ARTICLES

Optimum Sample Size for SSR-based Estimation of Representative Allele Frequencies and Genetic Diversity in Scots Pine Populations

DARIUS DANUSEVICIUS[1,2*], DARIUS KAVALIAUSKAS[1] AND BARBARA FUSSI[3]

[1]Institute of Forest Biology and Silviculture, Faculty of Forest and Ecology, Aleksandras Stulginskis University, Studentų str. 11, LT-53361 Akademija, Kaunas distr., Lithuania

[2]Institute of Forestry, Lithuanian Research Centre for Agriculture and Forestry, Liepų str. 1, Girionys, Kaunas distr., Lithuania

[3]Bavarian Office for Forest Seeding and Planting, Am Forstamtsplatz 1, D-83317, Teisendorf, Germany, E-mail: Barbara.Fussi@asp.bayern.de

Corresponding author: Darius Danusevicius; e-mail: darius.danusevicius@asu.lt, phone: +370 37 752232; fax: +370 37 752397

Danusevičius, D., Kavaliauskas, D. and Fussi, B. 2016. Optimum Sample Size for SSR-based Estimation of Representative Allele Frequencies and Genetic Diversity in Scots Pine Populations. *Baltic Forestry* 22(2): 194-202.

Abstract

We used the random subsampling approach based on the empirical data to identify the representative sample size for accurate estimates of allele frequencies within a population. The empirical data consisted of 12 nuclear microsatellite marker scores for 400 individuals sampled within 1 ha area in a representative natural stand of Scots pine. For each sample size, 100 resampled subsets were randomly drawn (without replacement). The sample size, at which 95 % of the resampled subsets contained all the alleles at a given frequency present in the empirical data set, was considered as a 95 % probability of sampling these alleles. The resampled subsets were also used to calculate main genetic diversity parameters and their variances to be used as a measure of accuracy of sampling. The results showed that at the 95 % probability level, the sample sizes of 20-25 and 65-80 individuals were large enough to capture all the alleles at frequency above 0.05 and 0.01-0.05, respectively. 300-350 individuals were required to sample the alleles at frequencies below 0.01 at the 95 % probability. The upper bound of the sample sizes was required for the loci exhibiting high He values (>0.80).

Keywords: differentiation, sampling, SSR, genetic diversity, *Pinus sylvestris*.

Introduction

It is generally accepted that the statistical power of molecular data can be boosted at three major levels (a) including more loci, (b) increasing sample size and (c) selecting better loci (Selkoe and Toonen 2006). In this study, we address the problem of optimum sample size for a microsatellite based study in Scots pine by using empirical genotypic data of a large sample size from a typical natural stand of Scots pine, representing a population of such natural stands. This is a particular feature of our study, because the earlier optimizations rarely focused on temperate and boreal forest tree species (with a rare exception in ash, Miyamoto et al. 2008) and usually used sampling from a series of populations, where the particular alleles were probably missing and a bias due to the population differentiation may have been introduced (e.g. Miyamoto et al. 2008, Hale et al. 2012). SSRs are among the most powerful and polymorphic markers in population genetic studies (Zane et al. 2002). Along with their advantages, SSRs also have a number of undesirable features such as null alleles, stuttering and homoplasy, all increasing the error of the genetic estimates (Selkoe and Toonen 2006). This results in difficulty in defining an optimal sample size, which can be specific to the loci and species used and even type of genetic entries sampled. Therefore, generalizing the optimizations of sample sizes from other species and markers may introduce an undesirable bias.

Commonly population genetic studies of plants based on DNA markers use sample sizes of 50 per population to estimate various genetic parameters (Nybom 2004). The relevant studies mainly used the rarefication, repeated random sampling or regression approaches to identify representative sample sizes for estimating the allelic richness or genetic distances, where the need of sampling of all the alleles in the population was assumed (Leberg 2002, Kalinowski 2002, Miyamoto et al. 2008, Bashalkhanov et al. 2009). Satisfying the significance thresholds accounting for all the alleles in the population requires large sample sizes, which depending on the variability of the loci and their discriminating power (e.g. the F_{ST} value) varied between 20 to 100 (Kalinowski 2002), 120 (Bashalkhanov et al. 2009) or at about 300 individuals (Miyamoto et al. 2008). These studies confirmed the tight relationship between the sample size and the allele frequency to be captured (e.g. Miyamoto et al. 2008, Kalinowski 2002, 2005, Hale et al. 2012). However, alleles occurring at variable frequency may have different evolutionary importance (Li et al. 2004). A large number of the SSR-based studies are aimed to identify population structure, where the informative alleles are those occurring at medium frequencies allowing to be shared within populations more than among populations (review by Selkoe and Toonen 2006). Thus, for a vast number of studies, sampling all the alleles in a population may be unnecessary. Therefore, for an efficient optimization, we need to set certain thresholds for allele frequency to be sampled (Hale et al. 2012). An approach to proceed is adapting the common acceptance that a locus is informative if the most common allele occurs at a frequency less than 0.95 or reversely less common alleles to be informative have to occur at a frequency above 0.05 (Hale at al. 2012, Hartl and Clark 1997). Thus, sampling of alleles with frequencies over 0.05 is the primary target of many population genetic studies. Significance of rare alleles for population differentiation is low and evolutionary meaning is meager because they represent recent recurrent mutations rather than evolutionary ancestry (Selkoe and Toonen 2006, Oliveira 2006). The alleles with frequencies of 0.01 to 0.05 are common for SSRs and are worth investigating separately as they may represent private alleles important to population differentiation and gene conservation. At the locus level, the optimum sample sizes are markedly affected by the evenness of allele frequencies usually expressed by the expected heterozygosity (H_e) suggesting a higher precision of separate optimization for loci with different H_e values (Hale et al. 2012). Separate optimization for loci with different H_e values is also more practical as most of the studies are planned for particular loci, for which then sample size can be adjusted based on the existing optimizations. Another aspect to be considered is that the sample size judged on visual scores of relative improvement of a parameter with increasing sample size is a subjective method sensitive to locus type and material sampled. Setting statistical thresholds for optimum sample size is much more useful approach. Moreover, there is a lack of studies genotyping large sample sizes from a single stand to be representative to the true ideal population on which the reliability of the modeling largely depends (Miyamoto et al. 2008).

The objective of our study was to identify the representative sample size required to accurately estimate allele frequencies within a population of Scots pine based on nuclear microsatellite markers scored within a single stand of Scots pine. We focused on finding optimum sample sizes for capturing less common and common alleles.

Materials and Methods

The empirical data set

The empirical data set used for the modeling of the sample size contains the nuclear SSR scores from 400 adult trees of Scots pine sampled within a natural mature Scots pine stand located in the north-eastern part of Lithuania. The Scots pine stand was even-aged composed of a single age class of ca. 60 years old and located in the middle of a large forest tract dominated by Scots pine, thereby representing the population of natural self-regenerating Scots pine forests. Such stands are commonly used for assessment of pine genetic structure. Our material is well suited for such an optimization of the sample sizes as it represents a large-scale genotyping of a single stand, not several stands, so it is not biased due to differentiation or genetic drift. The 400 trees were sampled systematically by a 4 × 4-metre grid within an area of 1 ha (no diameter-based or any other selection was applied for the sampling). The total stocking was ca. 600 trees per ha in this stand, which is a relatively higher number due to management restrictions because of a reserve status given two decades ago.

We genotyped the trees at the following 12 nuclear SSR loci Psyl57, Psyl2, Psyl18, Psyl42, Psyl25, Psyl16 (Sebastiani et al. 2012), Spag7.14, Spac12.5, Spac11.4 (developed for *Pinus sylvestris*, Soranzo et al. 1998) and PtTX4011, PtTX4001 (developed for *Pinus taeda*, Elsik et al. 2000), combined into 3 multiplexes for the PCR and capillary electrophoresis by using an automated sequencer (CEQ GeXP Beckman-Coulter) (Table 1). The DNA was extracted according to the modified ATMAB-method after Dumolin et al. (1995). The details for the DNA analyses are given in (Danusevicius et al. 2015). After the allele scoring, the data set was checked for scoring errors due to stuttering, large allele dropouts and null alleles with the approach estimating the excess of homozygotes implemented in the Microchecker software ver. 2.2.3 (Van Oosterhout et al. 2006). We used the GenAlex ver. 6.5 software to calculate the genetic diversity parameters for the loci.

Table 1. The description of the loci studied. The loci were grouped by the expected heterozygosity (H_e) value into three groups. N_a is the observed allele number. N_e is the effective allele number. H_o is observed heterozygosity. H_e is unbiased to sample size expected heterozygosity. F_{IS} is the inbreeding coefficient

Locus	N	N_a	N_e	H_o	H_e	uH_e	F_{IS}
Low H_e loci (<0.50)							
Psyl25	394	2	1.0	0.03	0.030	0.030	-0.02
Psyl44	388	6	1.1	0.05	0.050	0.051	0.08
Psyl2	395	6	1.5	0.31	0.330	0.330	0.06
Psyl18	389	8	1.1	0.07	0.076	0.076	0.08
Medium H_e loci (0.50-0.80)							
Psyl57	387	8	2.0	0.47	0.506	0.506	0.08
Psyl42	395	5	3.2	0.70	0.689	0.689	-0.01
PtTX4011	384	8	2.9	0.51	0.659	0.660	0.22
PtTX4001	396	17	3.9	0.72	0.744	0.745	0.04
High H_e loci (>0.80)							
Spac7.14	395	36	20.5	0.91	0.951	0.953	0.04
Psyl16	396	13	5.2	0.77	0.808	0.809	0.05
Spac11.4	396	17	6.2	0.86	0.839	0.841	-0.02
Spac12.5	395	33	17.7	0.93	0.944	0.945	0.02

The simulations

To estimate the minimum sample size required to capture alleles occurring within certain frequency threshold with a 95 % or 100 % probability, we used the repeated random sampling approach as follows. For each sample size of 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 65, 80, 100, 150, 200, 250, 300, 350, we simulated 100 data subsets by randomly resampling the individuals from the empirical data set without replacement (no entry could be sampled more than once) by the aid of Visual Basic programming in Excel. The percentage of the resampled subsets containing all the alleles at frequencies (a) above 0.05, (b) 0.01 to 0.05 and (c) less than 0.01 was estimated and used as the probability for detecting alleles with certain frequencies at variable sample sizes. Loci were grouped for simulation analysis according to their variability with high H_e (>0.8), medium H_e (0.5 to 0.8) and low H_e (<0.5). In addition, to find the minimum sample size required to capture the alleles at certain frequencies, we pooled those alleles from all the loci and identified the size of the resampled data set at which the corresponding allele occurred in more than 95% of the resampled subsets to be illustrated in a response plot (Figure 3).

The resampled subsets for each sample size were submitted to GenAlex ver. 6.5 to calculate the genetic diversity parameters (H_o, H_e, F_{IS}, N_e) for each resampled subset and the pairwise F_{ST} values between each resampled subset and the empirical data set (for each sample size separately). Genetic diversity parameters and pairwise F_{ST} values were averaged for each sample size to obtain the mean estimates and their variability as indicators of sampling accuracy at each sample size.

To investigate the overall accuracy of sampling, we pooled all the resampled subsets for each sample size into one dataset by using the Tukey LSD test from ANOVA (SAS software) to investigate the effect of sample size on the pairwise F_{ST} values between each sampled replicate and the empirical data set. Our assumption with the ANOVA was that the groups of sample sizes with no significant differences between each other may indicate the thresholds for minimum representative sample size for the overall allele frequencies in the population. To improve normal distribution of the response variable for the ANOVA, the pairwise F_{ST} values between each sampled replicate and the empirical data were transformed by raising it to the second power. None of the arcsine square, square root, inversion and log transformation improved normal distribution of the F_{ST}.

Results

Description of the loci

All the loci were polymorphic with 2 to 36 alleles (Table 1). Based on the variability, the loci can be grouped as (1) highly variable (17 to the 36 alleles) with high H_e values (>0.8), (2) medium variability (8 to 17 alleles) with medium H_e values (0.5-0.8) and (3) low variability (2 to 8 alleles) with low H_e values (<0.5) (Table 1). The SPAC loci were highly variable with the allele frequencies approximately following the SMM model (Figure 1). Except for Psyl16 and Psyl42, the remaining Psyl loci were least informative, e.g. for Psyl18 and Psly44, the most common allele reached the frequency over 0.96 (Figure 1). Inbreeding coefficient F_{IS} was close to 0 indicating low inbreeding levels in mature trees of the natural Scots pine stand.

Sample sizes required to detect alleles at 95 % probability level

The sample size of the resampled subsets containing all alleles with frequency above 0.05 was markedly affected by the evenness of allele frequencies at the particular loci (H_e, Figure 2). At the 95 % probability level, the sample size required for sampling the above 0.05 frequency alleles at any of the loci ranged from 20 and 25 individuals for medium and high H_e loci, respectively (dashed vertical lines in Figure 2). At the 100 % probability, the corresponding sampling sizes ranged among 35 to 40 for the medium and high H_e loci (solid lines in Figure 2). The loci with lower H_e values required relatively lower sample sizes to sample the alleles at the frequencies over 0.05, e.g. 10 individuals for Psyl16 with H_e of 0.81 (Figure 2). To capture the low frequency alleles at frequencies 0.01 to 0.05, markedly higher sample sizes were needed: 65 and 80 individuals for medium and high H_e loci, respective-

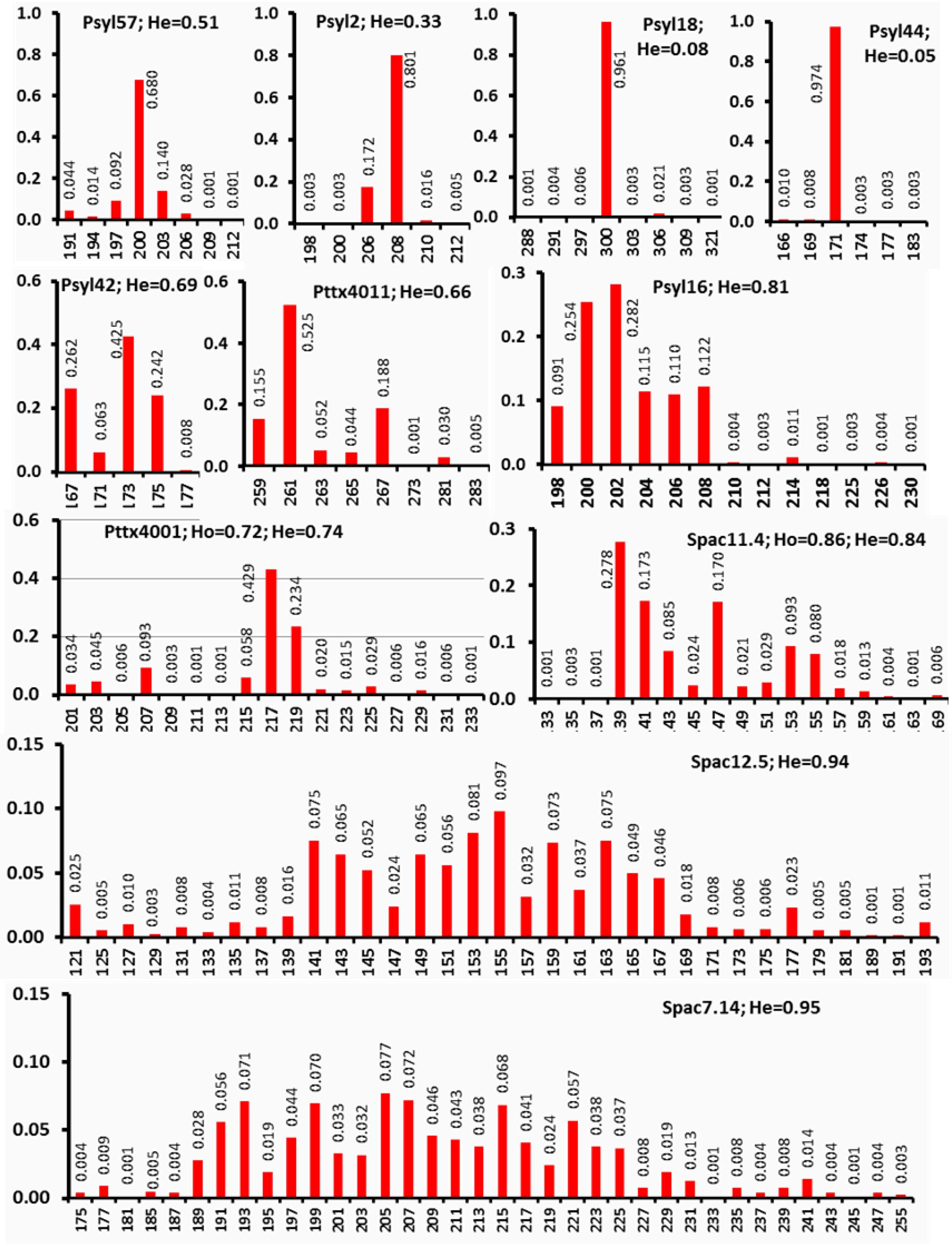

Figure 1. Allele frequencies of the empirical data set (*n* = 400) given for each locus. The expected heterozygosity is shown at the locus name

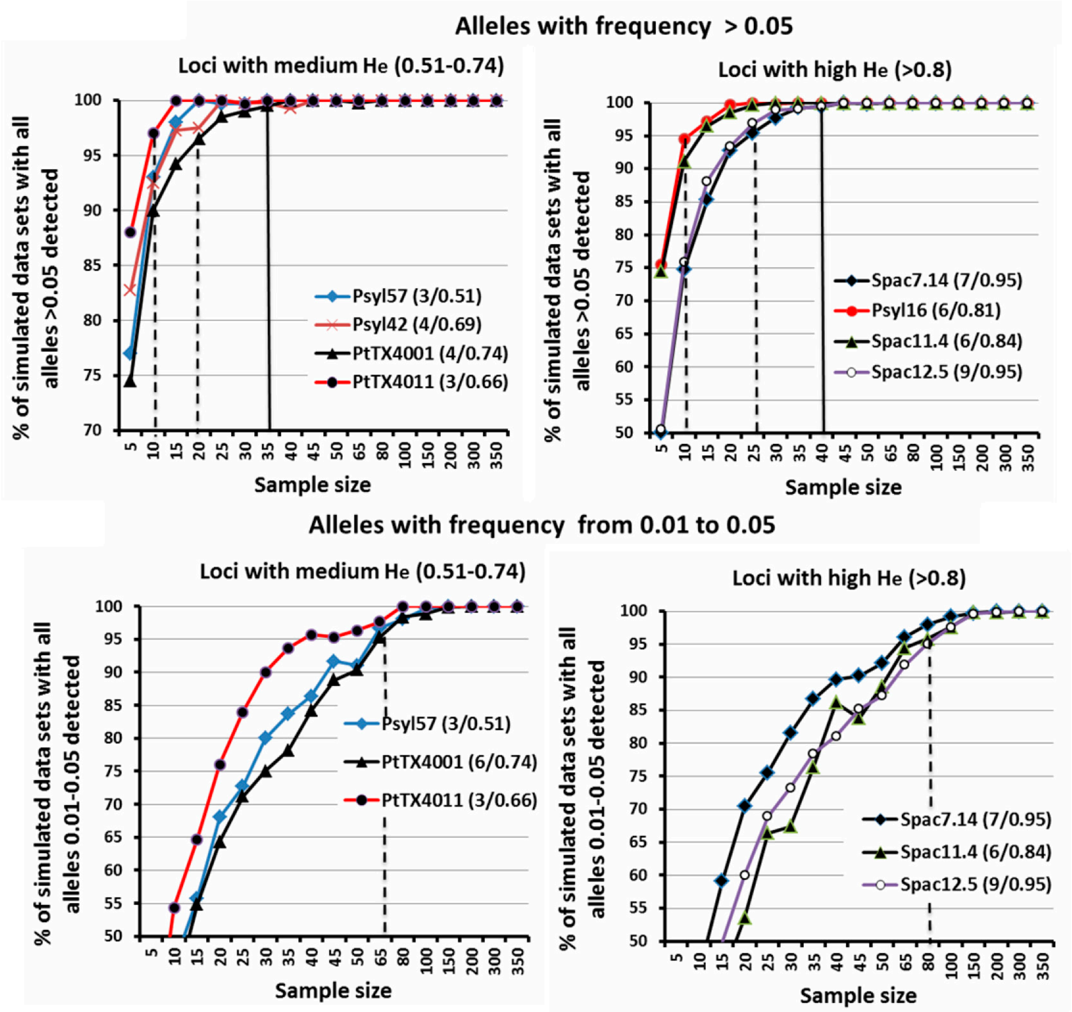

Figure 2. The estimates of the sample size required to capture the allelic variation at the loci with variable H_e scores in Scots pine stands given for frequent (frequency > 0.05, the two upper plots) and rare (frequency 0.01 to 0.05, the two lower plots) alleles. The percentage of the resampled data sets that contained all the alleles with the corresponding frequency present in the empirical data set (*Y* axis) is given for each sample size of the simulated data sets (*X* axis). The number of alleles at the respective locus at a frequency of >0.05 or 0.01-0.05 and the H_e value are indicated at each locus in the legend. The solid vertical lines indicate the sample sizes for which all the loci data sets contained the 100 % of the alleles in the empirical data sets. The dashed lines indicate the lower and upper margins for sample sizes for which 95 % of the sample data sets contained the frequency alleles of the two frequency classes. Loci with less than 2 alleles at a corresponding frequency class are not shown

ly (Figure 2). For rare alleles at frequencies below 0.01, the required sample sizes to achieve 95 % probability of sampling success reached 300 to 350 individuals, depending on the loci (Table 2).

In agreement with the above, the relationship between the allele frequency and the sample size required to capture the alleles at the corresponding frequency with the 95% probability indicated the sample size of 30 as sufficient to capture the alleles at the frequencies above 0.05 (Figure 3). It also summarizes the results with all the loci pooled and provides the guidelines for sample size thresholds, e.g. alleles at frequencies over 0.25 could be captured with sample sizes of 5 (Figure 3). Reducing sample size to 20 individuals would provide 95% probability of sampling alleles at frequencies over 0.1 (Figure 3).

Accuracy of sampling at variable sample size

The mean estimates of the unbiased expected heterozygosity (uH_e) and the observed heterozygosity (H_o) of the resampled subsets were independent on the sample size: as low sample sizes as 10 individuals returned similar uH_e and H_o values as sample sizes of 200-350 (Figure 4). This observation was true for both medium and high H_e loci (Figure 4). For high H_e loci, the estimates of H_o and uH_e varied less among the resampled subsets than for medium H_e loci (standard deviations in Figure 5). The standard de-

Table 2. Locus mean number of the resampled subsets containing all the alleles at frequencies below 0.01 present in the empirical data set. The numbers in the table heading are sample sizes of the resampled data sets (5 to 350). The loci were grouped by the H_e values. The allele number and mean allele frequency are indicated at each locus name in the leftmost column. The sample size for which over 95 % of the simulated data sets contained the corresponding allele is considered as statistically sufficient to capture the alleles at the corresponding frequencies

Locus (allele number/avg. freq.)	Sample size of the simulated data sets																
	5	10	15	20	25	30	35	40	45	50	65	80	100	150	200	300	350
	Number of the resampled subsets (low H_e loci <0.50)																
Psyl44 (4/0.004)	3	4	11	11	16	21	22	23	29	30	35	47	50	66	75	89	98
Psyl18 (6/0.003)	3	4	9	11	15	16	20	21	24	24	32	41	45	61	69	90	96
Psyl2 (3/ 0.003)	3	7	8	14	14	15	21	25	22	28	35	34	51	60	70	89	95
	Number of the resampled subsets (medium H_e loci 0.50-0.80)																
Psyl57 (2/0.001)	1	0	3	5	8	11	7	8	7	14	22	20	24	41	48	72	93
PtTX4011 (2/0.003)	6	5	8	8	8	15	23	19	24	23	27	35	38	55	66	91	93
Psyl42 (1/0.008)	8	17	23	21	37	44	47	52	56	52	64	74	85	94	100	100	100
PtTX4001 (8/0.004)	3	7	10	14	16	18	22	24	28	32	39	42	48	62	74	88	94
	Number of the resampled subsets (high H_e loci >0.80)																
Psyl16 (6/0.003)	2	5	8	8	13	13	16	19	21	24	30	35	41	56	73	89	97
Spac11.4 (6/0.003)	3	5	8	11	9	16	18	17	23	25	26	36	43	58	68	87	93
Spac12.5 (12/0.005)	4	11	13	18	23	25	30	34	38	39	49	54	61	78	86	95	98
Spac7.14 (15/0.004)	4	8	12	16	19	25	26	29	33	36	42	50	57	72	83	94	97

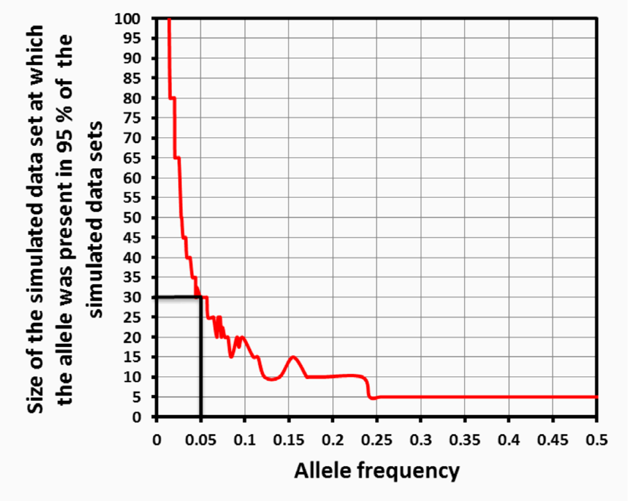

Figure 3. Relationship between the allele frequency and the size of the resampled data sets at which the alleles of that frequency were present in 95 % of the simulated data sets. The relationship provides estimates of the sample size required to capture alleles occurring at variable frequencies. The solid black line point at sample size needed to sample alleles at the frequency of 0.05

viations of the uH_e estimates stabilized at sample sizes exceeding 15-20 and 100 individuals for the high and medium H_e loci, respectively (Figure 4). The mean estimated F_{IS} values (from the resampled subsets) increased markedly with increasing sample size. F_{IS} stabilized at sample sizes above 20 and varied little above a sample size of 35 (Figure 4). The estimates of F_{IS} were more accurate for the loci with high than with medium H_e (lower standard errors in Figure 4). For the effective number of alleles at medium H_e loci, the increase of sample sizes beyond 15 had little effect on the estimates of effective number of alleles (N_e). However, for high H_e loci, N_e varied markedly over all sample sizes, though for sample sizes of more than 25-30, the increase was less drastic (Figure 4).

The pairwise F_{ST} values between each resampled data set of a variable sample size and the empirical data set (all loci) were decreasing with increasing sample size, but the decrease was much less below sample sizes of 25 individuals (Figure 5). Below the sample size of 25, the F_{ST} values were lower than 0.005 (Figure 5). The standard deviations of the F_{ST} mean values in the resampled data sets were below 0.001 and 0.0004 for sample sizes over 25 and 50 individuals, respectively (Figure 5). Note, however, this decrease of the standard deviations of the F_{ST} means depends on the size of the original data set so that with large data sets the decrease is less drastic.

The ANOVA-based pairwise Tukey LSD test of the transformed pairwise F_{ST} values (between each replicate and the empirical data, 100 F_{ST} values for each sample size) between the different sample sizes revealed no significant differences in the pairwise F_{ST} values among the sample sizes over 25 (not shown). Below the sample size of 25, the pairwise F_{ST} values of the sample sizes were significantly different between each other and all the remaining sample sizes, except that sample sizes of 15 and 20 were not significantly different from each other.

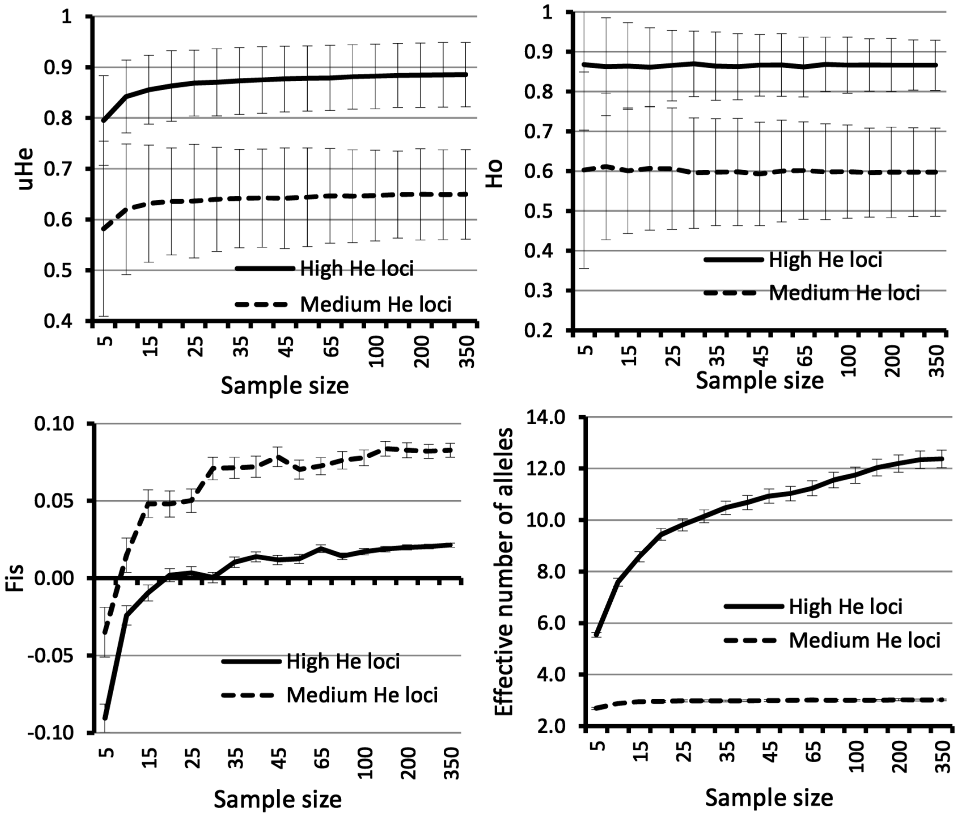

Figure 4. Effect of variable sample size on the accuracy of estimates of expected (H_e) and observed (H_o) heterozygosity (the two upper plots), the inbreeding coefficient (F_{IS}, lower left) and the effective number of alleles (N_e, lower right). For the heterozygosity, the error bars indicate standard error

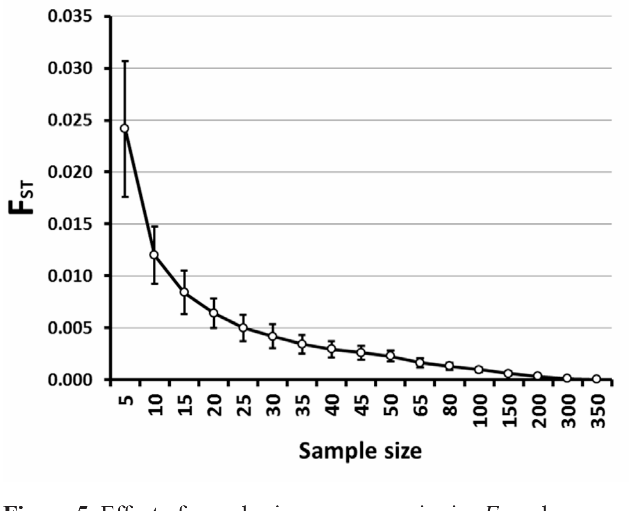

Figure 5. Effect of sample size on mean pairwise F_{ST} value between the resampled and the empirical data set with all the alleles included. The error bars indicate standard deviation

Discussion

The main finding of our simulations is that the nuclear SSR-based studies addressing such problems as population structure robust to the effects of rare alleles may plan for sample sizes 20-25 from a single Scots pine stand (alleles over 0.05 captured with 95% probability, Figure 2). The upper limits of this interval applies to highly polymorphic SSR loci representing perhaps the upper limit of the information obtainable for a SSR locus (e.g. Spac7.14 with 36 alleles, H_e = 0.95). Obviously, the loci with several dominant alleles (lower H_e) require relatively lower sample sizes. Hale et al. (2012) reported 25-30 individuals as representative sample size from a similar sample size optimization study based on resampling from empirical animal genomic SSR data sets at frequencies over 0.05 with a 95% probability. In comparison with our study, Hale et al. (2012) used markedly smaller empirical data sets (100 individuals) representing several populations that could have introduced a small bias requiring relatively large sample size of 30-35. The comparably large size of the empirical data set is a strong point of our study markedly exceeding the empirical data set sizes in similar resampling-based optimization studies (200 in Gapare et al. 2008, 100 to 180 in Bashalkanov et al. 2009). Another factor that may help choosing lower or upper bound of the sample sizes found by us is the discriminating power of particular loci. Kalinowski (2005) showed that loci with high F_{ST} values require small sample sizes, e.g. 20 if F_{ST} is over 0.05. This is because strong differentiation occurs at the cost of more frequent alleles and there is no need for large sample sizes to capture alleles with frequencies above 0.05 as shown by us. The sample sizes of 20-25 could be attractive to range-wide studies of geographical patters of DNA polymorphism requiring sampling of many populations, where the SSR loci usually display high differentiation (e.g. Buchovska et al. 2012).

In support to Hale et al. (2012), we consider the statically well-defined 95 % or 100 % probability thresholds for minimum sample sizes as an advantage of the approach based on resampling. These thresholds allow avoiding a subjective judgement for the optima according to the increase in precision of a parameter estimate with increasing sample size, rarely having a clear threshold (e.g. allelic richness in Bashalkanov et al. 2009). However, how robust are these representative sample sizes of 20 to 25 for sampling the informative alleles? The analysis pooled over loci of relationship between the allele frequency and sample size required to capture these alleles at 95 % probability indicates a similar sample size of 30 is needed to sample the alleles with frequency over 0.05 (Figure 3). These numbers are fairly consistent with probability theory indicating that number of diploid individuals needed to capture the alleles at frequency of 0.05 is equal to $1-(0.95)^{nm}$ (where n is the number of individuals and m is the ploidy level), that equals to 30 (cf. Hale et al. 2012). With increasing sample size, the reduction of the pairwise F_{ST} values between the resampled subsets (averaged over sample size and containing all the alleles) and the empirical data set is much less drastic beyond sample sizes of 25-30 ($F_{ST} < 0.005$) (Figure 4). The later estimate of 25-30 was lower than 50 obtained for the corresponding F_{ST} values after 50 replicate resampling from the empirical SSR data of *Picea rubens* and *Pinus strobus* of size 100 and 180, respectively (Bashalkanov et al. 2009). This indicates that the optimizations using a low sampling replicate number and small empirical data sets tend to overestimate sample sizes based on the differentiation between the resampled and the empirical data sets. Accuracy of estimating effective number of alleles (N_e) points at sample size of 20 for medium H_e loci and 25-30 for high H_e loci as possible optima (Figure 5). For the estimation of F_{IS}, the cumulative increase in F_{IS} with increasing sample size is also much lower above sample sizes of 20-25 (Figure 4). The variation among the resampled data sets in the estimated of H_e also stabilizes above the sample sizes of 15 (high H_e loci) and 35 (medium H_e loci) (Figure 4). A weak point of our study is absence of replications, i.e. we used data from one stand (not several). However, the stand sampled in our study is a typical self-regenerated Scots pine stand on optimum site type for Scots pine located in the middle of a large forest tract and managed as all commercial Scots pine stands in temperate and boreal zone. This provides a strong likelihood for the stand studied to represent the population of such stands and low risk to obtain significantly different SSR allele frequencies sampled in another representative Scots pine stand.

Much larger sample sizes of 65 to 80 are needed to sample rare alleles at frequencies 0.01-0.05 that were fairly abundant in the high H_e loci (95 % sampling probability, Figure 1). Sampling all the alleles would require as large samples as 250-300 (lower right, Figure 5, Table 2) that is in agreement with a more thorough estimation of sample size for SSR-based allelic richness in ash (Miyamoto et al. 2008) or trout (Banks et al. 2000). Such high sampling intensity may be required for gene conservation aiming at preservation of private alleles usually occurring at low frequency (Yanchuk 2001, Kalinowski 2004, 2005). Rare alleles are useful for parentage and relatedness analysis, where sample sizes also have to exceed 200 (Cavers et al. 2005).

The H_e and H_o estimates are insensitive to sample size as shown by many similar studies (Bashalkanov et al. 2009, Rajora et al. 2000, Buchert et al. 1997). This is attributable to the nature of the genetic material such as highly outbreed and diverse forest trees with large population sizes: if a locus exhibits high heterozygosity, it will be similarly reflected by 10 and by 200 individuals.

In conclusion, our results indicate that sample sizes of 20 to 25 individuals per natural stand of Scots pine is a safe margin for sampling the alleles at frequencies above 0.05 considered as informative for genetic structure studies in Scots pine. The upper boundary should be used if the loci are more variable with high H_e (> 0.8) with a lower discriminating power among populations. If the intention is to capture rare alleles the recommended sample sizes rise to 65-80 (alleles at frequencies 0.01-0.05) or 300-350 (alleles at frequencies below 0.01).

Acknowledgements

The study was carried out within the framework of the national project No VP1-3.1-ŠMM-08-K-01-025 entitled "Specific, genetic diversity and sustainable development of Scots pine forest to mitigate the negative effects of increased human pressure and climate change" supported by the EU Social Fund. Furthermore, the study was supported by the seventh framework infrastructure project No. 284181 "Trees for Future-Designing trees for the future" within a transnational access activity.

References

Banks, M. A., Rashbrook, V. K., Calavetta, M. J., Dean, C. A. and Hedgecock, D. 2000. Analysis of microsatellite DNA resolves genetic structure and diversity of chinook salmon (*Oncorhynchus tshawytscha*) in California's Central Valley. *Canadian Journal of Fisheries and Aquatic Sciences* 57(5): 915-927.

Bashalkhanov, S., Pandey, M. and Rajora, O. P. 2009. A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genetics* 10: 84.

Buchert, G. P., Rajora, O. P., Hood, J. V. and Dancik, B. P. 1997. Effects of Harvesting on Genetic Diversity in Old Growth Eastern White Pine in Ontario, Canada. *Conservation Biology* 11: 747-758.

Buchovska, J., Danusevičius, D., Baniulis, D., Stanys, V., Šikšnianienė, J. B. and Kavaliauskas D. 2013. The Location of the Northern Glacial Refugium of Scots Pine Based on Mitochondrial DNA markers. *Baltic Forestry* 19 (1): 2-12.

Cavers, S., Degen, B., Caron, H., Lemes, M. R., Margis, R., Salgueiro, F., Lowe, A. J. 2005. Optimal sampling strategy for estimation of spatial genetic structure in tree populations. *Heredity* 95: 281-289.

Danusevicius, D., Kavaliauskas, D. and Fussi, B. 2015. DNA markers reveal a genetic association between the sea-side Lithuanian and Bavarian Scots pine populations. (submitted manuscript) (unpubl.).

Dumolin, S., Demesure, B. and Petit, R. J. 1995. Inheritance of chloroplast and mitochondrial genomes in pedunculate oak investigated with an efficient PCR method. *Theoretical and Applied Genetics* 91: 1253-1256.

Elsik, C. G., Minihan, V. T., Hall, S. E., Scarpa, A. M. and Williams, C. G. 2000. Low-copy microsatellite markers for *Pinus taeda* L. *Genome* 43(3): 550-555.

Gapare, W., Yanchuk, A. and Aitken, S. 2008. Optimal sampling strategies for capture of genetic diversity differ between core and peripheral populations of *Picea sitchensis* (Bong.) Carr. *Conservation Genetics* 9 (2): 411-418.

Hale, M. L., Burg, T. M. and Steeves, T. E. 2012. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PloS one* 7 (9), e45170.

Hartl, D. L., Clark, A. G. and Clark, A. G. 1997. Principles of population genetics. Sunderland: Sinauer associates. Vol. 116. 635 pp.

Kalinowski, S.T. 2002. How many alleles per locus should be used to estimate genetic distances? *Heredity* 88: 62–65.

Kalinowski, S.T. 2005. Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity* 94: 33–36.

Kalinowski, S.T. 2004. Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conservation Genetics* 5(4): 539-543.

Leberg, P. L. 2002. Estimating allelic richness: effects of sample size and bottlenecks. *Molecular Ecology* 11(11): 2445-2449.

Li, Y.C., Korol, A.B., Fahima, T and Nevo, E. 2004. Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution* 21(6): 991-1007.

Miyamoto, N., Fernández-Manjarrés, J. F., Morand-Prieur, M. E., Bertolino, P. and Frascaria-Lacoste, N. 2008. What sampling is needed for reliable estimations of genetic diversity in *Fraxinus excelsior* L. (Oleaceae)? *Annals of Forest Science* 65(4): 403p1-7.

Nybom, H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology* 13(5): 1143-1155.

Oliveira, E. J., Pádua, J.G. and Zucchi, M.I. 2006. Origin, evolution and genome distribution of microsatellites. *Genetic and Molecular Biology* 29: 294-307.

Rajora, O. P., Rahman, M. H., Buchert, G. P. and Dancik, B. P. 2000. Microsatellite DNA analysis of genetic effects of harvesting in old growth eastern white pine (*Pinus strobus*) in Ontario, Canada. *Molecular Ecology* 9(3): 339-348.

Sebastiani, F., Pinzauti, F., Kujala, S.T., Gonzalez-Martinez, S.G. and Vendramin, G.G. 2012. Novel polymorphic nuclear microsatellite markers for *Pinus sylvestris* L. *Conservation Genetic Resources* 4: 231-234.

Selkoe, K.A. and Toonen, R. J. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* 9 (5): 615-629.

Soranzo, N., Provan, J. and Powell, W. 1998. Characterization of microsatellite loci in *Pinus sylvestris* L. *Molecular Ecology* 7: 1260-1261.

Van Oosterhout, C., Weetman, D. and Hutchinson, W. F. 2006. Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes* 6(1): 255-256.

Yanchuk, A. D. 2001. A quantitative framework for breeding and conservation of forest tree genetic resources in British Columbia. *Canadian Journal of Forest Research* 31(4): 566-576.

Zane, L., Bargelloni, L. and Patarnello, T. 2002. Strategies for microsatellite isolation: a review. *Molecular Ecology* 11(1): 1-16.

Received 06 April 2016
Accepted 30 September 2016