

Optimal sampling strategy for estimation of spatial genetic structure in tree populations

S Cavers¹, B Degen², H Caron³, MR Lemes⁴, R Margis^{5,6}, F Salgueiro⁶ and AJ Lowe⁷

¹Centre for Ecology and Hydrology-Edinburgh, Bush Estate, Penicuik, Midlothian EH26 0QB, Scotland, UK; ²BFH Institut fuer Forstgenetik und Forstpflanzenzuechtung, Sieker Landstrasse 2, D-22927 Grosshansdorf, Germany; ³Institut National de la Recherche Agronomique, UMR BIOGECO, 69 Route d'Arcachon, 33612 Cestas Cedex, France; ⁴Instituto Nacional de Pesquisas da Amazonia, Laboratorio de Genetica e Biologia Reprodutiva de Plantas (LabGen) Avenida Andre Araujo 2936, 69083-000 Manaus, AM, Brazil; ⁵Universidade Federal do Rio de Janeiro, LGMV – Laboratorio de Genetica Molecular Vegetal, CCS – Ilha do Fundao, Instituto de Biologia, CEP 21944-270 Rio de Janeiro, Brasil; ⁶Departamento de Bioquímica, Rua Ramiro Barcelos 9800, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil; ⁷School of Life Sciences, University of Queensland, St Lucia, Brisbane, Queensland 4170, Australia

Fine-scale spatial genetic structure (SGS) in natural tree populations is largely a result of restricted pollen and seed dispersal. Understanding the link between limitations to dispersal in gene vectors and SGS is of key interest to biologists and the availability of highly variable molecular markers has facilitated fine-scale analysis of populations. However, estimation of SGS may depend strongly on the type of genetic marker and sampling strategy (of both loci and individuals). To explore sampling limits, we created a model population with simulated distributions of dominant and codominant alleles, resulting from natural regeneration with restricted gene flow. SGS estimates from subsamples (simulating collection and analysis with amplified fragment length polymorphism (AFLP) and microsatellite markers) were correlated with the 'real' estimate (from the full model population). For both marker types, sampling ranges were

evident, with lower limits below which estimation was poorly correlated and upper limits above which sampling became inefficient. Lower limits (correlation of 0.9) were 100 individuals, 10 loci for microsatellites and 150 individuals, 100 loci for AFLPs. Upper limits were 200 individuals, five loci for microsatellites and 200 individuals, 100 loci for AFLPs. The limits indicated by simulation were compared with data sets from real species. Instances where sampling effort had been either insufficient or inefficient were identified. The model results should form practical boundaries for studies aiming to detect SGS. However, greater sample sizes will be required in cases where SGS is weaker than for our simulated population, for example, in species with effective pollen/seed dispersal mechanisms.
Heredity (2005) **95**, 281–289. doi:10.1038/sj.hdy.6800709; published online 20 July 2005

Keywords: spatial genetic structure; trees; sampling; spatial autocorrelation; dominant; codominant

Introduction

Plant species develop strong genetic structure, that is nonrandom distribution of genotypes (Vekemans and Hardy, 2004), at a variety of spatial scales due to their sedentary nature (Silvertown, 2001). In certain circumstances, for example, colonisation of new habitat, spatial genetic structure (SGS) may develop very quickly (<10 generations) and be highly persistent (Epperson, 1990) although patterns may be dynamic, changing with population age as phenomena such as dispersal independent selection, self-thinning and succession begin to act (Hamrick *et al.*, 1992, 1993; Epperson, 1993; Epperson and Alvarez-Buylla, 1997; Chung *et al.*, 1998; Jensen *et al.*, 2003).

The strength and spatial magnitude of population structuring may influence and be influenced by a variety of factors, including historical processes (vicariance, dispersal) and selection (Epperson and Li, 1996, 1997).

At a population scale, interspecific differences in the partitioning of variation are due largely to life form and breeding system, and several syntheses (Hamrick *et al.*, 1992; Degen *et al.*, 2001a; Vekemans and Hardy, 2004; Ward *et al.*, 2005) have identified generalisable trends. For example, selfing species generally maintain strong genetic structure, while among outcrossing species, animal-mediated pollen and gravity-mediated seed dispersal mechanisms create stronger patterns. However, in general, at the population level, despite the potential influence of highly localised factors such as spatial variation in the distribution of species and selection for microhabitat variation (Levin and Kerster, 1974; Epperson, 1993; Doligez *et al.*, 1998; Degen *et al.*, 2001a), SGS is predominantly a consequence of limited seed and pollen dispersal (Epperson and Li, 1997; Doligez *et al.*, 1998; Degen *et al.*, 2001a; Epperson, 2004; Vekemans and Hardy, 2004).

Conservation of forest genetic resources and the development of forest management plans that account for intraspecific genetic diversity are of significant contemporary interest, as part of global efforts to preserve biodiversity and ensure environmental sustainability (Lowe *et al.*, 2005; UN, 2000; Kanashiro *et al.*, 2002).

Correspondence: S Cavers, Biosystem Management, Centre for Ecology and Hydrology, Edinburgh Research Station, Bush Estate, Penicuik, Midlothian EH26 0QB, Scotland, UK. E-mail: scav@ceh.ac.uk
Received 21 February 2005; accepted 1 June 2005; published online 20 July 2005

Natural forests that come under management for production, sustainable or otherwise, are likely to experience considerable disruption of SGS (Young and Merriam, 1994; Degen *et al*, 2001a; Lowe *et al*, 2003). It should be a key aim for management plans to tailor extraction such that this disturbance is minimised and that remnant genetic structure is sufficient to promote regeneration and maintenance of genetic diversity (Lowe *et al*, 2005). To advance these efforts, several recent studies have taken advantage of new, highly variable genetic markers to conduct detailed analysis of tree populations and explore the link between limitations to seed and pollen dispersal and patterns of spatial genetic variation observed on the ground (Doligez and Joly, 1997; Geburek *et al*, 1998; Strieff *et al*, 1998; Degen *et al*, 2001a; Cottrell *et al*, 2003; Latouche-Halle *et al*, 2003).

Most commonly, analysis of SGS is approached using spatial autocorrelation methods (Sokal and Oden, 1978), comparing patterns of genetic variation with geographical distribution. In contrast to population genetic estimators (F_{ST} and related statistics), which require averaging across populations or hierarchical levels, spatial autocorrelation uses data from all pairs of individual locations across the sample surface and therefore accesses much more of the available information at the population scale (Epperson and Li, 1997). In addition, spatial autocorrelation makes no assumptions about the spatial scale of structuring in populations (Epperson, 1989; Heywood, 1991; Chung *et al*, 1998).

Multilocus measures using genetic distances have been shown to be very sensitive in detecting SGS (Smouse and Peakall, 1999). However, the statistical power of the technique depends on actual population structure, size of sample, and aspects of the scale, orientation and distribution of locations across the population surface (Kremer *et al*, 2005; Epperson and Li, 1996). In other words, the pattern and magnitude of sampling relative to the population are critical. In addition, for population genetic questions, the selection of molecular marker is also of great importance. In this analysis we aim to determine, for a dominant (amplified fragment length polymorphism, AFLP) and a codominant (microsatellite) marker, an optimal sampling strategy, that is numbers of markers and individuals to be sampled, for reliable estimation of SGS. We use a simulated population based on actual field data, to determine, for a variety of sampling strategies and for dominant and codominant markers, limits for meaningful estimation of SGS and use these limits to explore and criticise some recent data sets.

Methods

The model Eco-Gene (Degen *et al*, 1996; Degen and Roubik, 2004) was used to generate two artificial data sets (dominant and codominant) from field data. Using diameter distribution and density data for the neotropical tree species, *Symphonia globulifera*, at a permanent sample plot at Paracou, French Guiana (Figure 1), a population of 1900 trees in a 1200 m × 1200 m area (144 ha) was simulated. Initial codominant (microsatellite) and dominant (AFLP) data sets were created by distributing genotypes across this population. Each tree was given an artificial genotype of (a) 100 microsatellite loci and (b) 100 AFLP loci. Microsatellite genotypes were

generated based on actual allele frequencies of three tropical tree species (*Symphonia globulifera*, *Dicorynia guianensis* and *Sextonia rubra*; Degen *et al*, 2001a). For the AFLP data set we created 100 loci with two alleles (1 and 2). The frequency of allele 1 was evenly distributed over the 100 loci from 5 to 95% (5% intervals). Initially, for both data sets, genotypes were in Hardy–Weinberg proportions and there was no SGS.

Eco-Gene was then used to simulate the SGS that would develop in this population after 1000 years given limited pollen and seed dispersal (for details of model functions see <ftp://ghd.dnsalias.net/degen/software.html>). Pollen and seed dispersal curves were based on data for relatively abundant tropical tree species, as measured at an experimental plot at Paracou, French Guiana (Figure 2). The SGS of this population at the end of the simulations was used as the ‘real’ SGS for comparison in subsequent analyses. Using the same input data sets, four repetitions of the 1000 year run were carried out, allowing between-repetition comparison of the pattern of SGS established at the end of the simulation. For each repetition, mean pairwise genetic distance was calculated for trees distributed in 10 distance classes of 50 m each (50–500 m, Figure 3a, b). A distance-based approach was selected for the analysis of SGS as it can be applied to both dominant and codominant multilocus data, with the qualities/limita-

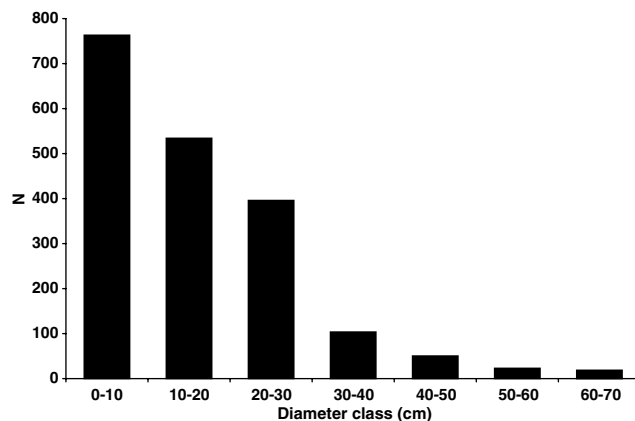


Figure 1 Diameter distribution of *Symphonia globulifera* in 144 ha of forest, from the experimental trial Paracou, French Guiana.

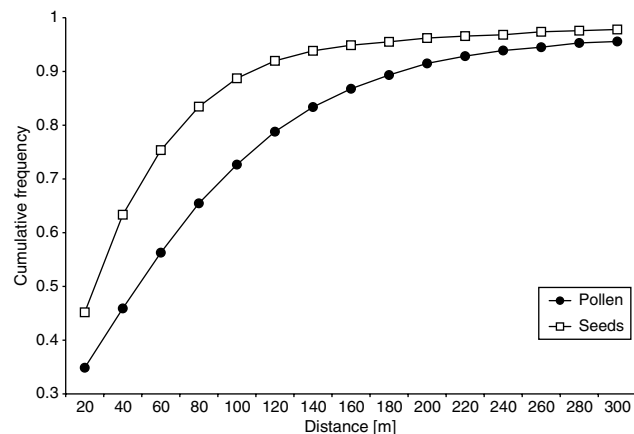


Figure 2 Simulated limited pollen and seed dispersal.

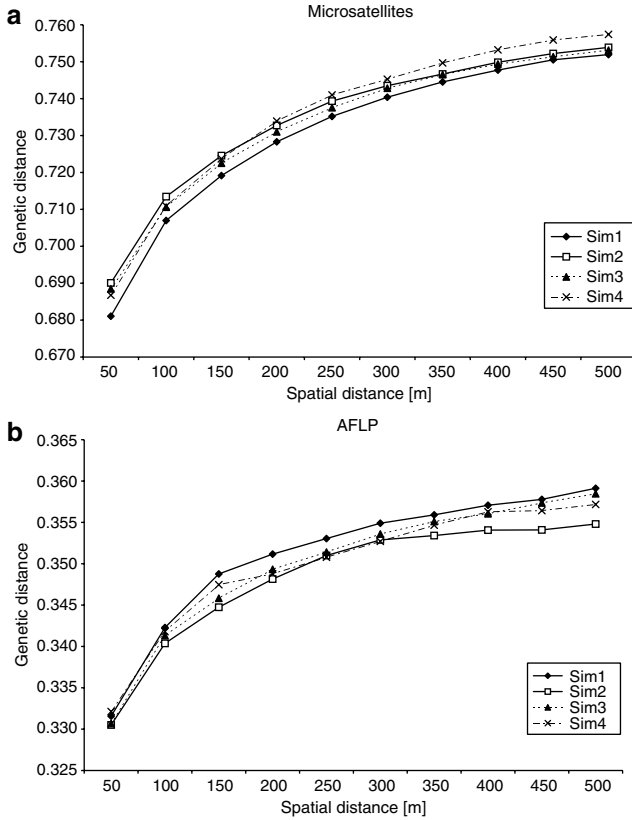


Figure 3 Distogram of the spatial genetic structure at (a, top) 100 microsatellite loci for four repetitions after 1000 simulated years; (b, bottom) 100 AFLP loci for four repetitions after 1000 simulated years.

tions of the marker type taken into account through selection of appropriate distance measures. For the microsatellite data sets, genetic distance was estimated using Gregorius' distance, D_G (Gregorius, 1978):

$$D_G(i, j) = \frac{1}{2} \sum_{k=1}^n |p_{ik} - p_{jk}|$$

where i and j represent two populations, n is the number of alleles or haplotypes, p_{ik} is the relative frequency of the k th allele or haplotype.

For AFLP data sets, allele 1 was assumed to be dominant over allele 2, hence the genotypes 11 and 12 were transformed to 1 and the genotype 22 to 0, creating a binary matrix of 1 and 0. Genetic distance was then estimated using Tanimoto's distance, D_{ij} (Degen et al, 2001b):

$$D_{ij} = 1 - \frac{v_{ij}}{v_{ij} + y_i + y_j}$$

where v_{ij} represents the number of loci scored as 1 in both individuals i and j , y_i and y_j are the numbers of loci that score 1 in only individual i or j , respectively.

Using the simulated population to determine 'real' SGS, a series of sampling strategies (ie variations of the numbers of individuals and loci used) were then tested for both codominant and dominant data sets. For the tests, the program SGS v1.0c (Degen et al, 2001b; ftp://ghd.dnsalias.net/degen/software.html) was used to analyse spatial autocorrelation in the data sets. Again,

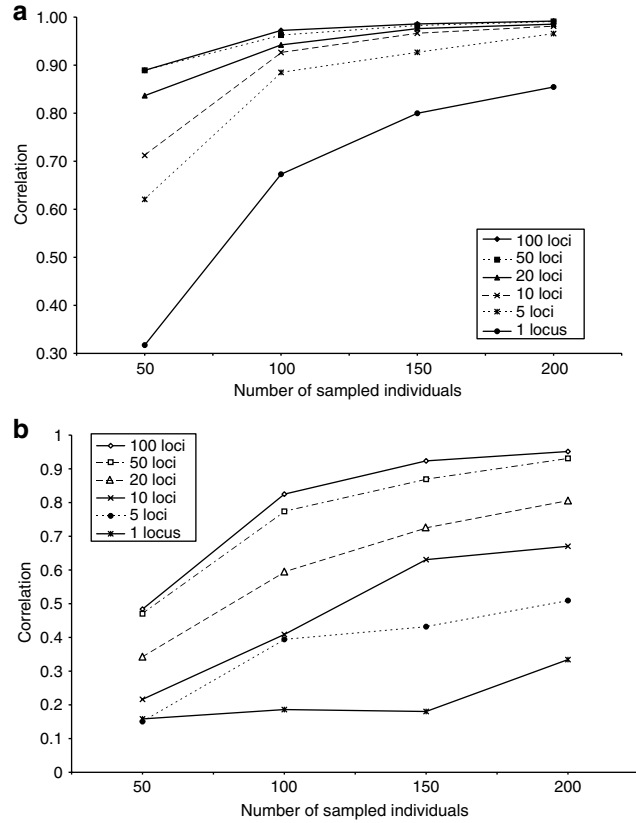


Figure 4 Mean correlation between 'real' distogram and the distogram drawn from series of subsamples for (a, top) microsatellites (number of sampled loci = 1, 5, 10, 20, 50, 100; number of sampled individuals = 50, 100, 150, 200), and (b, bottom) AFLPs (number of sampled loci = 1, 5, 10, 20, 50, 100; number of sampled individuals = 50, 100, 150, 200).

mean pairwise genetic distances for microsatellite and AFLP data sets were computed using Gregorius' and Tanimoto's distances, respectively. For each marker type, random samples of 50, 100, 150 and 200 individuals were drawn from the simulated population. At each sample size, a series of data sets were generated with increasing numbers of loci and used to estimate SGS (1, 5, 10, 20, 50, 100 loci). In each case, the estimated SGS, as determined from the sampled data set, was correlated with the 'real' SGS as determined for the full simulated population (Figure 4a,b). Each sampling strategy (number of individuals, number of loci) was repeated 100 times and a mean correlation coefficient calculated.

For both codominant and dominant data, the simulated results were used to make recommendations on the minimum sample size and number of loci necessary for meaningful determination of SGS. For a number of data sets drawn from published and new studies (Table 1), the relationship between the number of individuals sampled and number of loci used in the simulated data were explored using a resampling approach. SGS was analysed in subsamples of loci or individuals and distograms for each subsample were correlated with that for the full data set. Each subsampling was repeated 100 times and mean correlation reported. While such resampling of data sets inevitably introduces error, the trends revealed are informative and permit an evaluation

Table 1 Real data sets used for resampling, including seed and pollen dispersal characteristics, data set size and sampling area and density as determined from supplied coordinates

Species	Location	Area (km ²)	Density trees/ha	Agg. index	Seed disp	Pollen disp	N	Loci	Ref
<i>Swietenia macrophylla</i> ^a	Costa Rica	0.12	3.83	0.755	Wind	Insect	46	215 ^b	Lowe <i>et al</i> (2003)
<i>Swietenia macrophylla</i> ^c	Brazil	4.57	0.20	0.681	Wind	Insect	93	8	Lemes <i>et al</i> (2003)
<i>Sextonia rubra</i> ^c	F. Guiana	3.79	0.49	0.360	Gravity	Insect	184	4	Hardy, unpublished
<i>Dicorynia guianensis</i> ^c	F. Guiana	0.26	5.92	0.703	Wind	Insect	154	6	Degen <i>et al</i> (2001a)
<i>Symphonia globulifera</i> ^c	F. Guiana	0.29	5.10	0.427	Animal	Bird	148	3	Degen <i>et al</i> (2004)
<i>Eugenia uniflora</i> ^a	Brazil	0.09	30.89	0.202	Animal	Insect	278	109	Salgueiro, unpublished

^aAFLP data.^bData set contained a high proportion of low frequency/monomorphic loci: these were removed for analysis of SGS (no. of polymorphic loci = 44).^cMicrosatellite data.Area = precise area over which samples taken. Density = $N/(\text{Area} \times 100)$. Aggregation index from SGS program.**Table 2** Mean correlation for microsatellites and AFLPs between the real distogram and the distogram drawn from different samples

N\loci	Microsatellites						AFLPs					
	100	50	20	10	5	1	100	50	20	10	5	1
50	0.889	0.889	0.837	0.712	0.621	0.317	0.484	0.470	0.342	0.216	0.150	0.158
100	0.972	0.963	0.942	0.926	0.885	0.673	0.825	0.774	0.594	0.409	0.394	0.186
150	0.986	0.983	0.976	0.966	0.927	0.800	0.924	0.869	0.725	0.631	0.432	0.180
200	0.992	0.991	0.985	0.981	0.966	0.855	0.951	0.931	0.806	0.670	0.509	0.334

Italic values indicate correlation of >0.9: sampling strategies using these numbers are recommended.

of the effort used during collection of the data set. Variation of correlation according to the numbers of loci used was examined using two data sets for Mahogany (*Swietenia macrophylla*): an AFLP data set of 215 markers, 46 individuals, $N = 46$ (Lowe *et al*, 2003) and a microsatellite data set of eight loci, $N = 93$ (Lemes *et al*, 2003). The Mahogany AFLP data set contained a high proportion of low frequency or monomorphic markers, such that only 44 of the 215 loci were polymorphic (frequency of >0.05). The full data set was used as published (Lowe *et al*, 2003) to examine variation in locus numbers but, for analysis of SGS, the data set was reduced to include only polymorphic loci. In addition, variation in correlation according to the numbers of individuals sampled was examined for four microsatellite data sets: *S. macrophylla* (8 loci, $N = 93$; Lemes *et al*, 2003), *Sextonia rubra* (4 loci, $N = 184$; Hardy unpublished), *Dicorynia guianensis* (6 loci, $N = 154$; Degen *et al*, 2001a) and *Symphonia globulifera* (3 loci, $N = 148$; Degen *et al*, 2004) and a single AFLP data set: *Eugenia uniflora* (109 loci, $N = 278$; Salgueiro, unpublished).

For each real microsatellite and AFLP data set, the pattern of SGS was determined using the program SGS v1.0c, in all cases using 1000 permutations of the data set to obtain 95% confidence intervals. These were assessed in the light of the results from the simulations and resamplings.

Results

Repeated simulations of the development of SGS over 1000 years in the model population produced highly consistent patterns (Figure 3a, b). In addition, the scale of SGS observed, that is, the distances at which significant spatial autocorrelation was detected, was similar to

experimentally determined values observed for other tropical tree species (Degen *et al*, 2001a).

The sampling strategies evaluated indicate some clear patterns (Figure 4a, b; Table 2). Firstly, in our simulations, AFLP data required much greater sampling effort compared to microsatellite data. For any given sample size many more AFLP loci and greater numbers of individuals were required to achieve the same degree of correlation as for microsatellites. The pattern is clearly evident in comparison of the trends observed for variation in the estimates from the 100 repetitions carried out (Figure 5). Estimates derived from microsatellite data achieve consistency much more rapidly than those from AFLP data sets. The contrast is a consequence of the lower information content and lower allele numbers per locus of dominant markers as compared to codominant markers (Lynch and Milligan, 1994).

For both marker types, it is a logical expectation that the more markers and individuals sampled, the better the correlation with 'real' SGS. However, for both data types, target sampling ranges are evident, with lower limits below which meaningful estimates of real SGS cannot be made but with upper limits above which the information gain per unit sampling effort declines rapidly.

If a mean correlation of at least 0.9 between real and sampled distogram is taken as a minimum target then, for microsatellites, this can be achieved with a sample of 100 individuals and 10 loci, although close to 0.9 is achievable with five loci. With 200 individuals sampled, five loci can provide 0.95 correlation and little is gained from increasing either locus or individual numbers. For AFLPs, >0.9 correlation can be achieved with a sample of 150 individuals and 100 loci. With 200 individuals sampled, 100 loci provide >0.95 correlation and, again, little is gained for greater effort. It should be noted, however, that this means 100 polymorphic loci (no fixed

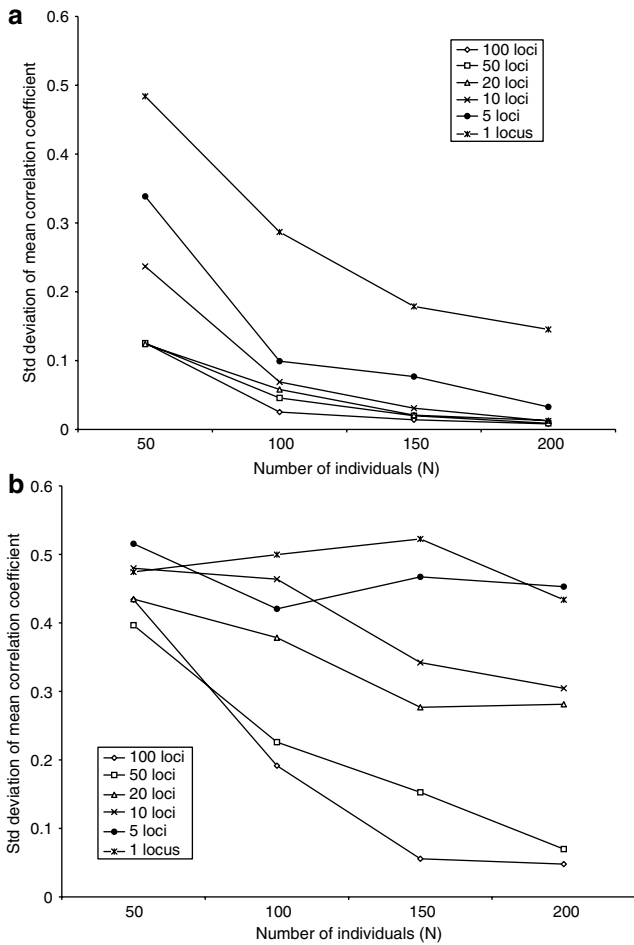


Figure 5 Variation in standard deviation of mean correlation between 'real' SGS and SGS estimated from series of subsamples (see Figure 4a, b). Each subsampling was repeated 100 times. Top: values for microsatellite subsample data sets. Bottom: values for AFLP subsample data sets.

loci were included in this analysis). With a sample of only 100 individuals, more than 100 polymorphic AFLP loci would be required to approach correlation of 0.9.

In this analysis we have assessed correlation by calculating, in each distance class, the mean Pearson correlation coefficient between the genetic distance in the real population and the values in the sampled population. This is a conservative approach in that the shape of whole distogram is considered, so the sampling ranges we identify can be considered stringent. In other words, sampling within the ranges identified is likely to allow efficient and accurate estimation of 'real' SGS. There is a limitation on the extent to which our results can be considered general, in that correlation between samples and the real population depends, to some extent, on the selection of distance classes. As greater numbers of distance classes are used, the number of data pairs per distance class decreases, introducing more stochastic variation and reducing correlation. Our analysis of the simulated data consistently used 10 distance classes so, within this framework, the patterns of correlation should be robust. As a natural outcome of the balance between physical sampling effort and ensuring that the numbers of data pairs per distance class is sufficient, selection of

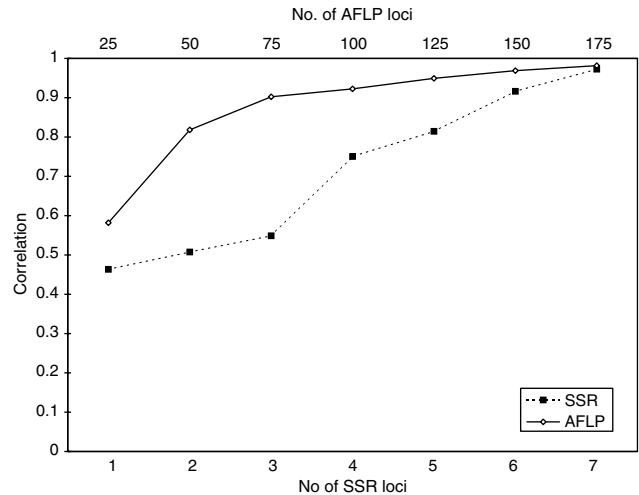


Figure 6 Resampling of real data for variation over number of loci. Mean correlation for SSR and AFLP markers between SGS from full data set and that derived from resampled data sets (no. of sampled loci: microsatellites – 1, 2, 3, 4, 5, 6, 7; AFLPs – 25, 50, 75, 100, 125, 150, 175). Note that full AFLP data set is used (no of polymorphic loci = 44).

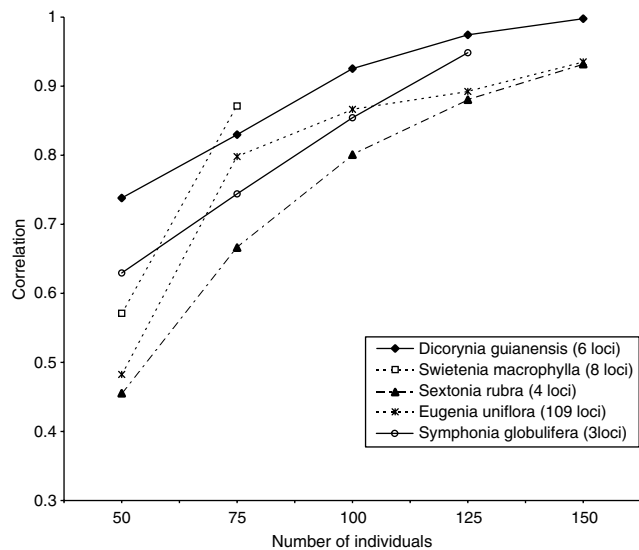


Figure 7 Resampling of real data for variation over number of individuals. Mean correlation for four SSR data sets and one AFLP data set, between SGS from full data set and that derived from resampled data sets (no of sampled individuals: 50, 75, 100, 125, 150).

around 10 distance classes is commonplace and as such our results should be broadly applicable.

The pattern of correlation observed in the simulations was mirrored in the resampling studies of real data sets (Figures 6 and 7). For both AFLP and microsatellite markers in *S. macrophylla*, >0.9 correlation is achieved with fewer markers than used in the published analyses (Figure 6; Lemes et al, 2003; Lowe et al, 2003). In the case of the AFLP study, >0.9 correlation with the final data set is achieved with 75 markers. As noted above, the AFLP data set used here was as published (Lowe et al, 2003) and contained a high proportion of low frequency/ monomorphic loci (44 polymorphic loci). So, the correla-

tion rapidly approaches 1 as the number of sampled loci approaches the number of polymorphic loci present. Comparing this data set with the simulations, all of the polymorphic loci would have to be included to make any estimation of SGS, and this would still have low correlation with the 'real' SGS. For microsatellites, estimation with six loci matched that made with the eight loci used in the published analysis (Lemes *et al*, 2003). For these data sets, if the scale of real SGS is of the same order as that observed for the simulated population, then even the full AFLP sample will only achieve a correlation of around 0.5 with the 'real' SGS, while for the microsatellite data, six loci and $N = 93$ would achieve nearly 0.9 correlation.

For both the AFLP and microsatellite data sets that were resampled for numbers of individuals (Figure 7), high levels of correlation with the 'real' data set were attained quickly when the number of loci was high. For microsatellites, the rate at which correlation was attained was not strictly dependent on the number of loci (*S. globulifera*, with three loci, approached full correlation faster than *S. rubra*, four loci), perhaps suggesting that qualities of the individual locus may become important (eg level of polymorphism), although we did not explore the relationship between allelic richness and SGS calculation efficiency.

Patterns of SGS for each species are shown in Figure 8. The mean pairwise genetic distance in each distance

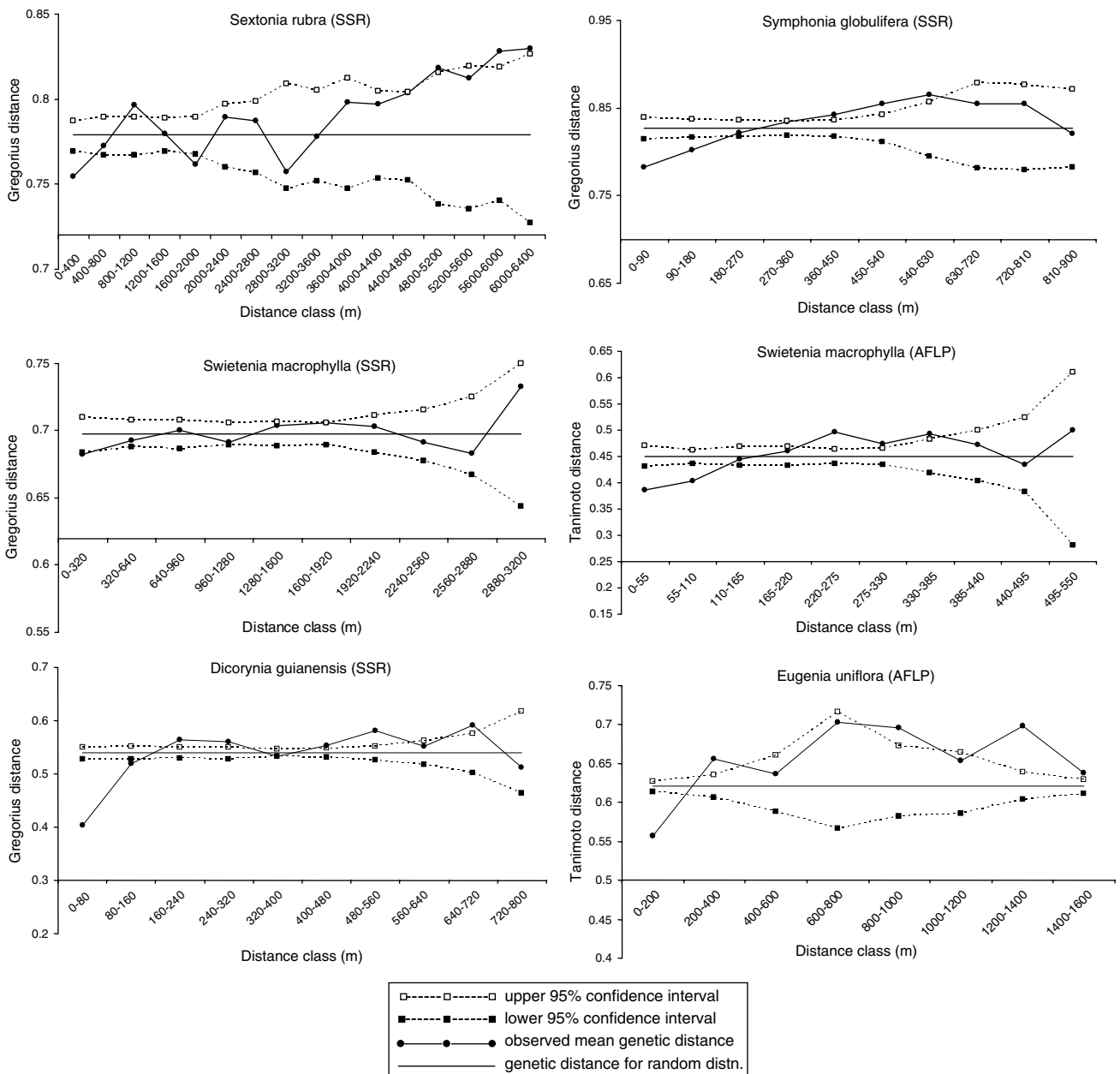


Figure 8 Distograms for real data sets used for resampling. All data sets were analysed using the program SGS. Solid central line indicates value at which there is no spatial autocorrelation. Solid line with filled circles indicates observed level of genetic distance. Dotted lines indicate 95% significance levels as determined using 1000 permutations of the actual data set: hollow squares – upper 95% confidence interval, filled squares – lower 95% confidence interval.

class is plotted together with the level at which genetic structure is random and upper and lower 95% confidence intervals generated by 1000 permutations of the data sets. Significant spatial structure is observed where the observed mean genetic distance is above or below the confidence interval, that is, where observed genetic distance is significantly greater or less than that expected from a random distribution respectively.

Discussion

Our simulations indicate that, where moderate SGS exists, there are clear target sampling ranges, of both numbers of individuals and loci, within which the effectiveness of a molecular marker type for estimating SGS is maximised. The sampling effort required (for both individuals and of loci) is much greater for AFLP markers than for microsatellite markers. Using microsatellites and for species with SGS of the same magnitude as that simulated, once five loci are available, it is much more effective to focus on increasing individual sample numbers than increasing numbers of loci. With five loci and 100 individuals, a correlation of close to 0.9 is achievable. Using dominant markers, the number of both loci and individuals required is much higher; at least 100 loci and 150 individuals. These recommendations are somewhat lower than previous predictions, for example (Geburek and Tripp-Knowles, 1994) recommend sampling of 300–400 trees, but it is important to bear in mind that the sampling scheme required will depend strongly on the particular characteristics of the species studied. The key question is how closely does the SGS estimated by the analysed sample reflect the actual SGS in the population? To successfully estimate real SGS, scale and distribution of sampled individuals and the number and type of molecular loci must be carefully considered.

Sampling of individuals

Theoretical expectations are that where the spatial scale of sampling is similar to the spatial scale of the pattern of SGS, the ability to make inferences based on autocorrelation statistics is limited (Slatkin and Arter, 1991). When the spatial scale of sampling is smaller than that of the SGS, autocorrelation can be powerful (Sokal and Oden, 1978; Epperson, 1990, 1993; Sokal *et al*, 1997). However, the number and distribution of individuals sampled from a population must be carefully considered, with respect to local species distribution, spatial density and expectations of SGS based on life form and breeding system (Hamrick *et al*, 1992). For example, species with effective long-distance dispersal mechanisms (eg wind-dispersed pollen or animal-dispersed seed) should be expected to show only weak SGS (Vekemans and Hardy, 2004). In this case it is likely to be more efficient to put effort into sampling individuals than increasing the number of markers. In general, if the SGS of a study species is expected to be weak, then sample sizes should be increased above our recommendations.

The orientation of sampling with respect to the distribution of individuals on the ground is also important. Our simulations used a random sampling of individuals across the whole population. In reality for tropical tree species, it is likely that sampling will be biased towards clustered individuals, due to the diffi-

culty of locating low-density target species in species-rich forest. To counter bias due to distribution, a mixed sampling strategy is probably best, balancing high-density local sampling with wider scale coverage, for example, using transects in multiple dimensions (Vekemans and Hardy, 2004). At the same time, however, the sampling strategy must ensure that sufficient numbers of pairwise comparisons are produced in each distance class to achieve statistical significance (a minimum of 30 pairs per class is recommended; Degen *et al*, 2001b). A further consideration for sample distribution is the age structure of the population. It is notoriously difficult to successfully estimate age in tree populations, particularly tropical tree populations, but diameter at breast height (DBH) measurements are often used (but see case of *Eugenia uniflora* below). Where possible these should be taken and incorporated into the data set. In natural stands, SGS is likely to be influenced by age: as populations age, self-thinning and succession will lead to changes in SGS, most likely increasing the spatial extent of patterns (eg Jensen *et al*, 2003). Therefore, it will be important to account for age when estimating SGS, particularly where comparative analysis is to be attempted.

Marker properties

Different molecular markers yield significantly different amounts of information and all require a critical minimum effort to provide a statistically meaningful picture of true SGS. Our simulations have clearly demonstrated the consequences of the lower information content of dominant markers relative to codominant markers. In addition, the markers themselves need to be critically evaluated, in particular the assumption of marker neutrality. Criticism has been made of the use of traditional autocorrelation statistics (eg Moran's *I*) to address population genetic questions (Slatkin and Arter, 1991), due to the risk that different loci experience different evolutionary forces, rendering averaged statistics meaningless (Hardy and Vekemans, 1999). However, if it is reasonable to assume linkage disequilibrium, and selective neutrality, then averaging over loci should not introduce bias (Epperson, 2004), although this should be explicitly tested (Kremer *et al*, 2005).

A further consideration is the allele frequency distribution. We based initial allele frequencies in our simulations on those of a series of neotropical trees (codominant data; Degen *et al*, 2001a) and on an even distribution of marker presence across 100 loci (dominant data). In both cases, it is possible (even likely) that allele frequency distributions for other species will differ and efficiency of SGS estimation may be affected.

Real data sets

Of the real data sets reanalysed here, most would give good estimates of SGS (ie >0.9 correlation), if the 'real' SGS for these species is of the same order as that simulated. For the *Swietenia macrophylla* AFLP data set, the number of individuals sampled is low ($N = 46$), thus, even with 215 loci, only a weak estimate of the 'real' SGS is achievable (likely to achieve around 0.6 correlation with 'real' SGS). Furthermore, there are only 44 polymorphic loci present in this data set. So while the inclusion of all loci in the original publication is justified

(analysis of genetic diversity), their inclusion for analysis of SGS is not and the extent of SGS estimated here (Figure 8) is much stronger than that found previously (Lowe *et al.*, 2003). In any case, for the purposes of estimating real SGS, the difference between the full and edited data sets is minimal as the number of individuals sampled is so low. Also, the inclusion of fixed loci in this case reduced the magnitude of the SGS detected, but the spatial pattern found was similar to that detected using only polymorphic loci. So, for the *S. macrophylla* AFLP data set, a greater sampling of both individuals and loci is necessary.

In the case of the *Symphonia globulifera* microsatellite data set, the number of loci sampled is low (three loci). However, due to the high information content of the marker a good estimate of the 'real' SGS is possible (>0.8 correlation with 'real' SGS). For this data set, the number of individuals sampled should be sufficient to successfully estimate SGS, and, if further effort were to be made, it would be most efficient to concentrate on obtaining additional loci (two more would give >0.9 correlation).

In terms of effort required to successfully estimate SGS, the microsatellite data sets for *S. macrophylla* (8 loci, $N=93$), *Dicorynia guianensis* (6 loci, $N=154$) and the AFLP data set for *Eugenia uniflora* (109 loci, $N=278$) all achieve a good balance (sufficient but not excessive). In the latter case, fewer individuals could theoretically have been sampled to achieve a successful estimate of SGS (with 100 loci, between 150 and 200 individuals should be sufficient). However, this case illustrates the necessity for considering individual species characteristics. *Eugenia uniflora* has a shrub form and identifying independent individuals can be difficult (Salgueiro, pers. comm.). As a result, many apparently independent samples may in fact be duplicates of single widespread individuals. Therefore, in this case, the extra sampling is justified in order to ensure a large enough sampling of independent trees. The form of *E. uniflora* also highlights the difficulty of using DBH to estimate age structure in populations: for this species DBH gives no real indication of individual age due to the shrub structure and high frequency of regrowing stems. For the species *Sextonia rubra* and *Symphonia globulifera*, the population sample is sufficient and additional loci would be the most efficient focus for further sampling effort.

Summary and extensions

The simulations and analysis presented provide indications of the numbers of individuals and loci for dominant and codominant markers, necessary for successful estimation of SGS in tree populations. These recommendations are qualified by the requirement that any study must consider the characteristics of its target species (mating system, seed and pollen dispersal mechanisms) and plan sampling and marker selection appropriately. In addition, SGS is a dynamic quality that changes over time with population aging, due to selection, density independent thinning and successional processes (eg Jensen *et al.*, 2003). Such considerations are particularly pertinent for comparative analyses that seek common biological factors responsible for patterns of genetic structure; studies that are being actively pursued to identify key considerations for forest management (Lowe *et al.*, 2005; Ward *et al.*, 2005). For these efforts, it will be

critical to ensure that sampling schemes for different species provide statistically meaningful outputs. Using the recommendations detailed here as a guide, sampling for each species can be designed such that estimates of SGS can be confidently expected to mirror real patterns.

Acknowledgements

The work reported in this paper was conducted as part of the project 'Assessment of levels and dynamics of intraspecific genetic diversity of tropical trees' (funded by the EC under FP5 – International Cooperation, Contract # ERBIC18CT970149). The project involved a consortium of six Institutions: CEH (UK), INRA (France), IPBO (Belgium), INPA (Brazil), CATIE (Costa Rica) & UFRJ (Brazil), and was coordinated by AJL. We thank Olivier Hardy for permission to use the *Sextonia rubra* data set, and two anonymous referees for their comments.

References

- Chung MY, Chung GM, Chung MG, Epperson B (1998). Spatial genetic structure in populations of *Cymbidium goeringii* (Orchidaceae). *Genes Genet Systems* **73**: 281–285.
- Cottrell J, Munro RC, Tabbener H, Milner AD, Forrest GI, Lowe AJ (2003). Comparison of fine-scale genetic structure using nuclear microsatellites within two British oakwoods differing in population history. *Forest Ecol Manage* **176**: 287–303.
- Degen B, Bandou E, Caron H (2004). Limited pollen dispersal and biparental inbreeding in *Symphonia globulifera* in French Guiana. *Heredity* **93**: 585–591.
- Degen B, Caron H, Bandou E, Maggia L, Chevallier MH, Leveau A *et al.* (2001a). Fine-scale spatial genetic structure of eight tropical tree species as analysed by RAPDs. *Heredity* **87**: 497–507.
- Degen B, Gregorius H-R, Scholtz F (1996). ECO-GENE, a model for simulation studies on the spatial and temporal dynamics of genetic structure of tree populations. *Silvae Genet* **45**: 323–329.
- Degen B, Petit RJ, Kremer A (2001b). SGS – Spatial Genetic Software: a computer program for analysis of spatial genetic and phenotypic structures of individuals and populations. *J Hered* **92**: 447–448.
- Degen B, Roubik DW (2004). Effects of animal pollination on pollen dispersal, self-pollination and effective population size of tropical trees: a simulation study. *Biotropica* **36**: 165–179.
- Doligez A, Baril C, Joly HI (1998). Fine-scale spatial genetic structure with non-uniform distribution of individuals. *Genetics* **148**: 905–919.
- Doligez A, Joly HI (1997). Genetic diversity and spatial structure within a natural stand of a tropical forest tree species, *Carapa procera* (Meliaceae), in French Guiana. *Heredity* **79**: 72–82.
- Epperson BK (1989). Spatial patterns of genetic variation within plant populations. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) *Plant Population Genetics, Breeding and Genetic Resources*. Sinauer Associates Inc.: Sunderland, MA. pp 229–253.
- Epperson BK (1990). Spatial autocorrelation of genotypes under directional selection. *Genetics* **124**: 757–771.
- Epperson BK (1993). Spatial and space-time correlations in systems of subpopulations with genetic drift and migration. *Genetics* **133**: 711–727.
- Epperson BK (2004). Multilocus estimation of genetic structure within populations. *Theoret Popul Biol* **65**: 227–237.

- Epperson BK, Alvarez-Buylla ER (1997). Limited seed dispersal and genetic structure in life stages of *Cecropia obtusifolia*. *Evolution* **51**: 275–282.
- Epperson BK, Li TQ (1996). Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proc Natl Acad Sci USA* **93**: 10528–10532.
- Epperson BK, Li TQ (1997). Gene dispersal and spatial genetic structure. *Evolution* **51**: 672–681.
- Geburek T, Mottinger-Kroupa S, Morgante M, Burg K (1998). Genetic variation of Norway Spruce (*Picea abies* (L.) Karst.) populations in Austria. II. Microspatial patterns derived from nuclear sequence tagged microsatellite sites. *Forest Genet* **5**: 231–237.
- Geburek T, Tripp-Knowles P (1994). Spatial stand structure of sugar maple (*Acer saccharum* Marsh) in Ontario, Canada. *Phyton-Ann REI Bot* **34**: 267–278.
- Gregorius HR (1978). The concept of genetic diversity and its formal relationship to heterozygosity and genetic distance. *Math Biosci* **41**: 253–271.
- Hamrick JL, Godt MJW, Sherman-Broyles S (1992). Factors influencing levels of genetic diversity in woody plant species. *New Forests* **6**: 95–124.
- Hamrick JL, Murawski DA, Nason JD (1993). The influence of seed dispersal mechanisms on the genetic structure of tropical tree populations. *Vegetatio* **107/108**: 281–297.
- 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.
- Heywood JS (1991). Spatial analysis of genetic variation in plant populations. *Annu Rev Ecol Syst* **22**: 335–355.
- Jensen JS, Olrik DC, Siegmund HR, Lowe AJ (2003). Population genetics and spatial autocorrelation in an unmanaged stand of *Quercus petraea* in Denmark. *Scand J Forest Res* **18**: 295–304.
- Kanashiro M, Thompson IS, Yared JAG, Loveless MD, Coventry P, Martins-da-Silva RCV *et al* (2002). Improving conservation values of managed forests: the Dendrogene project in the Brazilian Amazon. *Unasylva* **53**: 25–33.
- Kremer A, Caron H, Cavers S, Colpaert N, Gheysen L, Gribel R *et al* (2005). Monitoring genetic diversity in tropical trees with multilocus dominant markers. *Heredity* **95**: 274–280.
- Latouche-Halle C, Ramboer A, Bandou E, Caron H, Kremer A (2003). Nuclear and chloroplast genetic structure indicate fine-scale spatial dynamics in a neotropical tree population. *Heredity* **91**: 181–190.
- Lemes M, Gribel R, Proctor J, Grattapaglia D (2003). Population genetic structure of mahogany (*Swietenia macrophylla* King, Meliaceae) across the Brazilian Amazon, based on variation at microsatellite loci: implications for conservation. *Mol Ecol* **12**: 2875–2883.
- Levin DA, Kerster HW (1974). Gene flow in seed plants. *Evol Biol* **7**: 139–220.
- Lowe AJ, Boshier D, Ward M, Bacles CFE, Navarro C (2005). Genetic resource loss following habitat fragmentation and degradation; reconciling predicted theory with empirical evidence. *Heredity* **95**: 255–273.
- Lowe AJ, Jourde B, Breyne P, Colpaert N, Navarro C, Wilson J *et al* (2003). Fine-scale genetic structure and gene flow within Costa Rican populations of mahogany (*Swietenia macrophylla*). *Heredity* **90**: 268–275.
- Lynch M, Milligan BG (1994). Analysis of population genetic structure with RAPD markers. *Mol Ecol* **3**: 1–9.
- Silvertown J (2001). Plants stand still but their genes don't: non-trivial consequences of the obvious. In: Silvertown J, Antonovics J (eds) *Integrating Evolution and Ecology in a Spatial Context*. Cambridge University Press: Cambridge. pp 347.
- Slatkin M, Arter HE (1991). Spatial autocorrelation methods in population-genetics. *Am Nat* **138**: 499–517.
- Smouse PE, Peakall R (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**: 561–573.
- Sokal RR, Oden N (1978). Spatial autocorrelation in biology, 1. Methodology. *Biol J Linn Soc London* **10**: 199–228.
- Sokal RR, Oden NL, Thomson BA (1997). A simulation study of microevolutionary inferences by spatial autocorrelation analysis. *Biol J Linn Soc London* **60**: 73–93.
- Strieff R, Labbe T, Bacilieri R, Steinkeller H, Glossl J, Kremer A (1998). Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Mol Ecol* **7**: 317–328.
- UN (2000). Millenium development goals. UN Resolution 55/2, United Nations Millenium Declaration.
- Vekemans X, Hardy OJ (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol Ecol* **13**: 921–935.
- Ward M, Dick CW, Gribel R, Lemes M, Caron H, Lowe AJ (2005). To self, or not to self... A review of outcrossing and pollen-mediated gene flow in neotropical trees. *Heredity* **95**: 246–254.
- Young AG, Merriam HG (1994). Effects of forest fragmentation on the spatial genetic-structure of *Acer saccharum* Marsh (Sugar Maple) populations. *Heredity* **72**: 201–208.