Conserv Genet (2010) 11:709–720
DOI 10.1007/s10592-009-9842-z

RESEARCH ARTICLE

General G_{ST} and θ inflation due to biased intra-population sampling, and its consequences for the conservation of the Canarian Flora

Juli Caujapé-Castells

Received: 12 February 2008 / Accepted: 3 February 2009 / Published online: 20 February 2009

Abstract While knowledge of the degree of inter-population genetic differentiation underlies the understanding of microevolutionary processes in any organism, its calculation through G_{ST}, F_{ST}, or θ (which, unlike the previous two, was designed to correct for unequal and small sample sizes) is often based in severely restricted intra-population samples, which are nonetheless tacitly assumed adequate to their accurate estimation. Empirical assessment of the influence of the number and intra-population distribution of samples on the values of G_{ST} and θ for several Canarian endemic plants compellingly suggests that (1) contrary to expectations based on simulated datasets, θ does not account for empirical sampling bias better than G_{ST}; (2) sample sizes being equal, collections scattered across each population's occupancy area entail significantly lower over-estimates of G_{ST} and θ than if they only consider one of the population extremes, especially in narrow allogamous taxa with small populations; (3) in small samples, a scattered sampling strategy is significantly less sensitive to G_{ST} inflation than sampling in one of the population extremes; and (4) a software-related component of bias should be considered when pooling values of G_{ST} from different studies to calculate averages. Thus, unlike the sampling methods used for many plant endemics from the Canaries and other regions, collections for a reliable estimation of inter-population differentiation using molecular markers should encompass the whole occupancy area of each population, and include a higher proportion of individuals respect to the total size in narrow endemics than in widespread congeners. Critically, the high average allozyme inter-population differentiation reported for the Canarian endemic Flora is possibly an over-estimate, and could be explained predominantly by the generally biased intra-population sampling associated with G_{ST} estimates, rather than by specific factors of insularity that restrict gene flow radically, as it has been hitherto assumed.

Keywords G_{ST} and θ inflation · Sample sizes · Intra-population sampling · Canary islands · Conservation

Introduction

Because most significant evolutionary changes are driven by surges in the degree of genetic connectivity among populations, estimating the apportionment of genetic variation is important from a number of perspectives that encompass different fields of comparative biology, particularly microevolution and conservation genetics.

Until now, G_{ST} (Nei 1973), F_{ST} (Wright 1951), or θ (Weir and Cockerham 1984) are the most frequently used parameters to estimate the proportion of genetic diversity that resides among populations. The values of these statistics usually range from 0 to 1, although θ can take negative values sometimes (Weir 1996). Whilst G_{ST} is equivalent to Wright's (1951) F_{ST} in di-allelic loci, it equates to the weighted average of F_{ST} for all alleles in the case of multiple alleles (Nei 1973). Unlike these two parameters, θ was designed to correct for small and unequal sample sizes (Weir and Cockerham 1984).

Both G_{ST} and (to a lesser extent) its analog F_{ST}, has been the metric of choice to quantify genetic subdivision in most Canarian plant endemics assessed with allozymes, the molecular population genetic technique with the largest

---

J. Caujapé-Castells (✉)
Departamento de Biodiversidad Molecular y Banco de DNA, Jardín Botánico Canario "Viera y Clavijo", Apartado de correos 14 de Tafira Alta, 35017 Las Palmas de Gran Canaria, Spain
e-mail: julicaujape@gmail.com

database available for this archipelago's Flora in the scientific literature. Although there are discrepancies on how to (1) calculate G_{ST} (Nei and Chesser 1983; Weir and Cockerham 1984; Cockerham and Weir 1986; Nei 1986), (2) use it in comparisons across studies (Hamrick and Godt 1989; Culley et al. 2002) or (3) translate its values into a quantitative indicator of gene flow in terms of number of immigrants per generation (Slatkin 1987; Whitlock and McCauley 1999), it is generally accepted that lower values of G_{ST} indicate higher levels of gene flow among the populations compared and, therefore, a tighter genetic cohesion among them. Disruption of gene flow beyond a given level (Slatkin 1987) is thought to trigger independent evolution of the population groups, eventually leading to speciation or extinction. Since the absolute values of G_{ST} are related to the number of populations (n) to capture a given percentage (P) of genetic variation for ex situ conservation by the expression $P = 1 - G_{ST}^{n}$ (Hamrick et al. 1991), this parameter is also being incorporated with increasing frequency in the theory and practice of biodiversity management.

The latest review for allozyme data of Canarian endemic plants (Francisco-Ortega et al. 2000) reported an average G_{ST} value of 0.281 for a pool of 23 species endemic to the archipelago, a value somewhat lower than the average $G_{ST} = 0.338$ found in the Juan Fernández endemics (Crawford et al. 2001), but still much higher than the average $G_{ST} = 0.179$ estimated by Hamrick and Godt (1997) for endemic outcrossing plant species in general. One of the basic biological entailments of these estimates is that inter-population genetic connectivity of endemic plant species from the Canaries and the Juan Fernández must be extremely fragile. As already highlighted by Francisco-Ortega et al. (2000), if a large proportion of the diversity resides among populations, then more of them must be represented in germplasm collections than if each population contains much of the species-level diversity (Hamrick et al. 1991; Hamrick and Godt 1997). In addition, the higher the G_{ST} values, the lower the levels of inter-population gene flow, which thereby poses the risk that an enhanced action of genetic drift and/or inbreeding decreases population fitness. Both these consequences would stress the need of establishing radical measures for the conservation of the endemic Canarian Flora. Furthermore, the result of higher G_{ST} values in the Canaries and the Juan Fernández is intriguing, as it intimates that there might be outstanding (though so far unspecified) factors of oceanic insularity in these volcanic archipielagos that are not so conspicuous, or simply do not happen, in other similar enclaves.

While attempts to explain the intra-and inter-population differentiation levels reported for Canarian plant endemics by combinations of biotic and a-biotic variables have been unsuccessful (Francisco-Ortega et al. 2000), the possible effect of the different size and intra-population distribution of samples on G_{ST} values has never been examined empirically. Consideration of this factor is relevant, because many of the investigations that make up the present allozyme database for Canarian endemic plants are based in severely restricted samplings of the wild populations involved, which are however tacitly assumed adequate for the estimation of inter-population differentiation with this parameter. While the abrupt Canarian topography often prevents access to parts of a population, this is not the case in all the endemic plants from this archipelago assessed with allozymes as, in many of them, populations are amenable to a more thorough spatial sampling.

Furthermore, neither Nei's (1973) G_{ST}, nor its analog Wright's (1951) F_{ST} correct for sampling error, so their application seems only fully justified when sampling size is close to the total population census. Critically, Weir and Cockerham's (1984) θ, which should be more suitable when sample sizes are small and unequal, has never been used to quantify the apportionment of allozyme genetic variation in Canarian plant endemics.

Assessing the eventual impact of intra-population sampling methods on G_{ST} and θ values is also in order from the general standpoint of nature management, because many papers using allozymes and other molecular techniques offer conservation guidelines based on their absolute values and, with increasing frequency, biodiversity policy makers comission "urgent" characterisations of population differentiation (often based on very restricted samplings compelled by the afore-quoted urgency) that use these and other parameters to substantiate conservation decissions.

The two main objectives of this paper are (1) to assess empirically the possible effect of the number and spatial distribution of samples on G_{ST} vs. θ values as estimated from allozyme data for the genera *Lotus* (Fabaceae) (Oliva-Tejera et al. 2005, 2006) and *Crambe* (Brassicaceae) (Soto et al. 2007a, b), and (2) to explore how the eventual variations affect congeneric species with narrow and restricted distributions, based on the pattern of geographic occurrence of the model taxa. The insights to these questions will be used to stress some general guidelines on sampling methods for the estimation of the apportionment of genetic variation in nature.

Methods

Data

The two *Lotus* databases used consist of (1) 17 loci for nine populations (359 individuals) that represent the Gran

Canarian pine forest endemics *L. holosericeus* (six populations), *L. genistoides* (one population) and *L. spartioides* (two populations) discussed in Oliva-Tejera et al. (2005), and (2) 11 loci for the easternmost islands endemic *L. lancerottensis*, examined in 219 individuals that represent ten populations (five from Fuerteventura and five from Lanzarote) and were discussed in Oliva-Tejera et al. (2006). The *Crambe* database consists of 12 loci for 12 populations (726 individuals) that represent the narrowly distributed *C. tamadabensis* (four populations, 191 individuals) and the more widespread *C. pritzellii* (eight populations, 535 individuals). The genetic variation of these two closely related taxa endemic to Gran Canaria has been discussed by Soto et al. (2007a, b). The cases of *Lotus* and *Crambe* have been chosen because all the populations investigated were sampled exhaustively and through their total occupancy areas, following Caujapé-Castells (2006). Moreover, they represent the pattern of high allozyme genetic variation levels featured by Canarian endemic plants, which cannot be a consequence of reduced sample sizes. In all matrices, missing genotypes were well under 10% of the total number of genotypes. The geographic distribution of all the populations considered is depicted in Fig. 1. The genotype matrices for *L. lancerottensis* can be obtained from the web resource http://www.exegen.org/sage/index.php.

Data analysis

For each of these general biological cases, the estimates of Nei's (1973) G_{ST} and of Weir and Cockerham's (1984) θ based on the total samples were compared with those associated with five sub-samples resulting from removing 10, 20, 50, 60 and 80% of the total individuals analysed for each population involved (see Appendix 1 for details on population sample and sub-sample sizes). To examine the effect of the intra-population distribution of samples on the values of G_{ST} and θ, each of these five sub-samplings was carried out in all the 31 populations included in this study considering three different spatial situations (see Fig. 1, below): removing the individuals from only one of the population extremes (extreme-1), removing the individuals from only the other population extreme (extreme-2), and scattering the removed individuals of each sub-sample across the whole population occupancy area (scatter). Calculations were carried out through the five considered sub-sampling removal proportions for each of these three spatial situations in (1) *L. lancerottensis sensu lato*, (2) *L. lancerottensis* from Fuerteventura, (3) *L. lancerottensis* from Lanzarote, (4) the Gran Canarian pine forest endemics *L. holosericeus*, *L. genistoides* and *L. spartioides* pooled, (5) *L. holosericeus* only, (6) *L. spartioides* only, (7) *Crambe tamadabensis* + *C. pritzelii*, (8) *C. pritzelii* only, and (9) *C. tamadabensis* only. Apart from assessing general aspects of the estimation of population subdivision bearing on sample representativity and distribution, the pattern of geographic occurrence of these species will allow to discuss the possible impact of these factors in the cases of (a) restricted versus more widespread congeneric endemics (*L. holosericeus* vs. *L. spartioides* and *C. tamadabensis* vs. *C. pritzelii*) and of (b) species distributed in a single island (*L. holosericeus* and *L. spartioides*, both from Gran

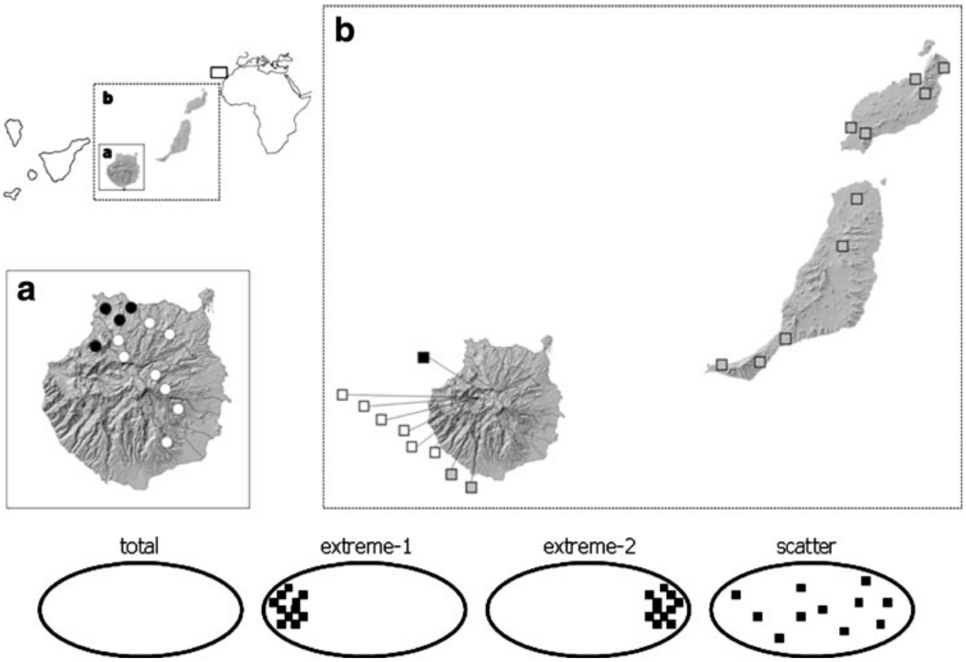


Fig. 1 *Above*: distribution maps of the Canarian endemic populations considered for the sub-sampling experiments. **a** *Crambe pritzelii* (*open circles*), and *C. tamadabensis* (*black circles*); **b** *Lotus genistoides* (*black square*), *L. holosericeus* (*open squares*), *L. lancerottensis* (*hatched squares*), and *L. spartioides* (*gray squares*). In all cases, the sampled populations represent exhaustively the known distribution of these endemic species. *Below*: scheme of the sub-sampling procedure for the three spatial situations considered. *Black squares* represent individuals removed from the total sample

Canaria) versus a congener distributed in different islands (*L. lancerottensis*, that occurs in Fuerteventura and Lanzarote). Comparisons between G_{ST} and θ estimates were obtained through the computer software FSTAT version 2.9.3.2 (Goudet 1995, 2002) after generating all the needed files from the corresponding genotype matrices using Transformer-3 (Caujapé-Castells and Baccarani-Rosas 2005). Because G_{ST} was estimated using GeneStat-PC 3.31 (Lewis and Whitkus 1993) in the papers where the data used are discussed (Oliva-Tejera et al. 2005, 2006 and Soto et al. 2007a, b), a comparison between values of this parameter estimated by that software and FSTAT version 2.9.3.2 was also undertaken. Non-parametric Mann–Whitney U tests were used to assess the significance of the differences in the values of G_{ST} and θ.

Results

Overall, estimates of G_{ST} and of θ (with respective averages of 0.116 and 0.137, see Table 1) were highly correlated ($r^2 = 0.920$, $P < 0.001$, Fig. 2a), though significantly different ($\theta > G_{ST}$, $P < 0.001$, one tailed Mann–Whitney U test). The estimates of G_{ST} obtained from Genestat-PC and FSTAT (respective averages of 0.156 and 0.116) were also significantly different [G_{ST} (Genestat) $>$ G_{ST} (Fstat), $P < 0.001$, one tailed Mann–Whitney U test], and exhibited a lower correlation than G_{ST} and θ ($r^2 = 0.823$, $P < 0.001$, Fig. 2b).

Out of the 270 G_{ST} and θ values calculated for the subsamplings with FSTAT (Table 1), 66 (24.4%) were lower, in most cases only slightly, than the respective value for the total sample (16 for θ and 50 for G_{ST}). Thirty-six of these were obtained in the scattered sub-samplings (11 for θ and 25 for G_{ST}), and 30 in one of the extreme sub-samplings (5 for θ and 25 for G_{ST}). The remaining 75.6% of the values associated with the sub-samples were variably higher than the G_{ST} and θ for the total sample, especially when removing 80% of the individuals (Figs. 3, 4).

While deviations corresponding to the removal of 10 and 20% of the individuals within the population did not entail major differences respect to the G_{ST} value for the whole sample, estimates for the removals of 50–60% of individuals were much more unreliable because of their high randomness (i.e., sometimes they ranked close to the G_{ST} value for the total sample, and sometimes they largely deviated from it). When removing 80% of individuals, most of the deviations of G_{ST} respect to the total sample were over-estimates, though in a few instances the value of G_{ST} and θ for the scattered sampling method decreased abruptly (see cases 1, 2 and 9 in Fig 3).

The minimum average G_{ST} increase for the scattered sub-samplings respect to the G_{ST} value for the total sample was observed in *Lotus lancerottensis* from Lanzarote (case 3, −22.9%) and, for the extreme sub-samplings, in *L. holosericeus* from Gran Canaria (−87.1%) [calculated from Table 1]. The highest positive increases in the values of G_{ST} (both on average and absolute) corresponded to one of the extreme sub-samplings of the Gran Canarian endemic *Lotus holosericeus* (respectively 187.1 and 344.1%, calculated from Table 1).

Differences between extreme sub-samplings were very sharp in some of the 9 cases assessed (see Fig. 3), though not significantly different from zero overall ($P = 0.248$, Mann–Whitney U test). However, extreme sub-samples (average $G_{ST} = 0.126$) did result in significantly higher over-estimates of G_{ST} than the scattered sub-samples ($P < 0.001$, one tailed Mann–Whitney U test Figs. 4 and 5). The average proportional increases of G_{ST} through the nine biological cases considered were 6.8% for the scattered sub-samplings and 21.1% for the extreme sub-samplings, and the maximum average G_{ST} increases for the extreme and the scattered sub-samplings were, respectively, 187.1% (in *L. holosericeus*) and 81.6% (in *L. spartioides*) [calculated from Table 1]. In the three groups involving *Crambe* populations (cases 7, 8, and 9 in Fig. 3) the differences between the values of G_{ST} estimated for either population extreme were not significant.

For the extreme sub-samplings, G_{ST} values resulting from removing 10% of the sample (average $G_{ST} = 0.168$) were significantly lower than when removing 80% of the individuals (average $G_{ST} = 0.233$) ($P < 0.05$, one tailed Mann–Whitney U test, see Fig. 5). By contrast, differences between G_{ST} values when removing 10 and 80% of the sample were not significant for the scattered subsamplings ($P = 0.389$, Mann–Whitney U test).

Considering all the spatial sub-samplings as a whole, average G_{ST} increments for narrow endemics from Gran Canaria (*Crambe tamadabensis* and *Lotus spartioides*) were higher than those for more widespread endemics from the same island (*C. pritzelii* and *L. holosericeus*) (18.6 and 30.2% respectively, Fig. 6), though these differences were not significant ($P = 0.108$, Mann–Whitney U test). In the case of *Lotus*, the average increments of G_{ST} for widespread versus narrow taxa distributed in the same island (33.5 and 47.8%, respectively) were significantly higher than those in *Crambe* (2.8 and 13.4%, respectively) ($P < 0.001$, one tailed Mann–Whitney U test). Single taxa distributed in the same island (*L. holosericeus* and *L. spartioides*, both endemic to Gran Canaria) displayed about twenty-fold higher average G_{ST} increments (40.5%) than congeneric taxa distributed in different islands (2.5%, *L. lancerottensis* from Fuerteventura and Lanzarote) ($P < 0.001$, one tailed Mann–Whitney U test, Fig. 6).

Table 1 Values of G_{ST} and θ, and increments respect to the value for for all samples in the extreme-1 (E1), extreme-2 (E2) and scatter (SC) sub-sampling proportions for the nine cases described in the methods

Groups assessed	E1					E2					SC				
	10	20	50	60	80	10	20	50	60	80	10	20	50	60	80
Nei's (1973) G_{ST}															
Lotus lancerottensis															
1. *L. lancerottensis* F + L (0.154, 0.164)	0.156	0.174	0.152	0.157	0.169	0.161	0.171	0.213	0.212	0.233	0.156	0.153	0.135	0.124	0.092
Increment (%)	1.3	13.0	−1.3	1.9	9.7	4.5	11.0	38.3	37.7	51.3	1.3	−0.6	−12.3	−19.5	−40.3
2. *L. lancerottensis* F (0.082, 0.099)	0.086	0.104	0.129	0.131	0.176	0.084	0.079	0.062	0.059	0.120	0.095	0.102	0.080	0.074	0.062
Increment (%)	4.9	26.8	57.3	59.8	114.6	2.4	−3.7	−24.4	−28.0	46.3	15.9	24.4	−2.4	−9.8	−24.4
3. *L. lancerottensis* L (0.160, 0.187)	0.161	0.183	0.139	0.149	0.142	0.168	0.179	0.260	0.264	0.221	0.152	0.141	0.121	0.110	0.093
Increment (%)	0.6	14.4	−13.1	−6.9	−11.3	5.0	11.9	62.5	65.0	38.1	−5.0	−11.9	−24.4	−31.3	−41.9
Three *Lotus* species from Gran Canaria															
4. All populations (0.125, 0.179)	0.129	0.134	0.181	0.198	0.215	0.118	0.113	0.110	0.113	0.110	0.124	0.128	0.151	0.137	0.141
Increment (%)	3.2	7.2	44.8	58.4	72.0	−5.6	−9.6	−12.0	−9.6	−12.0	−0.8	2.4	20.8	9.6	12.8
5. *L. holosericeus* (0.034, 0.048)	0.038	0.043	0.109	0.151	0.147	0.025	0.016	−0.007	0.001	−0.013	0.034	0.036	0.067	0.031	0.031
Increment (%)	11.8	26.5	220.6	344.1	332.4	−26.5	−52.9	−120.6	−97.1	−138.2	0.0	5.9	97.1	−8.8	−8.8
6. *L. spartioides* (0.025, 0.048)	0.029	0.035	0.037	0.026	0.033	0.026	0.026	0.031	0.018	0.024	0.031	0.036	0.046	0.038	0.076
Increment (%)	16.0	40.0	48.0	4.0	32.0	4.0	4.0	24.0	−28.0	−4.0	24.0	44.0	84.0	52.0	204.0
Two *Crambe* species from Gran Canaria															
7. All populations (0.159, 0.166)	0.162	0.167	0.188	0.191	0.221	0.158	0.155	0.175	0.180	0.192	0.157	0.156	0.155	0.153	0.156
Increment (%)	1.9	5.0	18.2	20.1	39.0	−0.6	−2.5	10.1	13.2	20.8	−1.3	−1.9	−2.5	−3.8	−1.9
8. *C. pritzelii* (0.153, 0.166)	0.154	0.154	0.155	0.157	0.194	0.153	0.150	0.174	0.181	0.201	0.151	0.148	0.145	0.140	0.153
Increment (%)	0.7	0.7	1.3	2.6	26.8	0.0	−2.0	13.7	18.3	31.4	−1.3	−3.3	−5.2	−8.5	0.0
9. *C. tamadabensis* (0.067, 0.075)	0.076	0.087	0.132	0.134	0.112	0.067	0.065	0.066	0.060	0.079	0.068	0.071	0.069	0.069	0.044
Increment (%)	13.4	29.9	97.0	100.0	67.2	0.0	−3.0	−1.5	−10.4	17.9	1.5	6.0	3.0	3.0	−34.3
Weir and Cockerham's (1984) θ															
1. *L. lancerottensis* F + L (0.154, 0.164)	0.164	0.172	0.165	0.171	0.207	0.175	0.182	0.231	0.228	0.257	0.167	0.165	0.147	0.137	0.102
Increment (%)	0.0	4.9	0.6	4.3	26.2	6.7	11.0	40.9	39.0	56.7	1.8	0.6	−10.4	−16.5	−37.8
2. *L. lancerottensis* F (0.082, 0.099)	0.103	0.122	0.152	0.157	0.236	0.102	0.094	0.082	0.074	0.134	0.111	0.114	0.098	0.096	0.078
Increment (%)	4.0	23.2	53.5	58.6	138.4	3.0	−5.1	−17.2	−25.3	35.4	12.1	15.2	−1.0	−3.0	−21.2

Table 1 continued

Groups assessed	E1					E2					SC				
	10	20	50	60	80	10	20	50	60	80	10	20	50	60	80
3. *L. lancerottensis* L (0.160, 0.187)	0.183	0.185	0.163	0.176	0.187	0.199	0.210	0.317	0.323	0.304	0.179	0.171	0.140	0.129	0.112
Increment (%)	−2.1	−1.1	−12.8	−5.9	0.0	6.4	12.3	69.5	72.7	62.6	−4.3	−8.6	−25.1	−31.0	−40.1
Three *Lotus* species from Gran Canaria															
4. All populations (0.125, 0.179)	0.187	0.190	0.193	0.208	0.228	0.177	0.174	0.192	0.197	0.179	0.178	0.178	0.190	0.183	0.182
Increment (%)	4.5	6.1	7.8	16.2	27.4	−1.1	−2.8	7.3	10.1	0.0	−0.6	−0.6	6.1	2.2	1.7
5. *L. holosericeus* (0.034, 0.048)	0.055	0.058	0.108	0.161	0.140	0.041	0.029	0.001	0.021	0.022	0.048	0.049	0.081	0.046	0.050
Increment (%)	14.6	20.8	125.0	235.4	191.7	−14.6	−39.6	−97.9	−56.3	−54.2	0.0	2.1	68.8	−4.2	4.2
6. *L. spartioides* (0.025, 0.048)	0.057	0.068	0.064	0.048	0.052	0.051	0.052	0.068	0.040	0.039	0.059	0.068	0.086	0.070	0.139
Increment (%)	18.8	41.7	33.3	0.0	8.3	6.3	8.3	41.7	−16.7	−18.8	22.9	41.7	79.2	45.8	189.6
Two *Crambe* species from Gran Canaria															
7. All populations (0.159, 0.166)	0.169	0.173	0.189	0.193	0.218	0.166	0.164	0.178	0.184	0.203	0.163	0.160	0.158	0.155	0.165
Increment (%)	1.8	4.2	13.9	16.3	31.3	0.0	−1.2	7.2	10.8	22.3	−1.8	−3.6	−4.8	−6.6	−0.6
8. *C. pritzelii* (0.153, 0.166)	0.167	0.167	0.168	0.172	0.217	0.165	0.162	0.189	0.197	0.219	0.161	0.157	0.152	0.145	0.165
Increment (%)	0.6	0.6	1.2	3.6	30.7	−0.6	−2.4	13.9	18.7	31.9	−3.0	−5.4	−8.4	−12.7	−0.6
9. *C. tamadabensis* (0.067, 0.075)	0.081	0.093	0.153	0.155	0.117	0.075	0.075	0.079	0.076	0.092	0.075	0.078	0.076	0.076	0.050
Increment (%)	8.0	24.0	104.0	106.7	56.0	0.0	0.0	5.3	1.3	22.7	0.0	4.0	1.3	1.3	−33.3

Pairs of values between brackets after the group names are, respectively, the G_{ST} and θ estimates for the total sample. Sample sizes for all populations involved in each case are given in Appendix 1

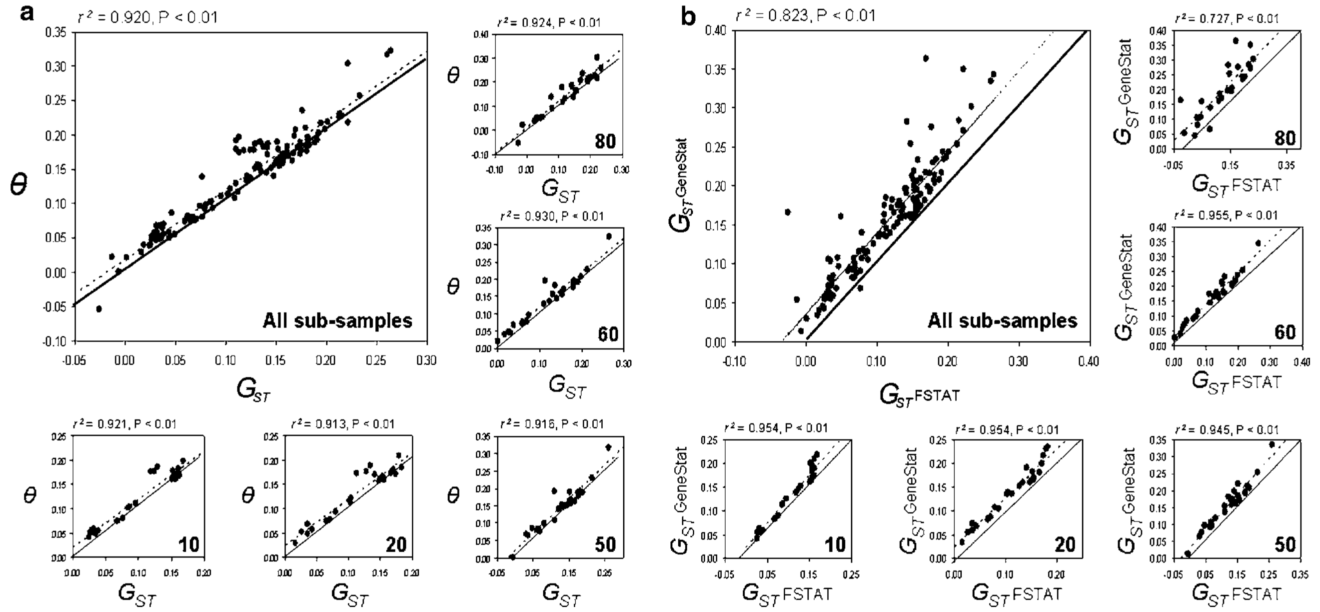


Fig. 2 Comparison plots between **a** G_{ST} and θ values obtained with FSTAT, and **b** G_{ST} values obtained with GeneStat and FSTAT. Square correlation values for each representation are given in the upper left corner, and the proportion of removed individuals in the bottom right corner. Discontinuous lines are the linear regressions between the corresponding pairs of values, and *solid lines* correspond to the expected relationships if the different estimates were always the same

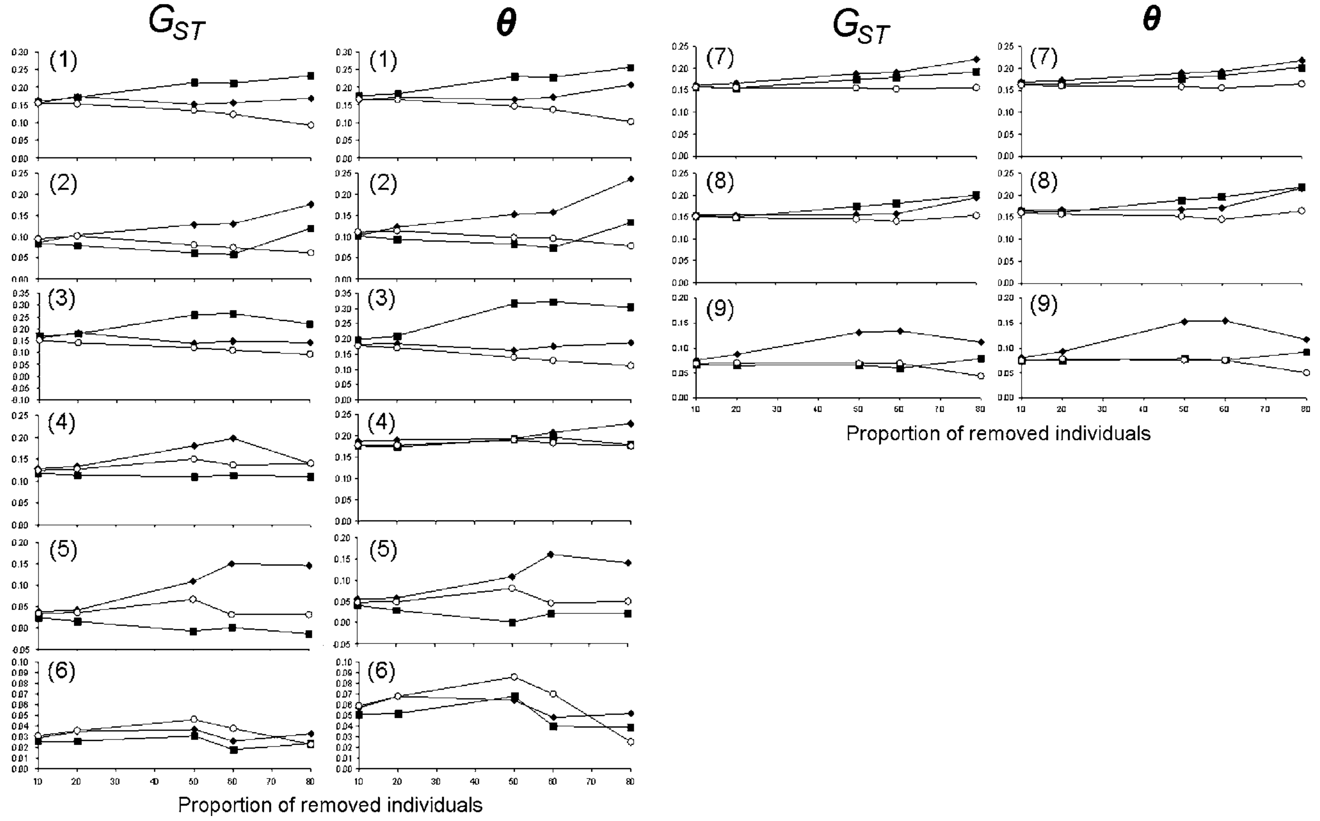


Fig. 3 Representation of the G_{ST} and θ values (y-axis) vs. the proportion of removed individuals (x-axis) through the nine cases described in the methods. *Empty circles* correspond to scattered subsamples, *black squares* and *diamonds* to each of the extreme sub-samples

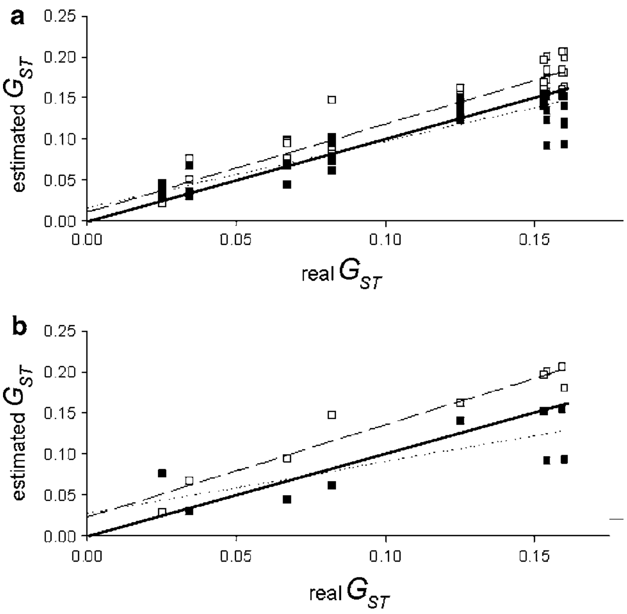


Fig. 4 Average variation of the extreme (*empty squares, hyphened line*) versus the scattered sub-samples (*black squares, dotted line*) respect to the real G_{ST} values. *Solid lines* correspond to the expected values if G_{ST} estimates were always the same, independent of the sub-samplings. Above: all sub-samples; below: only sub-samples that removed 80% of the sample sizes

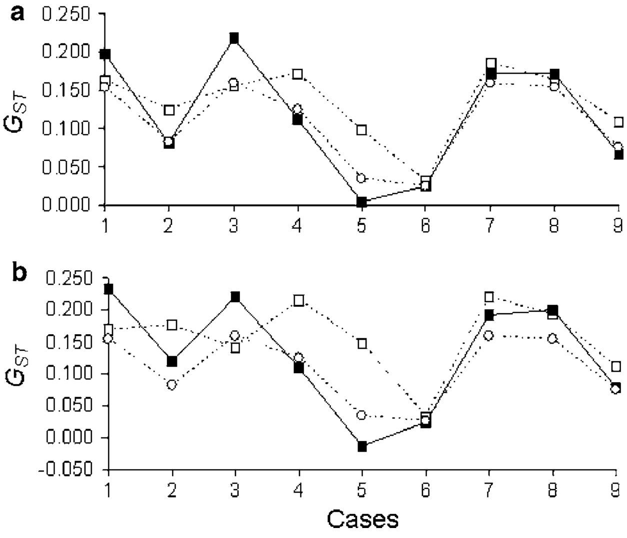


Fig. 5 Differences between the G_{ST} estimates from either population extreme (*black* and *empty squares*) vs. the value for the whole sample (*empty circles*) through the nine cases examined (coded as in methods). **a** average for all sub-samples; **b** only sub-samples that removed 80% of the individuals

Discussion

General implications for the estimation of genetic structure

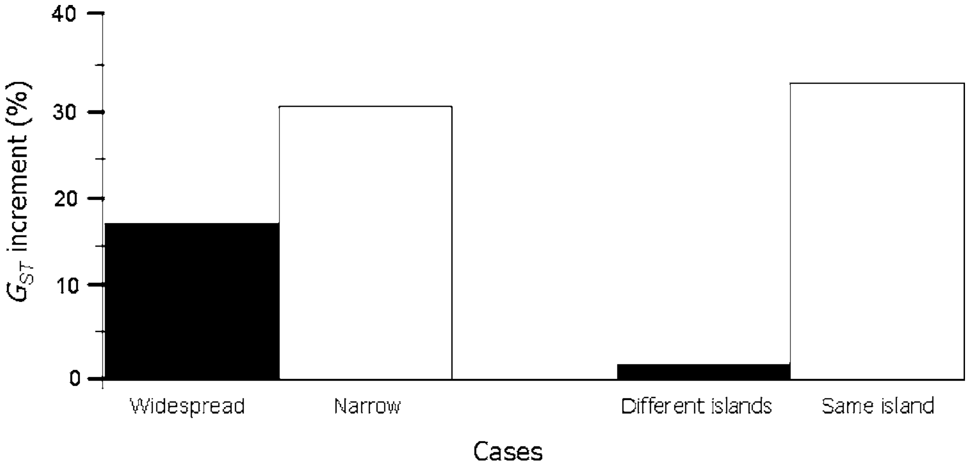
The accurate estimation of inter-population differentiation underlies the understanding of microevolutionary processes in all living organisms. Therefore, assessing the impact of sampling strategies and sizes on the estimated genetic subdivision is important for any researcher interested in questions impinging on evolutionary or conservation biology.

Unlike Nei's (1973) G_{ST} or Wright's (1943) F_{ST}, Weir and Cockerham's (1984) θ was designed to account for the error associated with small and unequal samples, and is theoretically more appropriate for most empirical studies. We should therefore expect a closer similarity of θ values associated with the subsamples respect to the value based on the total sample than in the corresponding G_{ST} values, in particular when sub-sample sizes are smaller. At odds with these expectations, the average deviation between the values of θ for the total sample and the corresponding subsamples (0.0275) is roughly three fold higher than that calculated for G_{ST} (0.0097), even when removing 80% of the sample (0.033 vs. 0.013, $P < 0.001$). Furthermore, the correlation between G_{ST} and θ is significantly high in all sub-samplings ($r^2 = 0.920$, $P < 0.001$), with θ values generally higher than the corresponding G_{ST} values (Fig. 2, Table 1). Both results highlight a small power of θ in removing sample size bias in empirical datasets, thus coinciding with Weicker et al. (2001) empirical tests with avian allozyme matrices. Comparison between the calculations for Canarian endemic *Lotus* and *Crambe* show that the effects of sub-sampling on G_{ST} and θ are much more evident in the former genus, though this could be due to each *Crambe* population being represented by more individuals (see Appendix 1), and hence less sensitive to the removal of samples.

The significance of the differences found between Nei's (1973) G_{ST} values estimated through GeneStat-PC version 3.31 (Lewis and Whitkus 1993) or FSTAT version 2.9.3.2 (Goudet 2002) (see results and Fig. 2b) brings out a software-related component of bias in the estimation of G_{ST}, and imposes extreme caution against pooling values calculated with different computer programs across different studies. However, assessing the causes underlying this kind of bias is not among the objectives of this paper; from this point on, the discussion will focus predominantly on the G_{ST} values obtained through FSTAT with regard to population sampling, as this parameter and its analog F_{ST} (and not θ), have been used to estimate the apportionment of genetic variation in all Canarian plant endemics assessed so far with allozymes. Nevertheless, since G_{ST} and θ are highly and significantly correlated (Fig. 2), the points raised for G_{ST} are widely applicable to θ.

Through revealing a consistent general pattern of higher estimated G_{ST} values with lower sample sizes (Figs. 3, 4), these sub-sampling experiments indicate that any sampling

Fig. 6 Average proportion of G_{ST} increments in comparisons between widespread and narrow endemics from the same island, and between endemic congeneric taxa distributed in different vs. the same island



process tends to over-estimate the values of G_{ST}, but especially when the number of wild individuals not considered for sampling is large (i.e., the situation in most real research practice). One feasible interpretation of this effect is that allele frequency changes induced by sampling mimic the effects of genetic drift, therefore resulting in a high positive correlation between the proportion of individuals removed from the sample and the inflation of the values of G_{ST}.

The results obtained also reveal a general and substantial influence of the intra-population distribution of samples on the estimation of inter-population genetic connectivity through G_{ST}. Thus, G_{ST} estimates corresponding to sub-samplings scattered across the population area result in significantly lower deviations than those calculated from either population extreme (respective average G_{ST} increments of 6.8 and 21.1%) (Fig. 3). Furthermore, only extreme sub-samplings result in significantly higher G_{ST} values when removing 80% of the individuals in the sample than when removing 10% ($P < 0.05$). This result compellingly suggests that G_{ST} estimates from scattered sub-samplings are much more accurate (i.e., less sensitive to sample size reductions). Hence, a correct estimation of the apportionment of genetic variation in nature is contingent upon a thorough sampling of the total occupancy area of all the populations included in the analyses. According to these empirical results, only a scattered distribution of samples agrees with the experiments with simulated datasets, where a small number of samples can produce reliable estimates of G_{ST} or F_{ST} (Holsinger et al. 2005).

Because lineage ascription imposes constraints on the levels and distribution of genetic variation (Gitzendanner and Soltis 2000), population genetic assessments for narrow endemics are more meaningful when widespread congeners are included in the comparisons. Therefore, it is important to know if the over-estimates of G_{ST} and θ associated with sampling are proportional in both categories of geographic distribution, or if they are differently affected.

The increment of G_{ST} for narrow and widespread congeners is similar when they occur in a single island (*L. holosericeus* vs. *L. spartioides*, and *C. pritzelii* vs. *C. tamadabensis*) (Fig. 7). However, differences are much more severe when the comparisons consider endemics from the same island versus congeners distributed in different islands, with the highest over-estimates associated with narrow endemics distributed in only one island (Fig. 7). A possible explanation to this effect is that a distribution in several islands is much proner to foster and reflect the action of drift, thereby inducing a higher default degree of inter-population differentiation than observed in congeners from a single island (G_{ST} for *L. lancerottensis s. l.* is 0.154, more than threefold higher than the value for either *L. holosericeus* or *L. spartioides*, with respective G_{ST} of 0.034 and 0.025 [Table 1]).

A biased intra-population sampling in endemics restricted to a single island seems thus more likely to result in higher G_{ST} over-estimates, because the component of drift induced by a biased sampling provokes a proportionally higher impact in these taxa than in taxa distributed in several islands, where drift is already a significant evolutionary factor. Table 1 bolsters this interpretation by showing that maximum G_{ST} increments induced by sampling in *L. holosericeus* and *L. spartioides* are of 187.7 and 81.6% (respectively), while that for *L. lancerottensis s. l.* is significantly lower, though still substantial (28.6%). A similar pattern of higher G_{ST} in widespread versus narrowly distributed taxa is detected in congeneric endemics from the same island (Gran Canaria), as G_{ST} values based on the total sample for the former class ($G_{ST} = 0.154$ and 0.034 for *C. pritzelii* and *L. holosericeus*, respectively) are roughly twofold higher than those associated with the latter ($G_{ST} = 0.076$ and 0.025 for *C. tamadabensis* and *L. spartioides*, respectively).

For similar reasons, the inflation of G_{ST} due to an insufficient intra-population sampling should be higher in allogamous species than in autogamous ones, because lower levels of genetic variation and inter-population gene flow associated with the latter category tend to induce much higher default values of G_{ST} as a result of the predominance of drift. Values of G_{ST} found for selfers by Hamrick and Godt (1989, 1997) are much higher than those for outcrossing or mixed-mating species. However, data on the Canarian endemics considered by Francisco-Ortega et al. (2000) did not indicate a consistent trend towards higher diversity in outcrossers vs. selfers, and Crawford et al. (2001) found no significant differences between outcrossing and highly selfing species from the Juan Fernández for any of the diversity measures that they considered (though they highlight that selfers exhibited less genetic variation in general).

Other factors notwithstanding, we would also expect more conspicuous G_{ST} increments in species exhibiting intra-population spatial structuring of genetic variation, because in this case (especially when collecting only at one of the population extremes) a biased intra-population sampling may misrepresent genetic variation more severely. This expectation is consistent with the result (Table 1) that the sub-samplings in *Lotus* (where seeds disperse by gravity, thereby fostering reproduction in small genetic neighborhoods Oliva-Tejera et al. 2005), give higher G_{ST} over-estimates than in the two species of *Crambe*, putative predominant outcrossers with a homomorphic sporophytic self-incompatibility system (Kroh 1964, Zuberi and Dickinson 1985) and a high seed dispersal capacity; accordingly, they exhibit low F_{IS} values (as opposed to the much higher F_{IS} values estimated in *Lotus* by Oliva-Tejera et al. 2005), and only moderate genetic structure (Soto et al. unpublished data).

Another possible source of inflation of G_{ST} not examined in this paper is that an inadequate number of populations are sampled. Therefore, future sub-sampling experiments should aim at assessing whether excluding different combinations of entire populations at different spatial localisations within the distribution range of a species may affect G_{ST} estimates differently.

Implications for the genetic cohesion of the Canarian endemic Flora

The results in this paper only refer to a few biological situations and lineages within the wide heterogeneity of the Canarian endemic flora, but they are consistent throughout the taxa and cases examined. If they prove to be representative of many Canary Island endemics, then they may have important general implications for the interpretation of genetic divergence and the design of conservation strategies to preserve the archipielago's Flora.

The largest G_{ST} over-estimates always correspond to removing 80% of the individuals in the extreme sub-samplings (Figs. 3, 4, 5, 6). Therefore, not considering the source of error associated with the use of different softwares (see results and Fig. 2b), many values of G_{ST} reported in the literature for different Canarian endemics are very possibly largely inflated, as they are based on samples that leave out much more than 80% of the individuals, and restrict to only one part of the target populations. While the abrupt Canarian topography very often prevents access to significant areas of a population, this cannot be the main explanation for most cases in the allozyme literature for endemic plants from this archipelago, where perhaps the urgency imposed by a tight collection schedule is more relevant to understand the sampling deficiencies. In certain instances, especially with very small population samples and very low allozyme diversity within populations, calculated G_{ST} values may be very high because of random fixation of different alleles at one or several loci, or very different allele frequencies at a few loci in different populations. Consideration of these possibilities might help understand some of the most extreme G_{ST} values calculated for Canarian endemics (e. g., G_{ST} of 0.773, 0.696 and 0.538 for *Sonchus hierrensis*, *S. canariensis* and *Babcockia platylepis* (Kim et al. 1999), or G_{ST} of 0.381 for *Lobularia canariensis* (Borgen 1996).

Thus, the contention that "…the mean G_{ST} value for 23 species endemic to the Canary Islands is much higher than the mean $G_{ST} = 0.179$ estimated by Hamrick and Godt (1997) for endemic outcrossing species in general" (Francisco-Ortega et al. 2000) is correct numerically, but the data onto which it is based might not reflect the underlying biological reality of the Canarian Flora. This is not to say that there are not important barriers to gene flow in the Canaries, but simply that their overall impact in the genetic cohesion of endemic plants from this archipelago could be much more limited than it has been traditionally assumed. Furthermore, there may be published G_{ST}/F_{ST} estimates in the present allozyme database that could be less inflated due to either a spatially robust sampling strategy, or to causes inherent to the lineages/populations investigated (e. g., very low overall genetic variation levels, or a homogeneous intra-population distribution thereof). Some of the taxa with lowest G_{ST} values in the review of Francisco-Ortega et al. (2000) might be affected by different combinations of these factors (e. g., *Lolium canariense* ($G_{ST} = 0.104$, Charmet and Balfourier 1994), or *Cistus osbaeckiaefolius* ($G_{ST} = 0.126$, Batista et al. 2001).

Conclusions

Two general considerations emerge that should be elementary intra-population sampling rules for the sake of an accurate estimation of inter-population parameters in conservation genetics. First, unlike the sampling strategies followed in many published studies with plant endemics from the Canaries and from other biodiversity hotspots, the thorough representation of each population's occupancy space is compulsory if we aim at estimating genetic subdivision accurately. Concomittantly, the necessary inclusion of widespread congeners in population genetic assessments with narrow endemics would be insufficient for comparative purposes if we did not also sample the narrowly distributed taxa more intensively always, but especially when they are allogamous. These conclusions should deter researchers from basing estimates of inter-population genetic parameters on other sampling schemes (i.e., collecting only in one part of the population), and then using these values to substantiate conservation efforts. Although the calculations underlying these conclusions were made with plant allozyme data only, it is feasible that they apply to other polymorphic molecular markers used in microevolutionary or genetic conservation research, and to other organisms.

If the results of these sub-sampling experiments are of wide applicability to the whole Canarian Flora, a significant factor to explain the high average value of $G_{ST} = 0.281$ calculated by Francisco-Ortega et al. (2000) is the over-estimation of this parameter arising from mainly (1) the limited and spatially biased intra-population sampling in many of the papers used by those authors to calculate that figure, but also from possibly (2) a source of error related to pooling values of G_{ST} or F_{ST} that were obtained using different softwares. Therefore, the genetic cohesion of the Canarian endemic Flora could be much tighter than presently assumed, despite the existence of geographic barriers that may make gene flow difficult in this archipelago.

Acknowledgments The author thanks Felicia Oliva-Tejera and Moisés Soto for the data on the biological cases of *Lotus* and *Crambe*, and for their committment with the thorough representation of population sizes and occupancy areas; Dr. Daniel J. Crawford (University of Kansas), Dr. Txema Iriondo-Alegría (Universidad Rey Juan Carlos I in Madrid), and two anonymous referees for insightful feedback on a previous draft; and the Spanish "Ministerio de Educación y Ciencia", for funding the complementary action CGL2006–27073-E/BOS to organize an international meeting of specialists in population genetics and biological databases that inspired the experiments discussed in this paper. The Cabildo de Gran Canaria and all the staff at the JBCVC are acknowledged for support to the various research lines conducted by the Departamento de Biodiversidad Molecular and DNA bank.

Appendix 1

See Table 2.

Table 2 Sample sizes for all the populations and sub-sampling proportions considered in the paper

Taxon/population	Island	Group	*N*	% of removed individuals				
				10	20	50	60	80
Lotus lancerottensis								
LLFEM	Lanzarote	1, 3	20	18	16	10	9	4
LLGOL	Lanzarote	1, 3	22	20	18	11	10	5
LLARR	Lanzarote	1, 3	15	13	12	8	7	3
LLFAM	Lanzarote	1, 3	23	21	19	12	11	5
LLTVH	Lanzarote	1, 3	21	19	17	10	9	4
LLCOR	Fuerteventura	1, 2	16	14	13	8	7	4
LLVMB	Fuerteventura	1, 2	37	33	30	18	16	7
LLJAB	Fuerteventura	1, 2	18	16	14	9	8	4
LLZAR	Fuerteventura	1, 2	22	20	18	11	10	5
LLJAN	Fuerteventura	1, 2	25	22	20	13	12	6
Lotus spartioides								
LSTA	Gran Canaria	4, 6	100	90	80	50	40	20
LSBE	Gran Canaria	4, 6	30	27	24	15	12	6
Lotus holosericeus								
LHPI	Gran Canaria	4, 5	30	27	24	15	12	6
LHRO	Gran Canaria	4, 5	25	22	20	13	10	5
LHAR	Gran Canaria	4, 5	75	67	61	38	30	15
LHPP	Gran Canaria	4, 5	25	22	20	12	10	5
LHCU	Gran Canaria	4, 5	15	13	11	8	6	3
LHTE	Gran Canaria	4, 5	10	9	8	5	4	2
Lotus genistoides								
LGCJ	Gran Canaria	4	49	44	39	25	20	10
Crambe pritzelii								
CPGA	Gran Canaria	7, 8	91	82	73	46	36	18
CPGY	Gran Canaria	7, 8	94	85	75	47	38	19
CPAN	Gran Canaria	7, 8	37	33	30	19	15	7
CPTE	Gran Canaria	7, 8	92	83	74	46	37	18
CPAG1	Gran Canaria	7, 8	54	49	43	27	22	11
CPAG2	Gran Canaria	7, 8	39	35	31	20	16	8
CPAZ	Gran Canaria	7, 8	95	86	76	48	38	19
CPJI	Gran Canaria	7, 8	33	30	26	17	13	7
Crambe tamadabensis								
CTAM	Gran Canaria	7, 9	32	29	26	16	13	6
CTGU	Gran Canaria	7, 9	41	37	33	21	16	8
CTCA	Gran Canaria	7, 9	29	26	23	15	12	6
CTTA	Gran Canaria	7, 9	84	76	67	42	34	17

'Group' stands for the groups where each population was included (see the "Methods" and Table 1). *N* is the total sample size

References

Batista F, Bañares A, Caujapé-Castells J, Carqué E, Marrero-Gómez M, Sosa PA (2001) Allozyme diversity in three endemic species of *Cistus* (Cistaceae) from the Canary Islands: Intraspecific comparisons and implications for genetic conservation. Am J Bot 88:1582–1592. doi:10.2307/3558402

Borgen L (1996) Genetic differentiation in endemic *Lobularia* (Brassicaceae) in the Canary Islands. Nord J Bot 16:487–503. doi:10.1111/j.1756-1051.1996.tb00263.x

Caujapé-Castells J (2006) Brújula para botánicos desorientados en la genética de poblaciones. EXEGEN ediciones (http://www.exegen.org/en/publications.php)

Caujapé-Castells J, Baccarani-Rosas M (2005) Transformer-3: a program for the analysis of molecular population genetic data. EXEGEN Software and Jardín Botánico Canario "Viera y Clavijo", Las Palmas de Gran Canaria, Spain (free downloads at http://www.exegen.org/en/software.php)

Charmet G, Balfourier F (1994) Isozyme variation and species relationships in the genus *Lolium* L. (ryegrasses, Graminaceae). Theor Appl Genet 87:641–649. doi:10.1007/BF00222888

Cockerham CC, Weir BS (1986) Estimation of inbreeding parameters in stratified populations. Ann Hum Genet 50:271–281. doi:10.1111/j.1469-1809.1986.tb01048.x

Crawford DJ, Ruiz E, Stuessy TF, Tepe E, Aqeveque P, Gonzalez F, Jensen RJ, Anderson GJ, Bernardello G, Baeza CM, Swenson U, Silva OM (2001) Allozyme diversity in endemic flowering plant species of the Juan Fernandez Archipelago, Chile: ecological and historical factors with implications for conservation. Am J Bot 88:2195–2203. doi:10.2307/3558381

Culley TM, Wallace LE, Gengler-Nowak KM, Crawford DJ (2002) A comparison of two methods of calculating G_{st}, a genetic measure of population differentiation. Am J Bot 89:460–465. doi:10.3732/ajb.89.3.460

Francisco-Ortega JA, Santos-Guerra A, Kim SC, Crawford DJ (2000) Plant genetic diversity in the Canary Islands: a conservation perspective. Am J Bot 87:909–919. doi:10.2307/2656988

Gitzendanner MA, Soltis PS (2000) Patterns of genetic variation in rare and endangered widespread plant congeners. Am J Bot 87:783–792. doi:10.2307/2656886

Goudet J (1995) Fstat version 1.2: a computer program to calculate Fstatistics. J Hered 86:485–486

Goudet J (2002) Fstat version 2.9.3.2: a computer program to calculate Fstatistics. http://www2.unil.ch/popgen/softwares/fstat.htm

Hamrick JL, Godt MJW (1989) Allozyme diversity in plant species. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) Plant population genetics, breeding and germplasm resources. Sinauer, Sunderland, pp 43–63

Hamrick JL, Godt MJW (1997) Effects of life history traits on genetic diversity in plant species. In: Sivertown J, Franco M, Harper JL (eds) Plant life histories: ecology, phylogeny and evolution. Cambridge University Press, Cambridge UK, pp 102–118

Hamrick JL, Godt MJW, Murawski DA, Loveless MD (1991) Correlations between species traits and allozyme diversity: implications for conservation biology. In: Falk DA, Holsinger K (eds) Genetics and Conservation of rare plants. Oxford University Press, New York, pp 75–86

Holsinger KE, Lewis PO, Dey D (2005) A Bayesian approach to inferring population structure from dominant markers. Technical report 01–30, Department of Statistics, University of Connecticut

Kim SC, Crawford DJ, Francisco-Ortega J, Santos-Guerra A (1999) Adaptive radiation and genetic differentiation in the woody *Sonchus* alliance (Asteraceae: Sonchinae) in the Canary Islands. Plant Syst Evol 215:101–118. doi:10.1007/BF00984650

Kroh M (1964) An electron microscopic study of the behaviour of Cruciferae pollen after pollination. In: Linskens HF (ed) Pollen physiology and fertilisation. North Holland, Amsterdam, pp 221–224

Lewis PO, Whitkus R (1993) GeneStat-PC, version 3.3. North Carolina State University, Raleigh

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323. doi:10.1073/pnas.70.12.3321

Nei M (1986) Definition and estimation of fixation indices. Evol Int J Org Evol 40:643–645. doi:10.2307/2408586

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. Ann Hum Genet 47:253–259. doi:10.1111/j.1469-1809.1983.tb00993.x

Oliva-Tejera F, Caujapé-Castells J, Naranjo-Suárez J, Navarro-Déniz J, Acebes-Ginovés JR, Bramwell D (2005) Population genetic differentiation in taxa of *Lotus* (Fabaceae: Loteae) endemic to the Gran Canarian pine forest. Heredity 94:199–206. doi:10.1038/sj.hdy.6800598

Oliva-Tejera F, Caujapé-Castells J, Navarro-Déniz J, Reyes-Betancort A, Scholz S, Baccarani-Rosas M, Cabrera-García N (2006) Patterns of genetic divergence of three Canarian endemic *Lotus* (Fabaceae): implications for the conservation of the endangered *L.kunkelii*. Am J Bot 93:1116–1124. doi:10.3732/ajb.93.8.1116

Slatkin M (1987) Gene flow and the geographic structure of natural populations. Science 236:787–792. doi:10.1126/science.3576198

Soto ME, Caujapé-Castells J, Roca Salinas A, Bramwell D (2007a) Selecting population targets for preserving the genetic variation of the narrow Canarian endemics *Crambe tamadabensis* and *C. pritzelii* in a seed bank. In: Proceedings of the 3rd International Biogeography Society Conference. Santa Cruz de Tenerife, Spain, 161 pp

Soto ME, Caujapé-Castells J, Roca Salinas A, Bramwell D (2007b) La variabilidad genética de dos especies de *Crambe* amenazadas en Canarias: implicaciones para su conservación. In: Proceedings of the III Congreso de la Sociedad Española de Biología de Conservación de Plantas. Santa Cruz de Tenerife, Spain, pp. 68–69

Weicker JJ, Brumfield RT, Winker K (2001) Estimating the unbiased estimator θ for population genetic survey data. Evol Int J Org Evol 55:2601–2605

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evol Int J Org Evol 38:1358–1370. doi:10.2307/2408641

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. Heredity 82:117–125. doi:10.1038/sj.hdy.6884960

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

Wright S (1951) The genetical structure of populations. Ann Eugen 15:323–354

Zuberi MI, Dickinson HG (1985) Pollen-stigma interaction in *Brassica* III. Hydration of the pollen grains. J Cell Sci 76:321–336