed by measuring pairwise relatedness between individuals. If individuals are separated by a distance r, then pairwise relatedness can be calculated as an extension of F_{ST} (which we refer to as pairwise F_{ij}, denoting the relatedness between individuals i and j) by setting F_S to be the probability of identity by descent and, correspondingly, $\overline{H_S}$ to be the probability of non-identity in state between genes which are at a distance r apart, thereby extending the idea of F_S from subpopulation to a set of pairs of individuals separated by any geographic distance class. $\overline{H_S}$ is calculated by finding the average pairwise heterozygosity between every pair of individuals which are within some interval {r, r+δr} of distance apart. This formulation is used to estimate F_{ij} between every pair of individuals relative to the total population, as a function of their geographic separation. Pairs of individuals are binned into distance classes of 20m each (i.e individuals within 20m, 21-40m, and so on) and the average pairwise F_{ij} and the distance corresponding to each bin is calculated. This was done for every year from 2009 to 2019, and the average calculated.

Variation in inbreeding

We calculated multilocus heterozygosity for each individual pooling across all years, denoted here by H, defined as the fraction of heterozygous loci in an individual. In this system "generations" cannot be clearly defined because of seed dormancy and perenniality. However, pooling data across years only reduced H by 0.08%.

We observed an excess of individuals with around half the mean heterozygosity (see Results). To check whether the pattern was consistent with rare selfing, we compared the likelihood of a single Gaussian to a mixture of two Gaussian distributions, one with the observed mean and variance and the other with half its mean and variance.

The variance in individual heterozygosity consists of two components. The first is due to the variance in whether an individual locus is heterozygous, and decreases in proportion to the number of SNP, n: it equals $(1-F)^2\overline{2pq(1-2pq)}/n$. The second is due to covariance in heterozygosity between loci, which is termed the identity disequilibrium (ID). For a given pedigree, unlinked genes flow independently. Thus, heterozygosity is independent across unlinked loci, and so this second component is proportional to the variance in inbreeding across individuals, var(F). The first component can be estimated from the allele frequencies, or simply by shuffling the data across individuals within loci, to eliminate ID. The excess variance is then proportional to the variance in F across individuals, and is measured by the statistic g_2:

$$g_2 = \frac{\Sigma_{i \neq j}\, cov[h_i, h_j]}{E[h]^2} = \frac{var(F)}{(1-E[F])^2}$$

(from Eq. 1 in [6]). Here, cov[h_i, h_j] is the ID between loci i and j, and the sum over all distinct i,j is the excess variance in H due to ID. Dividing by the square of the mean heterozygosity E[h]^2 eliminates dependence on allele frequency, such that g_2 estimates the variance in F across individuals.

To describe the variance of inbreeding across individuals, we first check if the variance in the distribution of individual heterozygosity is significantly greater than the average variance obtained from 100 replicates. This was done by shuffling heterozygous status randomly across individuals within loci, which would eliminate correlations between loci generated by ID. We then computed g_2 using the g2_snps function from the R package InbreedR (in R version 3.6.1 [20]), which implements a modified formula for large data sets to estimate g_2, and provides confidence intervals via bootstrapping to account for the finite number of individuals sampled [21]. We decomposed ID into components due to linked and unlinked SNPs by comparing correlations of H for all individuals to those with low H, at several scales: across all pairs, within linkage groups, and between adjacent SNPs (SM1.2: Table S1).

Additionally, g_2 can be used to estimate selfing rate within a population [10]. Using the software SPAGeDi [22], which implements the g_2-based selfing rate calculation described in [10], the selfing rate was estimated for the full population using the 91 SNP data.

Effects of pollen dispersal on heterozygosity

With isolation by distance, the distribution of heterozygosity is expected to depend on the distance between parents: heterozygosity of offspring from nearby parents will have a lower mean and higher variance compared to offspring from distant parents. To test this prediction, we simulated offspring using all field individuals as mothers and choosing fathers from a given distance away (detail in SM1.3). Then we compared the distribution of H between the field data and offspring simulated from matings with three models of pollen dispersal: the nearest neighbor to the mother, a Gaussian distribution (σ = 300 m), and a leptokurtic dispersal kernel sampled from 1463 empirical measurements of pollen dispersal, estimated as the distances between assigned parents (electronic supplementary material; D. Field, unpublished data). A CDF of the latter distribution (SM1.3: Fig. S3) shows that 75% of the matings occur within 60m and has a kurtosis of 16.5 showing that the distribution is indeed leptokurtic. The genotype of the offspring was assigned using Mendelian inheritance, either without linkage between markers, or using the known linkage map (electronic supplementary material; courtesy of Yongbiao Xue, Beijing Institute of Genomics). Including linkage did not substantially change results, so we mainly show results for simulations without linkage. We compared distributions, means, and variance of H using Kolmogorov-Smirnov tests, t-tests, and F-tests, respectively. For the leptokurtic pollen dispersal simulation, we checked for an excess of low-heterozygosity individuals generated by mating between close relatives by asking whether a mixture of two Gaussian distributions is more likely than a single Gaussian distribution.

Heterozygosity in a simulated spatial pedigree

In order to compare the actual distribution of heterozygosity with that expected for a spatially structured population, we simulated a continuous two-dimensional population, conditioned on the known locations of the individuals and the empirically measured seed and pollen dispersal distances (electronic supplementary material; D. Field, unpublished data), using

Mathematica 12.0 [23]. Our simulation differs from commonly used models (e.g., island [18], stepping stone [24] and continuous Wright-Malécot model [3], [25]) in that we include heterogeneity in density by specifying actual locations to determine relationships in the pedigree. Thus, our simulation parameters should be seen as "effective" values, analogous to the traditional N_e. Additionally, we also validated our simulation by comparing pairwise relatedness directly from the simulated pedigree and from replicate genotypes, and compared the realized and proposed dispersal kernels (SM1.4: Fig. S6, S7).

First, we simulated a population with uniform density (the continuous Wright-Malécot model) as a null model, to compare expected heterozygosity with and without heterogeneous spatial structure. We simulate a region of ~1.1 x 1.8 km that was sampled consistently in the *A. majus* focal population (SM1.4: Fig. S4). Locations were assigned by randomly sampling N points from a uniform distribution each generation, for 1000 generations. Genetic diversity is shaped over the coalescent timescale ($2N_e$, ~170,000 generations in *A. majus* [16]), which is far longer than the 1000 generations that we simulate. However, we are concerned here with the *local* population structure that determines the variation in inbreeding amongst individuals within an area of a few km^2, which will equilibrate rapidly [25]. The spatial pedigree was generated by choosing parents for each individual according to a backwards dispersal distribution measured empirically. The seed and pollen dispersal distances are estimated respectively as the distance between offspring and nearest parent (assumed to be the mother) and between parents (electronic supplementary material; D. Field, unpublished data). For every offspring, the mother and father are chosen from randomly drawn distances from the seed and pollen dispersal distributions. To choose a parent from a distance r, 6 points are assigned randomly on a circle of radius r centred at the focal individual and the nearest individual to each of them are found. The closest individual to any of these points is then chosen as the parent. The accuracy of our algorithm is verified by comparing the specified and realised seed and pollen dispersal distributions for the simulated pedigrees (SM1.4: Fig. S7, Table S4). The same procedure is repeated for the father, taking the mother as the starting point. Since *A. majus* is self-incompatible, the mother and father are not allowed to be the same individual.

Once the spatial pedigree is generated, 10 replicate sets of genotypes are assigned by dropping genes down the pedigree, starting with equal expected frequencies of both alleles at each of 91 loci. In fact, one could start with any initial frequencies, since F_{ST}-like measures are independent of them. Population size was adjusted so that F_{ST} matched the empirical data for the simulated sampling area. This was done by first simulating the population with an initial population size (N) and then repeating the process with higher or lower N until the desired F_{ST} is attained.

Next, we simulated a population with realistic heterogeneous spatial structure by using the individual locations available for the years 2009 to 2019 in the *A. majus* focal population (SM1.4: Fig. S5). There were fewer individuals from 2017-2018, so these were merged, giving distribution data for 10 time points. We randomly sample from the ten consecutive time points, and repeat for 100 cycles, thus iterating for 1000 generations. We sub-sample from these locations to maintain a constant population size (N). If N is greater than the number of plants available in a given time point, say k, all k plants are first included and the remaining N-k locations were re-sampled from the same time point, displaced at a random angle on a circle of radius 3m to avoid having plants in the same location. This naïve approach allows us to simulate a spatial structure that is realistic over at least small scales. We then generated a pedigree following the procedure used for the uniform population, again

adjusting population size to match the empirically observed F_{ST}. Ten replicate sets of genotypes were run for each of five replicate pedigrees.

Patterns of isolation by distance, heterozygote deficit (F_{IS}) and identity disequilibrium were compared between the two simulation types and the field data (calculated from the simulated sub-area of the field site). As the fitted population sizes were large (see Results), obtaining direct estimates of identity by descent and thus F_{ST} from the pedigrees was not feasible. Instead, F_{ST} was obtained for a pedigree as the average of replicate genotype sets generated from that pedigree. F_{IS} was calculated from the observed and expected heterozygosity. Values of g_2 were calculated for each replicate from each pedigree using InbreedR (in R version 3.6.1 [20]).

Results

Isolation by distance

If we consider pairs of individuals within 20m of each other, the average F_{ST} over the eleven years is 0.0244; however, this is an average over a quantity that depends strongly on distance. The average pairwise F_{ij} was calculated each year for individuals separated by different distance classes and then averaged across years. Pairwise relatedness (pairwise F_{ij}) between individuals decreased rapidly with geographic distance, showing isolation by distance (Fig. 2A). The sharp decline in pairwise identity over short spatial scales corresponds precisely to a rapid increase in H with distance between parents (SM1.3: Fig. S1), since heterozygosity is determined by the probability of identity by descent between the genes from each parent. Note that over large separations (>1Km), pairwise F_{ij} values are necessarily negative, because distant individuals are less closely related than the average for the whole population.

Variation in inbreeding

Excess variance in the distribution of individual heterozygosity (H) in the field data shows that there is variance in inbreeding in the population (Fig. 2B). Furthermore, there is an excess of individuals with around half the mean heterozygosity (i.e., with H~0.22, rather than 0.44; Fig. 2B, blue, lower left). These might be due to a low rate of selfing, and using the g_2 estimator calculated with SPAGeDi, the selfing rate for the population is estimated to be 1.2%. Indeed, a mixture between two Gaussian with means ~ 0.22 and 0.44, and variances in the same ratio, fits significantly better than a single Gaussian (Fig. 2B, compare red and black to blue) with an increased likelihood of 11.3. However, we shall see in the next section that this excess is also consistent with matings between close relatives, without the need to invoke a breakdown in self-incompatibility.

To examine whether the observed distribution of heterozygosity is significantly different to a distribution taken from a population with zero identity disequilibrium (ID), we compared the field data with heterozygous values shuffled across individuals, which eliminates ID by removing correlations between loci. We found greater variance in heterozygosity in the observed compared to the randomly shuffled field data (Fig. 2B, gray). For both data sets, the mean heterozygosity (0.44602) necessarily remains the same, but the observed variance in the field data (var(H) = 0.00336) was significantly higher than the average variance in 100 shuffled replicates (mean var(H) = 0.00282, s.d. 0.000029). This excess variance between the observed and shuffled data implies that the mean standardized ID is g_2 = 0.0029 (95% CI: 0.0026-0.0033), representing a significant variance in inbreeding between individuals.

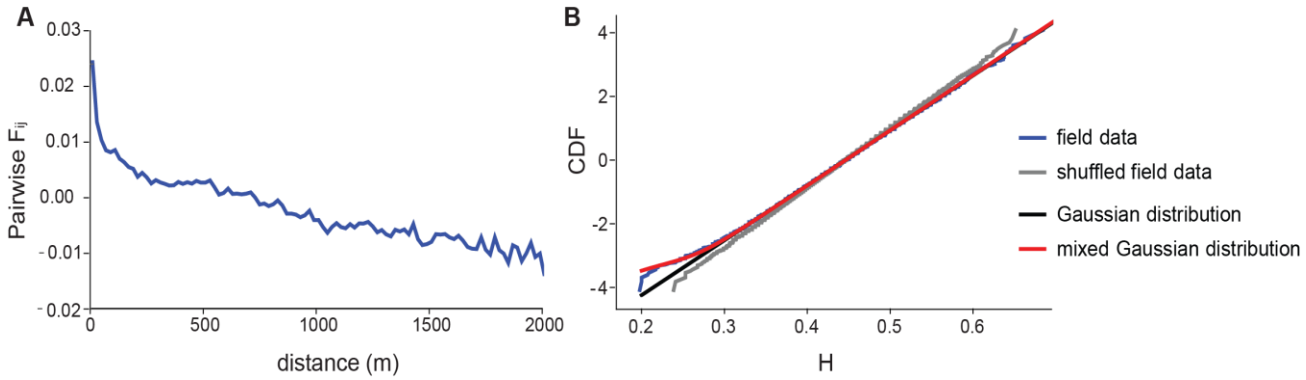

Figure 2. A: Pairwise relatedness (pairwise F_{ij}) between individuals decreases rapidly with geographic distance showing isolation- by- distance in the field data. **B:** Probit transform of the cumulative distribution function (CDF) of the distribution of individual heterozygosity (H). A Gaussian appears as a straight line on a probit scale, and the y-axis is the number of standard deviations of the standard normal distribution.

The overall ID, as measured by g_2, is due to correlations in heterozygosity between all pairs of loci, most of which are unlinked. We expect stronger correlations between linked loci, because relatives will share blocks of genome. We found that the mean covariance in heterozygosity between SNP on the same linkage group is substantially stronger than the overall mean (0.00265 vs. 0.00056). If we restrict attention to those individuals with H<0.3, we find that the covariance in heterozygosity between SNP on the same linkage group is still higher (0.00649), as expected if close relatives share long blocks of genome IBD. This higher covariance in heterozygosity translates to higher mean g_2, which is seen within linkage groups compared to the overall value (SM1.2: Table S1).

Effects of pollen dispersal on heterozygosity

The heterozygosity of simulated offspring depends on distance between their parents, with a rapid increase in mean H with distance (SM1.3: Fig. S1). We compared the observed distribution of heterozygosity with three alternative scenarios for pollen dispersal. There was no significant difference between the mean and variance of heterozygosity between the field data and offspring simulated from the observed leptokurtic dispersal . However, the mean and variance of heterozygosity differed between the field data and simulated matings with either nearest neighbors, or with Gaussian dispersal (Fig. 3A, SM1.3: Tables S2-S3). While all three dispersal schemes differed in the distribution tail as assessed by Kolmogorov-Smirnov tests, Gaussian and nearest neighbour matings are very different from the field data compared to the leptokurtic distribution (SM1.3: Table S3). These comparisons were made for a single replicate, but because each involves 22,353 individuals, there was little variation in the mean and variance between replicates.

We next examined deviations in the left tail of the distribution, where an excess of low heterozygosity individuals might arise from selfing or from matings between close relatives. We focused on the leptokurtic dispersal curve, which was the distribution closest to the field data. We estimated the increase in likelihood between fitting a single versus mixed Gaussian distribution (see “Variation in inbreeding”) for 100 replicate simulations. We found that the mixed Gaussian was a better fit than a single Gaussian, with an increase in log likelihood

greater than 2 for 69 of 100 replicates. The estimated fraction of putatively "selfed" individuals was 0.00043, averaged over replicates, which is about half the estimate from the actual data, 0.00086. In comparison, only 4/100 replicates gave higher estimates than that observed (SM1.3: Fig. S2). This suggests that the excess of individuals with low heterozygosity can to a large extent be explained by matings between relatives under leptokurtic pollen dispersal. Nevertheless, there is a marginally significant excess of such individuals, with twice as many being seen as expected from our simulations. There is considerable variation in fit between replicates, simply because deviations in the tail involve few individuals.

The coefficient g_2 reflects excess variation due to identity disequilibrium, and showed similar patterns as the variance in H. Here, we found no significant difference between g_2 from field data and offspring from simulated matings with leptokurtic pollen dispersal. However, g_2 from Gaussian and neighbor matings were 80% higher than g_2 from field data and leptokurtic matings. This nominally represents a significant difference given that the 95% confidence intervals between these groups do not overlap (Fig. 3B). However, as we discuss below, these confidence intervals only include sampling error, and not the additional variance due to random evolutionary realizations.

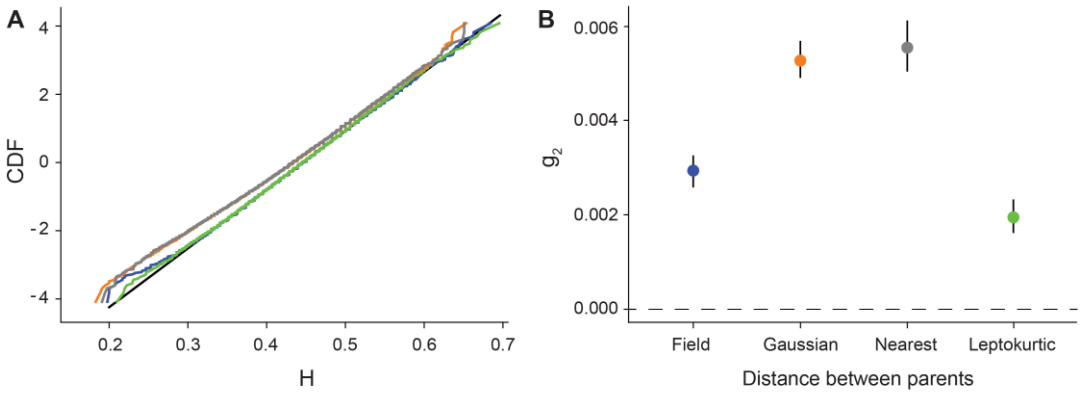

Figure 3. A: Probit transform of the CDF of multilocus heterozygosity, H, for the field data (blue) versus a single replicate of offspring simulated from Gaussian pollen dispersal (orange), nearest neighbor matings (gray), and leptokurtic pollen dispersal (green). A normal distribution (black) with the same mean and standard deviation as the field data is included for comparison **B:** Identity disequilibrium (g_2) for the same data as above indicating mean and 95% CI.

Heterozygosity in a simulated spatial pedigree

In the previous section, we simulated offspring across one generation. To examine whether the observed heterozygosity is consistent with a spatially structured model, we simulated pedigrees over 1000 generations with uniform and heterogeneous density, conditioned on the locations of individuals observed over ten years, repeated over 100 cycles for the latter case. The realized seed and pollen dispersal matched the empirical seed and pollen dispersal distribution for both density types (SM1.4: Fig. S7, Table S4). We required N = 15500 individuals for the heterogeneous density model and 40000 individuals for the uniform density, in order to match the observed $F_{ST} \sim 0.022$ calculated over a 20m scale from the simulated sub-area of the field site (SM1.4: Table S5). Up to distances of 1km, the decline in pairwise identity with distance matched between the field data and the five replicate pedigrees simulated with heterogeneous density (Fig. 4A, SM1.4: Fig. S8A). High variation

among replicates suggests that many more SNPs would be needed to match the pattern from the pedigree (SM1.4: Fig. S8B); moreover, linkage would increase this variance to some extent. We also compared the pattern of isolation by distance from the field data to that from the pedigrees generated for both the heterogeneous and uniform density scenarios (Fig 4B, SM1.4: Fig. S9); the heterogeneous density is a much better fit than the uniform density (SM1.4: Table S5).

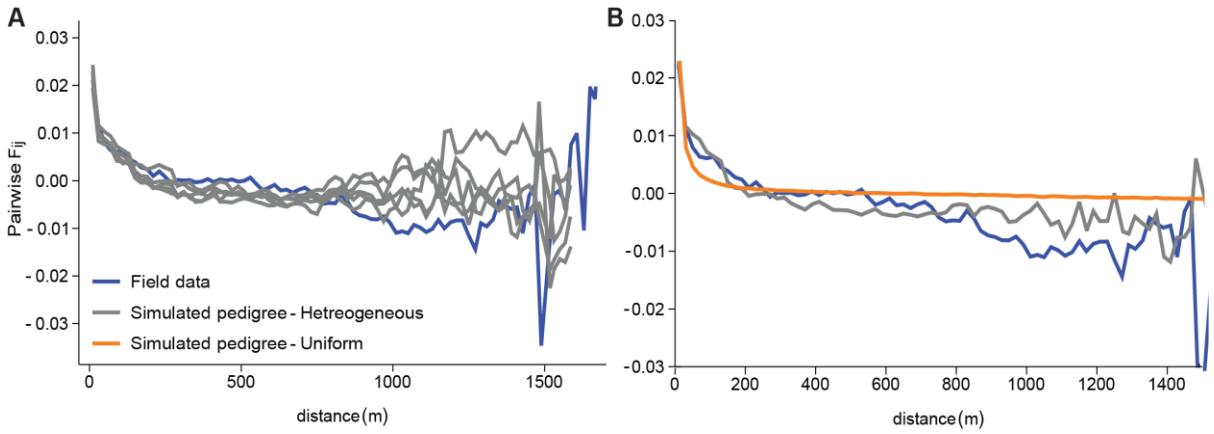

Figure 4. A:. Isolation by distance compared between the field data (blue) and five replicate simulated pedigrees (gray) based on a heterogeneous population density. **B:** Isolation by distance from the field data (blue) compared between the simulated pedigree with a heterogeneous (gray) and uniform (orange) population density.

Identity disequilibrium (g_2) estimates from the genotypes from pedigrees simulated with heterogeneous density showed substantial variation between the five simulated pedigrees, and between the ten draws of 91 SNPs from each pedigree (Fig. 5). The average g_2 estimated from the five pedigrees (each with 10 replicates) is 0.00264, which is consistent with the observed mean annual g_2 from the field of 0.00262. On the other hand, when assuming a uniform density, the average g_2 of 0.00171 is significantly lower than the field data. Note that the confidence limits for the field data, generated by InbreedR, only include error due to sampling a limited number of individuals. These errors do not account for sampling a limited number of SNPs, or the random variation between evolutionary realizations (see Discussion).

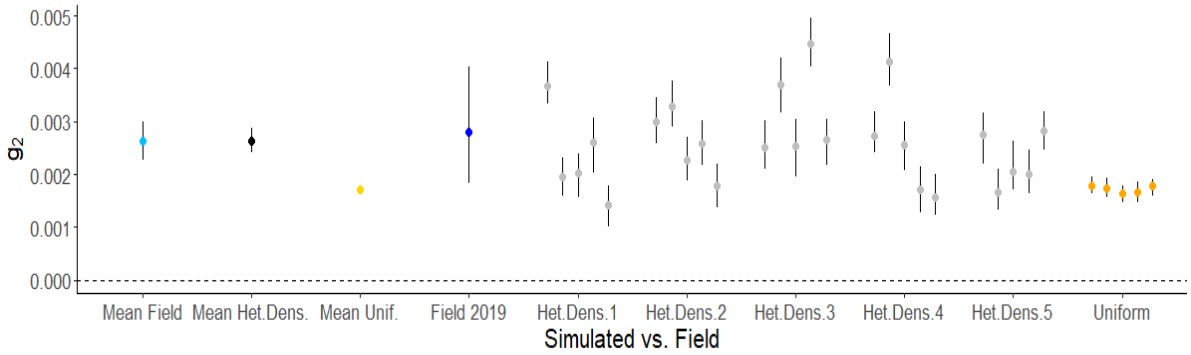

Figure 5. Identity disequilibrium (g_2) calculated from field data versus simulated pedigrees. Five of ten replicates per pedigree are shown (gray: heterogeneous density, with five simulated pedigrees; orange: uniform density, with one simulated pedigree). Mean from the field (light blue) is across 2009-2019, while mean from the heterogeneous (black) and uniform (yellow) simulations is across all replicates. The final year of field data (dark blue) is comparable to g_2 calculated from the final year of pedigree replicates (gray and orange).

Discussion

An enduring problem in evolutionary biology is understanding how demographic processes, such as heterogeneous density and dispersal, interact with spatial structure to determine the distribution of heterozygosity within populations. In this study of a long-term dataset, including more than 20,000 plants sampled over 11 years, we combine field data and simulations to address questions central to understanding how demography can influence patterns of heterozygosity. Namely, can we predict the distribution of heterozygosity for an outcrossing species from key demographic parameters? To address this question, we first confirmed that there was significant correlation in heterozygosity between markers (g_2, a measure of identity disequilibrium), which implies variation in inbreeding. By simulating offspring from matings between geo-referenced, genotyped individuals, we show that the mean heterozygosity increases, and the variance of heterozygosity decreases, with increasing distance between parents; strikingly, these changes occur over very short scales (~10m, SM1.3: Fig. S1). We found that the observed distribution of heterozygosity is consistent with the known leptokurtic distribution of pollen dispersal. We also simulate the population over 1000 generations using the actual seed and pollen dispersal kernels, and the observed heterogeneous density. We found that this model matches the observed identity disequilibrium, whereas a model with uniform density substantially underestimates the observed patterns. Thus, we explain the distribution of heterozygosity (mean, variance and tails) using known features of the population. Moreover, our results also highlight the limitations of making theoretical predictions from simulations that only assume simple demographies. Taken together, our findings highlight the potential for using the observed demography to explain the distribution of genetic diversity, and specifically the variance in inbreeding in spatially continuous populations.

Variation in heterozygosity within populations provides the potential for selection to reduce the frequency of less fit, inbred individuals. The association between inbreeding and fitness is often tested through heterozygosity-fitness correlations (HFC), which quantify inbreeding depression in natural populations by correlating measures of fitness with heterozygosity [6]. Many studies that test for HFCs find that the excess variation in heterozygosity, g_2, which arises from identity disequilibrium, is low and rarely significant [26]. In our study, we estimate a significant g_2 of 0.0029 (95% CI: 0.0026-0.0033). Although low, this estimate is of the same order as most of the g_2 values found across 105 vertebrate populations in a meta-analysis of 50 HFC studies (average of 0.007) [26], and on the same order as ~60% of the local populations surveyed in a long-lived tree [27]. Our estimate of significant variation in heterozygosity provides the opportunity to examine potential drivers of this variance and examine how density, spatial structure and dispersal contribute to a non-uniform distribution of heterozygosity.

In our study, beyond simply estimating identity disequilibrium, we use two types of simulation to explore how demography shapes variation in inbreeding. The first simulation shows how the spatial pattern of pollen dispersal affects the distribution of heterozygosity. Simulated matings with the empirically measured leptokurtic pollen dispersal curve were consistent with the actual g_2, compared to matings with nearest neighbors or a Gaussian pollen dispersal. This result is somewhat surprising because we did not include the complexities of the mating system of *A. majus*. *Antirrhinum majus* has a gametophytic self-incompatibility system (GSI [28]), whereby the pollen detoxifies secretions from the style unless the pollen and style genotypes share alleles at the S-locus [29]. This system not only

prevents selfing, but also reduces mating among relatives (i.e., biparental inbreeding) because related plants are more likely to share S-alleles [4], [30]. Thus, we might expect that our simulated matings would have lower mean heterozygosity than the empirical measurement; yet we found no evidence for such an effect. Indeed, we found that the excess of individuals with low heterozygosity, around half the mean, can be explained largely by a small amount of bi-parental inbreeding with leptokurtic pollen dispersal (Fig. 3, SM1.3: Fig. S2). However, we have little statistical power to distinguish this from rare selfing, which can occur in self-incompatible species. In fact, using the g_2 estimator of selfing rate from eq. 9 in [10], our significant g_2 value would imply a selfing rate of 1.2% for this population. However, as shown by [11], this estimate could be within the bounds of the upward bias of the estimator if strong biparental inbreeding is present, hence, this does not necessarily imply a breakdown of self-incompatibility. We believe that our method, fitting a model of two Gaussians, is a more robust way to estimate selfing than using g_2, since it focuses on the low-H individuals rather than the whole variance. However, it is still challenging to distinguish selfing from close inbreeding.

Our second simulation approach asked whether heterogeneous density promotes variation in inbreeding, given strong fine-scale population structure indicated by a rapid decay in pairwise F_{ij} (over a few metres, Fig. 2A). We only provide a proof-of-principle, by asking whether a plausible model of spatial structure can explain the observed heterozygosity. We do not include all features of the actual population – in particular, we extrapolate by repeatedly sampling ten years of spatial distributions; we ignore linkage; we simplify the self-incompatibility system; and we assume an annual life cycle (no perenniality or seed bank). Indeed, simulated pedigrees with uniformly distributed plants gave less identity disequilibrium than we observed. In contrast, simulated pedigrees conditioned on the actual, heterogeneous density of plants were consistent with identity disequilibrium measured in the field. This indicates that patchiness combined with leptokurtic dispersal shapes the distribution of heterozygosity. Simulations with heterogeneous density also better capture empirical isolation-by-distance patterns than those with a uniform density (Fig 4B, SM1.4: Fig. S9). However, the effective population size of 15,500 individuals in the heterogeneous-density simulations is an order of magnitude larger than the average number of plants observed in a year (~2500). We believe that most plants are sampled each year, so that this discrepancy is more likely to be due to a seed bank, which is expected to substantially increase the effective population size [31]. Nevertheless, despite simplifications such as non-overlapping generations, no seed bank, and a simple SI system, the heterogeneous-density simulation accurately captures patterns of identity disequilibrium and isolation-by-distance.

Our estimation of identity disequilibrium illustrates a general problem with statistical comparisons in evolutionary biology. There are three sources of error in estimating g_2: firstly, error generated from sampling a limited number of individuals, secondly, from sampling a limited number of SNPs, and thirdly from random variation between evolutionary realizations or trajectories. In our study, the first source (a limited number of individuals) is shown by the confidence intervals in Fig. 5, obtained by bootstrapping across individuals [21]. The second source of error (a limited number of SNPs) is shown by the substantial variation in g_2 of the ten replicates of each of five pedigrees. Here, variation is generated by random meiosis amongst unlinked markers on a fixed pedigree. This variation could be reduced by increasing the number of SNPs, but the effective number of segregating sites that can be included in the analysis is fundamentally limited by the length of the genetic map. Finally, there is additional variation between pedigrees, due to the random assignment of parents in the simulations, which generates a random pedigree. The wide variation in estimates of g_2 due to random

meiosis, and to the random generation of the pedigree (Fig. 5) is an important reminder that estimates of parameters are typically limited by the randomness of evolution. The stochasticity of evolution can potentially generate error variance far higher than that due to the limited number of individuals or SNPs sampled.

In addition to analyzing the effect of population structure on the distribution of heterozygosity, our study highlights the potential of utilizing multiple statistics to estimate population structure. We have shown that the variance of heterozygosity due to identity disequilibrium can distinguish alternative dispersal and density distributions, which implies that in combination with pairwise F_{ij} as a function of distance, g_2 can help estimate the demography. Genetic data contain far more information than is described by F_{ST} and g_2; for example, the mean squared disequilibrium can be used to estimate effective population size [32], [33], and this extends naturally to the covariance of pairwise linkage disequilibrium as a function of distance. We could simply use a set of such statistics to inform demographic inference via ABC [34]. However, our preference would be to first develop a theoretical understanding of how realistic demographies influence statistical measures of spatial covariance in allele frequency, identity disequilibria, and linkage disequilibria.

The distribution of heterozygosity has often been measured to estimate inbreeding depression and examine correlation with fitness. Yet, this type of data has rarely been used to investigate population structure *per se* and as a complement to the more widely used pairwise identity, F_{ST}. By bringing together local inbreeding and isolation-by-distance, our study provides a novel assessment of how dispersal and population density can explain both pairwise identity and the distribution of heterozygosity in spatially continuous populations. However, we have only begun to investigate how the distribution of heterozygosity can be shaped by population structure and demographic parameters. Our future work will focus on understanding how other features such as a seed bank influence genetic diversity, with the ultimate goal of deriving information about demographic history from the distribution of heterozygosity in populations that have fewer measured parameters. New models that include these complexities, as well as ecological, mating system and life history factors are required to extend our understanding of the drivers of population structure in natural populations.

Data availability

All data and code used to generate simulated data and carry out analysis is available at: https://doi.org/10.15479/AT:ISTA:11321. Data includes processed field data for 11 years of *Antirrhinum majus* sample collection, including SNP values, GPS locations and trait measurement values for each plant. Also included are dispersal data and a linkage map of 91 SNPs.

Acknowledgments

We thank the many volunteers and friends who have contributed to data collection in the field site over the years, in particular those who have managed field seasons: Barbora Trubenova, Maria Clara Melo, Tom Ellis, Eva Cereghetti, Lenka Matejovicova, Beatriz Pablo Carmona. Frederic Ferrer and Eva Salmerón Mateu have been immensely helpful with logistics at our informal field station, El Serrat de Planoles. We thank Sean Stankowski for technical help in producing figure 1. This research was also supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Funding

Part of this work was funded by Marie Curie COFUND Doctoral Fellowship and Austrian Science Fund FWF (grant P32166).

Conflict of interest

The authors declare that there is no conflict of interest.

References

[1] S. Wright, "Isolation by distance under diverse systems of mating.," *Genetics*, vol. 31, no. 1, pp. 39–59, Jan. 1946, doi: 10.1093/genetics/31.1.39.

[2] X. Vekemans and O. J. Hardy, "New insights from fine-scale spatial genetic structure analyses in plant populations," *Mol. Ecol.*, vol. 13, no. 4, pp. 921–935, Apr. 2004, doi: 10.1046/J.1365-294X.2004.02076.X.

[3] S. Wright, "Isolation by Distance," *Genetics*, vol. 28, no. 2, p. 114, Feb. 1943, doi: 10.1016/B978-0-12-374984-0.00820-2.

[4] D. Charlesworth and B. Charlesworth, "Inbreeding depression and its evolutionary consequences.," *Annu. Rev. Ecol. Syst. Vol. 18*, vol. 18, pp. 237–268, 1987, doi: 10.1146/annurev.es.18.110187.001321.

[5] M. Lynch and B. Walsh, "Inbreeding Depression," in *Genetics and Analysis of Quantitative Traits*, Sunderland, Massachusetts: Sinauer Associates, Inc, 1998, pp. 251–291.

[6] M. Szulkin, N. Bierne, and P. David, "Heterozygosity-fitness correlations: A time for reappraisal," *Evolution (N. Y).*, vol. 64, no. 5, pp. 1202–1217, 2010, doi: 10.1111/j.1558-5646.2010.00966.x.

[7] S. Y. W. Sin, B. A. Hoover, G. A. Nevitt, and S. V. Edwards, "Demographic history, not mating system, explains signatures of inbreeding and inbreeding depression in a large outbred population," *Am. Nat.*, vol. 197, no. 6, pp. 658–676, Jun. 2021, doi: 10.1086/714079.

[8] A. A. Winn *et al.*, "Analysis of inbreeding depression in mixed-mating plants provides evidence for selective interference and stable mixed mating," *Evolution (N. Y).*, vol. 65, no. 12, pp. 3339–3359, Dec. 2011, doi: 10.1111/j.1558-5646.2011.01462.x.

[9] M. D. Loveless and J. L. Hamrick, "Ecological determinants of genetic structure in plant populations.," *Annu. Rev. Ecol. Syst. Vol. 15*, pp. 65–95, Nov. 1984, doi: 10.1146/annurev.es.15.110184.000433.

[10] P. David, B. Pujol, F. Viard, V. Castella, and J. Goudet, "Reliable selfing rate estimates from imperfect population genetic data," *Mol. Ecol.*, vol. 16, no. 12, pp. 2474–2487, Jun. 2007, doi: 10.1111/J.1365-294X.2007.03330.X.

[11] O. J. Hardy, "Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate," *Mol. Ecol. Resour.*, vol. 16, no. 1, pp. 103–117, Jan. 2016, doi: 10.1111/1755-0998.12431.

[12] B. G. Milligan, F. I. Archer, A. L. Ferchaud, B. K. Hand, E. M. Kierepka, and R. S. Waples, "Disentangling genetic structure for genetic monitoring of complex populations," *Evol. Appl.*, vol. 11, no. 7, pp. 1149–1161, Aug. 2018, doi: 10.1111/eva.12622.

[13] G. S. Bradburd and P. L. Ralph, "Spatial Population Genetics: It's About Time," *Annu. Rev. Ecol. Evol. Syst.*, vol. 50, pp. 427–449, Nov. 2019, doi: 10.1146/annurev-ecolsys-

110316-022659.
[14] C. Andalo, M. B. Cruzan, C. Cazettes, B. Pujol, M. Burrus, and C. Thébaud, “Post-pollination barriers do not explain the persistence of two distinct Antirrhinum subspecies with parapatric distribution,” *Plant Syst. Evol.*, vol. 286, no. 3, pp. 223–234, May 2010, doi: 10.1007/s00606-010-0303-4.
[15] A. C. Whibley *et al.*, “Evolutionary paths underlying flower color variation in Antirrhinum,” *Science (80-.).*, vol. 313, no. 5789, pp. 963–966, Aug. 2006, doi: 10.1126/science.1129161.
[16] H. Tavares *et al.*, “Selection and gene flow shape genomic islands that control floral guides,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 43, pp. 11006–11011, Oct. 2018, doi: 10.1073/pnas.1801832115.
[17] H. Ringbauer, A. Kolesnikov, D. L. Field, and N. H. Barton, “Estimating barriers to gene flow from distorted isolation-by-distance patterns,” *Genetics*, vol. 208, no. 3, pp. 1231–1245, Mar. 2018, doi: 10.1534/genetics.117.300638.
[18] S. Wright, “Evolution in Mendelian Populations,” *Genetics*, vol. 16, no. 2, p. 97, Mar. 1931, doi: 10.1093/genetics/16.2.97.
[19] M. Jakobsson, M. D. Edge, and N. A. Rosenberg, “The relationship between FST and the frequency of the most frequent allele,” *Genetics*, vol. 193, no. 2, p. 515, 2013, doi: 10.1534/genetics.112.144758.
[20] R. C. Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria, 2014.
[21] M. A. Stoffel *et al.*, “inbreedR: an R package for the analysis of inbreeding based on genetic markers,” *Methods Ecol. Evol.*, vol. 7, no. 11, pp. 1331–1339, Nov. 2016, doi: 10.1111/2041-210X.12588.
[22] O. J. Hardy and X. Vekemans, “spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels,” *Mol. Ecol. Notes*, vol. 2, no. 4, pp. 618–620, Dec. 2002, doi: 10.1046/J.1471-8286.2002.00305.X.
[23] W. R. Inc., “Mathematica.” Wolfram Research, Inc., Champaign, Illinois, 2019.
[24] M. Kimura and G. H. Weiss, “The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance,” *Genetics*, vol. 49, no. 4, p. 561, Apr. 1964, doi: 10.1093/genetics/49.4.561.
[25] G. Malécot, *The Mathematics of Heredity (English translation, 1969)*. San Francisco: WF Freeman, 1948.
[26] J. M. Miller and D. W. Coltman, “Assessment of identity disequilibrium and its relation to empirical heterozygosity fitness correlations: a meta-analysis,” *Mol. Ecol.*, vol. 23, no. 8, pp. 1899–1909, Apr. 2014, doi: 10.1111/MEC.12707.
[27] I. Rodríguez-Quilón *et al.*, “Local effects drive heterozygosity-fitness correlations in an outcrossing long-lived tree.,” *Proceedings. Biol. Sci.*, vol. 282, no. 1820, p. 20152230, Dec. 2015, doi: 10.1098/rspb.2015.2230.
[28] A. McCubbin, R. Carpenter, E. Coen, and H. Dickinson, “Self-incompatibility in Antirrhinum,” in *Angiosperm Pollen and Ovules*, New York, NY: Springer, New York, NY, 1992, pp. 104–109.
[29] S. Fujii, K. I. Kubo, and S. Takayama, “Non-self- and self-recognition models in plant self-incompatibility,” *Nat. Plants*, vol. 2, no. 9, 2016, doi: 10.1038/nplants.2016.130.
[30] R. A. Cartwright, “Antagonism between local dispersal and self-incompatibility systems in a continuous plant population,” *Mol. Ecol.*, vol. 18, no. 11, pp. 2327–2336, Jun. 2009, doi: 10.1111/J.1365-294X.2009.04180.X.
[31] L. Heinrich, J. Müller, A. Tellier, and D. Živković, “Effects of population- and seed bank size fluctuations on neutral evolution and efficacy of natural selection,” *Theor. Popul. Biol.*, vol. 123, pp. 45–69, Sep. 2018, doi: 10.1016/J.TPB.2018.05.003.

[32] W. G. Hill, “Estimation of effective population size from data on linkage disequilibrium1,” *Genet. Res. (Camb).*, vol. 38, no. 3, pp. 209–216, 1981, doi: 10.1017/S0016672300020553.
[33] R. Vitalis and D. Couvet, “Estimation of effective population size and migration rate from one- and two-locus identity measures,” *Genetics*, vol. 157, no. 2, pp. 911–925, Feb. 2001, doi: 10.1093/genetics/157.2.911.
[34] M. A. Beaumont, “Approximate Bayesian computation in evolution and ecology,” *Annu. Rev. Ecol. Evol. Syst.*, vol. 41, pp. 379–406, Nov. 2010, doi: 10.1146/annurev-ecolsys-102209-144621.