MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 2743–2760 doi: 10.1111/j.1365-294X.2012.05562.x

Recent colonization of the Galápagos by the tree *Geoffroea spinosa* Jacq. (Leguminosae)

S. CAETANO,*[1] M. CURRAT,†[1] R. T. PENNINGTON,‡ D. PRADO,§ L. EXCOFFIER¶** and Y. NACIRI*

*Plant Systematics and Biodiversity Laboratory, Molecular Phylogeny and Genetics Unit, Conservatoire et Jardin botaniques, 1 Chemin de l'Impératrice, CP 60, CH-1292 Chambésy, Genève, Switzerland, †Laboratory of Anthropology, Genetics and Peopling History, Anthropology Unit, Department of Genetics and Evolution, University of Geneva, 12, Rue Gustave-Revilliod, CH-1227 Carouge, Geneva, Switzerland, ‡Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, UK, §CONICET & Cátedra de Botánica Morfológica y Sistemática, Facultad de Ciencias Agrarias, UNR, Casilla de Correo No.14, S2125ZAA, Zavalla, Argentina, ¶Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland, **Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland*

Abstract

This study puts together genetic data and an approximate bayesian computation (ABC) approach to infer the time at which the tree *Geoffroea spinosa* colonized the Galápagos Islands. The genetic diversity and differentiation between Peru and Galápagos population samples, estimated using three chloroplast spacers and six microsatellite loci, reveal significant differences between two mainland regions separated by the Andes mountains (Inter Andean vs. Pacific Coast) as well as a significant genetic differentiation of island populations. Microsatellites identify two distinct geographical clusters, the Galápagos and the mainland, and chloroplast markers show a private haplotype in the Galápagos. The nuclear distinctiveness of the Inter Andean populations suggests current restricted pollen flow, but chloroplast points to cross-Andean dispersals via seeds, indicating that the Andes might not be an effective biogeographical barrier. The ABC analyses clearly point to the colonization of the Galápagos within the last 160 000 years and possibly as recently as 4750 years ago (475 generations). Founder events associated with colonization of the two islands where the species occurs are detected, with Española having been colonized after Floreana. We discuss two nonmutually exclusive possibilities for the colonization of the Galápagos, recent natural dispersal vs. human introduction.

Keywords: ABC analysis, chloroplast, founder events, genetic structure, long-distance dispersal, microsatellites

Received 20 August 2010; revision received 9 February 2012; accepted 16 February 2012

Introduction

The Galápagos Archipelago is one of the most emblematic subjects of insular biogeography (Grehan 2001), and understanding the colonization of these isolated oceanic islands has been the focus of many studies (e.g. Bisconti *et al.* 2001). Great attention has been given to the way the unique fauna of the Archipelago has evolved. This is the case of the marine iguanas (Rassman 1997), the giant tortoises (Beheregaray *et al.* 2004; Milinkovitch *et al.* 2007) and the Galápagos petrel (Friesen *et al.* 2006), for which the diversification within and between the islands has been analysed. Other authors have been interested in understanding the relationships between Central and South America and the Islands for several other emblematic organisms, such as the Darwin's finches (Sato *et al.* 1999, 2001), or the lava lizards (Kizirian *et al.* 2004). On the other hand, and despite the high number of native species (approximately 560; McMullen 1999), plants have received much less attention. Only a few studies have evaluated the phylogeographical patterns of plants in

Correspondence: Yamama Naciri, Fax: +41 22 418 5101; E-mail: yamama.naciri@ville-ge.ch

[1]These two authors have equally contributed to this manuscript.

the Galápagos Archipelago (Yeakley & Weishampel 2000; Willerslev *et al.* 2002; Weeks & Tye 2009), and the ones that analysed the potential continental origins of native species are even less numerous (Schilling *et al.* 1994; Moore *et al.* 2006).

The Galápagos Archipelago lies in the eastern equatorial Pacific, 960 km west of Ecuador's coast and is formed by true oceanic islands (Kurz & Geist 1999) that were created between 0.5 and 5 million years ago (Ma; White *et al.* 1993). From a biogeographical perspective, the origin and evolutionary history of the archipelago's flora is still an open issue (Grehan 2001). These remote islands have never been connected to other landmasses, which suggests that colonization by long-distance dispersal might have occurred quite frequently given the taxonomic diversity of their present biota (de Queiroz 2005). A long-distance dispersal origin for much of the native flora of the Galápagos is also suggested by its composition of families displaying good long-distance dispersal abilities (Carlquist 1966, 1967). Avifaunal transport is assumed for a majority of Galápagos plant species, but passive mechanisms such as wind and oceanic drift are also suggested (Porter 1976). Human introductions have also had great influence on the Galápagos flora, as nonnative plants currently represent 45% of total species (Mauchamp 1997) and constitute a potential major threat to native species (Caujapé-Castells *et al.* 2010).

The establishment of any species in the islands is likely to have relied on the successful foundation of a new population by a rather small number of individuals (i.e. a founder event; Nei *et al.* 1975). Differing genetic signatures of such founder events are nevertheless expected, depending on colonization time. Whereas a recent establishment of a species in the islands should still be potentially perceived from a signature of low diversity that represents a subset of that of mainland populations, the signal of the founder effect associated with an early colonization is expected to be erased by the regeneration of diversity through time (Austerlitz *et al.* 2000).

In this study, we address the unresolved status of the tree *Geoffroea spinosa* in the Galápagos by focussing on the timing of colonization of the islands. We examine the population genetic architecture of continental and island populations using three intergene chloroplast (cp) spacers and six microsatellite markers (Naciri-Graven *et al.* 2005). We further use an Approximate Bayesian Computation (ABC; Beaumont *et al.* 2002) framework to infer the parameters of a plausible colonization scenario of the islands. Recent applications of the ABC approach have allowed an improved understanding of past demographic events for different species (e.g. Chan *et al.* 2006; François *et al.* 2008; Neuenschwander *et al.* 2008; Ross-Ibarra *et al.* 2009), and the method is used here to obtain more precise inferences of the colonization of the Galápagos Islands by *G. spinosa*.

Geoffroea spinosa is a taxonomically well-defined, deciduous species (Ireland & Pennington 1999) that grows as single isolated trees on floodable ground in riverbanks or next to stagnant water in xerophytic woodland areas (Nascimento *et al.* 2003). This implies that seeds are able to survive immersion and dispersal in fresh water (Ireland & Pennington 1999). The fruit of *G. spinosa* is a drupe with a hard woody endocarp containing a single seed, used for animal feed and occasionally eaten by humans (Sanchez *et al.* 2006). *Geoffroea spinosa*'s wood has also been traditionally used in construction, for carpentry and furniture making, as well as for fuel (Lucena *et al.* 2007). The species has a scattered distribution pattern over five disjunct areas in South America (Ireland & Pennington 1999): north-eastern Brazil (northern Ceará, Pernambuco and eastern Bahia); Colombia, Venezuela and the Dutch Antilles; north-eastern Argentina (east of Gran Chaco), Paraguay and Bolivia (north-west of Gran Chaco); northern Peru and Ecuador; and the Galápagos (Floreana and Española islands). The status of *G. spinosa* is still unresolved in the Galápagos, as botanists do not clearly assess or do not agree on whether it is introduced or native (Ireland & Pennington 1999; Vargas *et al.* 2012; A. Tye, personal communication).

The genetic information obtained for the population samples from Peru and the Galápagos Islands, combined with the ABC simulations, allows us to draw a more accurate history for *G. spinosa* by (i) identifying three geographical regions defined by major biogeographical barriers; (ii) establishing the timing of the islands colonization; and (iii) assessing the possible introduced status of the tree in the Galápagos. This is the first case, to our knowledge, that ABC is used to infer the doubtful introduced vs. native status of a species, although other studies recurrently used molecular data for the same purpose (Bagnoli *et al.* 2009; Lorenzo *et al.* 2009; Martín-Bravo *et al.* 2009).

Materials and methods

Population sampling

A total of 160 *Geoffroea spinosa* adult individuals were sampled in Peru and in the Galápagos (Fig. 1), young leaves or buds being stored in air-tight plastic bags with silica gel. The five population samples of Peru (81 individuals) are located in the seasonally dry tropical forests of the Pacific Coast (Pacific Coast-1, 2 and 3) and in the Inter Andean valleys (Inter Andean-1 and 2). In the Galápagos, samples were obtained from the two islands

Fig. 1 Geographic localization of the sampled populations in South America (Peru) and the Galápagos.

where the species occurs, Floreana (five population samples; 73 individuals) and Española (six individuals from a single population sample). The sampling of *G. spinosa* in South America was restricted to Peru, because available data on the species showed that populations from all areas, but Peru, were very divergent from the Galápagos ones for both nuclear microsatellites and chloroplast haplotypes (Caetano 2008; Caetano & Naciri 2011). Moreover, some microsatellites developed by Naciri-Graven *et al.* (2005) did work on Peru and the Galápagos whereas they were not successfully amplified in other regions. This constituted an additional argument to restrict the analysis to Peru as a potential source for the Galápagos populations.

Molecular protocols

Genomic DNA was extracted using the DNeasy Plant Kit (Qiagen). For each individual, three chloroplast spacers were amplified using the following primers: *trn*H and *psb*A for HA, *trn*S and *trn*G for SG (Hamilton 1999) and *trn*L_e and *trn*F_f for LF (Taberlet *et al.* 1991). Both strands of the PCR products were sequenced using BIGDYE® TERMINATOR v3.1 Cycle Sequencing Kit, on an ABI 377 automated sequencer (Applied Biosystems). The nucleotide sequences were assembled with SEQUENCHER™ 4.8 and aligned using CLUSTAL W (Thompson *et al.* 1994) implemented in the BIOEDIT program (Hall 1999). A haplotype distribution map was constructed using ARCMAP GIS (Environmental Systems Research Institute Inc., Redlands, CA, USA).

Each individual was genotyped at six microsatellite loci (Naciri-Graven *et al.* 2005), namely *Gspi*.B458, *Gspi*.B331, *Gspi*.A149, *Gspi*.I168, *Gspi*.B264 and *Gspi*.B284. PCR amplifications were performed using an optimized multiplex mix (Qiagen), adapting the manufacturer's instructions to a total reaction volume of 5 μL. Multiplex PCR products were electrophoresed on a 4.75% denaturing polyacrylamide gel on an ABI 377 automated sequencer. Allele sizes were scored relative to an internal known size standard, Genescan-400 Rox (Applied Biosystems), using the GENESCAN software.

Genetic data analyses

For microsatellite markers, linkage disequilibrium (LD; 10 000 permutations) and Hardy–Weinberg equilibrium (HWE; 1 000 000 steps of Markov Chain and 100 000 dememorization steps) were tested for each locus-population combination, using ARLEQUIN (Excoffier *et al.* 2005a).

The genetic structure of the populations was identified with the program STRUCTURE (Pritchard *et al.* 2000), and the most likely number of clusters (K) was inferred using the ΔK statistic (see Evanno *et al.* 2005). STRUCTURE assigns individuals probabilistically to clusters on the sole basis of the genetic data and without prior information on sample location. To avoid detecting local optimum, ten independent runs for each K value (ranging from one to eleven) were performed, assuming the admixture model with correlated allele frequencies and using burn-in and MCMC lengths of 50 000 and 100 000 iterations, respectively.

Levels of genetic diversity, as measured by the observed number of alleles (N_A), observed heterozygosity (H_O) and expected heterozygosity (H_E), were determined per locus and population using ARLEQUIN. The allelic richness (R_S) per locus and population was calculated using FSTAT (Goudet 2001). Genetic divergence between populations was quantified by computing the traditional F_{ST} pairwise measure (Weir & Cockerham 1984). Additionally, the distribution of the genetic variation within *G. spinosa* was quantified by means of analyses of molecular variance (AMOVA; Excoffier *et al.* 1992) using a pairwise genetic distance.

The significance level for multiple tests was adjusted according to the modified false discovery rate method (Benjamini & Yekutieli 2001) as follows: $\alpha / \sum_{i=1}^{k} 1/i$, where k is the number of tests performed.

ABC outline

Three evolutionary models for the colonization of the Galápagos Islands by *G. spinosa* were simulated, and the parameters of interest were estimated by means of

the ABC method (Beaumont *et al.* 2002; Bertorelle *et al.* 2010). The general idea is to compare simulated data sets to the observed one by computing Euclidian distances between sets of simulated and observed summary statistics (see below for details). The most probable value for each parameter is then extracted from the simulations closest to the original data (smallest Euclidian distances). Our model assumes a colonization of the Galápagos Islands from Peru (Fig. 2), and for this purpose, two populations were considered: Peru (where we merged data from five populations; 81 individuals) and Floreana (where we merged data from five populations; 73 individuals). Unidirectional, bidirectional or absence of gene flow between Peru and Floreana were further assumed, depending on the scenario. Española was removed from the simulations due to its small sample size (only six individuals sampled for this island). All ABC analyses were performed using microsatellite data.

Parameter estimation

The ABC method obtains, for a given scenario, the marginal posterior distribution of all parameters, from which point estimates and credible intervals can be extracted. The procedure includes, for each of the three scenarios: 1, simulation of one million genetic data sets with the same number of loci and sample sizes as in the observed data set. This step was performed using the program SIMCOAL2 (Laval & Excoffier 2004). For each simulation, parameter values were randomly drawn from prior distributions (see below); 2, computation of the summary statistics (see below) for each simulation using the program ARLEQUIN (Excoffier *et al.* 2005a); 3, comparison between the simulated and the observed data sets on the basis of the Euclidean distances between simulated and observed summary statistics, and retention of the 2000 simulations (0.2%) closest to the observation (smallest distances). The remaining simulations were rejected at this step; 4, estimation of the posterior distributions of all parameters by a locally weighted linear regression of the summary statistics computed from the retained simulations (Beaumont *et al.* 2002). The parameters were transformed before regression, as $y = \log[\tan(x)^{-1}]$ to restrict the posterior parameter distribution within the range of their prior distribution (Excoffier *et al.* 2005b). Steps 3 and 4 were both performed using the program ABCEST (Excoffier *et al.* 2005b).

Prior information

Prior information for the different parameters was either defined on the basis of direct observations or using information from the literature. The parameters were all drawn randomly from uniform prior distributions, except for the mutation rate for which a gamma distribution was used.

The population of Floreana was assumed to have grown exponentially after the colonization of the island, with an initial effective size (N_0) of [1–1000] and a current size (N_1) of [500–5000]. The effective size in Peru was set to [10 000–100 000] individuals. On the basis of a floristic study in the Peruvian seasonally dry tropical forests (Linares-Palomino 2006), a mean density of 3–5 trees per hectare (ha) has been considered here for *G. spinosa*. Given that within the northern Peru and Ecuador region, around 2.5 million ha are suitable for this species, we estimate the total census size in the whole region to be between 7.5 and 12.5 million individuals. We thus considered that the five mainland populations used in this study are representative of a range of 10 000–100 000 individuals corresponding to about 0.1–1% of the total population estimated to grow in the whole region of northern Peru and Ecuador. The colonization time (t) necessarily occurred after the emergence of Floreana, 1.5 Ma (White *et al.* 1993). Assuming a very conservative mean generation time of 25 years for *G. spinosa*, the colonization time is thus necessarily between 1 and 60 000 generations. The migration rate corresponds to the proportion of migrants sent from Peru to Floreana (and from Floreana to Peru depending on the scenario) at each generation after the colonization event. It represents the

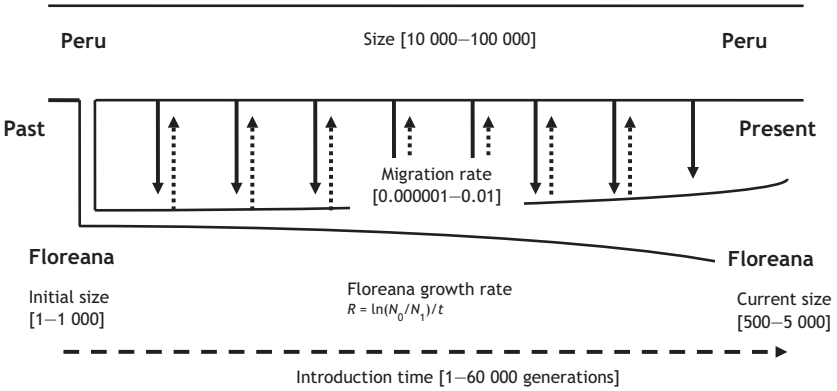

Fig. 2 Simulated scenario of the colonization of the Galápagos by *Geoffroea spinosa*. Scenario 1 assumes migration from Peru to Floreana, scenario 2, no migration after colonization and scenario 3, migration in both directions.

probability for each lineage to have migrated from Peru to Floreana (and *vice versa*) at each generation. Migration rate was considered constant over time and its prior distribution was set to [0.000001–0.01]. Information on the mutation rate (μ) was obtained indirectly from published studies that adressed the rate at which microsatellites mutate within different plant species (Udupa & Baum 2001; Thuillet *et al.* 2002; Vigouroux *et al.* 2002; O'Connell & Ritland 2004). For each locus *i*, a mutation rate μ_i was randomly drawn from a gamma distribution, with a mean randomly drawn between 0.0001 and 0.01, and a shape parameter *k* set to 10. The generalized stepwise mutation model (GSM; di Rienzo *et al.* 1994) was used, assuming that the number of repeats by which a new allele differs from its ancestral state follows a geometric distribution with a parameter p varying between [0 and 0.3]. Note that $P = 0$ corresponds to a strict stepwise mutation model (SMM). Three different scenarios were simulated using these parameters: the first one assumes migration from Peru to Floreana after the first colonization event (S1); in the second, there is no migration after the colonization event (S2); and the third one allows for migration in both directions (S3).

Summary statistics

A total of 10 summary statistics was used to capture different aspects of the data: number of alleles within each population (K_G and K_P); total number of alleles (K_tot); gene diversity within each population (H_G and H_P); total gene diversity (H_tot); allelic range in each population (R_G and R_P); total allelic range (R_tot); and finally F_{ST} as an estimator of genetic differentiation between the two pooled regions, the Galápagos (G) and Peru (P).

Comparison of the three scenarios

The relative probabilities of the three scenarios simulated with SIMCOAL were performed using a direct approach (Cornuet *et al.* 2008). This method takes the proportion $n\delta$ of data sets among the N simulated that are the closest to the observed data set. Here, N is equal to 3 000 000 simulations (1 000 000 per scenario), and we used three different values of $n\delta$ to ensure the robustness of the comparison: $n\delta$ = 10 000 (0.33%), 5000 (0.17%) and 1000 (0.03%).

Validation of the estimation procedure

Following Neuenschwander *et al.* (2008), the quality of our estimation was first assessed by examining the coefficient of determination (R^2) computed for each parameter over the first 10 000 simulations. This coefficient corresponds to the proportion of the parameter variance explained by summary statistics and reveals the potential of a given parameter to be correctly estimated. Still following Neuenschwander *et al.* (2008), we considered that parameters with coefficients of determination below 10% were unreliable.

The performance of our model was further validated using two arbitrarily fixed sets of parameters that differed by the value of colonization time (100 and 800 generations). The values of the remaining parameters were fixed to: Floreana initial size = 15; Floreana current size = 1000; Peru size = 10 000; migration rate per generation = 0.0045; mutation rate = 0.0015; geometric parameter of the GSM model = 0.15. These parameters were chosen because they fall within or not far from the ranges of parameters found from S1 to S3. For each colonization time (100 and 800), we generated 1000 pseudo-observed data sets based on the fixed parameters. The parameters were then estimated using 1 million simulations, and the precision of the estimation was evaluated relative to the known (true) values of the pseudo-observed data sets (Tables S1 and S2, Supporting information). This procedure allowed us to evaluate the power of the methodology used to estimate parameters and additionally to validate the most confident point estimator to be used (mean, median or mode of the posterior distribution). The performance evaluation corresponds to the assessment of the similarity between the 'true values' (averaged over the 1000 data sets) and the estimated values (based on the 1 million simulations). The quality of the point estimators was also assessed by computing several statistics: the relative bias, $\text{bias} = \frac{1}{\theta}\frac{1}{n}\sum_{i=1}^{n}\theta_i - \theta$; the relative root mean square error, $\text{RMSE} = \frac{1}{\theta}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \theta)^2}$, where θ_i is the estimator of the parameter θ and n is the number of test data sets (here 1000) and the factor 2 statistics (Excoffier *et al.* 2005b), which represents the proportion of the estimated values lying in the interval comprised between 50% and 200% of the 'true value'. Moreover, to assess the quality of the posterior distribution, the 50% and 90% coverage were also computed (which correspond to the proportion of simulations in which the 'true value' lies within the respective 50% and 90% credible intervals around the estimate). Note that the factor 2 statistics gives information on the absolute estimator precision, which is not provided by coverage properties.

Results

Molecular data

The three cp loci pooled together resulted in a fragment of 1525 bp long and four combined haplotypes. The HA

locus was monomorphic (GenBank-EF564430), and both SG and LF displayed two haplotypes each. The SG haplotypes corresponded to two variants of a polyT (haplotypes K and R, GenBank-EF564432 and EF564439), whereas the two LF haplotypes differed by a single substitution (haplotypes Z and Y, GenBank-EU234508 and EU234509). Two populations in Floreana displayed the four haplotypes, whereas the Inter Andean-1 and Española populations were monomorphic (Fig. 3). The haplotype KRZ was absent from Peru.

A total of 72 alleles were detected among the six microsatellite loci. With the exception of *Gspi*.B284, all loci were polymorphic, and the number of alleles varied between 10 and 22. Significant heterozygote deficits were observed in two populations, Pacific Coast-1 and Floreana-2 (F_{IS} = 0.188 and 0.246, respectively; $P < 0.001$). Among the 110 LD tests computed, significance for the corrected $P = 0.0095$ was found for seven pairs of loci (5.5%), six of them occurring within Floreana populations. We therefore assumed that the five polymorphic loci were unlinked.

Although the number of individuals sampled varied among populations (5–27 individuals/population), there was no relationship between sample size and average heterozygosity (Pearson's correlation test; $n = 11$, $r = 0.62$, $P = 0.52$). Levels of average heterozygosity ranged between 0.459 ± 0.253 (Española) and 0.823 ± 0.044 (Pacific Coast-3; Table 1). The highest diversities were reported in the Pacific Coast populations.

The most probable number of clusters inferred from the ΔK statistic was two. The two clusters were supported geographically, with the first one harbouring 90% of the Peruvian individuals and the second one including 81% of individuals from Galápagos (Fig. 4). The remaining individuals were either assigned to the other 'geographical' cluster (eleven Galápagos individuals were assigned to the Peru cluster, and two Peruvian individuals were assigned to the Galápagos cluster) or could just not be assigned at all, as they could equally belong to the one or the other cluster (six individuals from Peru and four from Galápagos).

According to the previous results, and despite the wrongly or unassigned individuals, two major geographical groups were considered, Peru and Galápagos. To identify potential founder effects in the islands, the expected heterozygosity (H_E) and the allelic richness (R_S) were computed for these two geographical groups (Table 6). Both measures were compared between the two groups with a one-tail paired Student's t test, and a significantly higher genetic diversity was observed in Peru (H_E, $P = 0.0222$ and R_S, $P = 0.0037$). Despite similar number of individuals within each cluster, 31 alleles were private to Peru, against only one in the Galápagos.

Among the 55 pairwise F_{ST} comparisons, 22% were nonsignificant ($P = 0.0109$; Table 2). No differentiation was found among the five Floreana populations, and Española reported the highest level of differentiation. The Inter Andean populations also appeared to be quite

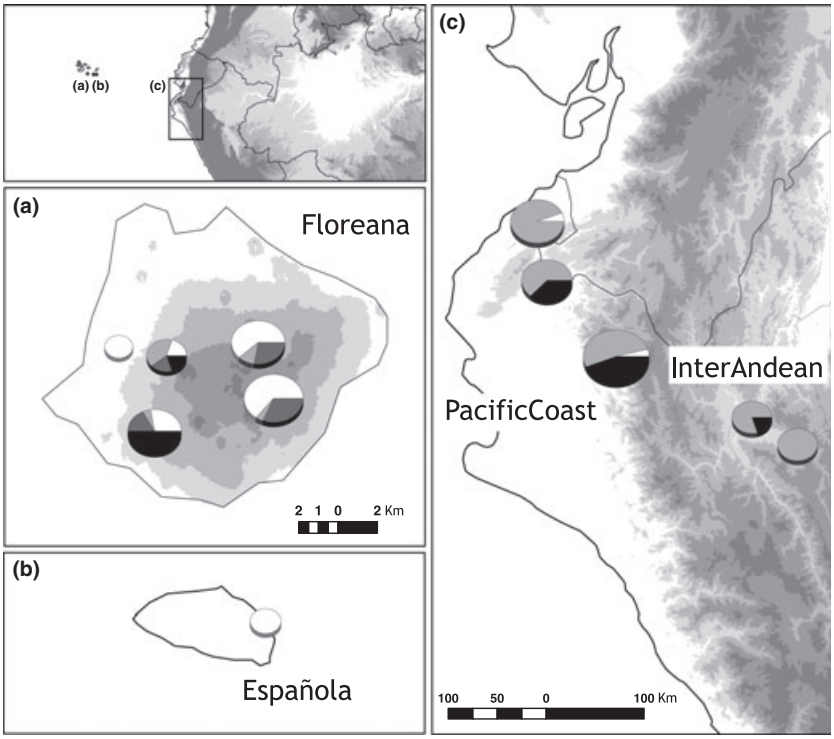

Fig. 3 Distribution of the four chloroplast combined haplotypes found in the eleven sampled populations of *Geoffroea spinosa*. Each circle corresponds to one population and different shadings were assigned to each haplotype: KKZ in white, KKY dark grey, KRZ light grey, and KRY in black. Only one haplotype (KKZ) is found on Española.

Table 1 Genetic diversity indices within eleven *Geoffroea spinosa* populations

	InterAndean-1	InterAndean-2	PacificCoast-1	PacificCoast-2	PacificCoast-3	Floreana-1	Floreana-2	Floreana-3	Floreana-4	Floreana-5	Española
N	10	10	18	27	16	5	18	18	10	22	6
Chloroplast											
HA											
K	10	10	18	27	16	5	18	18	10	22	6
SG											
K	10	8	18	15	10	5	13	6	6	15	6
R	–	2	–	12	6	–	5	12	4	7	–
LF											
Z	–	–	1	1	–	5	16	8	4	21	6
Y	10	10	17	26	16	–	2	10	6	1	–
Overall											
KKZ	–	–	1	1	–	5	11	5	2	14	6
KKY	10	8	17	14	10	–	2	1	4	1	–
KRZ	–	–	–	–	–	–	5	3	2	7	–
KRY	–	2	–	12	6	–	–	9	2	–	–
Microsatellites											
Gspi.I168											
N_A	9	9	8	11	12		6	6	3	6	3
R_S	5.7	6.4	4.4	4.5	6.3	4.0	4.2	3.9	2.9	3.7	2.8
H_O	0.800	0.900	0.444	0.481	0.625	0.400	0.278	0.333	0.700	0.636	0.500
H_E	0.832	0.889	0.671	0.669	0.887	0.733	0.759	0.725	0.616	0.691	0.621
F_{IS}	0.040	−0.013	0.345*	0.284*	0.302*	0.484*	0.641*	0.548*	−0.145	0.081	0.211
Gspi.A149											
N_A	3	6	7	7	7	3	5	5	5	4	2
R_S	3.0	5.2	5.2	4.8	5.3	3.0	3.5	3.4	3.8	3.1	2.0
H_O	0.600	0.900	0.611	0.815	0.750	0.800	0.556	0.667	0.500	0.500	0.833
H_E	0.668	0.847	0.838	0.820	0.845	0.600	0.678	0.646	0.621	0.663	0.530
F_{IS}	0.107	−0.066	0.277	0.006	0.115	−0.391	0.185	−0.033	0.204	0.250	−0.667
Gspi.B331											
N_A	3	2	8	14	8	6	8	6	6	9	2
R_S	2.7	2.0	5.5	6.5	5.2	6.0	5.3	4.6	5.1	5.4	2.0
H_O	0.500	0.600	0.722	0.815	0.938	1.000	0.778	0.722	0.900	0.818	0.167
H_E	0.484	0.442	0.843	0.890	0.825	0.889	0.806	0.806	0.826	0.840	0.409
F_{IS}	−0.034	−0.385	0.147	0.086	−0.142	−0.143	0.036	0.107	−0.095	0.027	0.615
Gspi.B458											
N_A	4	3	9	8	8	4	7	7	6	7	4
R_S	3.4	2.9	5.4	5.6	4.9	4.0	4.7	5.0	5.0	4.5	4.0
H_O	0.700	0.500	0.722	0.852	0.625	1.000	0.778	0.944	0.800	0.773	0.600
H_E	0.605	0.616	0.833	0.859	0.752	0.711	0.805	0.825	0.832	0.784	0.733
F_{IS}	−0.167	0.196	0.137	0.008	0.174	−0.481	0.034	−0.149	0.040	0.015	0.200

Table 1 *Continued*

	InterAndean-1	InterAndean-2	PacificCoast-1	PacificCoast-2	PacificCoast-3	Floreana-1	Floreana-2	Floreana-3	Floreana-4	Floreana-5	Española
Gspi.B264											
N_A	5	4	8	7	6	4	5	4	4	6	1
R_S	3.9	3.2	5.3	4.4	4.6	4.0	4.3	3.3	3.2	3.7	1.0
H_O	0.600	0.300	0.778	0.852	0.813	0.600	0.500	0.611	0.500	0.727	–
H_E	0.726	0.553	0.829	0.766	0.809	0.733	0.759	0.605	0.595	0.632	–
F_{IS}	0.182	0.471	0.063	−0.115	−0.005	0.200	0.348*	−0.011	0.167	−0.155	–
Overall											
N_A	4.8 ± 2.2	4.8 ± 2.5	8.0 ± 0.6	9.4 ± 2.7	8.2 ± 2.0	4.2 ± 1.0	6.2 ± 1.2	5.6 ± 1.0	4.8 ± 1.2	6.4 ± 1.6	2.4 ± 1.0
R_S	3.7 ± 1.2	3.9 ± 1.2	5.2 ± 0.4	5.2 ± 0.9	5.3 ± 0.6	4.2 ± 1.1	4.4 ± 0.6	4.0 ± 0.8	4.0 ± 1.0	4.1 ± 0.9	2.4 ± 1.1
H_O	0.640 ± 0.102	0.640 ± 0.233	0.656 ± 0.119	0.763 ± 0.142	0.750 ± 0.119	0.760 ± 0.233	0.578 ± 0.188	0.656 ± 0.197	0.680 ± 0.160	0.691 ± 0.113	0.420 ± 0.300
H_E	0.663 ± 0.116	0.670 ± 0.172	0.803 ± 0.066	0.801 ± 0.078	0.823 ± 0.044	0.733 ± 0.092	0.761 ± 0.047	0.722 ± 0.086	0.698 ± 0.107	0.722 ± 0.078	0.459 ± 0.253
F_{IS}	0.037	0.046	0.188*	0.048	0.092	−0.041	0.246*	0.094	0.027	0.044	0.094

N_A=number of alleles, R_S=allelic richness, H_O=observed heterozygosity, H_E=expected heterozygosity, F_{IS}=coefficient of heterozygote deficit, *=significance level based on the 5% confidence level.

distinct from the other ones. AMOVAS were performed independently considering two and three groups of populations: the Galápagos and the Peruvian populations that were either grouped together or divided into Inter Andean and Pacific Coast (Table 3). The analyses best supported the three-group arrangement, for which the regional differences accounted for 10% of the total variation (F_{CT} = 0.103; P = 0.0040 ± 0.0002; F_{SC} = 0.029; $P \leq 0.0001$). When considering populations divided into two groups, F_{CT} (0.064; P = 0.0014 ± 0.0004) was similar to F_{SC} (0.063; $P \leq 0.0001$), which implies that differences observed between populations within groups were equivalent to the differentiation between groups. Most of the variation within *Geoffroea spinosa* was attributed to differences between individuals within populations (Table 3) with F_{ST} ranging between 0.123 and 0.128 depending on the grouping used.

Performance evaluation and ABC estimations

The results of the performance tests are given as supplementary data (Tables S1 and S2, Supporting information for 100 and 800 generations, respectively). When compared to the results obtained with the mean and the median, the mode was clearly identified as the most reliable point estimator. Indeed, closer estimates and lower deviations to the 'true values' are found with the mode (mode, average bias = 32% for 100 generations and 9.5% for 800 generations; median, 71% for 100 and 60% for 800; mean, 123% for 100 and 144% for 800; Tables S1 and S2, Supporting information). Focusing on the mode, the relatively high RMSEs found for 100 and 800 generations ([0.237–1.430] and [0.250–0.915], respectively) suggested that the estimation of some parameters should be taken cautiously. In both cases, the GSM p-parameter seems to be the best estimated parameter (smallest bias, maximum 90% coverage, high Factor 2, but see below for a detailed comment on that result), and Floreana current size the worst estimated (highest bias and RMSE, minimum 90% coverage). According to Factor 2, the colonization time was well estimated with a value closer to the fixed prior for simulations at 100 generations than for the ones at 800 generations (Tables S1 and S2, Supporting information).

ABC simulations for the three scenarios

In all scenarios, two parameters are consistently badly estimated: the GSM p-parameter with fraction of the variance explained (R^2) <1% and the current size of Floreana with values of R^2 ranging between 3% and 9%, in keeping with the performance evaluation. Because there is almost no information about GSM p-parameter, the posterior distribution is very similar to the prior

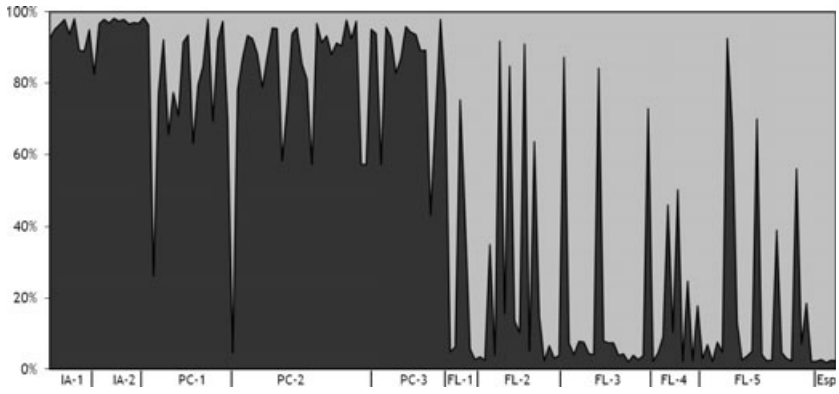

Fig. 4 Summary plot of estimates of Q. Each individual is represented by a single vertical peak separated into two colored segments, with lengths proportional to each of the two inferred clusters. The codes IA-1, Ia-2, PC1, PC-2, PC-3, FL-1, FL-2, FL-3, FL-4, FL-5 and Esp respectively correspond to the predefined population samples InterAndean1-2, PacificCoast1-3, Floreana1-5 and Española.

(data not shown), and consequently its mean and median are also close to the middle value of the prior distribution. This was confirmed by simulating various values of the GSM p-parameter (not shown) that does not change the mode (hereafter designated as *M*), mean and median values of its posterior distribution.

In S1, four parameters out of seven are assumed to be satisfactorily estimated ($R^2 > 10\%$): mutation rate (M = 0.0003, R^2 = 56.5%), Peru size (M = 19 533, R^2 = 29.2%), migration rate from Peru to Floreana (M = 0.0014, R^2 = 21.4%) and Floreana initial size (M = 23, R^2 = 18.1%). The colonization time is less satisfactorily estimated (M = 475, R^2 = 8.9%). In S2, the best estimated parameter is the colonization time (M = 76, R^2 = 75.3%), followed by the mutation rate (M = 0.0004, R^2 = 57.6%), the Peru size (M = 15 410, R^2 = 32.0%) and the Floreana initial size (M = 77, R^2 = 21.0%). In S3, all parameters excepting the Floreana current size and the GSM P-parameter are satisfactorily estimated. The mutation rate and the colonization time are the best estimated parameters (M = 0.0006, R^2 = 37.8% and M = 721, R^2 = 31.5%, respectively), followed by the Floreana initial size, the Peru size and the two migration rates ($11.0\% < R^2 < 23.0\%$, Table 4).

As shown in Table 4 and Fig. 5, the estimation of the colonization time is very recent with all three alternative scenarios. This result points to a colonization that occurred between 76 and 721 generations ago (from 1900 to 18 000 years ago, using the very conservative maximum generation time of 25 years). Such a window of time corresponds to very narrow posterior distributions of the parameter, when compared to the prior range of 1–60 000 generations. Scenario S2 is the one for which the coefficient of determination R^2 is the highest for colonization time (R^2 = 75.3%). This is also the scenario that gives the most recent estimate (M = 76, with the 90% Bayesian Credible Interval (CI) equal to [1–542]; Table 4). The scenario with the worst estimate for the colonization time (S1; R^2 = 8.9%) gives an intermediate estimation of 475 generations (90% CI equal to [1–16 208]; Table 4). In all cases, the more recent times within the 90% CI intervals seem the most likely, as clearly shown by the posterior distributions for the three scenarios (Fig. 5) and the point estimates.

Assuming a constant rate of gene flow since the first colonization of the island (S1), the estimated migration rate indicated that each Floreana individual has a probability of 1.4‰ of being issued from Peru at each generation (Table 4). When assuming constant bidirectional gene flow since the first colonization event (S3), the estimations showed that the migration rate was slightly higher from Floreana to Peru (1.1‰ probability that a Peruvian individual is issued from Floreana) than from Peru to Floreana (0.7‰ probability for the reverse), although the two figures are not significantly different.

Comparison of the three scenarios

The direct approach favoured the two scenarios that included gene flow (S1 and S3; Table 5). With the most stringent conditions ($n\delta$ = 1000), S1 (unidirectional gene flow from Peru to Floreana) gave nearly half of the retained simulations among $n\delta$, whereas with less stringent conditions ($n\delta$=10 000), it is S3 (bidirectional gene flow) that seems more probable. Scenario S2, which gives the best estimated time of colonization, is therefore the least likely but cannot be excluded however, as a proportion of 15–19% of the retained simulations were obtained using this model.

Discussion

Genetic differentiation and founder dynamics

The microsatellite-based analyses show strong genetic differentiation between Peruvian and the Galápagos

Table 2 Pairwise F_{ST} values among eleven *Geoffroea spinosa* populations. The significance level (α=0.05) was corrected using the modified false discovery rate method to p=0.0109.

	InterAndean-1	InterAndean-2	PacificCoast-1	PacificCoast-2	PacificCoast-3	Floreana-1	Floreana-2	Floreana-3	Floreana-4	Floreana-5	Española
InterAndean-1	0.000										
InterAndean-2	0.067*	0.000									
PacificCoast-1	0.191*	0.176*	0.000								
PacificCoast-2	0.164*	0.149*	0.008	0.000							
PacificCoast-3	0.123*	0.136*	0.037*	0.039*	0.000						
Floreana-1	0.116*	0.169*	0.071*	0.072*	0.058*	0.000					
Floreana-2	0.175*	0.209*	0.069*	0.075*	0.067*	−0.011	0.000				
Floreana-3	0.195*	0.232*	0.067*	0.069*	0.084*	−0.007	0.004	0.000			
Floreana-4	0.200*	0.240*	0.066*	0.073*	0.094*	−0.004	0.045	0.017	0.000		
Floreana-5	0.195*	0.225*	0.075*	0.081*	0.097*	−0.016	0.007	0.006	0.005	0.000	
Española	0.340*	0.373*	0.218*	0.219*	0.220*	0.118*	0.076	0.116*	0.155*	0.096*	0.000

populations, most likely due to a limited gene flow between the mainland and the islands. The two genetic clusters obtained with STRUCTURE are geographically well supported, with only few individuals from the Galápagos being assigned to the Peruvian group and the reverse for two individuals from coastal Peruvian populations (Fig. 4). A second level of differentiation appears when considering the pairwise F_{ST} values, which separate the mainland populations into Inter Andean and Pacific Coast geographical regions (0.123–0.191; Table 2). Accordingly, the AMOVA supports well the separation of the populations into three groups (Galápagos–Inter Andes–Pacific Coast), with regional differences accounting for 10% of the total variation (Table 3). The differentiation of these three groups of populations is also established with cp markers, which show a private haplotype in the Galápagos (KRZ) and the absence of the haplotype KKZ in the Inter Andean populations (Fig. 3). The three groups of populations identified above match the biogeographical regions delimited by two major barriers: the approximately 900 km sea distance that isolates the Galápagos from the continent and the Andes that separate the Inter Andes from the Pacific Coast.

Nuclear diversity values indicate that the Galápagos populations had undergone greater genetic drift than mainland populations. Microsatellite analyses support the idea of founder events having occurred at the time of the colonization of the islands as five of the six significant pairwise LD coefficients between unlinked markers are reported within Floreana. In addition, the theory predicts that whenever new populations are founded from a reduced number of individuals, the number of alleles is more drastically reduced than the heterozygosity (gene diversity) level (Nei *et al.* 1975), and Table 6 indeed shows an average decrease of 44% for the allelic richness within the Galápagos, against a loss of only 13% for the expected heterozygosity. This founder effect is most probably related to the colonization process, the restricted gene flow between source continental populations and the islands and the low effective sizes at the origin of the island populations. Española seems to have experienced a more drastic or recent founder event, as reflected by the low diversity indices (Table 1) and its high pairwise F_{ST} values recorded with the other populations (Table 2). This result matches the observed increased population differentiation after a strong bottleneck (Wade & McCauley 1988). The few other studies that compared islands and continental populations in plant species also revealed reduced genetic diversities in the islands (Ledig & Conkle 1983; Glover & Barrett 1987; Affre *et al.* 1997; Rivera-Ocasio *et al.* 2002).

Conversely, the highest cp diversity is found in Floreana. Four populations of the island display one

Table 3 AMOVA results for *Geoffroea spinosa* populations divided into two and three groups. The probability of obtaining a more extreme estimate by chance was determined by 10000 permutations.

Source of variation	d.f.	SSD	Variance component	% Total	*P*-value
Peru vs. Galápagos					
Between regions	1	27.86	0.1354	6.4	0.0014 ± 0.0004
Among populations within regions	9	47.71	0.1240	5.9	<0.0001
Among individuals within populations	309	573.47	1.8559	87.7	<0.0001
InterAndean vs. PacificCoast vs. Galápagos					
Between regions	2	48.47	0.2176	10.2	0.0004 ± 0.0002
Among populations within regions	8	27.10	0.0546	2.6	<0.0001
Among individuals within populations	309	573.47	1.8559	87.2	<0.0001

Table 4 Estimated parameters of the colonization history of Floreana Island by *Geoffroea spinosa*, following three scenario. Three point estimates are listed (mean, median and mode), as well as the 50 and 90% Bayesian credible intervals and the fraction of variance explained (R^2).

	Point estimator			Credible interval		
Parameters	Mode	Mean	Median	50%	90%	R^2
Colonization time	475	6052	2129	[1–2190]	[1–16 208]	0.0887
Floreana initial size	23	123	54	[1–55]	[1–345]	0.1808
Floreana current size	855	1841	1499	[500–1503]	[505–4418]	0.0897
Migration rate	0.0014	0.0035	0.0029	[0.0000–0.0100]	[0.0000–0.0100]	0.2135
Mutation rate	0.0003	0.0008	0.0006	[0.0001–0.0039]	[0.0001–0.0039]	0.5645
GSM P-parameter	0.0342	0.1398	0.1327	[0.0001–0.2993]	[0.0001–0.2993]	0.0003
Peru size	19 533	46 169	41 749	[10 740–42 591]	[10 022–83 505]	0.2916
	Point estimator			Credible interval		
Parameters	Mode	Mean	Median	50%	90%	R^2
Colonization time	76	239	139	[1–141]	[1–542]	0.7531
Floreana initial size	77	405	335	[1–894]	[1–970]	0.2097
Floreana current size	4512	2905	2958	[2419–4809]	[1019–4876]	0.0309
Mutation rate	0.0004	0.0011	0.0008	[0.0001–0.0043]	[0.0001–0.0043]	0.5760
GSM P-parameter	0.0684	0.1393	0.1354	[0.0004–0.2998]	[0.0004–0.2998]	0.0004
Peru size	15 410	31 061	23 971	[10 023–24 065]	[10 023–62 572]	0.3203
	Point estimator			Credible interval		
Parameters	Mode	Mean	Median	50%	90%	R^2
Colonization time	721	8416	2839	[1–2880]	[1–27 231]	0.3148
Floreana initial size	18	100	41	[1–42]	[1–283]	0.2267
Floreana current size	788	1809	1371	[500–1373]	[500–3728]	0.0474
Migration rate (F to P)	0.0011	0.0033	0.0027	[0.0000–0.0100]	[0.0000–0.0100]	0.1104
Migration rate (P to F)	0.0007	0.0021	0.0014	[0.0000–0.0099]	[0.0000–0.0099]	0.1169
Mutation rate	0.0006	0.0017	0.0013	[0.0001–0.0100]	[0.0001–0.0100]	0.3781
GSM P-parameter	0.2636	0.1578	0.1630	[0.1508–0.2840]	[0.0284–0.2873]	0.0002
Peru size	18 018	44 255	38 438	[10 000–38 466]	[10 000–92 065]	0.1351

combined haplotype that is missing from the mainland populations (KRZ). It should be noted that this haplotype is characterized by a mutation in a polyT also found in the mainland, but in a different combination. PolyT are known to be more subject to homoplasy than substitutions, because they evolve at higher rates, with a specific addition and deletion pattern (Provan *et al.* 2001; Ingvarsson *et al.* 2003). The existence of the KRZ haplotype in the Galápagos can be seen either as a sign of a sufficiently

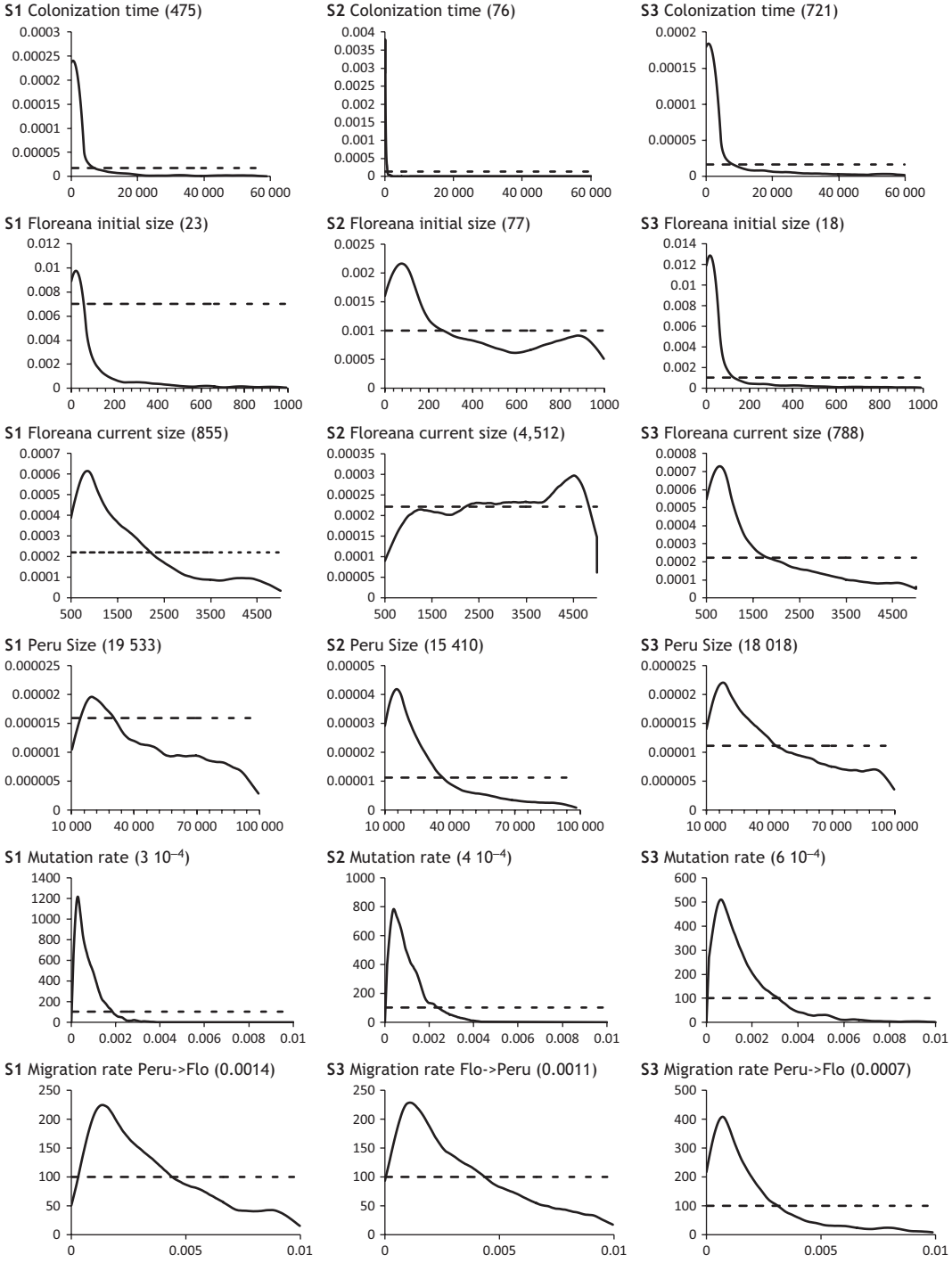

Fig. 5 Posterior (solid line) and prior (dashed line) density curves for the six parameters used to simulate the colonization of the Galápagos by *Geoffroea spinosa* with scenario S1, S2 and S3. For each parameter, the *x*-axis extends over the full range of its prior distribution. The modal value of each estimated parameter is provided between brackets.

ancient settlement for a new PolyT mutation to appear or to the effect of a founder event in the islands that would have increased the frequency of this haplotype in the Galápagos compared to the mainland.

Our nuclear results clearly indicate a deep genetic divergence between the two regions on both sides of the Andean divide (Tables 2 and 3). Judging from the shared cp haplotypes in both Inter Andean and Pacific Coast regions (Fig. 3), the isolation detected with

Table 5 Comparison of the three scenarios using the direct approach, with different values of nδ, the number of retained simulations among 3 million ones (1 million per scenario).

	Nδ		
Scenario	10 000	5000	1000
S1 (migration P → F)	0.35	0.40	0.48
S2 (no migration)	0.15	0.16	0.19
S3 (migration P↔F)	0.50	0.44	0.34

Table 6 Genetic diversity indices of *Geoffroea spinosa* in Peru (81 individuals) and Galapagos (79 individuals), based on microsatellite data.

Locus name	**Peru**	**Galápagos**
Gspi.I168		
N_{PA}	10	–
H_E	0.849	0.723
R_S	20.8	10.0
Gspi.A149		
N_{PA}	6	–
H_E	0.844	0.641
R_S	10.9	5.0
Gspi.B331		
N_{PA}	6	1
H_E	0.840	0.831
R_S	14.9	10.0
Gspi.B458		
N_{PA}	5	–
H_E	0.837	0.800
R_S	13.0	8.0
Gspi.B264		
N_{PA}	4	–
H_E	0.814	0.630
R_S	10.0	6.0

N_{PA}=number of private alleles, H_E=expected heterozygosity, R_S=allelic richness.

microsatellites seems, however, relatively recent. This is probably due to the Andes in northern Peru and southern Ecuador being at their lowest elevation along the entire cordillera in South America (*c.* 2500 m). Whilst this area, the 'Huancabamba Depression', has been proposed being a dispersal barrier for highland taxa, it does not offer much of a dispersal barrier for lowland species. Support for the cross-Andean dispersal in the Huancabamba-Amotape region has already been inferred for the palm tree *Ceroxylum echinulatum* (Trénel *et al.* 2008). In the case of *Geoffroea spinosa*, cross-Andean dispersal is particularly plausible. Indeed, the seasonally dry tropical forest in which *G. spinosa* grows occurs on either side of the mountains in some areas of Peru and it is separated by only a narrow belt of mesic cloud forest at the highest elevations (Bridgewater *et al.* 2003).

Comments on the ABC model

By definition, a model is a simplified representation of reality that includes and combines multiple processes in a quantitative way. Because knowledge about these processes is often incomplete, models are necessarily imperfect and therefore subject to critical appraisal. Adequate historical information is fundamental to elaborate accurate scenarios (Estoup *et al.* 2001; Estoup & Clegg 2003), but such information is not available so far for *G. spinosa*. Although the ABC method allows the analysis of complex demographic scenarios (e.g. Estoup *et al.* 2004), more complex models do not necessarily mean better models if there are large uncertainties on additional parameters (Wakeley 2004). In this line, scenarios with and without gene flow were tested, and those allowing for gene flow between Peru and Floreana produced, on average, a higher proportion of simulations close to the real data. Finally, the genetic data that supply the simulations have a great influence on the precision of the estimations. Genetic data can now be produced at an unprecedented scale without much difficulty, but gathering such huge data sets remains expensive. The data set used in the present study was based on five microsatellites, although additional loci were tested but with unsuccessful results. We are however aware that increasing the number of markers could improve the estimations of some parameters. Despite this drawback, the parameter of interest in our model, namely the colonization timing of Floreana, appears very well estimated, and the ABC inference procedure therefore provides new and valuable insights into the history of *G. spinosa* in the Galápagos.

Colonization of the Galápagos

The current study presents strong evidence for the recent establishment of *G. spinosa* in the Galápagos Islands. Using an ABC approach, we find that the colonization of Floreana occurred within a range of 76–721 generations ago, with the 90% CI for the three tested scenarios that overlapped over the first 542 generations. This low number of generations was moreover confirmed by IMa2 analyses using an MCMC approach (Hey, 2005; Hey & Nielsen, 2007). Indeed, the colonization time was estimated to have occurred 267 generations ago, which falls within the range time found with ABC simulations (Appendix S2, Table S3 and Fig. S1, Supporting information). The estimation of the current size of Floreana is furthermore concordant with the one found with Scenario S1 (728 vs. 855 individuals, respectively), and the migration rates obtained were lower (m = 0.00002 from Peru to Floreana and no migration for the reverse). Using IMa2, Peru current size was

estimated to be lower than with ABC (5095 vs. 15 533 individuals).

Depending on which generation time is used, 10 years for the first flowering trees in specific conditions (R.T. Pennington, personal observation; A. Tye, personal communication) or 25 years for the mean flowering time, ABC estimates of colonization time translate into a period lying between 760 and 7200 years ago or between 1900 and 18 025 years ago, respectively. This does not fundamentally change the main conclusion of a recent colonization time of the Galápagos. We indeed emphasize that our model assumes an enormous prior range for the colonization time (which corresponds to the island's age; White *et al.* 1993), and the ABC method therefore helps reducing the probable period of time for island colonization. With both the most, but not exclusively, supported ABC scenario (S1) and IMa2, the point estimates of the colonization period are very close (475 and 267 generations, respectively) and they do not exclude that the colonization might have occurred even more recently.

Converting this ABC result into the true story of *G. spinosa* in the Galápagos is not straightforward, and two, nonmutually exclusive scenarios are possible. The Galápagos archipelago sits at the confluence of an important system of coastal currents that could have served as natural dispersion routes out of the American continent (Lea *et al.* 2006). *G. spinosa* fruits can float and their seed is surrounded by an indehiscent woody endocarp. Seed survival after a long period in salt water is currently unknown. However, Vargas *et al.* (2012) recently considered *G. spinosa* to be a native Galápagos species, emphasizing that many species have successfully colonized the Galápagos without any specific ability for long-distance dispersal. In any case, a natural colonization of the Galápagos, if it ever occurred, seems to have been rare and stochastic, as illustrated by the presence of the species on only Floreana and Española, despite the availability of suitable habitat on other islands of the archipelago. A distribution restricted to a small number of islands is however a common rule in the Galápagos, and very few species are found throughout the Archipelago (A. Tye, personal communication). A detailed colonization scenario with long-distance oceanic dispersal would necessarily take into account factors such as the changing conditions in the ocean circulation system over the past 160 000 years (Rahmstorf 2002; Voelker 2002). The El Niño/Southern Oscillation (ENSO) is one of the most studied events, and it has major effects on oceanic dynamics within the Pacific basin (Holmgren *et al.* 2001). Interestingly, climate models suggest an increase in strength and frequency of ENSO in the late Holocene (during the last approximately 3000 years ago; Riedinger *et al.* 2002), which coincides quite well with the date inferred for the most probable arrival date of *G. spinosa* in Floreana. For the two scenarios (S1 and S3) that include gene flow, we find very low levels of gene flow with CI including zero between Floreana and Peru, suggesting an almost complete isolation of Floreana after the colonization. Similarly, very low levels of gene flow were confirmed using IMa2 (Appendix S1, Supporting information).

A second explanation compatible with our results is that *G. spinosa* was introduced to the Galápagos by man. This hypothesis would better fit with S2 for which no migration was simulated after the first colonization as it could have happened if the human introduction occurred once. Interestingly, this is the scenario for which the more recent time is estimated (76 generations ago [1–141]), a time compatible with humans' history in Floreana. Floreana's history is often associated with whalers, pirates, convicts and immigrants, even before the permanent human settlement in the 18th century (McMullen 1999), but nothing is known about previous 'non-European' settlements. Moreover, within Floreana, *G. spinosa* occurs as large patches of 50–100 trees, often near farms, especially the two populations located in the centre of the island that harbour the four cp haplotypes. It is difficult to explain why *G. spinosa* does occur on the islands as monospecific stands, unlike other native trees. This situation might be the result of an extensive planting by the first human colonists. Moreover, a human introduction scenario is more compatible with the large number of chloroplast haplotypes found on Floreana, as this high diversity is difficult to explain by a recent natural colonization from a small number of seeds. Similar arguments have been proposed for the likely introduced cork oak (*Quercus suber*, Lorenzo *et al.* 2009) in Minorca and for the introduced *Ligustrum robustum* in the Mascarene Islands (Milne & Abbott 2004). Finally, *G. spinosa* has additional properties that support a man-mediated introduction scenario, as it can be used as timber and its fruits are used for animal feeding and are occasionally eaten by people (Sanchez *et al.* 2006).

A way to definitely assess the introduced (or native) status of the species would be to use pollen records as was done for other problematic species of the Galápagos Archipelago (Coffey *et al.* 2011), the Azores Islands (van Leeuwen *et al.* 2005) or the Iberian Peninsula (Bagnoli *et al.* 2009). By now, *G. spinosa* is far from being as invasive as other recently introduced trees (e.g. *Cinchona pubescens*; Jager *et al.* 2007 or *Psidium guajava*; Walsh *et al.* 2008), but nothing is known about its capacity to compete and threaten endemic and native plants. The species is in fact missing from published guides of the Galápagos flora (e.g. McMullen 1999), as well as from the published lists of introduced species

(Pacific Island Ecosystems at Risk (PIER), US Forest Service). This lack of information needs to be promptly remedied, and the potential alien status of *G. spinosa* in the Galápagos deserves further attention.

Acknowledgements

We wish to thank the Darwin's Foundation and the Galápagos National Park for the facilities provided in the Galápagos, Dr. A. Tye for the information he provided on *Geoffroea spinosa* in the Galápagos and the help for the permit procedures, the Galápagos Diving for the transport, the Family Cruz for the accommodation in Floreana, E. Rosero and A. Daza for their help in sampling in Floreana and Peru, respectively, L. Schneider and R. Niba for help at the CJBG laboratory, Dr. N. Wyler for the map and Drs. S. Neuenschwander and N. Ray for sharing parts of programs and scripts. This work was supported by the Swiss National Foundation, grants 3100A0/100806-1 & 2 to YN and 31003A-112651 to MC and by the Conservatoire & Jardin Botaniques of Geneva (CJBG). Plant material was collected and exported to the CJBG laboratory in agreement with each country's law in force at the time.

References

Affre L, Thompson J, Debussche M (1997) Genetic structure of continental and island populations of the Mediterranean endemic *Cyclamen balearum* (Primulaceae). *American Journal of Botany*, **84**, 437–451.

Austerlitz F, Mariette S, Machon N, Gouyon P-H, Godelle B (2000) Effects of colonization processes on genetic diversity: differences between annual plants and tree species. *Genetics*, **154**, 1309–1321.

Bagnoli F, Vendramin GG, Buonamici A *et al.* (2009) Is *Cupressus sempervirens* native in Italy? An answer from genetic and paleobotanical data. *Molecular Ecology*, **18**, 2276–2286.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Beheregaray LB, Gibbs JP, Havill N *et al.* (2004) Giant tortoises are not so slow: rapid diversification and biogeographic consensus in the Galápagos. *Proceedings of the National Academy of Sciences United States of America*, **101**, 6514–6519.

Benjamini Y, Yekutieli D (2001) The control of false discovery rate dependency. *Annals of Statistics*, **29**, 1165–1188.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, **19**, 2609–2625.

Bisconti M, Landini W, Bianucci G *et al.* (2001) Biogeographic relationships of the Galápagos terrestrial biota: parsimony analyses of endemicity based on reptiles, land birds and *Scalesia* land plants. *Journal of Biogeography*, **28**, 495–510.

Bridgewater S, Pennington RT, Reynel CA *et al.* (2003) A preliminary floristic and phytogeographic analysis of the woody flora of seasonally dry forests in northern Peru. *Candollea*, **58**, 129–148.

Caetano S (2008) Insights on the history of Seasonally Dry Tropical Forests in South America: Inferences from the genetic structure of the trees *Astronium urundeuva* (Anacardiaceae) and *Geoffroea spinosa* (Fabaceae). PhD Thesis no 3946 of the University of Geneva, 268 p.

Caetano S, Naciri Y (2011) The biogeography of seasonally dry tropical forest in South America. In: *Seasonally Dry Tropical Forests: Ecology and Conservation* (eds Dirzo R, Young HS, Mooney HA, Ceballos G), pp. 23–44. Island Press, Stanford, CA.

Carlquist S (1966) The biota of long-distance dispersal. IV. Genetic systems in the floras of oceanic islands. *Evolution*, **20**, 433–455.

Carlquist S (1967) The biota of long-distance dispersal. V. Plant dispersal to Pacific Islands. *Bulletin of the Torrey Botanical Club*, **94**, 129–162.

Caujapé-Castells J, Tye A, Crawford DJ *et al.* (2010) Conservation of oceanic island floras: present and future global challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, **12**, 102–129.

Chan Y, Anderson C, Hadly E (2006) Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *PLoS Genetics*, **2**, 451–460.

Coffey EED, Froyd C, Willis KJ (2011) When is an invasive not an invasive? Macrofossil evidence of doubtful native plant species in the Galápagos Islands. *Ecology*, **92**, 805–812.

Cornuet J-M, Santos F, Beaumont MA *et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

de Queiroz A (2005) The resurrection of oceanic dispersal in historical biogeography. *Trends in Ecology and Evolution*, **20**, 68–73.

di Rienzo A, Peterson AC, Garza JC *et al.* (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences United States of America*, **91**, 3166–3170.

Estoup A, Clegg SM (2003) Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*. *Molecular Ecology*, **12**, 657–674.

Estoup A, Wilson I, Sullivan C *et al.* (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.

Estoup A, Beaumont M, Sennedot F *et al.* (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad *Bufo marinus*. *Evolution*, **58**, 2021–2036.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

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

Excoffier L, Laval G, Schneider S (2005a) Arlequin, ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Excoffier L, Estoup A, Cornuet J-M (2005b) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.

François O, Blum MG, Jakobsson M, Rosemberg NA (2008) Demographic history of european populations of *Arabidopsis thaliana*. *PLoS Genetics*, **4**, e1000075.

Friesen VL, González JA, Cruz-Delgado F (2006) Population genetic structure and conservation of the Galápagos petrel (*Pterodroma phaeopygia*). *Conservation Genetics*, **7**, 105–115.

Glover D, Barrett S (1987) Genetic variation on continental and island populations of *Eichhornia paniculata* (Pontederiaceae). *Heredity*, **59**, 7–17.

Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3) Available from http://www.unil.ch/izea/softwares/fstat.htlm.

Grehan J (2001) Biogeography and evolution of the Galápagos: integration of the biological and geological evidence. *Biological Journal of the Linnean Society*, **74**, 267–287.

Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hamilton MB (1999) Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. *Molecular Ecology*, **8**, 513–525.

Holmgren M, Scheffer M, Ezcurra E *et al.* (2001) El Niño effects on the dynamics of terrestrial ecosystems. *Trends in Ecology and Evolution*, **16**, 89–94.

Ingvarsson PK, Ribstein S, Taylor DR (2003) Molecular evolution of insertions and deletions in the chloroplast genome of *Silene*. *Molecular Biology and Evolution*, **20**, 1737–1740.

Ireland H, Pennington RT (1999) A revision of *Geoffroea* (Leguminosae-Papilionoideae). *Edinburgh Journal of Botany*, **56**, 329–347.

Jager H, Tye A, Kowarik I (2007) Tree invasion in naturally treeless environments: Impacts of quinine (*Cinchona pubescens*) trees on native vegetation in Galápagos. *Biological Conservation*, **140**, 297–307.

Kizirian D, Trager A, Donnelly MA, Wright JW (2004) Evolution of Galápagos island lava lizards (Iguania: Tropiduridae: *Microlophus*). *Molecular Phylogenetics and Evolution*, **32**, 761–769.

Kurz MD, Geist D (1999) Dynamics of the Galápagos hotspot from helium isotope geochemistry. *Geochimica et Cosmochimica Acta*, **63**, 4139–4156.

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics Applications Note*, **20**, 2485–2487.

Lea DW, Pak DK, Belanger CL *et al.* (2006) Paleoclimate history of Galápagos surface waters over the last 135,000yr. *Quaternary Science Reviews*, **25**, 1152–1167.

Ledig F, Conkle M (1983) Gene diversity and genetic structure in a narrow endemic Torrey pine (*Pinus torreyana* Parry ex Carr). *Evolution*, **37**, 70–85.

Linares-Palomino R (2006) Phytogeography and floristics of seasonally dry tropical forests in Peru. In: *Neotropical Savannas and Dry Forests: Plant Diversity, Biogeography, and Conservation* (eds Pennington RT, Lewis GP, Ratter JA), pp. 417–432. CRC Press, New York City, New York.

Lorenzo Z, Burgarella C, López de Heredia U *et al.* (2009) Relevance of genetics for conservation policies: the case of Minorcan cork oaks. *Annals of Botany*, **104**, 1069–1076.

Lucena R, Albuquerque U, Monteiro J *et al.* (2007) Useful plants of the semi-arid northeastern region of Brazil – a look at their conservation and sustainable use. *Environmental Monitoring and Assessment*, **125**, 281–290.

Martín-Bravo S, Vargas P, Luceño M (2009) Is *Oligomeris* (Resedaceae) indigenous to North America? Molecular evidence for a natural colonization from the Old World. *American Journal of Botany*, **96**, 507–518.

Mauchamp A (1997) Threats from alien plant species in the Galápagos Islands. *Conservation Biology*, **11**, 260–263.

McMullen C (1999) *Flowering Plants of the Galápagos*. Cornell University Press, New York City, New York.

Milinkovitch M, Monteyne D, Russello M *et al.* (2007) Giant Galápagos tortoises; molecular genetic analyses identify a trans-island hybrid in a repatriation program of an endangered taxon. *BioMed Central Ecology*, **7**, 2, doi: 10.1186/1472-6785-7-2.

Milne RI, Abbott RJ (2004) Geographic origin and taxonomic status of the invasive Privet, *Ligustrum robustum* (Oleaceae), in the Mascarene Islands, determined by chloroplast DNA and RAPDs. *Heredity*, **92**, 78–87.

Moore M, Tye A, Jansen R (2006) Patterns of long-distance dispersal in *Tiquilia* subg. *tiquilia* (Boraginaceae): implications for the origins of amphitropical disjuncts and Galápagos Islands endemics. *American Journal of Botany*, **93**, 1163–1177.

Naciri-Graven Y, Caetano S, Prado DE, Pennington RT, Spichiger R (2005) Development and characterization of 11 microsatellite markers in a widespread Neotropical seasonally dry forest tree species, *Geoffroea spinosa* Jacq. (Leguminosae). *Molecular Ecology Notes*, **5**, 542–545.

Nascimento CES, Rodal MJN, Cavalcanti AC (2003) Phytosociology of the remaining xerophytic woodland associated to an environmental gradient at the banks of the São Francisco river – Petrolina, Pernambuco, Brazil. *Revista Brasileira de Botânica*, **26**, 271–287.

Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. *Evolution*, **29**, 1–10.

Neuenschwander S, Largiadèr C, Ray N, Currat M, Vonlanthen P, Excoffier L (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Molecular Ecology*, **17**, 757–772.

O'Connell LM, Ritland K (2004) Somatic mutations at microsatellite loci in western redcedar (*Thuja plicata*: Cupressaceae). *Journal of Heredity*, **95**, 172–176.

Pacific Island Ecosystems at Risk (PIER). US Forest Service, Available from http://www.hear.org/pier/ accessed on 18.05.2010.

Porter D (1976) Geography and dispersal of the Galápagos Islands vascular plants. *Nature*, **264**, 745–746.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology and Evolution*, **16**, 142–147.

Rahmstorf S (2002) Ocean circulation and climate during the past 120,000 years. *Nature*, **419**, 207–214.

Rassman K (1997) Evolutionary age of the Galápagos iguanas predates the age of the present Galápagos Islands. *Molecular Phylogenetics and Evolution*, **7**, 158–172.

Riedinger M, Steinitz-Kannan M, Last W, Brenner M (2002) A ~6100 ^{14}C yr record of El Niño activity from the Galápagos Islands. *Journal of Paleolimnology*, **27**, 1–7.

Rivera-Ocasio E, Aide T, McMillan W (2002) Patterns of genetic diversity and biogeographical history of the tropical wetland tree, *Pterocarpus officinalis* (Jacq.), in the Caribbean basin. *Molecular Ecology*, **11**, 675–683.

Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the genus *Zea*. *Genetics*, **181**, 1397–1409.

Sanchez O, Aguirre Z, Kvist L (2006) Timber and non-timber uses of dry forests in Loja Province. *Lyonia*, **10**, 73–82.

Sato A, O'hUigin C, Figueroa F *et al.* (1999) Phylogeny of Darwin's finches as revealed by mtDNA sequences. *Proceedings of the National Academy of Sciences United States of America*, **96**, 5101–5106.

Sato A, Tichy H, O'hUigin C *et al.* (2001) On the origin of Darwin's finches. *Molecular Biology and Evolution*, **18**, 299–311.

Schilling EE, Panero JL, Eliasson UH (1994) Evidence from chloroplast DNA restriction site analysis on the relationships of *Scalesia* (Asteraceae: Heliantheae). *American Journal of Botany*, **81**, 248–254.

Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Molecular Biology*, **17**, 1105–1109.

Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Thuillet A-C, Bru D, David J *et al.* (2002) Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, *Triticum turgidum* (L.) Thell. ssp *durum* Desf. *Molecular Biology and Evolution*, **19**, 122–125.

Trénel P, Hansen M, Normand S, Borchsenius F (2008) Landscape genetics, historical isolation and cross-Andean gene flow in the wax palm, *Ceroxylom echinulatum* (Arecaceae). *Molecular Ecology*, **17**, 3528–3540.

Udupa S, Baum M (2001) High mutation rate and mutational bias at $(TAA)_n$ microsatellite loci in chickpea (*Cicer arietinum* L.). *Molecular Genetics and Genomics*, **265**, 1097–1103.

van Leeuwen JFN, Schäfer H, van der Knaap WO, Rittenour T, Björck S, Ammann B (2005) Native or introduced? Fossil pollen and spores may say. An example from the Azores Islands. *Neobiota*, **6**, 27–34.

Vargas P, Heleno R, Traveset A, Nogales M (2012) Colonization of the Galápagos Islands by plants with no specific syndromes for long-distance dispersal: a new perspective. *Ecology*, **35**, 33–43.

Vigouroux Y, Jaqueth JS, Matsuoka Y *et al.* (2002) Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution*, **19**, 1251–1260.

Voelker AHL (2002) Global distribution of centennial-scale records for Marine Isotope Stage (MIS) 3: a database. *Quaternary Science Reviews*, **21**, 1185–1212.

Wade M, McCauley D (1988) Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution*, **42**, 995–1005.

Wakeley J (2004) Recent trends in population genetics: More data! More Math! Simple models? *Journal of Heredity*, **95**, 397–405.

Walsh SJ, McCleary AL, Mena CF *et al.* (2008) QuickBird and Hyperion data analysis of an invasive plant species in the Galápagos Islands of Ecuador: implications for control and land use management. *Remote Sensing of Environment*, **112**, 1927–1941.

Weeks A, Tye A (2009) Phylogeography of palo santo trees (*Bursera graveolens* and *Bursera malacophylla*; Burseraceae) in the Galápagos archipelago. *Botanical Journal of the Linnean Society*, **161**, 396–410.

Weir B, Cockerham C (1984) Estimation *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

White WM, McBirney AR, Duncan RA (1993) Petrology and geochemistry of the Galápagos islands: portrait of a pathological mantle plume. *Journal of Geophysical Research*, **98**, 19533–19564.

Willerslev E, Hansen AJ, Klitgaard K, Adsersen H (2002) Number of endemic and native plant species in the Galápagos Archipelago in relation to geographical parameters. *Ecography*, **25**, 109–119.

Yeakley JA, Weishampel JE (2000) Multiple source pools and dispersal barriers for Galápagos plant species distribution. *Ecology*, **81**, 893–898.

S.C. is a population geneticist. She worked on this study as part of her PhD, which dealt with the phylogeography and population genetics of two Seasonally Dry Tropical Forests trees, *Astronium urundeuva* and *Geoffroea spinosa*. M.C. is a population geneticist, specialized in the development and the use of spatially-explicit simulation approaches. His main themes of research are related to the evolution of genes in populations and especially to the combined effects of past demographic events and evolutionary factors on the genetic structure of populations. D.P. is interested in biogeography, phytosociology, plant taxonomy and plant reproductive biology with a special focus on grasslands and dry forest vegetation of the Neotropics. R.T.P.'s research focuses on systematics and evolution of the legume family, and on biogeography and conservation of seasonally dry tropical forests. L.E. is a population geneticist with interests in estimating demographic parameters from genetic data and retracing the history of populations under complex evolutionary models, with a focus on range expansions. Y.N.'s research focuses on population genetics of different organisms; among them plants, with a special interest in phylogeography species boundaries and speciation patterns.

Data accessibility

GenBank Accession numbers: EF564430, EF564432, EF564439, EU234508 and EU234509. IMa2 (microsatellites data) and Arlequin input files (microsatellite and chloroplast data): DRYAD entry doi: 10.5061/dryad.87v741v0.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Accuracy of the estimated parameters assessed using the mean, the median and the mode of the posterior distribution, by simulating a 1000 test datasets with known parameter values, time of colonization being fixed to 100 generations.

Table S2 Accuracy of the estimated parameters assessed using the mean the median and the mode of the posterior distribution, by simulating a 1000 test datasets with known parameter values, time of colonization being fixed to 800 generations.

Table S3 Estimation of population sizes, colonization time and migration parameters using IMa2, with a mutation rate fixed at $\mu = 0.0005$.

Fig. S1 Posterior density curves for the four parameters used to simulate the colonization of the Galápagos by *Geoffroea spinosa* with IMa2.

Appendix S1 ABC validation procedure.

Appendix S2 IMa2 simulations.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.