Does the seed fall far from the tree? - weak fine scale genetic structure in a continuous Scots pine population

Alina K. Niskanen[1,3*], Sonja T. Kujala[2], Katri Kärkkäinen[2], Outi Savolainen[1], Tanja Pyhäjärvi[3]

[1]*Ecology and Genetics Research Unit, University of Oulu, 90014 University of Oulu, Finland*

[2]*Natural Resources Institute Finland (Luke), University of Oulu, 90014 University of Oulu, Finland*

[3]*Department of Forest Sciences, University of Helsinki, 00014 University of Helsinki, Finland*

*Corresponding authors:

Alina Niskanen, alina.niskanen@gmail.com

Tanja Pyhäjärvi, tanja.pyhajarvi@helsinki.fi

Abstract

Knowledge of fine scale genetic structure, *i.e.*, the distribution of genetic diversity at short distances, is important in evolutionary research and practical applications such as conservation and breeding programs. In trees, related individuals often grow close to each other due to limited seed and/or pollen dispersal. The extent of seed dispersal also limits the speed at which a tree species can spread to new areas.

We studied the fine scale genetic structure of Scots pine (*Pinus sylvestris*) in two naturally regenerated sites 20 km from each other located in continuous south-eastern Finnish forest. We genotyped almost 500 adult trees for 150k SNPs using a custom made Affymetrix array. While we detected some pairwise relatedness at short distances, the relatedness decreased with increasing distance, as expected. Despite the clustering of related individuals, the sampling sites were not differentiated (F_{ST} = 0.0005). According to our results, Scots pine has a large neighborhood size (Nb = 1680–3120), but relatively short gene dispersal distance (σ_g = 36.5–71.3 m). Knowledge of Scots pine fine structure can be used to define suitable sampling distances for evolutionary studies and practical applications. Detailed empirical estimates of dispersal are necessary both in studying post-glacial recolonization and predicting the response of forest trees to climate change.

Introduction

Understanding the fine scale genetic structure of species is important in a wide range of fields. It builds up, through gene flow, from an interplay between: i) external factors, such as fragmented habitat and differences in growth conditions, which affect selection, population size and density, and ii) internal factors, such as dispersal ability and mating patterns of the species (Loveless & Hamrick 1984; Vekemans & Hardy 2004). Gene flow takes place between individuals in physical space (Bradburd & Ralph 2019) often leading to isolation–by–distance pattern (Wright 1943; Málecot 1967). Thus, individuals close to each other are commonly expected to be more closely related than a random sample from within the species. In sedentary species, like trees, spatial aggregation of related individuals often results from limited seed and/or pollen dispersal (Hardy & Vekemans 1999). In wind pollinated species, pollen can disperse long distances from hundreds of meters to hundreds of kilometers (e.g., Kremer *et al*. 2012; Desilva & Dodd 2021), whereas seed dispersal has shorter average distances (Kremer *et al*. 2012). Pollen dispersal distances in animal pollinated species are commonly short—from few meters to few kilometers (e.g., Levin & Kerster 1974; Kremer *et al*. 2012; but see Ahmed *et al*. 2009). As a result of the differences in dispersal distances, strong fine–scale structure is common in animal pollinated and rare in wind pollinated tree species (Vekemans & Hardy 2004; Hardy *et al*. 2006; Born *et al*. 2008; Vakkari *et al*. 2020). Thus, the distribution of genotypes at a short distance is usually transient, largely impacted by life–history traits and can be evaluated through estimating relatedness between individuals.

In evolutionary research, spatial genetic information is valuable for inferring the strength of evolutionary forces—genetic drift, selection, and gene flow—that participate in forming the genetic structure (Rousset 2003; Slatkin 1985). The genetic structure of populations needs to be considered in many applications. For example, in association analyses, spurious associations may arise due to correlation of allele and trait frequencies in different

populations but also within a population if the underlying genetic structure is not corrected for (Pritchard & Rosenberg 1999; Persyn *et al*. 2018). Also in practical applications, when individuals are chosen for conservation, population management or breeding programs from natural populations, it is essential to know how genetic diversity and, for example, rare alleles are distributed in space to maintain high genetic diversity, avoid inbreeding and, on the other hand, unnecessary mixing of differentially adapted populations (Desilva & Dodd 2021; Escudero *et al*. 2003; Smith *et al*. 2018). When the span of spatial autocorrelation is known, sampling can be adapted to the needs of each application.

Knowledge of fine scale genetic structure can recursively be used to infer dispersal distances (Málecot 1967; Rousset 1997 & 2003). The ability to disperse becomes increasingly important in the light of climate change when plant species and their locally adapted populations may become maladapted to their current locations (Gougherty *et al*. 2021). Dispersal information can be used, for example, in predicting species' potential for adaptation (Kuparinen *et al*. 2010; Kremer *et al*. 2012; Barton 1979; Slatkin 1973). If the natural dispersal rate is estimated to be too slow to respond to the challenges of climate change, human assisted migration is one possible way to aid adaptation (Marris 2009; Aitken & Whitlock 2013).

Scots pine (*Pinus sylvestris*) is a keystone conifer species in large parts of the forests of Northern Eurasia and, thus, important for the ecosystem function (Pyhäjärvi *et al*. 2020 and references therein). As a major source of timber, paper and pulp, Scots pine holds also high economic value. Due to its continuous and wide distribution, wind-pollination and predominantly outcrossing mating system (Muona & Harju 1989), Scots pine shows very weak population structure at the global scale; largely in the form of subtle isolation–by–distance across its distribution range (Tyrmi *et al.* 2020). Less is known about within-population structure as only small fragmented populations have been studied for this aspect so far (Robledo-Arnuncio & Gil 2005; Sofletea *et al*. 2020). Although artificial regeneration with genetically improved seedlings is becoming more common in forestry,

natural regeneration has been the predominant regeneration method in forests, making the patterns described here common in Fennoscandia. Since Scots pine has such a dominant role in the boreal forest ecosystems, even small changes in its distribution or adaptation ability may have large roll-down effects.

Here we investigate the fine scale spatial genetic structure of Scots pine in two naturally regenerated sites that are located in south-eastern Finland. We use a genome-wide 400k single nucleotide polymorphism (SNP) array (Kastally *et al*. 2022), which allows us to estimate relatedness in a large sample of 469 trees. We estimate the parameters of the isolation-by-distance model and use this information to derive estimates of gene dispersal distance. Then we investigate the spatial spread and sharing of rare alleles that can have a distinct profile of, for example, fitness effects and average allele age compared to more common alleles. The knowledge of fine-scale genetic structure and dispersal distance are useful in practical applications of Scots pine breeding and associating genetic and trait variation in natural populations, but also for modeling adaptation and making predictions on how widely-distributed wind pollinated trees respond to climate change.

Material and Methods

Samples and genotypes

Our study population at Punkaharju intensive study site (ISS, http://www.evoltree.eu/index.php/intensive-study-sites/sites/punkaharju) is a naturally regenerated site in south–eastern Finland and includes two sampling sites, Mäkrä (61°50'N, 29°23'E) and Ranta–Halola (61°39'N, 29°17'E) located 21 km apart (Figure 1). We sampled 469 adult (33–145 years) Scots pines approximately at 20 m distance, with shortest within sampling site distance of 10 and 14 m and longest 464 and 1164 m in Mäkrä and Ranta-Halola, respectively.

Needle samples from the adult trees were genotyped on a custom–made Affymetrix SNP array including 407 540 markers. Development of the SNP array and genotyping of the samples is described in detail in Kastally *et al*. (2022). In short, we used a dataset of 157 325 polymorphic SNPs with the ThermoFisher conversion types Poly High Resolution and No Minor Homozygote as a starting point. We filtered for Mendelian errors and errors likely arising from paralogy common in conifers (Neale *et al*. 2014). Further filtering of the genotype data was done according to the requirements of each analysis as described below.

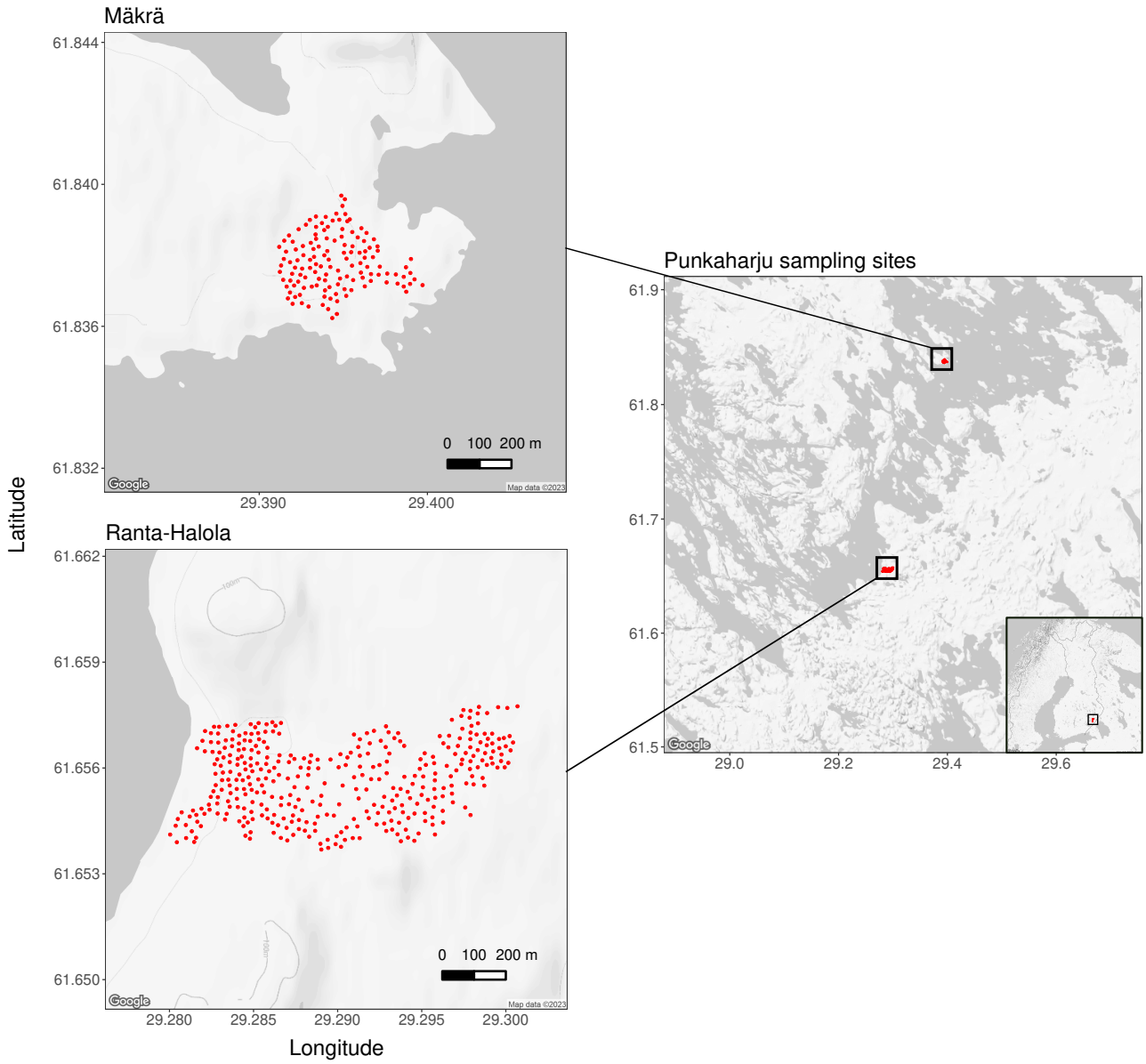

Figure 1. Maps of the sampling sites. Sampled trees are indicated as red dots.

Spatial data and pairwise distances

We recorded the coordinates for each tree using a portable GPS locator in August-October 2020. The initial coordinates in the ETRS-TM35FIN geodetic coordinate system were transformed to the coordinates in the EUREF-FIN-GRS80 geodetic coordinate system using a Finnish map service on the web (https://kartta.paikkatietoikkuna.fi/). One of the study trees had died between the needle sample collection and coordinate recording and was excluded from the spatial analyses. We estimated the pairwise spatial distance matrix

for 468 adult trees with coordinates using the R (v. 3.6.3, ref) package *fields* (Nychka *et al*. 2017) function “rdist.earth”.

Population structure

To get an overall picture of the genetic structure of our study population, we conducted principal component analysis (PCA) using the R package *pcadapt* (Privé *et al*. 2020). We ran PCA for two sets of individuals, first for all 469 individuals and second for 332 individuals excluding: i) individuals used in the SNP discovery (Kastally *et al*. 2022), ii) individuals related to the SNP discovery individuals (pairwise relatedness, GRM ≥ 0.044; Yang *et al*. 2011; see details below), and iii) one individual from each pair with pairwise relatedness ≥ 0.044. We excluded closely related individuals and SNP discovery individuals from the PCA to detect the underlying population structure without the signal of family structure or SNP ascertainment effects. We used a set of 65 498 SNPs with the following characteristics: minor allele frequency (MAF) ≥ 0.05, close to Hardy Weinberg equilibrium (HW; exact test *p* value ≥ 0.001), and relatively low linkage disequilibrium (LD) $r^2 < 0.9$ with other SNPs in 10 kb windows within a contig (using PLINK v. 1.9; Purcell *et al*. 2007), unless stated otherwise. We used four embryo samples and their parents (described in Kastally *et al*. 2022) to estimate Mendelian errors for each locus in PLINK and excluded SNPs with more than one Mendelian error. Since Scots pine has a high proportion of repetitive and paralogous genome sequence (Wegrzyn *et al*. 2014), we excluded SNPs with more than one seemingly heterozygous genotype in haploid megagametophyte samples (described in Kastally *et al*. 2022) to avoid SNPs in potentially paralogous genomic regions.

We estimated pairwise F_{ST} (Weir & Cockerham 1984) between the two study sites using R package *StAMPP* (Pembleton *et al*. 2013) and performed 1000 bootstraps to estimate its 95% confidence interval (CI). We also estimated ϕ_{ST} (Excoffier *et al*. 1992), which measures population differentiation at molecular level, using Nei’s *D* pairwise genetic

distances (Nei 1972) between individuals and 10 000 permutations to gain a *p*-value in *StAMPP.*

Fine-scale genetic structure

We estimated pairwise relatedness between all 469 samples as a genomic relationship matrix (GRM; Yang *et al*. 2011) using PLINK command "–make-grm". GRM was estimated between individuals *j* and *k* over SNP loci from *i* to *N* using the formula

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)},$$

where x_{ij} and x_{ik} are the numbers of reference (major) alleles in an individual *j* and *k*, and p_i is the reference allele frequency. Since inclusion of closely related individuals may inflate the relatedness estimates, we used allele frequencies estimated for 387 unrelated adult individuals (GRM < 0.0625, the mean of relationship class for, e.g., first cousins once removed) as reference allele frequencies in the estimation of GRM. We classified each pair of individuals into family relationship classes (e.g., Manichaikul *et al*. 2010) based on pairwise GRM: second degree between [0.177,0.354), e.g., half-siblings, third degree between [0.088, 0.177), e.g., first cousins, fourth degree between [0.044, 0.088), e.g., first cousins once removed, and unrelated below 0.044. We also estimated genomic inbreeding coefficient (F_{GRM}) for each individual based on genomic relationship matrix in GCTA (Yang *et al*. 2011).

Mantel test (Mantel 1967) is a traditional test of spatial autocorrelation where the relationship of two dissimilarity (i.e., distance) matrices is investigated. In spatial genetics, the null hypothesis of a Mantel test is that genetic distance, or similarity when measured as relatedness or kinship, and spatial distance are not correlated. To study the relationship between relatedness and spatial distance, we estimated their correlation using Spearman's correlation coefficient (ρ) separately for the two sampling sites Mäkrä and Ranta–Halola. We conducted the Mantel test separately for both sites using the R package

ecodist (Goslee & Urban 2007) with 10 000 permutations. The use of the Mantel test in spatial genetics has been criticized because of the lack of homoscedasticity and linear correlation between genetic and spatial distances (Legendre *et al*. 2015). These problems are, however, less severe in Mantel correlogram analysis where samples are divided into pre–defined spatial distance classes and each distance class is compared separately to joint data from other distance classes. We therefore used also Mantel correlograms (10 000 permutations) to evaluate correlation between relatedness and distance within distance classes.

Neighborhood size and dispersal distance

We estimated neighborhood size (*Nb*), the number of individuals belonging to a panmictic breeding unit, and gene dispersal distance (σ_g) using an iterative approach implemented in SPAGeDi (Hardy & Vekemans 2002). To attain σ_g, SPAGeDi first estimates neighborhood size using the formula

$$Nb = -(1/F_N)/blog,$$

where F_N is the mean kinship (Loiselle *et al.* 1995) of the first distance class (0–50 m in Mäkrä and 0–60 m in Ranta–Halola) and *blog* is the regression slope of kinship on the natural logarithm of spatial distance over all distance classes (Rousset 2000; Hardy & Vekemans 2002). Kinship was estimated with 28 378 SNPs with MAF ≥ 0.20 (due to computational limitations), using the formula

$$F_{ij} = \sum_l [\sum_a (p_{ila} - p_{la})(p_{jla} - p_{la}) + \sum_a \frac{p_{la}(1-p_{la})}{(n_l - 1)}] / \sum_l \sum_a (p_{la}(1-p_{la})),$$

where p_{ila} is the frequency of allele *a* at locus *l* in individual *I* and n_l is the number of gene copies defined in the sample at locus *l* (Loiselle *et al*. 1995; Hardy & Vekemans 2002). Then gene dispersal distance was estimated using the formula

$$\sigma_g = [Nb/(4\pi D_e)]^{1/2},$$

where D_e is the effective population density that accounts for differences in the reproductive success of individuals (Hardy & Vekemans 2002). The *Nb* estimation procedure was repeated using *blog* from kinship~distance regression up to the distance of the chosen maximum σ value. Based on the census density (*D*) estimate of 2 000 trees per hectare (200 000 trees per km^2) in commercial forests (Fahlvik *et al*. 2005; Väisänen *et al*. 1989), we estimated σ_g assuming ratios of effective to census density of 0.25 (D_e = 500 trees/ha) and 0.5 (D_e = 1 000 trees/ha). Census density estimates of unmanaged Finnish Scots pine forests vary between 608-4470 trees/ha (Lönnroth 1926); moist forests have lower densities than dry forests at the same age due to faster growth. Gene dispersal distance can reliably be estimated within a distance that is assumed to be in mutation–drift–equilibrium that should be reached in a few generations within the distance $\sigma_g/(2\mu)^{1/2}$ (μ = mutation rate; Rousset 1997 & 2000). Based on a 10^{-3} mutation rate of microsatellites, this distance is approximately $20*\sigma_g$. Since the SNP mutation rate (10^{-9}; Willyard *et al*. 2007) is lower than the microsatellite mutation rate, we estimated *Nb* (and thus also sigma) using both $20*\sigma_g$ and a higher value of $60*\sigma_g$. However, since the maximum distances within our study sites are shorter than the estimated $20*\sigma_g$ distance, increasing the maximum distance does not affect our estimates of sigma nor *Nb*. To measure the strength of the fine genetic structure, we estimated the intensity of spatial genetic structure (*Sp*; Vekemans & Hardy 2004) using the formula

$$Sp = -blog/(1 - F_N).$$

Rare alleles

To study the spatial spread of rare alleles, we investigated sharing of rare alleles with MAF < 0.01 (23 623 SNPs after removing singletons) between individuals and correlated (Spearman's ρ) this with pairwise relatedness and spatial distance. To avoid the SNP discovery ascertainment bias in rare allele sharing, we estimated the correlation using only 426 individuals excluding: i) individuals used in the SNP discovery and ii) individuals with

pairwise relatedness (GRM) >= 0.044 with the SNP discovery individuals. We used within sampling site Mantel correlograms to study the correlation between rare allele sharing and spatial distance within distance classes. *P*-values for the correlations were constructed with Mantel's test using 10 000 permutations. For illustrative purposes, we fitted a local (LOESS) regression in R for the proportion of shared rare alleles on relatedness and included sampling site as a fixed predictor.

Results

Spatial genetic analyses

On the population scale with all 469 adult trees, we identified very weak structure between Mäkrä and Ranta–Halola based on PCA (Figure 2). The majority of the samples in both Mäkrä and Ranta–Halola clustered together along principal components (PC) 1 and 2 (Figure 2a). However, we detected one outlying cluster of samples from Ranta–Halola on PC1 and one on PC2 and a group of samples from Mäkrä dispersed along PC2. These outlying samples consisted of related individuals, and the outliers disappeared when we conducted PCA only for 332 individuals that were unrelated (GRM < 0.044) and not used in the SNP discovery (Figure 2b). Concordantly, we found low F_{ST} of 0.0005 (95% CI: 0.0004–0.0005) and a low but significant between-study site variance component (ϕ_{ST}) of 0.0019 ($p < 0.001$).

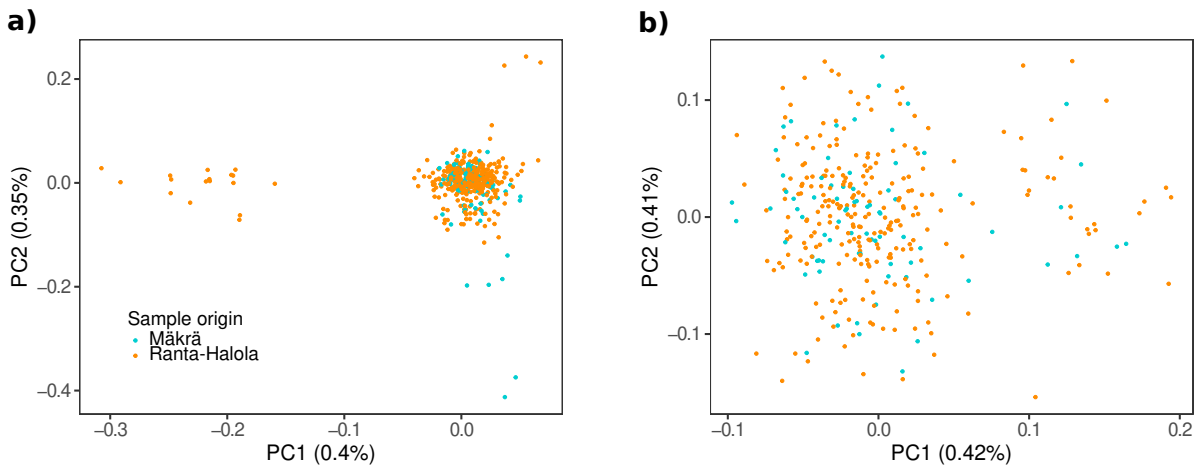

Figure 2. Principal component analysis (PCA) for Mäkrä and Ranta–Halola for a) all 469 adult trees and b) only 332 trees; excluding individuals used in the SNP discovery (Kastally *et al*. 2022), individuals with higher than 0.044 pairwise relatedness with the SNP discovery individuals, and one individual of each pair with pairwise relatedness >= 0.044.

We found that pairwise relatedness was low within study sites (Figure 3). Even in the class with shortest distance between individuals, the mean GRM was 0.0004 in Mäkrä and 0.0018 in Ranta–Halola. However, relatedness decreased with increasing distance in both sampling sites with similar rate (Figure 3). The individuals in closer distance classes were

more related to each other compared to individuals in other distance classes until ~65 m in Mäkrä and ~200 m in Ranta–Halola. In these distance classes, there was negative correlation between relatedness and distance as shown by Mantel correlogram (Figure 4). Mantel test indicated subtle but significant decay of relatedness by spatial distance at Ranta–Halola (Spearmann's ρ = -0.044, one-tailed $p < 0.0001$) but not at Mäkrä (Spearmann's ρ = -0.010, one-tailed $p = 0.216$). We also estimated the intensity of spatial genetic structure and found that, despite there is spatial structure, its intensity is low in both study sites (Ranta–Halola Sp = 0.0008, Mäkrä Sp = 0.0005). Low intensity is caused by mild decrease in the pairwise kinship with distance over each sampling site and by low average kinship in the first distance class.

Among the 69 163 pairwise comparisons, 24 closely related pairs (GRM > 0.177) were identified (Figure 5). All closely related individuals were growing on the same sampling site. The highest relatedness (GRM) for a between sampling site pair was 0.031, whereas the highest within site relatedness was 0.332 in Mäkrä and 0.349 in Ranta–Halola. When the trees were divided into family relationship classes, the median distance of the most related family relationship class found here (GRM = 0.177–0.354) was 51 m in Mäkrä and 59 m in Ranta–Halola, compared to the respective median distances 166 m and 357 m of unrelated individuals (GRM < 0.044; Figure 5). This illustrates that the spatial aggregation of closely related individuals is comparable in both the smaller and the larger study area. We did not find any sign of close inbreeding in our sample; the highest inbreeding coefficient (F_{GRM}) was 0.035 in Mäkrä and 0.080 in Ranta–Halola.

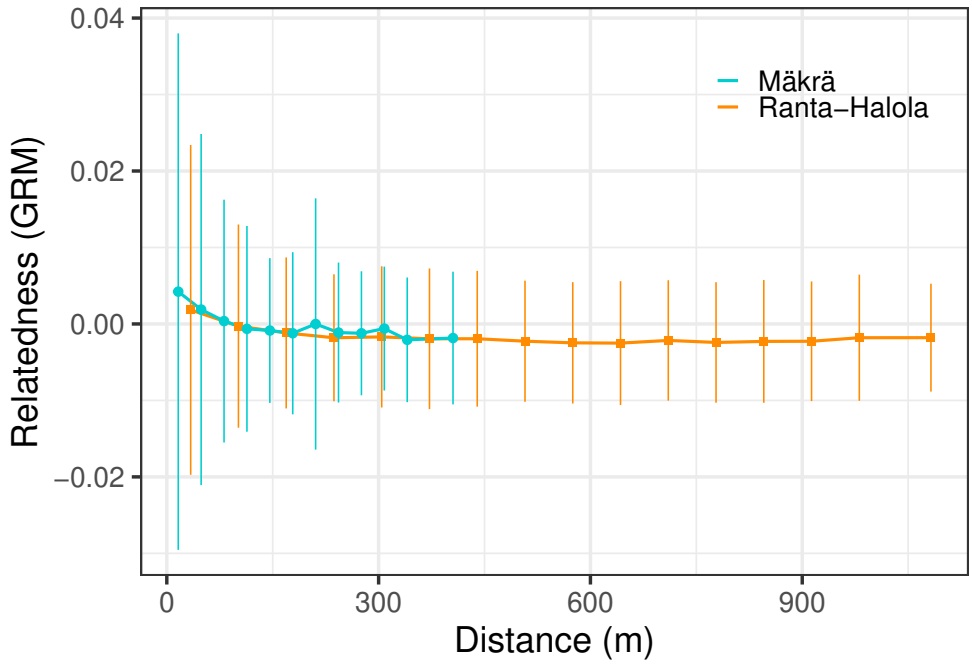

Figure 3. Decay of pairwise relatedness with distance in Mäkrä (turquoise) and Ranta–Halola (orange) sampling sites. Mean (circles and squares) and standard deviation (vertical lines) of relatedness is plotted for each distance class; number of individuals and the mean distance of each distance class are shown in Table S1.

Depending on the effective population density estimate we used, the mean neighborhood size (*Nb*) over the two Scots pine sites was 3210 (with D_e = 500 trees/ha; Table 1) or 1680 trees (with D_e = 1 000 trees/ha). We estimated the mean gene dispersal distances (σ_g) together with *Nb* and found that the mean σ_g were 71.3 m (D_e = 500 trees/ha; Table 1) and 36.5 m (D_e = 1 000 trees/ha).

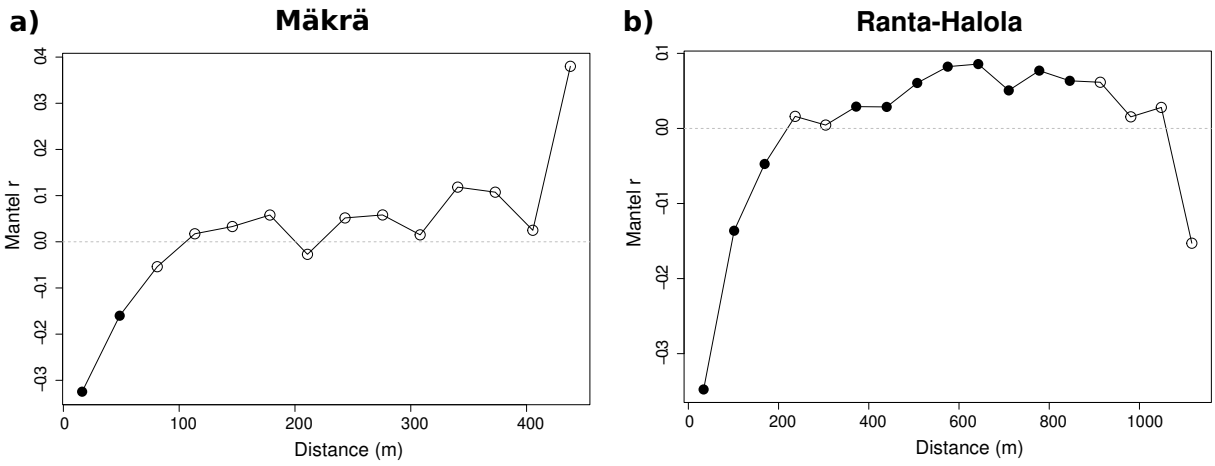

Figure 4. Correlation (measured as Mantel *r*) between pairwise relatedness and distance within distance classes estimated as Mantel correlogram in a) Mäkrä and b) Ranta–Halola. Filled circles indicate *p*-value smaller than 0.05.

Table 1. Gene dispersal distance (σ_g) and neighborhood size (*Nb*) estimates for Mäkrä and Ranta–Halola for two different effective (D_e) and census (D) population densities. In this study, D_e/D ratio of 0.5 equals D_e = 1000 trees/ha and 0.25 equals D_e = 500 trees/ha. Examples of *Nb* and σ_g for animal and wind pollinated tree species from previous studies.

	D_e / D = 0.5		D_e / D = 0.25			
Study site (*Pinus sylvestris*)	***Nb***	**σ_g (m)**	***Nb***	**σ_g (m)**	**Pollinator**	**Reference**
Mäkrä	1404	33	3673	76	wind	this study
Ranta-Halola	1957	39	2747	66	wind	this study
Species						
Dicorynia guianensis	116	222		203	insects	Hardy *et al*. 2006
Moronobea coccinea	20	134		195	birds	Hardy *et al*. 2006
Milicia excelsa			370	3755	wind	Bizoux *et al*. 2009
*Pinus pinaster**	97	30			wind	De-Lucas *et al.* 2009
Araucaria angustifolia	381	140			wind	Sant'Anna *et al*. 2013
*Thuja occidentalis**	65	55			wind	Pandey & Rajora 2012

*Only core or continuous populations included.

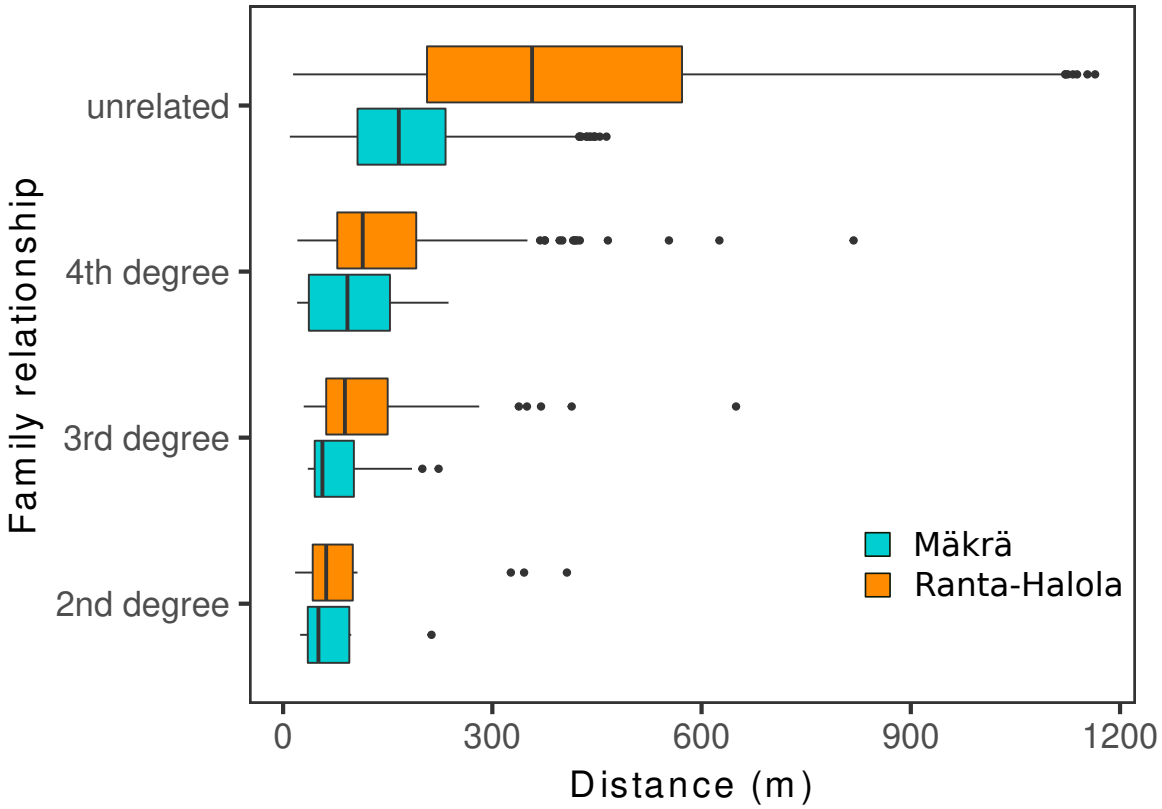

Figure 5. The relation between family relationship classes and spatial distance. The pairwise distances of the second degree related individuals in a) Mäkrä and b) Ranta-Halola are shown as red lines. The distribution of all pairwise distances of individuals in different family relationship classes are shown in c) for Mäkrä (turquoise) and Ranta-Halola (orange). Boxplots show the median (central vertical line), the lower and upper quantiles (boxes), and up to 1.5 interquartile range (whiskers) distances. The family relationships are classified based on pairwise GRM: second degree between 0.177–0.354 (e.g., half-sibling), third degree between 0.088–0.177 (e.g., first cousin), fourth degree between 0.044–0.088 (e.g., first cousin once removed), and unrelated below 0.044.

Rare alleles

In line with overall relatedness, we found that more related individuals also shared a higher proportion of rare alleles than unrelated individuals (n = 426, ρ = 0.049, one-tailed p < 0.0001 estimated using Mantel's test; Figure 6), but this relationship is only visible in higher relatedness values. Rare allele sharing and distance had a weak negative relationship within study site (ρ = -0.010, p = 0.057 in Ranta–Halola and ρ = -0.028, p = 0.084 in Mäkrä; Figure S1). Mantel correlograms showed that pairwise sharing of rare alleles decreased with increasing spatial distance - similarly to the decrease of relatedness with spatial distance (Figure S2).

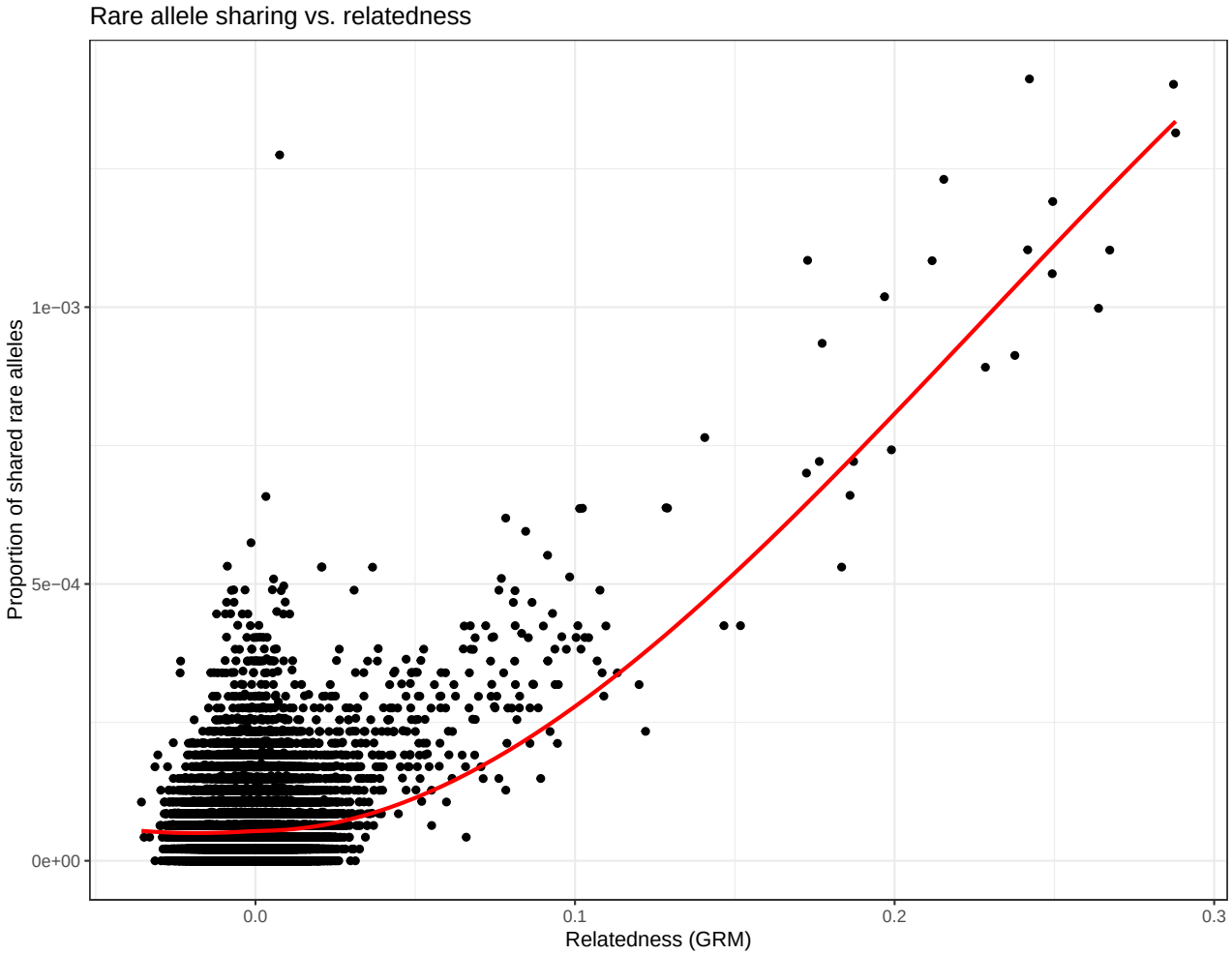

Figure 6. The relationship between proportion of shared rare alleles of all rare alleles and relatedness (GRM). Pairwise comparisons between individuals (dots) and LOESS curve fitted to the proportion of shared rare alleles on relatedness (GRM).

Discussion

Large continuous population of Scots pine has weak fine scale genetic structure

Species with very subtle structure at the between population level are often referred as panmictic, which indicates completely random mating patterns also within populations. While Scots pine's genotypic frequencies follow HW equilibrium at adult stage similarly to a panmictic species (Muona & Harju 1989, Pyhäjärvi *et al*. 2020), its mating patterns are not completely random in space (Robledo-Arnuncio & Gil 2005; Torimaru *et al*. 2012). In concordance with spatially restricted mating patterns, we showed that fine scale genetic structure—albeit very weak—is maintained in adult Scots pine stands (Figures 3 & 4). This was evident in the spatial proximity of individuals with higher pairwise relatedness (Figure 5). Scots pine has partial selfing, but only a part of selfed offspring survive to mature seed stage (Koski 1971, Kärkkäinen & Savolainen 1993), and even fewer survive to the adult stage (Koelewijn *et al*. 1999). Thus, relatedness patterns of adult trees found here are not a result of selfing. Fine scale genetic structure has previously been found in smaller fragmented Scots pine populations (Robledo-Arnuncio & Gil 2005; Sofletea *et al*. 2020) but here we show this pattern within a large continuous population for the first time.

Long-distance pollination and continuous distribution range are likely major contributors to the extremely low intensity of spatial structure we detected in our study (*Sp* from 0.0005 to 0.0008). In addition, high population density increases the neighborhood size so that fine scale genetic structure stays weak even with relatively short gene dispersal distance. In contrast, when populations are far from each other and non-continuous, the chance of pollination from nearby trees is higher, which leads to stronger spatial structure. This effect of population fragmentation is evident in previous Scots pine studies where *Sp* has ranged from up to 0.0098 in Scotland (González-Díaz *et al*. 2017) to the mean of 0.01 on the Carpathian Mountains (Sofletea et al. 2020). Our study sites show low intensity of spatial

structure also when compared to other conifers (*Sp* = 0.001–0.0349; Desilva & Dodd 2021; Kitamura *et al*. 2018; Sant'Anna *et al*. 2013; De-Lucas *et al*. 2009; Vekemans & Hardy 2004) and other tree species (*Sp* = 0.002–0.075; Bizoux *et al*. 2009; Hardy *et al*. 2006; Vekemans & Hardy 2004). Low intensity of population structure is in line with strong gene flow through pollen and indicates that Scots pine populations, for example, at the northern distribution edge receive ample gene flow, and do not suffer from decreased genetic variation as long as they are connected to the main population. Gene flow from southern to northern populations (Varis *et al*. 2009) also aids adaptation to climate change. However, the situation is different in populations that are more fragmented and isolated, such as the Spanish and Scottish populations, where the spatial structure is stronger (Hampe & Petit 2005).

Relatively short dispersal distance of Scots pine and its implications for selection

We estimated that the average gene dispersal distance in our Scots pine stands was 54 meters, which is relatively short compared to the gene dispersal distance estimates of other wind–pollinated trees typically between 30 and 3755 meters (Table 1; Pandey & Rajora 2012; Sant'Anna *et al*. 2013; Bizoux *et al*. 2009; De-Lucas *et al*. 2009), and of animal–pollinated trees up to 1296 meters (Hardy *et al*. 2006), although population density plays a large role in estimating the dispersal distances. Earlier Scots pine seed (10–20 m; Debain *et al*. 2007) and pollen dispersal (47.6–53 m; Koski 1970; Robledo-Arnuncio & Gil 2005) estimates are relatively close to our dispersal distance estimate, bearing in mind that our estimate includes both seed and pollen components. Most dispersal distance estimates—including ours—assume that the distribution of dispersal distances (i.e., the dispersal kernel) is Gaussian. However, especially pollen is able to disperse very long distances (Lindgren *et al*. 1995; Robledo-Arnuncio 2011) leading to potentially more

leptokurtic dispersal kernels that are challenging to estimate and would result in higher mean dispersal distances (Robledo-Arnuncio & Gil 2005; Debain *et al*. 2007).

Understanding of the species' dispersal distance is crucial for understanding how quickly it can spread to new suitable habitats but also for estimating its ability to respond to selection and adapt. The balance between the amount of gene flow and the strength of selection defines the probability of local adaptation as a response to spatially diversifying selection (Lenormand 2002). When the environment is heterogeneous, locally adapted individuals are more likely to produce offspring which succeed in the proximity of their parents. However, gene flow from differentiated populations causes migration load and hinders local adaptation (gene swamping; Lenormand 2002). A population can only respond to selection if the underlying environmental heterogeneity occurs over a distance longer than the characteristic length ($L = \sigma / \sqrt{s}$; Slatkin 1973) that is determined by the length of offspring mean dispersal distance (σ) and the strength of selection (s). With our refined estimates of mean gene dispersal distance (36.5-71.3 m, depending on the effective population density used in estimation), it is possible to estimate the size of the selection coefficient needed for a particular characteristic length. For instance, in order to a population to adapt on a very fine local scale (environmental change over 100 meters), the selection coefficient should be 0.29 ($s = (\sigma / L)^2$, where L = 100 m and σ = 53.9 m). Adaptation to a very small spatial scale would thus require rather strong selection (Thurman & Barrett 2016), which is possible in steep ecological gradients (Scotti *et al*. 2023) but unlikely in a homogeneous landscape such as our study area.

Fitness and practical implications of the fine scale spatial structure

Aggregation of relatives leads to a higher chance of inbreeding and inbreeding depression. The Scots pine carries a high number of lethal equivalents (Koski 1971; Savolainen *et al*. 1992), which makes selfing and more distant forms of inbreeding detrimental. Our results indicate that related individuals carry an excess of shared rare alleles. A large number of

loci in Scots pine have very low minor allele frequencies (Tyrmi *et al*. 2020). Rare alleles are typically young and also enriched for recessive deleterious variants, thus spatial structure may lead to more homozygosity and fitness reduction than expected in a totally panmictic population, where these alleles would rarely appear as homozygotes.

Fine structure also has practical implications, e.g., in tree breeding. When closely located trees are also more likely to be related, determining a suitable collection distance of potential breeding individuals is very important in order to avoid introduction of related individuals into breeding programs and production populations. Due to strong inbreeding depression, accidental selection of related individuals to seed orchards would likely lead to lowered yield of viable seeds. Furthermore, information on spatial genetic structures can help to define a minimum distance between trees to be used for collecting seed or cuttings in practical gene conservation work. Knowledge on the extent and intensity of fine spatial structure is of importance also to forest management on naturally regenerating sites as it can quide the optimization of distance between spared seed trees during harvesting.

Conclusions

Here, we carefully described the extent of spatial genetic structure and average dispersal distance in a large population of Scots pine from a continuous part of the distribution. We demonstrated that even a wind pollinated widely distributed species with large effective population size can have detectable, although weak, fine scale population structure. Our estimates of dispersal distance are relevant for understanding the balance between gene flow and other evolutionary factors, especially selection, as well as in predicting responses to environmental changes.

Acknowledgements

This project was funded by: the European Union's Horizon 2020 research and innovation program, under grant agreement no. 773383 (to UOULU and Luke); the Seventh framework program for research and development, under grant agreement no. 211868 (to UOULU and Luke); the EU Network of Excellence EVOLTREE grant no. 016322 (to UOULU and Luke); the Academy of Finland grants 287431, 293819 and 319313 (to TP), 307582 (to OS), 307581 (to KK), and 309978 (to STK); NoE EVOLTREE (to STK). We thank Natural Resources Institute Finland staff in Loppi and Punkaharju for collecting samples and recording the GPS positions of the trees, and Soile Alatalo for molecular laboratory work at the Ecology and Genetics Unit of the University of Oulu.

Conflicts of interest

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflict of interest: Tanja Pyhäjärvi is a managing board member and a recommender of PCI Evol Biol.

Supplementary Data

The following supplementary data is available as a separate document:

Figure S1. Decay of the proportion of shared rare alleles with spatial distance in Mäkrä (turquoise) and Ranta–Halola (orange) sampling sites. Mean (circles and squares) and standard deviation (vertical lines) of relatedness is plotted for each distance class.

Figure S2. Mantel correlogram for rare allele sharing and pairwise distance in a) Mäkrä and b) Ranta-Halola. Filled circles indicate *p*-value smaller than 0.05.

Data availability statement

The data supporting the findings of this study are available in Figshare (DOI: 10.6084/m9.figshare.23531142; the data will be open to public after this manuscript has been accepted for publication).

References

Ahmed A, Compton SG, Butlin RK & Gilmartin PM (2009) Wind-borne insects mediate directional pollen transfer between desert fig trees 160 kilometers apart. Proceedings of the National Academy of Sciences of the United States of America 106: 20342-20347. https://doi.org/10.1073/pnas.0902213106

Aitken SN & Whitlock MC (2013) Assisted gene flow to facilitate local adaptation to climate change. Annual Review of Ecology, Evolution, and Systematics 44: 367-388. https://doi.org/10.1146/ANNUREV-ECOLSYS-110512-135747

Barton NH (1979) The dynamics of hybrid zones. Heredity 43: 341-359. https://doi.org/10.1038/hdy.1979.87

Bizoux JP, Daïnou K, Bourland N, Hardy OJ, Heuertz M, Mahy G & Doucet JL (2009) Spatial genetic structure in *Milicia excelsa* (Moraceae) indicates extensive gene dispersal in a low-density wind-pollinated tropical tree. Molecular Ecology 18: 4398-4408. https://doi.org/10.1111/j.1365-294X.2009.04365.x

Born C, Hardy OJ, Chevallier MH, Ossari S, Attéké C, Wickings EJ & Hossaert-Mckey M (2008) Small-scale spatial genetic structure in the Central African rainforest tree species *Aucoumea klaineana*: A stepwise approach to infer the impact of limited gene dispersal, population history and habitat fragmentation. Molecular Ecology 17: 2041–2050. https://doi.org/10.1111/j.1365-294X.2007.03685.x

Bradburd G & Ralph P (2019) Spatial population genetics: it's about time. Annual Review of Ecology, Evolution, and Systematics 50: 427-449. https://doi.org/10.1146/annurev-ecolsys-110316-022659

Debain S, Chadœuf J, Curt T, Kunstler G & Lepart J (2007) Comparing effective dispersal in expanding population of *Pinus sylvestris* and *Pinus nigra* in calcareous grassland. Canadian Journal of Forest Research 37: 705-718. https://doi.org/10.1139/X06-265

De-Lucas AI, González-Martínez SC, Vendramin GG, Hidalgo E & Heuertz M (2009) Spatial genetic structure in continuous and fragmented populations of *Pinus pinaster* Aiton. Molecular Ecology 18: 4564-4576. https://doi.org/10.1111/j.1365-294X.2009.04372.x

Desilva R & Dodd RS (2021) Patterns of fine-scale spatial genetic structure and pollen dispersal in giant sequoia (*Sequoiadendron giganteum*). Forests 12: 1–14. https://doi.org/10.3390/f12010061

Escudero A, Iriondo JM & Torres ME (2003) Spatial analysis of genetic diversity as a tool for plant conservation. Biological Conservation 113: 351–365. https://doi.org/10.1016/S0006-3207(03)00122-8

Excoffier L, Smouse PE & Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131: 479-491.

Fahlvik N, Ekö P-M & Pettersson N (2005) Influence of precommercial thinning grade on branch diameter and crown ratio in *Pinus sylvestris* in southern Sweden. Scandinavian Journal of Forest Research 20: 243-251. https://doi.org/10.1080/02827580510008266

González-Díaz P, Jump AS, Perry A, Wachowiak W, Lapshina E & Cavers S (2017) Ecology and management history drive spatial genetic structure in Scots pine. Forest Ecology and Management 400: 68-76. https://doi.org/10.1016/J.FORECO.2017.05.035

Goslee SC & Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software 22: 1-19.

Gougherty AV, Keller SR & Fitzpatrick MC (2021) Maladaptation, migration and extirpation fuel climate change risk in a forest tree species. Nature Climate Change 11: 166-171. https://doi.org/10.1038/s41558-020-00968-6

Hampe A & Petit RJ (2005) Conserving biodiversity under climate change: the rear edge matters. Ecology Letters 8: 461-467. https://doi.org/10.1111/j.1461-0248.2005.00739.x

Hardy OJ & Vekemans X (1999) Isolation by distance in a continuous population: Reconciliation between spatial autocorrelation analysis and population genetics models. Heredity 83: 145–154. https://doi.org/10.1046/j.1365-2540.1999.00558.x

Hardy OJ & Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618-620. https://doi.org/10.1046/j.1471-8286.2002.00305.x

Hardy OJ, Maggia L, Bandou E, Breyne P, Caron H, Chevallier MH, Doligez A, Dutech C, Kremer A, Latouche-Hallé C, Troispoux V, Veron V & Degen B (2006) Fine-scale genetic structure and gene dispersal inferences in 10 Neotropical tree species. Molecular Ecology 15: 559–571. https://doi.org/10.1111/j.1365-294X.2005.02785.x

Kastally C, Niskanen AK, Perry A, Kujala ST, Avia K, Cervantes S, Haapanen M, Kesälahti R, Kumpula TA, Mattila TM, Ojeda DI, Tyrmi JS, Wachowiak W, Cavers S, Kärkkäinen K, Savolainen O & Pyhäjärvi T (2022) Taming the massive genome of Scots pine with PiSy50k, a new genotyping array for conifer research. The Plant Journal 109: 1337-1350. https://doi.org/10.1111/TPJ.15628

Kitamura K, Nakanishi A, Lian C & Goto S (2018) Distinctions in fine-scale spatial genetic structure between growth stages of *Picea jezoensis* Carr. Frontiers in Genetics 9: 490. https://doi.org/10.3389/fgene.2018.00490

Koelewijn HP, Koski V & Savolainen O (1999) Magnitude and timing of inbreeding depression in Scots pine (*Pinus sylvestris* L.). Evolution 53: 758-768. https://doi.org/10.1111/J.1558-5646.1999.TB05370.X

Koski V (1970) A study of pollen dispersal as a mechanism of gene flow in conifers. Metsätieteellisen tutkimuslaitoksen julkaisuja 70(4).

Koski V (1971) Embryonic lethals of *Picea abies* and *Pinus sylvestris*. Metsätieteellisen tutkimuslaitoksen julkaisuja 75(3).

Kremer A, Ronce O, Robledo-Arnuncio JJ, Guillaume F, Bohrer G, Nathan R, Bridle JR, Gomulkiewicz R, Klein EK, Ritland K, Kuparinen A, Gerber S & Schueler S (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change. Ecology Letters 15: 378-392. https://doi.org/10.1111/J.1461-0248.2012.01746.X

Kuparinen A, Savolainen O & Schurr FM (2010) Increased mortality can promote evolutionary adaptation of forest trees to climate change. Forest Ecology and Management 259: 1003-1008. https://doi.org/10.1016/j.foreco.2009.12.006

Kärkkäinen K & Savolainen O (1993) The degree of early inbreeding depression determines the selfing rate at the seed stage: Model and results from *Pinus sylvestris* (Scots pine). Heredity 71: 160-166. https://doi.org/10.1038/hdy.1993.120

Legendre P, Fortin MJ & Borcard D (2015) Should the Mantel test be used in spatial analysis? Methods in Ecology and Evolution 6: 1239–1247. https://doi.org/10.1111/2041-210X.12425

Lenormand T (2002) Gene flow and the limits to natural selection. Trends in Ecology & Evolution 17: 183-189. https://doi.org/10.1016/S0169-5347(02)02497-7

Levin DA & Kerster HW (1974) Gene flow in seed plants. In *Evolutionary Biology*, eds. Th. Dobzhansky, M. K. Hecht, W. C. Steere, New York: Plenum 7: 139-220.

Lindgren D, Paule L, Xihuan S, Yazdani R, Segerström U, Wallin J-E & Lejdebro ML (1995) Can viable pollen carry Scots pine genes over long distances? Grana 34: 64-69. https://doi.org/10.1080/00173139509429035

Loiselle BA, Sork VL, Nason J & Graham C (1995) Spatial genetic structure of a tropical understorey shrub, *Psychotria officinalis* (Rubiaceae). American Journal of Botany 82: 1420–1425.

Loveless MD & Hamrick JL (1984) Ecological determinants of genetic structure in plant populations. Annual review of ecology and systematics 15: 65-95.

Lönnroth E (1926) Untersuchungen über die innere Struktur und Entwicklung gleichaltriger naturnormaler Kiefernbestände: Basiert auf Material aus der Südhälfte Finnlands. Acta Forestalia Fennica 30: 1-269.

Malécot G (1967) Identical loci and relationship. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. L.M. LeCam and J. Neyman, University of California Press, Berkeley 4: 317–332.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M & Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867-2873. https://doi.org/10.1093/bioinformatics/btq559

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Research 27: 209–220.

Marris E (2009) Planting the forest of the future. Nature 459: 906–908. https://doi.org/10.1038/459906a

Muona O & Harju A (1989) Effective population sizes, genetic variability, and mating system in natural stands and seed orchards of *Pinus sylvestris*. Silvae Genetica 38: 221–229.

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW *et al*. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology 15: R59. https://doi.org/10.1186/gb-2014-15-3-r59

Nei M (1972) Genetic distance between populations. The American naturalist 106: 283-292.

Nychka D, Furrer R, Paige J & Sain S (2017) fields: Tools for spatial data. R package version 11.6. https://doi.org/10.5065/D6W957CT

Pandey M & Rajora O (2012) Higher fine-scale genetic structure in peripheral than in core populations of a long-lived and mixed-mating conifer - Eastern white cedar (*Thuja occidentalis* L.). BMC Evolutionary Biology. 12: 48. https://doi.org/10.1186/1471-2148-12-48

Pembleton L, Cogan N & Forster J (2013) StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Molecular Ecology Resources 13: 946-952. https://doi.org/10.1111/1755-0998.12129

Persyn E, Redon R, Bellanger L & Dina C (2018) The impact of a fine-scale population stratification on rare variant association test results. PLOS One 13: e0207677. https://doi.org/10.1371/journal.pone.0207677

Pritchard JK & Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65: 220-228. https://doi.org/10.1086/302449

Privé F, Luu K, Vilhjálmsson BJ, Blum MGB & Rosenberg M (2020) Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. Molecular Biology and Evolution 37: 2153–2154. https://doi.org/10.1093/molbev/msaa053

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81: 559–575. https://doi.org/10.1086/519795

Pyhäjärvi T, Kujala ST & Savolainen O (2020) 275 years of forestry meets genomics in *Pinus sylvestris*. Evolutionary Applications 13: 11-30. https://doi.org/10.1111/eva.12809

Robledo-Arnuncio JJ & Gil L (2005) Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. Heredity 94: 13-22. https://doi.org/10.1038/sj.hdy.6800542

Robledo-Arnuncio JJ (2011) Wind pollination over mesoscale distances: an investigation with Scots pine. New Phytologist 190: 222-233. https://doi.org/10.1111/J.1469-8137.2010.03588.X

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145: 1219-1228. https://doi.org/10.1002/ajmg.c.30221

Rousset F (2000) Genetic differentiation between individuals. Journal of Evolutionary Biology 13: 58–62. https://doi.org/10.1046/J.1420-9101.2000.00137.X

Rousset F (2003) Inferences from spatial population genetics, in: *Handbook of statistical genetics,* eds. D. J. Balding, M. Bishop, & C. Cannings. John Wiley & Sons.

Sant'Anna CS, Sebbenn AM, Klabunde GHF, Bittencourt R, Nodari RO, Mantovani A & dos Reis MS (2013) Realized pollen and seed dispersal within a continuous population of the dioecious coniferous Brazilian pine [*Araucaria angustifolia* (Bertol.) Kuntze]. Conservation Genetics 14: 601-613. https://doi.org/10.1007/S10592-013-0451-5

Savolainen O, Kärkkäinen K & Kuittinen H (1992) Estimating numbers of embryonic lethals in conifers. Heredity 69: 308-314.

Scotti I, Lalagüe H, Oddou-Muratorio S, Scotti-Saintagne C, Ruiz Daniels R, Grivet D, *et al*. (2023) Common microgeographical selection patterns revealed in four European conifers. Molecular Ecology 32: 393-411. https://doi.org/10.1111/mec.16750

Slatkin M (1973) Gene flow and selection in a cline. Genetics 75: 733-756.

Slatkin M (1985) Gene flow in natural populations. Annual Review of Ecology and Systematics 16: 393-430.

Smith JR, Ghazoul J, Burslem DFRP, Itoh A, Khoo E, Lee SL, Maycock CR, Nanami S, Ng KKS & Kettle CJ (2018) Are patterns of fine-scale spatial genetic structure consistent between sites within tropical tree species? PLOS ONE 13: e0193501. https://doi.org/10.1371/journal.pone.0193501

Sofletea N, Mihai G, Ciocîrlan E & Curtu AL (2020) Genetic diversity and spatial genetic structure in isolated Scots pine (*Pinus sylvestris* L.) populations native to Eastern and Southern Carpathians. Forests 11: 1047. https://doi.org/10.3390/f11101047

Thurman TJ & Barrett RDH (2016) The genetic consequences of selection in natural populations. Molecular Ecology 25: 1429-1448. https://doi.org/10.1111/mec.13559

Torimaru T, Wennström U, Lindgren D & Wang X-R (2012) Effects of male fecundity, interindividual distance and anisotropic pollen dispersal on mating success in a Scots pine (*Pinus sylvestris*) seed orchard. Heredity 108: 312-321. https://doi.org/10.1038/hdy.2011.76

Tyrmi JS, Vuosku J, Acosta JJ, Li Z, Sterck L, Cervera MT, Savolainen O & Pyhäjärvi T (2020) Genomics of clinal local adaptation in *Pinus sylvestris* under continuous environmental and spatial genetic setting. G3: Genes, Genomes, Genetics 10: 2683-2696. https://doi.org/10.1534/g3.120.401285

Vakkari P, Rusanen M, Heikkinen J, Huotari T & Kärkkäinen K (2020) Patterns of genetic variation in leading-edge populations of *Quercus robur*: genetic patchiness due to family clusters. Tree Genetics and Genomes 16: 73. https://doi.org/10.1007/s11295-020-01465-9

Varis S, Pakkanen A, Galofré A & Pulkkinen P (2009) The extent of South-North pollen transfer in Finnish Scots pine. Silva Fennica 43: 717-726.

Vekemans X & Hardy O (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. Molecular Ecology 13: 921-935. https://doi.org/10.1046/j.1365-294X.2004.02076.x

Väisänen H, Kellomäki S, Oker-Blom P & Valtonen E (1989) Structural development of *Pinus sylvestris* stands with varying initial density: A preliminary model for quality of sawn timber as affected by silvicultural measures. Scandinavian Journal of Forest Research 4: 223-238. https://doi.org/10.1080/02827588909382560

Weir BS & Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370. https://doi.org/10.2307/2408641

Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, *et al*. (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196: 891-909. https://doi.org/10.1534/genetics.113.159996

Willyard A, Syring J, Gernandt DS, Liston A & Cronn R (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus.* Molecular Biology and Evolution 24: 90-101. https://doi.org/10.1093/MOLBEV/MSL131

Wright S (1943) Isolation by distance. Genetics 28: 114–138.

Yang J, Lee SH, Goddard ME & Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics 88: 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011