Impact of Sampling Schemes on Demographic Inference: An Empirical Study in Two Species with Different Mating Systems and Demographic Histories

K. R. St. Onge,*,[1,2] A. E. Palmé,* S. I. Wright,[†] and M. Lascoux*,[‡]

*Department of Plant Ecology and Genetics, Uppsala University, SE-752 36 Uppsala, Sweden, [†]Department of Ecology and Evolutionary Biology and Centre for Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada, and [‡]Laboratory of Evolutionary Genomics, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai 200031, China

ABSTRACT Most species have at least some level of genetic structure. Recent simulation studies have shown that it is important to consider population structure when sampling individuals to infer past population history. The relevance of the results of these computer simulations for empirical studies, however, remains unclear. In the present study, we use DNA sequence datasets collected from two closely related species with very different histories, the selfing species *Capsella rubella* and its outcrossing relative *C. grandiflora*, to assess the impact of different sampling strategies on summary statistics and the inference of historical demography. Sampling strategy did not strongly influence the mean values of Tajima's D in either species, but it had some impact on the variance. The general conclusions about demographic history were comparable across sampling schemes even when resampled data were analyzed with approximate Bayesian computation (ABC). We used simulations to explore the effects of sampling scheme under different demographic models. We conclude that when sequences from modest numbers of loci (<60) are analyzed, the sampling strategy is generally of limited importance. The same is true under intermediate or high levels of gene flow ($4Nm > 2–10$) in models in which global expansion is combined with either local expansion or hierarchical population structure. Although we observe a less severe effect of sampling than predicted under some earlier simulation models, our results should not be seen as an encouragement to neglect this issue. In general, a good coverage of the natural range, both within and between populations, will be needed to obtain a reliable reconstruction of a species's demographic history, and in fact, the effect of sampling scheme on polymorphism patterns may itself provide important information about demographic history.

KEYWORDS
population structure
Tajima's D
frequency spectrum
Capsella

doi: 10.1534/g3.112.002410
Manuscript received February 29, 2012; accepted for publication May 10, 2012

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.112.002410/-/DC1
[1]Present address: Plant Ecophysiology, Institute of Environmental Biology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.
[2]Corresponding author: Plant Ecophysiology, Institute of Environmental Biology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.
E-mail: K.R.St.Onge@uu.nl

Since the seminal work of Sewall Wright (1931), one of the main concerns of population genetics has been the description and explanation of genetic variation within and between populations and species (Lewontin 1991). To a large extent, this is still true, and the already immense body of literature on the subject continues to grow. Many recent studies have somewhat neglected population structure, concentrating instead on inference of past demographic events, such as bottlenecks or population expansion, through advanced coalescent simulations in single populations or assuming a single population (*e.g.* Thornton and Andolfatto 2006). These inferences have often been based on "pooled" data. Typically, a rather limited number of chromosomes (at least compared with the number often employed in "classical" population genetics studies) are sampled in a few popula-

tions across the range of the species of interest and sequenced at a group of loci. When sequences are collected across the range, they are then pooled and analyzed. Generally, the analysis starts by calculating some summary of the site frequency spectrum such as Tajima's D (Tajima 1989), Fu and Li's D (Fu and Li 1993), Fay and Wu's H (Fay and Wu 2000), or of the joint frequency spectrum, such as Wright fixation indices (Wright 1965), when many populations are considered. Among these summaries of the site frequency spectrum, Tajima's D was the first to be developed and has certainly been the most popular. It belongs to a large family of "neutrality" tests that compare different estimates of the population mutation rate, $\theta = 4N_e\mu$ (Zeng *et al.* 2006), where N_e is the effective population size and μ is the mutation rate. The key idea behind these tests is that some estimators of $\theta = 4N_e\mu$ will be more sensitive than others to specific departures from the standard neutral model. Comparing two estimators of θ should then indicate the type of departure experienced by the population. For example, Tajima's D is calculated by taking the difference between $\theta\pi$, the nucleotide diversity, which is based on pairwise differences between sequences, and θ_W, which is based on the number of segregating sites, the latter being more sensitive than the former to an excess of rare variants (Tajima 1989), resulting in a negative value when there is an excess of rare variants and a positive value when there is an excess of intermediate frequency variants. Tajima's D is versatile and generally has more statistical power to detect a selective sweep than do other single tests based on the site frequency spectrum (Simonsen *et al.* 1995; Przeworski 2002), though as most other such tests, it tends to have low power when considering a single locus. Today with the coming of age of powerful coalescent simulation tools, single-summary statistics such as Tajima's D are rarely used alone to infer demographic history, but they remain important as input for approximate Bayesian computation (ABC) analysis and for the insights they provide on the underlying genealogical process.

Recent papers (Städler *et al.* 2009; Chikhi *et al.* 2010) building on earlier theoretical work by Wakeley (1999; 2001), Wakeley and Aliacar (2001), Ray *et al.* (2003), and De and Durrett (2007) drew attention to the fact that pooling of data across populations can be a problem for demographic inferences. Briefly, the question is whether one can reasonably expect to recover the true species history from a sample of sequences originating from a single deme (local sampling) or pooled from a limited number of demes across the natural range (pooled sampling) when the species is geographically and genetically structured. Simulation studies of simple models of population structure, such as the island model and the stepping-stone model, suggest that local sampling or pooled sampling tends to result in a higher frequency of intermediate frequency polymorphisms than expected in a panmictic population, leading to positive Tajima's D (De and Durrett 2007; Städler *et al.* 2009). Collection of a single sample from each population (scattered sampling) or random sampling would, on the other hand, result in a frequency spectrum more similar to expectations under the standard neutral model, with a Tajima's D around zero (De and Durrett 2007; Städler *et al.* 2009).

However, in many empirical datasets, pooled samples have a lower Tajima's D than local samples (*e.g.* Ptak and Przeworski 2002). Ptak and Przeworski (2002) observed a trend of a more negative Tajima's D as more ethnicities are included in human datasets. They interpreted this as the result of a combination of population structure and growth and confirmed with limited simulations that such a pattern can be created in a model including both population structure and expansion. Using more extensive simulations, Städler *et al.* (2009) verified that such a model can result in negative estimates of Tajima's D in scattered and pooled samples under a wide range of migration rates and that the estimate based on local samples is either positive or at least less negative ($4N_0m = 0.5$–100, where N_0 is the population size and m is the migration rate). This clearly shows that sampling can have an effect on the estimation of Tajima's D and therefore can influence the interpretation of the data. For example, a positive Tajima's D across genes is often interpreted as a sign of a population size decrease, but if local sampling in a structured population of stable size can result in positive values, it is easy to draw the wrong conclusion. This problem is not restricted to the estimation of Tajima's D; sampling can also affect the inference of population history in full Bayesian methods or ABC methods (Chikhi *et al.* 2010; Peter *et al.* 2010). Population structure, if not properly accounted for, can, for example, lead to the detection of spurious bottlenecks (Chikhi *et al.* 2010).

Simulation studies thus indicate that sampling can in some circumstances have a serious impact on the inference of population history. However, by necessity, the demographic models implemented in coalescent simulations remain fairly simple. The basic model in Städler *et al.* (2009), which we build upon in this paper, involves an ancestral population that splits instantly into I demes at time τ in the past. Individual demes all have the same constant population size through time and are connected through migration. Both an n-island model and a stepping-stone model are considered. In the equilibrium model, τ is assumed to be very large, and the total size of the subdivided population is equal to the size of the ancestral population. In the global expansion model in which range expansion is simulated by fragmenting an ancestral population instantaneously into I demes, the global population size over the demes is larger than the ancestral size, thereby simulating global expansion; however, it does not include population expansion within demes. We note that this procedure can therefore result in an initial bottleneck of each single deme; the same is true for the equilibrium model of Städler *et al.* (2009). Additionally, the study did not consider hierarchical population structure, a rather common feature of natural populations with important implications for population genetics inferences (Robertson 1975). Städler *et al.* (2009) tested the effect of the pooled and local sampling schemes on estimates of Tajima's D and Fu and Li's D from multilocus data from wild tomatoes, but unfortunately, the data were too limited to implement the scattered sampling scheme.

Although the studies by Städler *et al.* (2009) and Chikhi *et al.* (2010) are both timely and important reminders of the importance of population structure and sampling schemes and have already had an impact on the way empirical studies are conducted (*e.g.* Li *et al.* 2010; Wheat *et al.* 2010), the general relevance of their results for empirical datasets remains untested. In the present study, our first goal is to explore the impact of sampling on demographic inferences within two empirical datasets with divergent population structure and history and with a number of loci in the same range as many sequence studies on natural populations of nonmodel organisms (13–16 loci). This contrasts with the simulations of Städler *et al.* (2009), which assumed 1000 independent loci. Second, we want to compare these results with comparable simulated datasets to explore the effect of hierarchical structure and expansion within demes, which was not considered previously. The two investigated species are *Capsella rubella* and *Capsella grandiflora*, two closely related species with contrasting population histories and population structures. The selfing species *C. rubella* derived recently from *C. grandiflora* (around 13,000 to 70,000 years ago) and underwent a severe bottleneck followed by a population expansion (Foxe *et al.* 2009; Guo *et al.* 2009; St. Onge *et al.* 2011). *C. rubella* shows high values of Wright fixation index F_{ST}, and Bayesian clustering methods give many clusters with a large amount of admixture (St. Onge *et al.* 2011). In contrast, F_{ST}

Table 1 Sampling schemes used in both *C. rubella* and *C. grandiflora*

Sampling Scheme	Number of Populations	Number of Individuals per Population	Number of Individuals	Number Subsamples in Each Sampling Scheme
Local	1	12	12	4
Pooled	4	3	12	4
Scattered	12	1	12	4

values in *C. grandiflora* are low. but the species exhibits a clear population genetic structure with three well-delineated clusters showing either evidence of weak population growth or no departure from the standard neutral model (St. Onge *et al.* 2011).

MATERIALS AND METHODS

Populations and subsampled datasets

Populations of *C. grandiflora* and *C. rubella*, previously described in St. Onge *et al.* (2011), were combined with additional populations: numbers 83, 86, 87, 133, 135 in *C. grandiflora* and numbers 32, 39, 4, 6, 8, 63 in *C. rubella* (supporting information, Table S1). The total number of individuals was 79 and 95 in *C. grandiflora* and *C. rubella*, respectively. Briefly, the original populations (core populations) consisted of seven populations of each species, with 6 to 12 individuals each. Five of our *C. grandiflora* core populations originated from northwestern Greece, and two originated from the Greek island of Corfu. Six of our *C. rubella* core populations originated from southern Italy and Sicily, and one originated from northeastern Spain. Five additional populations were added to the *C. grandiflora* populations with 2 individuals each, 3 from northwestern Greece and two from southern Albania. Six additional populations, from Italy, Spain, France, and Luxembourg, were added to the *C. rubella* populations with 1 to 3 individuals each.

These datasets were subsampled into 3 different datasets corresponding to three different sampling strategies (Table 1), and this subsampling was replicated four times within each species, giving a total of 24 subsampled datasets for analysis. Following Wakeley (1999; 2001) and Städler *et al.* (2009), we defined three basic sampling schemes: local, pooled, and scattered (Figure 1). Local samples contain *n* individuals from a single deme. Pooled samples contain several individuals sampled from several demes. Finally, scattered samples include single individuals from each of *n* different demes. In each species, the populations were randomly subsampled to create 4 random samples in each of the three sampling schemes, resulting in a total of 12 subsamples in each species. Each local subsample consisted of 12 individuals sampled from the same geographic population. Due to the limit in the size of the some populations, the 4 populations with the highest number of individuals were used as the local subsamples for each species. The pooled subsamples were obtained by sampling the 12 individuals from 4 populations (three individuals from each), and finally, the scattered subsamples consisted of 12 individuals sampled in as many populations. Individuals and populations were chosen randomly for both the pooled and scattered subsamples, except for populations with fewer than 3 individuals, which were excluded from the pooled subsamples to keep equal numbers of sampled individuals per population. In total for each species, we analyzed 12 subsamples each with sequence data from 12 individuals (Table 1). Because of the high levels of inbreeding in *C. rubella*, the data from this species was treated as haploid and the same was done for *C. grandiflora* for comparison of summary statistics.

Sequence data

Sequence data were collected for 16 and 13 genes in *C. grandiflora* and *C. rubella*, respectively (Table S2). Details of PCR and sequencing of these genes can be found in St. Onge *et al.* (2011). Briefly, PCR primers were designed using the *A. thaliana* reference genome to amplify 500–700 bp fragments, including both exons and introns. The forward and reverse strands of these fragments were sequenced directly at the Macrogen sequencing facility in Korea. The sequences were aligned and edited using CodonCode Aligner v2.0.6 (CodonCode, Dedham, MA). The software PHASE 2.1 (Stephens *et al.* 2001; Stephens and Donnelly 2003), implemented in DNAsp v4.50 (Librado and Rozas 2009) was used to generate haplotypes. We also cloned and sequenced several PCR-products of each gene, except for AT3G62890 and AT4G08920, and this information was used to improve phase determination by including known haplotypes in PHASE.

Population structure

A previous analysis of genetic structure in the two species showed that 3 distinct genetic clusters are present in the seven *C. grandiflora* core populations, and that at least 5 clusters with some admixture are present in the seven *C. rubella* core populations (St. Onge *et al.* 2011). For this study, we reran the STRUCTURE analysis described in St. Onge *et al.* (2011) with the newly sequenced individuals from the additional populations. Briefly, we used the Bayesian clustering algorithm in the program STRUCTURE 2.2 (Pritchard *et al.* 2000) to investigate the probability of 1 to 12 clusters (K). For this analysis, we used the complete sequences of each sequenced locus as haplotypes, excluding singletons. For each value of K, the program was run 10 times using the admixture model with uncorrelated allele frequencies, burn-in period of 100,000, and run length of 500,000, at which stage convergence was achieved (Figure S3). To evaluate the optimal number of genetic clusters, delta K was calculated according to Evanno *et al.* (2005). We also report the LnP(D) values in File S1. Because they are not informative, singletons were excluded as well as individuals with more than 50% missing data.

To be able to compare our results with those reported in Städler *et al.* (2009), we also estimated Wright's fixation indices (Wright 1951, 1965) using an AMOVA based on pairwise differences (Excoffier *et al.*

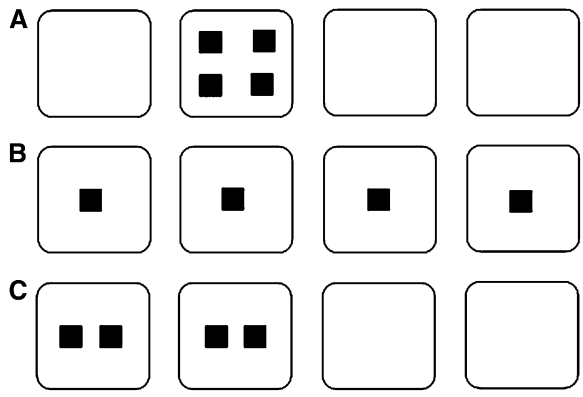

Figure 1 Schematic representation of the three sampling schemes: (A) local, (B) scattered, and (C) pooled. Boxes represent sampled populations and dots represent sampled individuals.

1992) implemented in Arlequin v3.5 (Excoffier *et al.* 2005). As there is some missing data, we followed the suggestion of the software authors and used a locus-by-locus AMOVA rather than a standard AMOVA. Significance of F_{ST} was tested by 1000 permutations of haplotypes among populations.

Summary statistics

Watterson's estimate of the population mutation rate (θ_W) (Watterson 1975), the number of segregating sites (S), the nucleotide diversity (π), Tajima's D (Tajima 1989), and Fu and Li's D (Fu and Li 1993) were calculated for all sites using the libsequence C++ library (Thornton 2003). In *C. grandiflora*, we also considered only the synonymous sites to avoid potential effects of selection. Synonymous site were also considered in *C. rubella* for comparison with the synonymous statistics of *C. grandiflora*; however, it should be noted that data are very limited in the haploid synonymous dataset of *C. rubella* due to the low variation in this species. All calculations were done separately for each of the 12 subsamples and for the whole dataset in each species. Averages and variances for each sampling scheme were calculated by first averaging each locus across the four replicates and then averaging over all loci.

ABC analysis

To infer the demographic history of species, one seldom relies only upon raw summary statistics of the frequency spectrum. Instead, one uses them within ABC (Lopes and Beaumont 2010) analysis or likelihood models. We therefore conducted an ABC analysis using the program SEQLIB v1.0 (De Mita *et al.* 2007) on all 12 subsamples of *C. rubella* and *C. grandiflora*. This analysis was conducted following St. Onge *et al.* (2011), using synonymous sites for *C. grandiflora* and total sites for *C. rubella*. Analysis was done with somewhat shorter runs; 250,000 samples were generated in both the initial and ABC steps. Because this number of samples gave similar results to our previous study using this program, 250,000 samples were considered adequate. The summary statistics used here are the nucleotide diversity (π), Watterson's estimator of the population mutation rate (θ_W), and the number of unique haplotypes (excluding singletons) (K), and these statistics were calculated for 16 loci in *C. grandiflora* and 13 in *C. rubella*. In the initial step, 5% or 10% of samples in *C. rubella* and *C. grandiflora*, respectively, were kept on which the new prior ranges were based. The model probabilities were based on the acceptance rate from the ABC step using an absolute acceptance threshold of a Euclidean distance of 0.05 or less from the real data. We tested the same demographic models tested in St. Onge *et al.* (2011): (i) a standard neutral model (SNM) that includes one parameter (θ, the population mutation rate); (ii) a population expansion model (PEM) with two parameters (θ and α, an exponential growth factor); (iii) an instant change model with three parameters [θ, the time of population size change (T), and the population mutation rate of the ancestral population (θA)]; and (iv) a bottleneck model (BNM) with four parameters [θ, time since end of bottleneck (T), duration of bottleneck (d), and size of the population during the bottleneck (f)]. See supplementary text in File S1 for further details.

Simulations

To understand the nature of the discrepancy between our results (minimal effects of sampling on summary statistics) and those of Städler *et al.* (2009) and thereby the robustness of the latter, we used computer simulations. The simulations and figures were produced with the program Sampling v0.5 (http://guanine.evolbio.mpg.de/sampling/), the program used by Städler *et al.* (2009) (Figures 4 and 5), or with slightly modified versions of it (Figure 6A, B). The model implemented in this program considers a large ancestral population, which instantaneously splits into subdivided populations at time τ with migration between them, 4Nm, and an expansion factor β. We first used Sampling with and without expansion to compare simulated results to our empirical results by assuming 20 demes and 16 loci, the number of loci used in the present study. Using Sampling without expansion, we varied the number of loci to assess the robustness of the results to the number of loci used. Then using Sampling with expansion, we drew contour plots of Tajima's D for different values of the time at which population expansion started (τ) and of the population expansion factor (β) under local and pooled samplings. Finally, the models implemented in Städler *et al.* (2009) are rather simplistic, with an instantaneous shift from an ancestral population to subdivided populations that likely exacerbates the differences between the different sampling schemes more than a progressive shift toward an n-island model would do. To test the robustness of the model to departures from the patterns seen in Städler *et al.* (2009), we made two simple modifications to the basic model implemented by Sampling. First their global expansion model does not consider growth of individual populations but only a relative increase of the global population size compared with the ancestral population. We therefore added exponential growth ($\alpha = 25$) to each subpopulation in the Wright island model with 100 demes without recombination. Twenty individuals were sampled. Second, we implemented a simple hierarchical model in which the ancestral population first splits into two main populations that then split into subpopulations, but otherwise use the same parameters as the Städler *et al.* (2009) global expansion model. Looking backward, demes 1 to 50 merge into a single deme and population 51 to 100 into another deme of the same size at time $\tau = 1$, and these two remaining demes merged into a single ancestral deme at time $\tau = 2$. At each stage, we assume a Wright island model with no recombination, and 20 individuals were sampled. The final number of subpopulations was the same as in the model of Städler *et al.* (2009), and migration was possible between the two initial subpopulations as well as among the final subpopulations. This model included global expansion but no expansion within local populations. The modified Sampling code is available from M. Lascoux.

RESULTS

Genetic diversity and population structure

Summary statistics of the total dataset for our two species are reported in Table S2. Although the present dataset is somewhat larger than that presented in St. Onge *et al.* (2011), diversity statistics are highly similar. Briefly, on average 151 chromosomes were sequenced over 12 populations of *C. grandiflora*, and 86 (haploid number) chromosomes were sequenced over 13 populations of *C. rubella*. Genetic diversity is reduced in *C. rubella* compared with *C. grandiflora* (total sites $\pi =$ 0.007 and 0.005, synonymous sites $\pi = 0.021$ and 0.013, and average number of haplotypes per locus = 38.3 and 4.23 for the total datasets of *C. grandiflora* and *C. rubella*, respectively). Tajima's D for the total datasets are −1.15 and 0.71 (−0.52 and 0.73 for synonymous) for *C. grandiflora* and *C. rubella*, respectively, suggesting population expansion in *C. grandiflora* and a population bottleneck in *C. rubella*.

The results from clustering analysis in STRUCTURE generally agreed well with our expectations based on geography and previous analysis on a subset of these populations (St. Onge *et al.* 2011). In *C. grandiflora*, the clustering analysis supports the presence of three genetic clusters and the populations that have been added cluster with

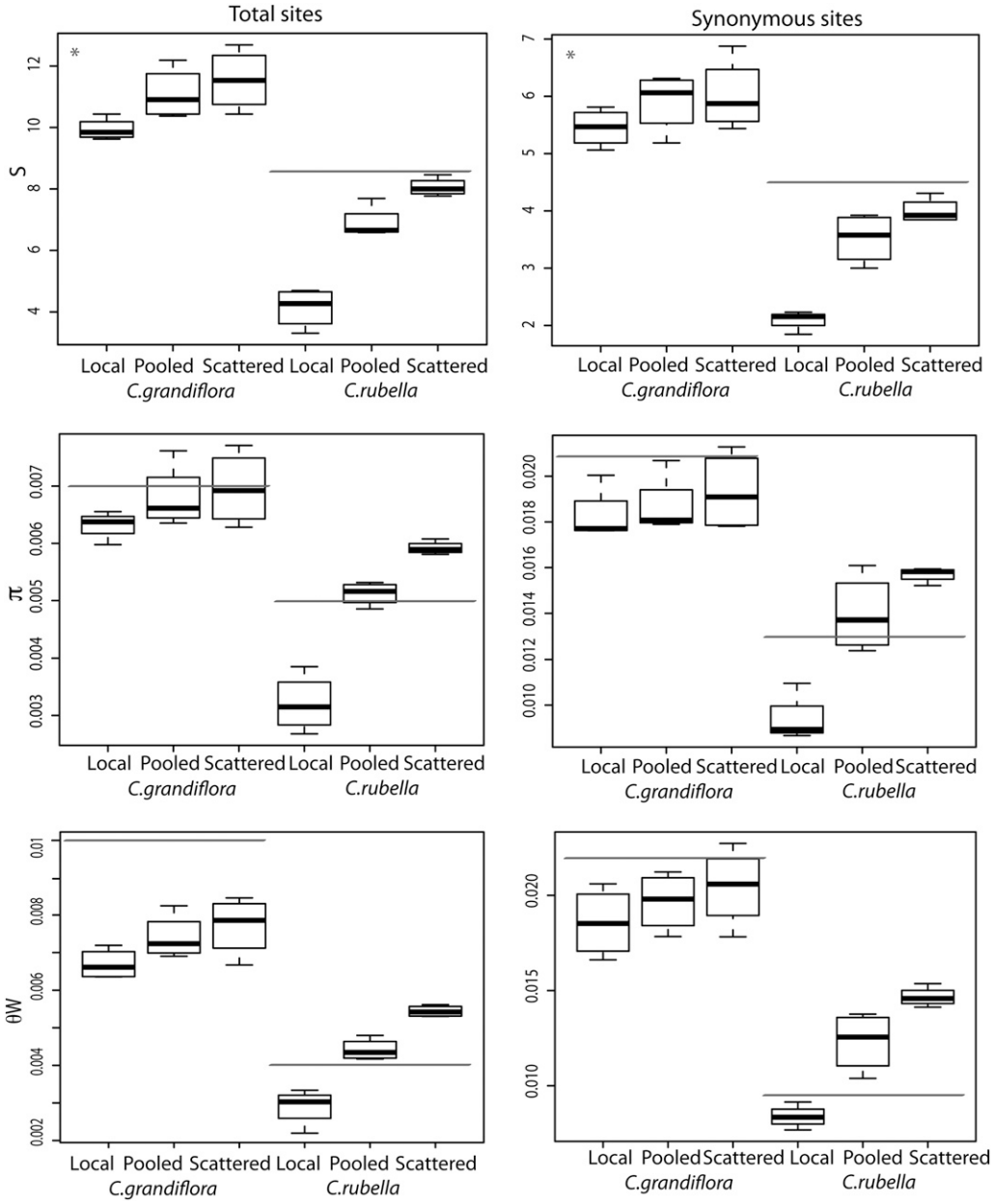

Figure 2 Box plots of summary statistics averaged over all loci for the different sampling strategies: number of segregating sites, S (top row); nucleotide diversity, π (middle row); and Watterson's estimate of the population mutation rate, θ_W (bottom row). Statistics are shown for both total site and synonymous sites. Gray lines indicate the average value for the total dataset. An asterisk (*) indicates total dataset values that lie far outside the range of the figure; these values are 30.5 and 13.7 for S in *C. grandiflora* for total sites and synonymous sites, respectively. Each box plot is calculated using n = 4, boxes range the upper and lower quartiles, the bold line indicates the median value, and whiskers indicate 1.5 times the quartile range.

one or two of the neighboring original populations (for clustering and the plot of ΔK over the number of clusters, see Figure S1 and Figure S2; for plots of LnP(D) over K, see File S1). In general, geographically close populations are inferred to belong to the same genetic clusters. However, as the sample sizes of some populations are low, we have limited power to detect additional genetic clusters containing only such populations. At values of K = 4 and greater, admixture is added (Figure S2).

As in St. Onge *et al.* (2011), the clustering analysis in *C. rubella* shows a large amount of admixture, and populations 22, 23, 49 and 50 show low levels of admixture (Figure S2). At the most likely number of genetic clusters (K = 6, Figure S2) and also for other values of K, population 6 is genetically similar to the geographically close population 50, whereas populations 3, 35, 39, and 4 show a high level of admixture.

The average F_{ST} over all loci and all populations were 0.14 and 0.54 in *C. grandiflora* and *C. rubella*, respectively. Permutation tests (1000 permutations) showed that both values are significantly different from zero ($P < 0.01$).

Summary statistics

In both species, there was an increase in the average number of segregating sites from local to pooled to scattered sampling (4.1, 6.8, and 8.0 in *C. rubella*, and 9.9, 11.1, and 11.5 in *C. grandiflora* for local, pooled, and scattered, respectively, Figure 2). A similar increase is also observed in Watterson's estimate of θ, which is a direct function of S, and for π. In *C. grandiflora*, the differences between the three sampling schemes were not very pronounced, with overlapping interquartile ranges between local and pooled or pooled and scattered, but not between the local and scattered sampling schemes (paired *t*-test, $P > 0.05$ for all test). Differences were greater between the *C. rubella* sampling types, where interquartile ranges were not overlapping between any of the sampling schemes (paired *t*-tests, $P > 0.05$ for all

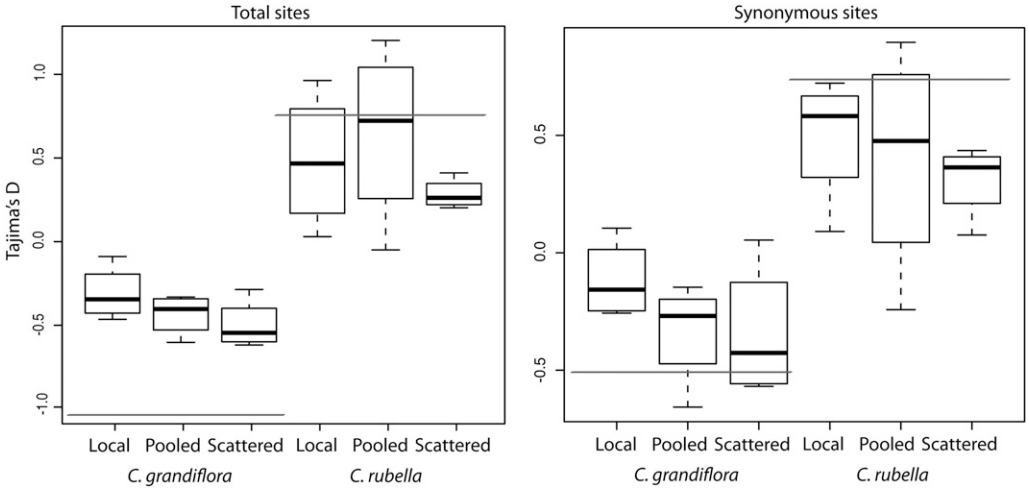

Figure 3 Box plots of Tajima's D averaged over all loci for different sampling strategies in *C. grandiflora* and *C. rubella*. Statistics are shown for both total site and synonymous sites. Gray lines indicate the average value for the total dataset. Each box plot is calculated using n = 4, boxes range the upper and lower quartiles, the bold line indicates the median value, and whiskers indicate 1.5 times the quartile range.

tests except for local sampling values of S, θ, and π, which are significantly different from those of scattered and pooled sampling).

The average number of singletons also tends to increase from local to pooled to scattered samples, although this trend is less apparent than the increasing trend of the diversity statistics discussed above. In *C. grandiflora*, the average number of singletons of the scattered samples is actually slightly lower than the estimate of the pooled sampling. No great differences are seen between sampling strategies in *C. rubella* (Figure S4). Note that, for comparison, the average number of singletons was calculated for total sites due to a lack of data in *C. rubella* synonymous dataset.

The average Tajima's D is not strongly affected by the sampling strategies in either species. In *C. rubella* average Tajima's D values are positive (0.53, 0.99, and 0.53 for local, pooled, and scattered samplings, respectively), whereas all estimates of average Tajima's D are negative (−0.36, −0.43, and −0.50 for local, pooled, and scattered samplings, respectively) in *C. grandiflora*. The distribution of Tajima's D estimates becomes narrower in the scattered sampling in *C. rubella* (Figure 3). Similar patterns were observed for Fu and Li's D (data not shown).

Also note that, because we used total diversity in these analyses, Tajima's D may be partly influenced by purifying selection on nonsynonymous sites (Slotte *et al.* 2010; Zeng and Charlesworth 2011). Indeed, although the same subtle trend of decreasing Tajima's D is apparent in *C. grandiflora* when considering just synonymous sites, the average values of Tajima's D tend to be closer to zero (Figure 3).

In comparing the summary statistics of the subsamples to the total dataset, which is a mixed dataset with many individuals sampled in some locations and few in others (Table S1), we find that for all statistics, the scattered sample of *C. grandiflora* is closest to the total dataset. Diversity is increased in the total dataset, and Tajima's D is more negative than under any of the three sampling schemes (−1.15). For *C. rubella*, the comparison is not so straightforward; the number of segregating sites is larger in the total dataset, whereas the number of singletons is smaller than in the subsamples. This seems intuitive, considering the selfing habit of the species: as we increase the number of sampled populations, we increase the amount of variation we sample, and as we increase the number of individuals sampled per population, we increase the chance of sampling the same haplotype multiple times, thereby reducing the number of singletons. It follows that the estimates of θ_W and π for the pooled samples are closest to those of the total dataset, whereas the estimates of Tajima's D in the local and scattered samples are closest to that of the total dataset (Table S2, Table S3 and Figures 2 and 3).

ABC analysis

To further investigate the robustness of the demographic inference, we conducted simulations using an Approximate Bayesian Computation (ABC, Lopes and Beaumont 2010) approach described in St. Onge *et al.* (2011) on all 12 subsampled datasets in both species. Four models were tested: (i) a standard neutral model (SNM), (ii) a population expansion model (PEM), (iii) an instant population size change model (ICM), and (iv) a bottleneck model (BNM).

In *C. rubella*, this analysis supports a recent and strong reduction in population size (ICM or BNM models) in all four subsamples of all three sampling strategies (Table 2), clearly demonstrating that the signature of the bottleneck at speciation in *C. rubella* is stronger than any effect that the sampling strategy may have. Because the acceptance rate was 0.0 for the SNM for all analyses, raw acceptance rates are presented rather than Bayes factors. From the instant change model, point estimates of effective population size, date of population size change, and the ancient population size ranged from 6000 to 26,000 individuals, from 8000 to 204,000 years ago, and from 5 to 31 times larger than the current size, respectively. Overall, these results suggest that major demographic events may simply overpower the effects of the sampling strategy. Posterior probability curves for these parameters are given in File S2.

In all 12 subsamples in *C. grandiflora*, probabilities of the three models are very similar (Table 3) and would give Bayes factors no larger than three in any subsample. This supports previous studies (Foxe *et al.* 2009; St. Onge *et al.* 2011) that find either no strong departure from the standard neutral model or evidence of weak expansion in *C. grandiflora*, and it also demonstrates that sampling strategy does not alter this conclusion. Posterior probability curves for this analysis are given in File S3.

Note that the main purpose of the ABC analysis here is to see whether different conclusions would be drawn under different sampling schemes and not to identify the model with the best fit to our data, as demography in these two species has previously been the focus of several independent studies (Foxe *et al.* 2009; Guo *et al.* 2009; St. Onge *et al.* 2011)

Simulations

Figure 4 gives the mean of simulated Tajima's D values under the three sampling schemes as a function of gene flow ($4N_0m$) and assuming an equilibrium stepping-stone model equivalent to Städler *et al.* (2009) (Figure 1A). The number of loci was equal to 16 in Figure 4A. In Figure 4B–F, the number of loci is doubled for each plot from

Table 2 Probabilities for the models tested in *C. rubella*

Model	Subsample 1	Subsample 2	Subsample 3	Subsample 4
		Local sampling		
Standard neutral model	0.0	0.0	0.0	0.0
Population expansion model	0.0	0.0	0.0	0.0
Instant size change model	0.001052	0.000004	0.000152	0.000120
Bottleneck model	0.000004	0.000004	0.0	0.0
		Pooled sampling		
Standard neutral model	0.0	0.0	0.0	0.0
Population expansion model	0.0	0.0	0.0	0.0
Instant size change model	0.000836	0.000260	0.001088	0.000056
Bottleneck model	0.0	0.0	0.0	0.000044
		Scattered sampling		
Standard neutral model	0.0	0.0	0.0	0.0
Population expansion model	0.0	0.0	0.0	0.0
Instant size change model	0.000292	0.001120	0.000436	0.000356
Bottleneck model	0.0	0.000020	0.000012	0.000096

ABC analysis (Lopes and Beaumont 2010) was used in each species on all four subsamples across all three sampling strategies. Probabilities correspond to the proportion of accepted simulations with an acceptance threshold of 0.05. Analyses are based on total sites.

panel B to panel F. Our estimate of F_{ST} is equal to 0.54 in *C. rubella*. An F_{ST} of 0.54 corresponds approximately to a value of $4Nm = 1$ under a Wright island equilibrium model [$F_{ST} = 1/(4Nm + 1)$]. However, *C. rubella* does not follow an equilibrium model and instead, like many other selfers, most likely went through bottlenecks during colonization (St. Onge *et al.* 2011). An F_{ST} of 0.54 will therefore likely be obtained for a proportionally higher number of migrants (as there is more drift within populations) than under an equilibrium model, making our estimate of migration upwardly biased. Therefore, we expect that for a species at equilibrium, the same F_{ST} value would be obtained for a lower level of migration and a 4Nm of slightly less than 1. If we then consider the 4Nm values between 0.5 and 1 in Figure 4A, we see that there is a good qualitative fit with the *C. rubella* data: the local and pooled samplings yield similar values of Tajima's D, and the scattered sampling yields a somewhat lower value although the difference is not significant. It is therefore clear that a simple model with divergence among subpopulations can explain at least part of the deviation from expectations under the standard neutral model (Tajima's $D = 0$ under neutral expectations) observed in Tajima's D in *C. rubella*. Note that the effect of sampling becomes clear only when the number of loci is approximately 100.

According to the results of Städler *et al.* (2009), it is possible to get a negative Tajima's D under some circumstances with local and pooled sampling using a model combining expansion and population structure. To explore which parameter values of the global expansion model will produce simulations with Tajima's D values similar to the observed values of Tajima's D in our *C. grandiflora* dataset, we used Sampling v.0.5 to draw contour plots of Tajima's D for different values of the time at which population expansion started (τ) and of the population expansion factor (β) under local and pooled samplings. We assumed an island model with 20 demes and 16 loci, $4Nm = 6$ and $\theta = 4$, which correspond roughly to the values per locus in *C. grandiflora*. The number of chromosomes sampled in single demes was 12, and the number of chromosome in the pooled sample was 48. The contour plots are given in Figure 5. The observed mean values of Tajima's D for *C. grandiflora* are -0.44 and -0.49 for the local and pooled sampling strategies, respectively. Such values can be obtained for relatively low values of τ (<10) and moderate values of β (>2 in pooled and >6 in local; Figure 5).

When local population growth was added to the global expansion model, negative Tajima's D values are obtained for both local and scattered sampling schemes under low migration and not only under the scattered sampling strategy (Figure 6A). Thus, if there is within-deme expansion, it can be detected by within-deme sampling, but this local expansion signal becomes obscured with pooled sampling. Conversely, this within-deme local sampling is not sufficient to detect global expansion if local samples are experiencing bottlenecking as a result of subdivision, as is the case in the Städler *et al.* (2009)

Table 3 Probabilities for the models in tested *C. grandiflora*

Model	Subsample 1	Subsample 2	Subsample 3	Subsample 4
		Local sampling		
Standard neutral model	0.023906	0.021943	0.024364	0.017462
Population expansion model	0.025716	0.021559	0.023288	0.024967
Bottleneck model	0.030185	0.034777	0.036350	0.035327
		Pooled sampling		
Standard neutral model	0.022127	0.019107	0.025184	0.018780
Population expansion model	0.021171	0.023279	0.024462	0.023276
Bottleneck model	0.021300	0.038277	0.032474	0.041330
		Scattered sampling		
Standard neutral model	0.022910	0.018927	0.030752	0.019591
Population expansion model	0.065575	0.021527	0.024750	0.024492
Bottleneck model	0.039082	0.017925	0.021514	0.022461

ABC analysis (Lopes and Beaumont 2010) was used in each species on all four subsamples across all three sampling strategies. Probabilities correspond to the proportion of accepted simulations with an acceptance threshold of 0.05. Analyses are based on synonymous sites.

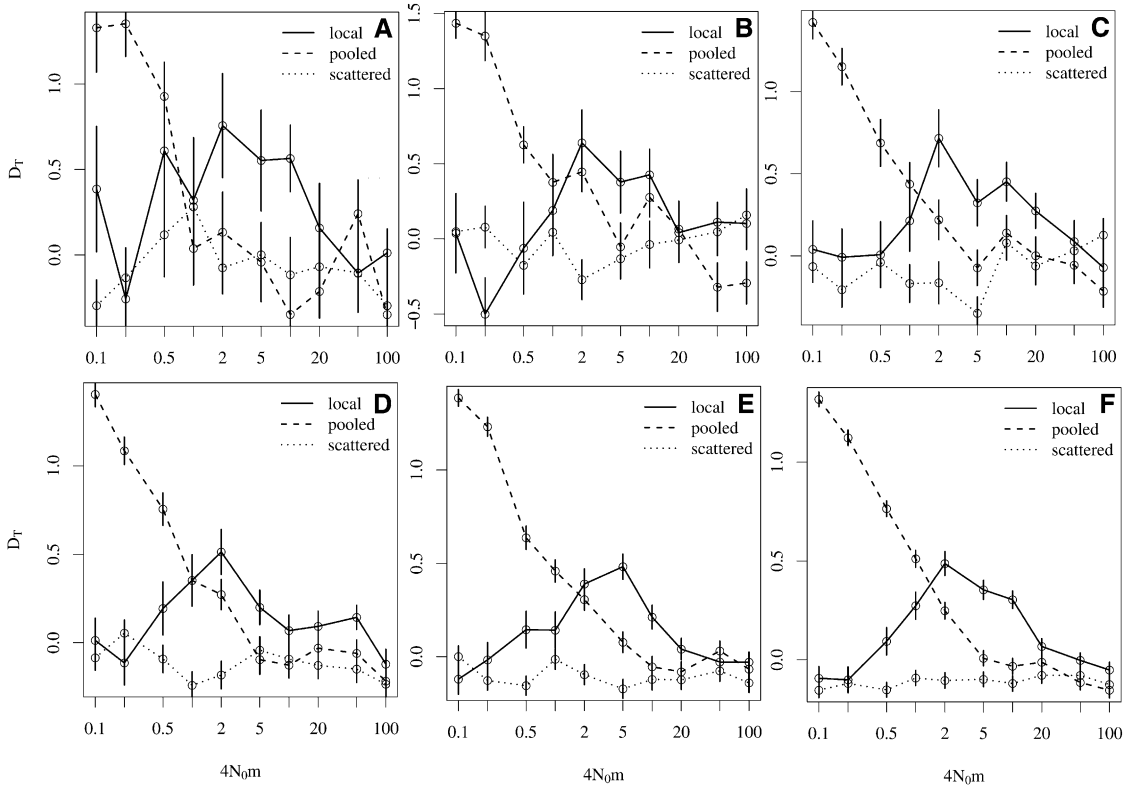

Figure 4 Tajima's D (D_T) as a function of gene flow ($4N_0m$) in an equilibrium stepping-stone model (without expansion). Simulations were run with SAMPLING v0.5 in which every plotted point is the average based on 1000 independently generated datasets and vertical lines are standard errors. In panel A, we assumed 20 demes and 16 loci, the number of loci used in the present study. Panels B–F are the same as panel A with increasing number of loci 32, 64, 128, 256, and 512. See text for further details.

expansion model. Importantly, this only holds for strongly subdivided populations, as the three sampling schemes lead to similar values of Tajima's D at intermediate and high migration rates (above $4Nm = 5$ or 10).

When a simple hierarchical structure is added to the model, the scattered and pooled sampling strategies lead to similar and positive values of Tajima's D, whereas for low migration rates, the local sampling strategy leads to values close to zero (Figure 6B). For higher levels of migration ($4Nm > 2$), again the three sampling strategies give similar results. Thus, at high or intermediate levels of migration, the estimation of Tajima's D is not largely affected by the sampling strategy if there is either expansion within subpopulations or a hierarchical structure.

DISCUSSION

In the present study, we evaluated the impact of different sampling strategies on summary statistics of the site frequency spectrum and on ABC analyses in two related species with different mating systems and demographic histories. *C. rubella* is a young species that experienced a severe bottleneck when it derived from *C. grandiflora* (Foxe *et al.*

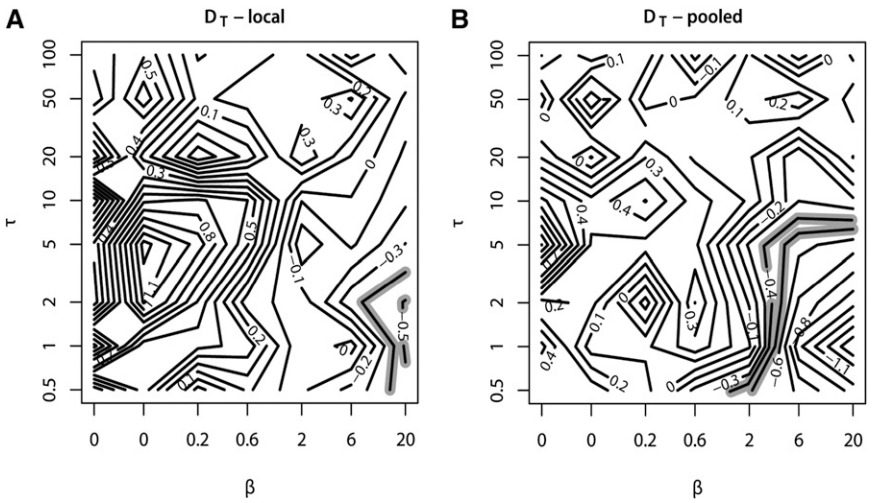

Figure 5 Contour plots showing average values of Tajima's D (D_T) obtained in simulations designed to mimic the empirical data from *C. grandiflora*. (A) Single deme samples (n = 12 sequences) and (B) pooled samples (n = 48 sequences). We assumed an island model with 20 demes, $4Nm = 6$, $\theta = 4$. The contours discussed in the text in relation to the *C. grandiflora* dataset are highlighted with gray.

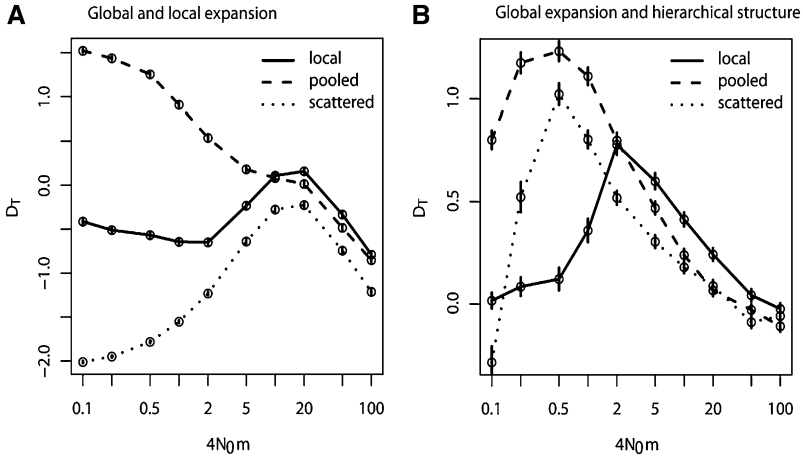

Figure 6 Averages of Tajima's D (D_T) as a function of migration rates between demes in a model including both species-wide expansion ($\beta = 10$, $\tau = 2$) and growth within demes (A) or hierarchical structure (B). As in Figure 2 in Städler *et al.* (2009), simulations were run with SAMPLING v0.5 and every plotted point is based on 1000 independently generated datasets, corresponding to 1000 unlinked loci; error bars represent standard errors. In panel A, the exponential growth rate within each deme was $\alpha = 25$.We simulated a Wright island model with 100 demes without recombination. In panel B, demes 1 to 50 merge into one deme and demes 51 to 100 into another one at time $\tau = 1$, and then the two remaining demes merge into a single ancestral deme at time $\tau = 2$. At each stage, we assume a Wright island model with no recombination.

2009; Guo *et al.* 2009; St. Onge *et al.* 2011), and it is geographically structured (St. Onge *et al.* 2011). In contrast, *C. grandiflora* experienced some limited population growth but showed no evidence of bottleneck (St. Onge *et al.* 2011). Its population genetic structure is clear, but the amount of population subdivision is low. The two species should therefore constitute a good empirical ground to test the relevance of the results presented by Städler *et al.* (2009), who showed that pooled sampling yielded values of the summary statistics that are intermediate between those of the local and scattered sampling schemes. Our interest was also in the possibility that, rather than simply viewing sampling as an "artifact," looking at the effects of different sampling schemes on polymorphism patterns may actually provide important insights into demographic history. We reached two main conclusions. First, the trends across the three sampling schemes in our results fit quite well with the predictions of Städler *et al.* (2009). Second, the effect of the sampling schemes on the summary statistics, and the conclusions drawn from them, seem limited in the range of population structure and demographic history encompassed by the two species studied here, even though the expected trends in summary statistics over sampling schemes are clear. Below we discuss both aspects, and in the last section, we attempt to give some general conclusions.

The overall trend observed in this dataset is that diversity increases from local to pooled to scattered sampling (Figure 2), even though the sample size remains the same. That the sampling of more subpopulations results in the identification of more polymorphisms and larger differences among sequences is, of course, not surprising. Per definition, population structure implies that individuals within a subpopulation are more similar to each other than to individuals in other subpopulations. This trend of increasing variation with the number of subpopulations sampled is also expected to be stronger in selfing than in outcrossing species as, at equilibrium, population structure is expected to be more pronounced in the former than in the latter and a larger part of the variation is therefore due to population differences. In concordance with this, we observe a stronger increase in diversity (S, θ_W, π) in *C. rubella* than in *C. grandiflora* (Figure 2). Both theoretical and empirical studies indicate that more singletons or rare alleles should be present in samples from many populations than from few (Ptak and Przeworski 2002; De and Durrett 2007). However, we did not observe a significant pattern in this direction for *C. rubella*, even though there is a significant trend in the other diversity statistics (Figure 2). This could perhaps be explained by the very low number of singletons in this species and a fairly large variance due to the severity of the population bottleneck.

Population structure and bottleneck explain Tajima's D in *C. rubella*

Simulations showed that a simple model with divergence among subpopulations can explain at least part of the deviation from expectations under the standard neutral model observed in Tajima's D in *C. rubella*. So, is the inference of a population bottleneck in *C. rubella* incorrect? It is quite clear from other evidence that this is not the case. Three studies that used different population genetic approaches and sampling strategies, all concluded that *C. rubella* has gone through a bottleneck (Foxe *et al.* 2009; Guo *et al.* 2009; St. Onge *et al.* 2011). However, they all used variations on the pooled sampling strategy and theoretical studies suggest that the best way to avoid, or at least minimize, the problem of sampling in a structured population is to use only a single sample from each population (Städler *et al.* 2009; Chikhi *et al.* 2010). We see in Figure 3 that even with the scattered sampling Tajima's D is different from zero. In addition, the ABC analysis with different sampling strategies led to the same conclusion strongly suggesting that the deviation from the standard neutral model is not only due to population structure.

Population expansions could explain part of the deviation in Tajima's D in *C. grandiflora*

In *C. grandiflora*, all three sampling strategies result in negative Tajima's D (Figure 3). A negative Tajima's D is not expected with any sampling strategy if the model only includes population structure (De and Durrett 2007; Städler *et al.* 2009). However, if a model includes both population structure and expansion, overall negative Tajima's D values are expected under some circumstances (Städler *et al.* 2009).

Simulations showed that relatively low values of the time since expansion τ (<10) and moderate values of the expansion factor β (>2 in pooled and >6 in local) can lead to values of Tajima's D observed in *C. grandiflora*. A $\tau = 2$, for example, corresponds to an expansion $8N_0$ generations ago, or an expansion starting about 2.5–4 Mya, assuming $N_0 = 300{,}000$-$500{,}000$ and $\beta = 6$, which corresponds to a 6-fold increase of the ancestral population size. This seems consistent with the results of St. Onge *et al.* (2011) that detected no departure from the standard neutral model or evidence of limited population expansion in the three *C. grandiflora* clusters. Also the ABC analysis (Tables 2 and 3) indicates no strong departure from the standard neutral model, as probabilities across models for each subsample are very similar and none would confer a Bayes factor above 3 (Kass and Raftery 1995), which normally leads, because of convenience, to the acceptance of the simplest model. Compared with *C. rubella*, *C. grandiflora* is clearly much closer to the standard neutral

model, and strong inferences about its demographic history are therefore more difficult to make. In such cases, it is possible that also a weak effect of sampling strategy could lead to erroneous conclusions. In our case, there is a trend toward higher probabilities for the bottleneck model under local and pooled sampling, which is absent with the scattered sampling (Tables 2 and 3). This is in line with the results of Chikhi *et al.* (2010) that population structure can result in false bottleneck signals. With higher power, the effect of sampling may overcome the weak demographic signal in *C. grandiflora*.

The effects of background selection in *C. grandiflora* are clear when comparing the synonymous and total sites Tajima's D values (Figure 3). Tajima's D in all three sampling schemes is closer to zero when considering only synonymous sites. This change is not seen in *C. rubella*; although the Tajima's D values change somewhat between the total sites data and synonymous data, the values remain near 0.5 (Figure 3). This is likely a result of the enormous difference in effective population sizes between the two species. The effective population size of *C. grandiflora* is estimated to be at least an order of magnitude higher than that of *C. rubella* (Foxe *et al.* 2009, St. Onge *et al.* 2011, 2012). Accordingly, studies of the difference in selective effects in these species have demonstrated efficient positive and purifying selection in *C. grandiflora* (Slotte *et al.* 2010) and a diminished efficiency of selection in *C. rubella* (Qiu *et al.* 2011). It is possible that part or all of the weak expansion signal we detect in *C.grandiflora* is a results of background selection on codon usage bias, which would affect synonymous sites (Qiu *et al.* 2011).

Sampling effect on Tajima's D is limited with few loci

A difference between the simulations described above and those in Städler *et al.* (2009) is the number of simulated loci. The simulations of Städler *et al.* (2009) were based on 1000 unlinked loci (independent replicates), whereas our empirical tests used only 13–16 loci, and in our simulations we varied the number of loci. We wanted to make the simulations comparable not only with our empirical data but also with other studies. Even though sequencing data are increasingly easy and inexpensive to accumulate, most studies on nonmodel species are still conducted using a limited number of loci. As a lower number of loci naturally results in lower power, what might be seen as an important significant effect at 1000 loci might not be so at 10 or 20. The lack of significant differences between the different sampling strategies observed in our study (Figure 4) and previous empirical studies may thus simply reflect the low number of loci used (Li *et al.* 2010; Wheat *et al.* 2010), suggesting that even though sampling can definitely have an impact in structured populations, it might not be crucial when the number of loci is low and will in several cases be overpowered by the effect of demography. This is demonstrated in our simulations in which we observe a large amount of variation in Tajima's D at lower numbers of loci, and the estimations of Tajima's D with different sampling strategies are largely overlapping. This pattern is found in simulations with 16 and 32 loci and, to a lesser extent, at a higher number of loci (Figure 4).

Hierarchical structure and expansion within subpopulations

We found that simple departures from the models implemented by Städler *et al.* (2009) altered the conclusions. For example, with the inclusion of expansion within demes as well as global expansion, the power to detect expansion with local samples increases and negative Tajima's D is expected at lower levels of migration ($4Nm < 5$; Figure 6A). In the simulations by Städler *et al.* (2009) that only included global expansion, Tajima's D is expected to be near zero or positive for the same levels of migration; in other words, local sampling is detecting signals only at a local level in strongly subdivided populations. In general, when we add local expansion or hierarchical structure to the global expansion model of Städler *et al.* (2009), the differences among the three different sampling strategies decrease at higher migration rates. At 4Nm above 10 in the local expansion model or 2 in the hierarchical model, the three sampling strategies result in very similar values of Tajima's D. The range of populations to which this could apply should be rather large, as a migration rate of $4Nm = 2$ or above should correspond to an F_{ST} between 0 and 0.33 at equilibrium. F_{ST}, or G_{ST}, in this range is very common among outcrossers (Hamrick and Godt 1996; Nybom and Bartish 2000; Nybom 2004).

In the same vein, we note that the many-deme model may not always be appropriate; for example, STRUCTURE analysis of *C. grandiflora* provides evidence for only three major regional clusters, with no evidence in the present case of further clustering within regions. If this is accurate, "scattered" samples combining more "populations" within a region may rather be the equivalent of increasing the sampling within populations. If we would want true scattered sampling from genetically differentiated populations in this species, we would have to be content with a very low sample size, in this case, three. Even if there are more than three genetic populations, perhaps in unsampled regions or in regions where low sample sizes limit the power to detect new genetic clusters, the number of clusters is most likely small, as we have sampled a rather large part of the geographic distribution of *C. grandiflora*. The case of *C. grandiflora* exemplifies two potentially problematic aspects of using scattered sampling in natural populations. First, without prior information, it can be easy to design a scattered sampling scheme that turns out to be a pooled sampling scheme if you consider genetic populations. Second, if you want to base your sampling scheme on population genetic structure, you need a good dense sampling to have the power to detect the population genetic structure, and only then can you pick a smaller subsample according to the scattered sampling scheme.

CONCLUSIONS

Both the empirical data, based on two datasets with markedly different population structure and demographic history, and the simulated data presented here suggest that the effect of sampling on the site frequency spectrum is limited in many cases. For instance, if populations have experienced large changes in demography, such as the recent large change in population size in *C. rubella*, the effect of demography overpowers the effect of sampling. In addition, the effect of sampling is expected to be small if migration is not too limited (F_{ST} between 0 and 0.33) and there is a hierarchical population structure or local expansion (Figure 6), which should be common in natural populations. Additionally, the number of sequenced loci is important; as our simulations show the effect of sampling is low in datasets with <60 loci (Figure 4).

However, it might not always be easy to determine if sampling strategy will have an impact, as many of the factors determining this are not known in advance. In some parts of the parameter space under the models explored here, sampling will definitely have a large effect on Tajima's D and, consequently, on inferences of population history or selection. Examples of what factors are important are not only the presence of expansion or population structure *per se*, but also the form of expansion (within deme or increase in the number of demes), the amount of population differentiation (migration rate), and type of population structure (island model/hierarchical model), and these factors are generally not prior knowledge. Thus. the results of Städler *et al.* (2009) may not invalidate the conclusions of earlier empirical

studies, but they certainly encourage caution when choosing a sampling strategy.

To conclude, even though we failed to detect an effect of sampling on the site frequency spectrum as strong as that suggested by earlier theoretical studies, we are not advocating the neglect of sampling issues. We would, however, like to caution against using only a scattered sampling strategy. Only a good coverage of the natural range, using both within- and between-population samples and a large number of loci, are likely to lead to reliable inferences of species population structure and demographic history and, thereby, to correct conclusions on adaptation and the evolutionary history of the species, especially in species having experienced recent range expansions (*e.g.* Colautti *et al.* 2010). In fact, our simulation analyses highlight that contrasting diversity patterns with different sampling strategies can be quite useful when distinguishing equilibrium population structure with gene flow from nonequilibrium patterns of divergence and population size change. Future studies should further investigate how the patterns from multiple sampling strategies might be built into coalescent approaches fitting demographic models. Also, in species that are likely to have experienced significant intensity of selection, it would be better to restrict the analysis to the sites that are less likely to have been under purifying selection, such as synonymous sites or noncoding sequences far from genes. With the falling cost of genomic surveys, it will be easier to reach those goals, even in nonmodel species.

ACKNOWLEDGMENTS

We thank Aldo Musacchio, Domenico Gargano, Salvatore Cozzolino, Santiago González Martínez, Alexis Ducousso, Myriam Heuertz, Fazia Krouchi, and Tanja Slotte for help with sampling. This work was supported by Erik Philip-Sörensens Stiftelse, Nilsson-Ehle-fonden, Carl Tryggers stiftelse, FORMAS, and the Royal Swedish Academy of Sciences. M. Lascoux thanks the Chinese Academy of Sciences and the Swedish Research Council for financial support and Li Haipeng for his hospitality in Shanghai.

LITERATURE CITED

Chikhi, L., V. C. Sousa, P. Luisi, B. Goossens, and M. A. Beaumont, 2010 The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics 186: 983–995.

Colautti, R. I., C. G. Eckert, and S. C. H. Barrett, 2010 Evolutionary constraints on adaptive evolution during range expansion in an invasive plant. Proc. Biol. Sci. 277: 1799–1806.

De, A., and R. Durrett, 2007 Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. Genetics 176: 969–981.

De Mita, S., J. Ronfort, H. I. McKhann, and C. Poncet, 2007 Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signalling in Medicago truncatula. Genetics 177: 2123–2133.

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14: 2611–2620.

Excoffier, L., P. E. Smouse, and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–491.

Excoffier, L., G. Laval, and S. Schneider, 2005 Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. 1: 47–50.

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Foxe, J. P., T. Slotte, E. A. Stahl, B. Neuffer, H. Hurka *et al.*, 2009 Recent speciation associated with the evolution of selfing in Capsella. Proc. Natl. Acad. Sci. USA 10: 5241–5245.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Guo, Y.-L., J. S. Bechsgaard, T. Slotte, B. Neuffer, M. Lascoux *et al.*, 2009 Recent speciation of Capsella rubella from Capsella grandiflora, associated with loss of self-incompatibility and an extreme bottleneck. Proc. Natl. Acad. Sci. USA 106: 5246–5251.

Hamrick, J. L., and M. J. W. Godt, 1996 Effects of life history traits on genetic diversity in plant species. Philos. T. Roy. Soc. B 351: 1291–1298.

Kass, R. E., and A. E. Raftery, 1995 J. Am. Stat. Assoc. 90: 773–795.

Lewontin, R. C., 1991 Twenty-five years ago in Genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone? Genetics 128: 657–662.

Li, Y., M. Stocks, S. Hemmilä, T. Kallman, H. Zhu *et al.*, 2010 Demographic histories of four spruce (Picea) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. Mol. Biol. Evol. 27: 1001–1014.

Librado, P., and J. Rozas, 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452.

Lopes, J. S., and M. A. Beaumont, 2010 ABC: a useful Bayesian tool for the analysis of population data. Infect. Genet. Evol. 10: 825–832.

Nybom, H., 2004 Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol. Ecol. 13: 1143–1155.

Nybom, H., and I. V. Bartish, 2000 Effects of life history traits and sampling strategies on genetic diversity estimates obtained with RAPD markers in plants. Perspect. Plant Ecol. Evol. Syst. 3: 93–114.

Peter, B. M., D. Wegmann, and L. Excoffier, 2010 Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Mol. Ecol. 19: 4648–4660.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics 160: 1179–1189.

Ptak, S. E., and M. Przeworski, 2002 Evidence for population growth in humans is confounded by fine-scale population structure. Trends Genet. 18: 559–563.

Qiu, S., K. Zeng, T. Slotte, S. I. Wright, and D. Charlesworth, 2011 Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and Capsella species. Genome Biol. Evol. 3: 868–880.

Ray, N., M. Currat, and L. Excoffier, 2003 Intra-Deme molecular diversity in spatially expanding populations. Mol. Biol. Evol. 20: 76–86.

Robertson, A., 1975 Gene frequency distributions as a test of selective neutrality. Genetics 81: 775–785.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413–429.

Slotte, T., J. P. Foxe, K. M. Hazzouri, and S. I. Wright, 2010 Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size. Mol. Biol. Evol. 27: 1813–1821.

St. Onge, K., T. Källman, T. Slotte, M. Lascoux, and A. E. Palmé, 2011 Contrasting demographic history and population structure in Capsella rubella and Capsella grandiflora, two closely related species with different mating systems. Mol. Ecol. 20: 3306–3320.

St. Onge, K., J. P. Foxe, L. Junrui, L. Haipeng, K. Holm *et al.*, 2012 Coalescent-based analysis distinguishes between allo- and autopolyploid origin in shepherd's purse (Capsella bursa-pastoris). Mol. Biol. Evol. 10.1093/molbev/mss024.

Städler, T., B. Haubold, C. Merino, W. Stephan, and P. Pfaffelhuber, 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182: 205–216.

Stephens, M., and P. Donnelly, 2003 A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. 73: 1162–1169.

Stephens, M., N. J. Smith, and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68: 978–989.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
Thornton, K., 2003 Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19: 2325–2327.
Thornton, K., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics 172: 1607–1619.
Wakeley, J., 1999 Nonequilibrium migration in human history. Genetics 153: 1863–1871.
Wakeley, J., 2001 The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. 59: 133–144.
Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. Genetics 159: 893–905.
Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.
Wheat, C. W., C. R. Haag, J. H. Marden, I. Hanski, and M. J. Frilander, 2010 Nucleotide polymorphism at a gene (Pgi) under balancing selection in a butterfly metapopulation. Mol. Biol. Evol. 27: 267–281.
Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.
Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354.
Wright, S., 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19: 395–420.
Zeng, K., and B. Charlesworth, 2011 The joint effects of background selection and genetic recombination on local gene genealogies. Genetics 189: 251–266.
Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431–1439.

Communicating editor: Y. Kim