



# PAICE: A new R package to estimate the number of inter-island colonizations considering haplotype data and sample size

Alberto J. Coello<sup>1,2</sup> | Mario Fernández-Mazuecos<sup>1,3,4</sup> | Ruben H. Heleno<sup>5</sup> | Pablo Vargas<sup>1</sup>

<sup>1</sup>Departamento de Biodiversidad y Conservación, Real Jardín Botánico (RJB-CSIC), Madrid, Spain

<sup>2</sup>Escuela Internacional de Doctorado, Universidad Rey Juan Carlos, Móstoles, Spain

<sup>3</sup>Departamento de Biología (Botánica), Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid, Spain

<sup>4</sup>Departamento de Biodiversidad, Ecología y Evolución, Facultad de Ciencias Biológicas, Universidad Complutense de Madrid, Madrid, Spain

<sup>5</sup>Centre for Functional Ecology, TERRA Associated Laboratory, Department of Life Sciences, University of Coimbra, Coimbra, Portugal

## Correspondence

Alberto J. Coello, Real Jardín Botánico de Madrid (RJB-CSIC), Plaza Murillo 2, 28014 Madrid, Spain.

Email: albjcoello@gmail.com

Pablo Vargas, Departamento de Biodiversidad y Conservación, Real Jardín Botánico (RJB-CSIC), Madrid, Spain  
Email: vargas@rjb.csic.es

## Funding information

Consejo Superior de Investigaciones Científicas; Fundação para a Ciência e a Tecnologia; Ministerio de Economía, Industria y Competitividad, Gobierno de España; Ministerio de Educación, Cultura y Deporte

**Handling Editor:** Isabel Sanmartín

## Abstract

**Aim:** Colonization is a central topic in ecology and one of the cornerstones of island biogeography. Although the evolutionary history of island species is widely studied, the quantification of colonization is particularly challenging because the same area may be colonized multiple times by the same species, whereas initially successful colonization events may eventually be followed by extinction. Nevertheless, an estimation of the actual number of within-archipelago colonization events can be achieved when using adequate sample size and genetic data, which are essential parameters in the inference of colonization success of any species.

**Location:** Canary Islands, Azores and Galápagos Islands.

**Taxon:** *Buteo galapagoensis*, *Croton scouleri*, *Setophaga petechia aureola* and *Xylocopa darwini* (Galápagos); *Canarina canariensis*, *Cistus monspeliensis*, *Juniperus cedrus* and *Olea europaea* subsp. *guanchica* (Canary Islands); and *Juniperus brevifolia* and *Picconia azorica* (Azores).

**Methods:** The new R package PAICE uses haplotype (from organelle DNA) sharing and haplotype relationships, and controls for sampling effort to estimate the number of within-archipelago colonization events in island-like systems. PAICE applies a sampling-effort correction based on rarefaction curves of field sampling (number of individuals or populations) and genetic sampling (number of DNA variable positions). The number of colonization events for the 10 insular species were estimated with PAICE and results compared with previous methods.

**Results:** PAICE estimates a number of inter-island colonization events up to an order of magnitude greater than previous methods. Furthermore, PAICE can quantify the colonization events of any study species, in multiple biogeographic contexts and considering sampling size, thus providing a standardized estimate of colonization success.

**Main conclusions:** The new package PAICE provides an estimation of the number of inter-island colonization events (regardless of dispersal routes or rates) based on haplotype data across islands. This new tool will allow gaining new insights on the intensity of long-distance-dispersal events, their drivers and consequences for the assembly of insular faunas and floras.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of Biogeography* published by John Wiley & Sons Ltd.

## KEYWORDS

archipelagos, genealogies, island biogeography, oceanic islands, R package, rarefaction curves

## 1 | INTRODUCTION

Since the 19th century, oceanic islands have been considered natural laboratories to study ecological and evolutionary processes (Darwin, 1859; Fernández-Mazuecos et al., 2020; Grant & Grant, 1995; Patiño et al., 2017; Valente et al., 2015). Oceanic islands offer fragmented land emerged from the oceanic floor that has never been connected with a continent (Darwin, 1859; Whittaker & Fernández-Palacios, 2007). Therefore, the occurrence of terrestrial organisms on these islands can only be explained by long-distance dispersal from other territories (i.e. dispersal processes that overcome large biogeographic barriers), a phenomenon that has long intrigued scientists (Carlquist, 1966; Darwin, 1859; Patiño et al., 2017).

The occurrence of lineages on an oceanic island inevitably results from the arrival of founder propagules (i.e. colonizers) that travelled from another territory—either a continent or another island. However, in order to result in effective colonization after arriving in a new territory, immigrants need to establish a self-sustaining population. Colonization is thus a two-stage process in which dispersal and establishment are equally important (Heleno & Vargas, 2015). For this reason, the study of species colonization success is based on the subset of propagules that resulted in currently persisting populations. On the one hand, genetic data and phylogenetic analyses are paramount to understand the past colonization history of any species or territory (Sanmartín et al., 2008). On the other hand, a fine-scale approach to detect elusive colonization events at the population level needs the use of highly variable genetic markers and phylogeographic analyses (e.g. Fernández-Mazuecos et al., 2020). In phylogeography, the analysis of uniparentally inherited and non-recombinant genetic markers (from the mitochondrial and plastid genomes; Avise, 2000) is particularly valuable to infer recurrent within-archipelago colonization events when haplotypes meet on the same island (Vargas et al., 2015). This approach not only helps infer the colonization events across islands within an archipelago, but also in other island-like systems such as discontinuous habitats (e.g. lakes) or sky islands (Gillespie & Clague, 2009). The method proposed by Vargas et al. (2015) uses parsimony-based analyses to reconstruct haplotype genealogical relationships (Clement et al., 2000; Templeton et al., 1992) and infer a minimum number of inter-island colonization events (i.e. the minimum number of colonizations required to explain the distribution of haplotypes in multiple islands), but genealogical relationships among haplotypes are not considered. Furthermore, it is also important to avoid homoplasy (i.e. loops) in the haplotype network reconstruction, as the minimum number of colonization events could be overestimated in the presence of homoplasy. An alternative is the use of model-based methods to account for homoplasy, but their statistical power is lower than that of parsimony methods when migration and mutation rates are low

(Bloomquist et al., 2010). In any case, the potential of haplotype data (from organelle DNA) to quantify colonization events appears to be limited by the lack of a more specific phylogeographic methodology.

In phylogeography, the inference of the number of inter-island colonization events is highly dependent on sampling effort, including both genetic sampling (measured as number of variable positions, assuming genomic data are not available) and field sampling (number of populations and individuals) (Björklund & Bergek, 2009; Coello et al., 2021). An increase in sample size can modify genealogies and thus phylogeographic patterns, which makes sample-size correction important to estimate the number of colonization events (see Coello et al., 2021). Phylogeographic methods based on non-recombinant DNA such as discrete phylogeographic analysis (DPA) can estimate the number of colonization events in archipelagos (Lemey et al., 2009). This method can estimate colonization routes and migration rates; however, it is vulnerable to sampling effort bias as shown by De Maio et al. (2015). Müller et al. (2017) developed the marginal lineage states approximation of the structured coalescent (MASCO), a method that is less sensitive to sampling bias and allows the inference of dispersal considering coalescence in a phylogenetic context. However, outside the rapidly evolving virus sequences (Duffy, 2018), phylogeographic reconstructions of animals, fungi and plants using Sanger sequencing frequently result in polytomies due to low mutation rates, hindering the application of this approach to estimate dispersal rates between islands of an archipelago. In contrast, the use of haplotype-based methods may enable the analysis of animal and plant data with a relatively low mutation rate. One method based on haplotypes was proposed by Manolopoulou et al. (2011) to identify geographical population clusters and estimate dispersal between them. However, this method does not estimate the degree of dispersal between predefined island-like areas and does not address sampling bias.

In this study, we aimed to develop a new analytical approach to infer the minimum number of inter-island colonization events in any archipelago or island-like system by using haplotypes from mitochondrial or plastid genealogies of a species. By incorporating a sampling effort correction, this tool would allow the direct comparison of inter-island colonization success across species and biogeographic contexts. In particular, we hypothesize that by combining species and haplotype distributions with haplotype genealogies and rarefaction curves (Gotelli & Colwell, 2001, 2011), we can improve the estimates of colonization events for low-variability genetic markers. A new integrative method herein described (PAICE, Phylogeographic Analysis of Island Colonization Events) will enable not only a more appropriate estimation of the number of colonization events within a species based on haplotype data but also a standardized comparison across taxa and archipelagos, facilitating future tests of multiple hypotheses that are central in the island biogeography theory.



Furthermore, the complementarity between PAICE (i.e. estimation of the number of colonization events) and other methods aimed at reconstructing colonization routes (e.g. Lemey et al., 2009) could considerably improve our understanding of island biogeography. Our specific objectives are to (i) analyse haplotype distributions (haplotype sharing) together with haplotype genealogical relationships (haplotype networking) in PAICE to infer the minimum number of inter-island colonization events; (ii) develop a correction for field (populations or individuals) and genetic (variable nucleotide positions) sampling; (iii) implement PAICE metrics in empirical case studies of insular species already published; and (iv) compare PAICE colonization estimates to those from DPA. The PAICE approach is implemented in a new R package available at <https://github.com/PAICEcode/PAICE>.

## 2 | MATERIALS AND METHODS

### 2.1 | Inference of colonization events

The method developed here to infer the minimum number of inter-island colonization events is based on the following considerations: (i) the study species is distributed in an island-like system, that is the range consists of multiple fragments (real islands or island-like habitats) that have always been separated by biogeographic barriers, and therefore its occurrence on more than one fragment can only be explained by colonization events surpassing these barriers (thus excluding systems where species distributions across fragments can be explained by vicariance; Gillespie & Clague, 2009); (ii) the sample (at both genetic and field levels) is an unbiased representation of the distribution of haplotype diversity of the study species across the archipelago; (iii) all island populations of the species form a monophyletic group (i.e. they originated from a single ancestor; as a result, we only consider inter-island colonization events of the same archipelago, which, in normal situations, tend to be more frequent than continent-island colonizations); (iv) ancestry is determined by the mitochondrial or plastid genealogy (haplotype network) of island populations, which is connected to an outgroup (Guzmán & Vargas, 2009); (v) all the haplotypes in the haplotype network are connected; (vi) DNA polymorphisms lack homoplasy, that is the parsimony-based haplotype network of the species in the archipelago does not show any loops; (vii) haplotypes are inherited without recombination, typically maternally inherited, for example plastid markers in most angiosperms (Corriveau & Coleman, 1988) and mitochondrial markers in animals (Avise et al., 1987); (viii) recurrent colonization of a single haplotype on the same island cannot be detected; and (ix) extinction of haplotypes in the archipelago is negligible and uniformly distributed across the phylogeographic distribution of the studied species (i.e. in PAICE, extinct populations may be interpreted as unsampled data, and therefore incorporated into the estimation as part of sampling bias, see below).

The minimum number of colonization events ( $c$ ) is inferred considering three components. Type 1 colonization events ( $c_1$ ) represent

the minimum number of colonizations required to explain the distribution of haplotypes in more than one islands (haplotype sharing). Component  $c_1$  does not consider the direction of colonization, as it only incorporates the number of islands where each haplotype is present (see below). In contrast, type 2 ( $c_2$ ) and type 3 ( $c_3$ ) colonization events are derived from the occurrence of related haplotypes (forming a haplotype lineage) across islands. Mathematically,  $c$  is calculated as the sum of the three types of colonization events:

$$c = c_1 + c_2 + c_3$$

More specifically, type 1 colonization events,  $c_1$ , are those simply inferred by haplotype sharing, that is the same haplotype is found on different islands. If  $\mathbf{p}_i$  is the vector of presences (1) and absences (0) that indicates the islands where haplotype  $i$  occurs (from the set of haplotypes,  $I$ , occurring in the archipelago), and  $\|\mathbf{p}_i\|^2$  is the squared norm of vector  $\mathbf{p}_i$ , which is equivalent to the total number of islands on which haplotype  $i$  occurs (Coello et al., 2021; Vargas et al., 2015), then the minimum number of colonization events considering only haplotype occurrences is calculated as (Figure 1a):

$$c_1 = \sum_{i \in I} (\|\mathbf{p}_i\|^2 - 1)$$

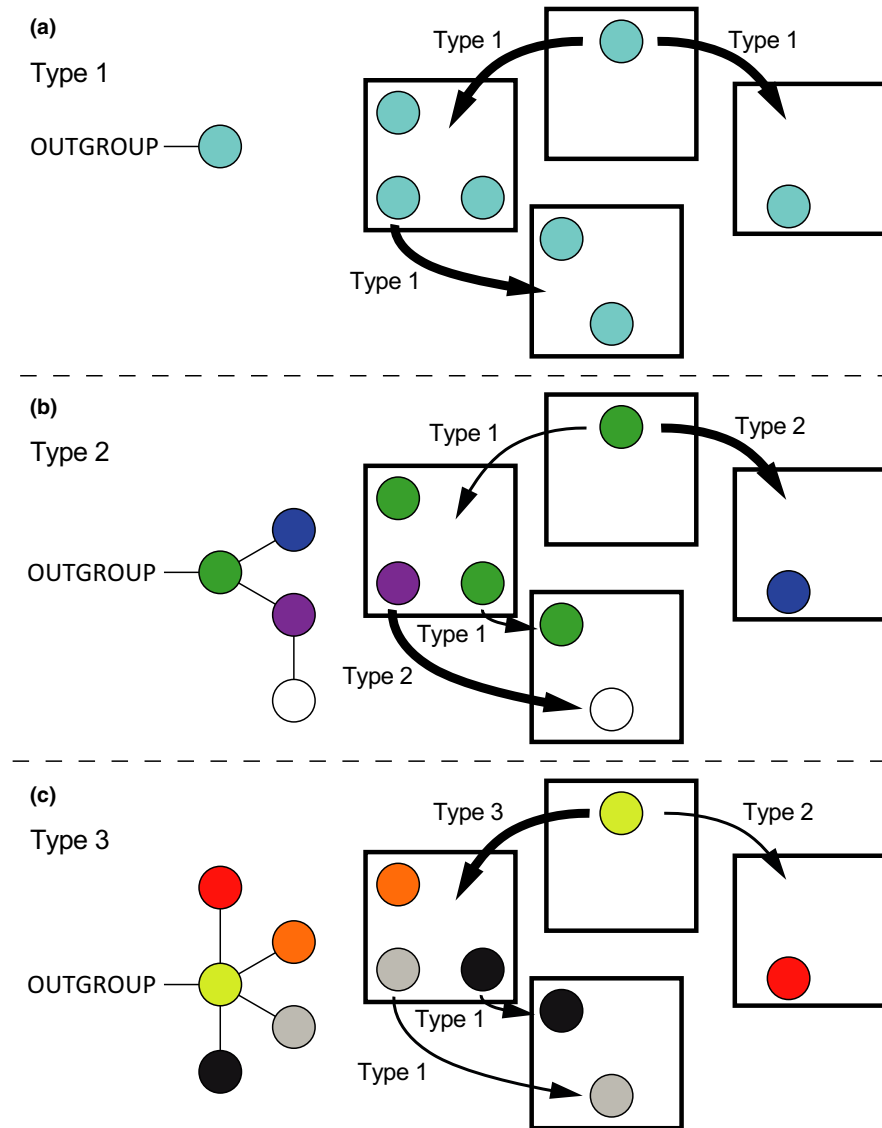
Type 2 colonization events,  $c_2$ , include colonizations that are inferred based on haplotype differentiation and genealogical relationships (not haplotype sharing) in such a way that colonization events can be assigned to direct connections between two haplotypes. If  $D$  is a set of derived haplotypes that do not occur on the same island as their ancestral haplotype and  $n(D)$  is the cardinal of  $D$ , that is the count of elements in set  $D$ , then the number of type 2 colonization events is:

$$c_2 = n(D)$$

More specifically,  $D$  is the set of haplotypes  $i$  from the total number of haplotypes,  $I$ , in the archipelago for which the scalar product between the vector of islands inhabited by haplotype  $i$ ,  $\mathbf{p}_i$ , and the vector of islands inhabited by its ancestor,  $\mathbf{a}_i$ , is zero; and the scalar product between the vector of islands inhabited by haplotype  $i$ ,  $\mathbf{p}_i$ , and the vector of island inhabited by its sister haplotype,  $\mathbf{b}_j$  (for each of its sister haplotypes,  $j$ , from the set of sister haplotypes,  $J$ , of haplotype  $i$ ), is also zero. 'Sister haplotypes' are defined as two or more haplotypes derived from the same parental haplotype. In other words,  $D$  represents the set of haplotypes that do not share inhabited islands neither with their parental haplotype nor with their sister haplotypes (Figure 1b):

$$D = \left\{ i \in I \mid \mathbf{p}_i \cdot \mathbf{a}_i = 0 \wedge \sum_{j \in J} \mathbf{p}_i \cdot \mathbf{b}_j = 0 \right\}$$

Finally, type 3 colonization events,  $c_3$ , are those also inferred from genealogical relationships (not haplotype sharing) when a group of derived haplotypes (more than one) are distributed on islands that are different from those inhabited by the parental haplotype.



**FIGURE 1** Inference of minimum number of colonization events under three scenarios considered in the mathematical expression of PAICE (see text). In particular, three circumstances (types of colonization events) are contemplated depending on haplotype differentiation and dispersal. (a) Three colonization events based exclusively on sharing of a single haplotype (type 1), that is the occurrence of a haplotype (light blue) on four islands can only be explained by three colonization events. (b) Four colonization events accompanied by haplotype differentiation; observe two colonizations based on haplotype sharing (green, type 1) and two additional colonizations based on parental-derived haplotypes (type 2, i.e. any haplotype occurring on different islands than its parental haplotype and sister haplotypes can only be explained by a colonization event between the parental and the derived haplotype): one inferred by a direct connection between the ancestral haplotype (green) and one of its derived haplotypes (blue), and the other between the purple haplotype and its derived haplotype (white). (c) Four colonization events based on two type 1 colonizations (black, grey), one type 2 colonization (red) and an additional colonization to account for the occurrence of a group of three derived haplotypes (orange, grey and black) (type 3, i.e. if a group of sister haplotypes occur on the same island or islands, and these are different from the one where the parental haplotype occurs, an additional colonization event is required). In all three panels, only one of several possible configurations of colonization routes is represented. In each panel, the new type of colonization is represented by thicker arrows

Although  $c_2$  can be explained as a particular case of  $c_3$ , there are two reasons for distinguishing them. First, each  $c_2$  colonization event can be assigned to a specific connection between two haplotypes, while  $c_3$  colonization events cannot be assigned to particular connections. Second, calculating  $c_3$  after  $c_2$  is computationally more efficient, given that, if there are no groups of derived haplotypes, calculation of  $c_3$  can be omitted ( $c_3 = 0$ ). Calculation of  $c_3$

is defined by the set of haplotype groups,  $G_i$ , derived from haplotype  $i$  that require a colonization event to connect them to their parental haplotype and explain their distribution—where  $n(G_i)$  is the cardinal of  $G_i$ :

$$c_3 = \sum_{i \in I} n(G_i)$$

The groups of descendants,  $G_i$ , from haplotype  $i$  that show additional colonization events are defined by the set  $h$  of groups of derived haplotypes (i.e. sister haplotypes that share distribution on the same island; Supplementary Material S1) within the set of groups of haplotypes directly descending from the parental haplotype  $i$ ,  $H_i$ , for which the scalar product between the vector of islands inhabited by group  $h$ ,  $\mathbf{g}_h$ , and the vector of islands inhabited by haplotype  $i$ ,  $\mathbf{p}_i$ , is zero; that is the set of groups of haplotypes that do not share islands with their parental haplotypes and thus require an additional colonization event (Figure 1c):

$$G_i = \{h \in H_i | \mathbf{g}_h \cdot \mathbf{p}_i = 0\}$$

The inference of the minimum number of inferred colonization events  $c$  resulting from the sum of these three lines of evidence ( $c_1$ ,  $c_2$ ,  $c_3$ ) is implemented in the function *colonization* of the R package PAICE.

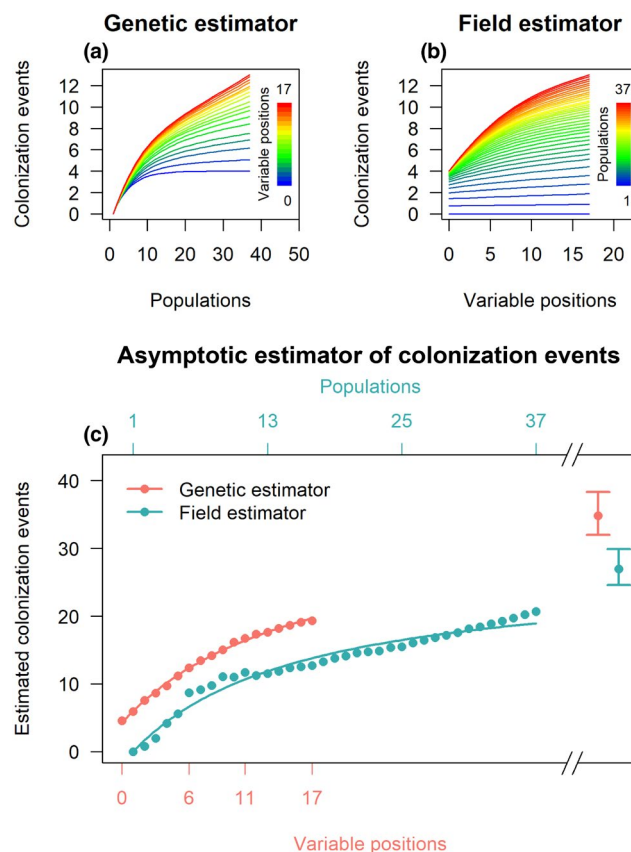
## 2.2 | Controlling for sampling effort when inferring colonization events

The inference of colonization events is conditioned by sampling effort (Supplementary Material S2; Coello et al., 2021), a pattern very similar to that found in studies of species richness (Bunge & Fitzpatrick, 1993). The effect of poor sampling effort is usually controlled by using rarefaction curves, which are produced by repeatedly resampling the data pool in order to derive an asymptotic estimator of the total species richness estimated by curve-fitting extrapolation (see below) (Gotelli & Colwell, 2001). Similarly, we propose to control for sampling effort in the estimation of the number of colonization events by using an approach based on rarefaction curves. In our approach, sampling effort depends on two sampling variables (field and genetic sample sizes; Björklund & Bergek, 2009; Coello et al., 2021). We first estimate the rarefaction curve for one of the sampling variables by keeping the sampling effect of the other variable constant in each level of this variable. Then, we use these estimates to control the other sampling variable. Subsequently, to avoid bias due to the order of the variables, the process is repeated in the opposite order (Figure 2; complete process described in Supplementary Material S3). Curve fitting for calculation of asymptotic estimators is explained in section 2.3.

Generation of rarefaction curves to obtain both genetic and field estimators is implemented in the *rarecol* function of the R package PAICE.

## 2.3 | Curve-fitting extrapolation of colonization events

Compared to species richness, asymptotic estimations of colonization events are more difficult to calculate. Nonparametric



**FIGURE 2** Calculation of asymptotic estimators of colonization events using the PAICE package and data from *Cistus monspeliensis* from the Canary Islands with populations as sampling unit in the field and five times the number of levels in each variable as number of replicates. A rarefaction curve of the number of colonization events as a function of field sampling (populations) is calculated for each level of genetic sampling (variable positions) (a). Then, for each level of genetic sampling, asymptotic estimators controlled by field sampling are used (c; in red) to infer the 'genetic estimator' of colonization events. The process is repeated by calculating a rarefaction curve of the number of colonization events as a function of genetic sampling (variable positions) for each level of field sampling (populations) (b). Then, for each level of field sampling, asymptotic estimators controlled by genetic sampling are used (c; in blue) to infer the 'field estimator' of colonization events. Genetic and field estimators are shown on the right in panel c, including means and 95% confidence intervals

estimators suggested by Gotelli and Colwell (2011) require that species be assigned to specific individuals, but colonization events are emergent properties inferred from the geographic distribution of genetic diversity and then they cannot be assigned to specific individuals. An alternative approach is to study the slope of the curve to obtain an asymptotic estimator (Phillips et al., 2020), but it may be difficult with insufficient data (Phillips et al., 2015). Therefore, for our rarefaction curves, we extrapolated the asymptotic estimator for the number of colonization events by fitting a modified Michaelis-Menten function (Johnson & Goody, 2011). When using field sampling as an independent variable, it is possible to detect colonization events only when at least two sampling

units are sampled, that is one population (or individual) cannot occur on more than one island. As a result, the curve cuts the x-axis in (0, 1) and the number of colonization events,  $c$ , as a function of  $f$  is calculated from the position of the asymptote of the function,  $m_f$ , the Michaelis constant for  $f$ ,  $k_f$ , and the field sampling size (populations or individuals),  $f$ , as follows:

$$c(f) = \frac{m_f \cdot (f - 1)}{k_f + f - 1}$$

In the case of the curve of colonization events as a function of field sampling,  $m_f$  is the asymptotic estimator for colonization events,  $c(f)_{\max}$ :

$$c(f)_{\max} = \lim_{f \rightarrow \infty} c(f) = m_f$$

In the case of the curve of colonization events as a function of genetic sampling, the minimum number of inferred colonization events varies depending on the number of colonization events inferred based on distribution only (i.e. without genetic information). The reason is that the number of islands where a species occurs contains information about the minimum number of colonization events between islands (Arjona et al., 2018).

Then, the number of colonization events as a function of genetic sampling effort ( $g$ ; measured as the number of variable DNA sequence positions) is calculated considering the position of the asymptote of the function,  $m_g$ , the Michaelis constant for  $g$ ,  $k_g$ , and the number of colonization events without considering genetic information,  $c_0$ . The latter parameter is also estimated when the curve is fitted. However, when the algorithm is unable to fit the curve because the data produces a hardly pronounced curve, the algorithm can try to fit the curve by using the average value from the distribution data (this option can be disabled by the user). The expression to calculate  $c(g)$  is:

$$c(g) = \frac{m_g \cdot g}{k_g + g} + c_0,$$

For the curve of the number of colonization events as a function of genetic sampling, the asymptotic estimator for detectable colonization events,  $c(g)_{\max}$ , is the sum of  $m_g$  and  $c_0$ :

$$c(g)_{\max} = \lim_{g \rightarrow \infty} c(g) = m_g + c_0.$$

For both types of colonization curves (as a function of field and genetic sampling effort), hardly pronounced curves make function fitting impossible and then the asymptotic estimator cannot be calculated. This happens when sampling size is too small and colonization events increase almost linearly.

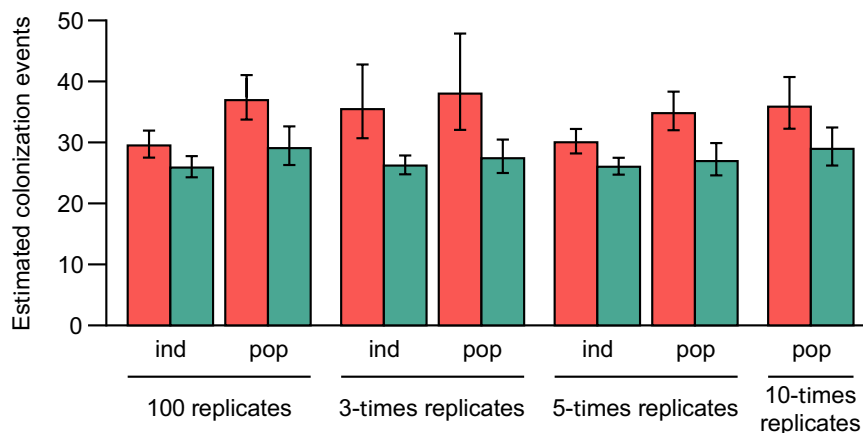
The calculation of asymptotic estimators considering genetic- and field-based equations described above is implemented in function *maxCol* (which uses the R function *nls* to fit curves to the equations above) in the R package PAICE.

## 2.4 | Estimation of colonization events in a case study

Phylogeography of *Cistus monspeliensis* has been studied in detail across the Mediterranean region and the Canary Islands in the last 12 years (Coello et al., 2021; Fernández-Mazuecos & Vargas, 2010, 2011). Data from the Canary Islands are analysed in this study to evaluate the performance of the PAICE method. In particular, the PAICE package includes two datasets: *CmonsData*, containing haplotype occurrences in Canarian populations; and *CmonsNetwork*, containing the haplotype genealogy reconstructed using a statistical parsimony method (Templeton et al., 1992) as implemented in TCS (Clement et al., 2000). This species was previously used to explore the effect of varying population and genetic sample sizes on the number of inferred colonization events (Coello et al., 2021). In the present study, we use the same dataset to develop the algorithm generating rarefaction curves to improve those estimates by removing the effect of sampling bias.

We also used *C. monspeliensis* to explore the number of replicates needed to properly fit the Michaelis–Menten equation to rarefaction curves. We originally considered 100 replicates per variable (for both field and genetic variables); and we then tested the numbers of replicates equal to three, five and 10 times the number of levels in each variable. For field sampling, each individual or population adds one level, for example, 10 sampled populations correspond to a field sampling of 10 levels. In the case of genetic sampling, each variable position adds one level, but starting at one level with zero variable positions (i.e. with no genetic data, distribution information is still available for the inference of a minimum number of colonization events; Arjona et al., 2018), for example 10 variable positions correspond to a genetic sampling of 11 levels. We repeated the analysis using populations or individuals as units for field sampling. In the case of *C. monspeliensis*, using populations as units is more appropriate because the sampling of individuals was stratified by population (Coello et al., 2021). However, in other studies, populations can be difficult to define, and thus individuals might be a more suitable sample unit.

For every analysis, we created rarefaction curves using the *rarecol* function and inferred the asymptotic estimator of colonization events between islands after deleting extreme values using the *maxCol* function (i.e. by deleting values below the 2.5% quantile and above the 97.5% quantile). The 95% confidence intervals (needed to compare results with those of DPA, see below) for the final estimates were calculated using the *maxCol* function. When rarefaction curves could not be fitted to the Michaelis–Menten equation, we tried to fit curves with one less parameter (i.e. by fixing the fitted number of colonization events without genetic information,  $c_0$ , to the specific value of the average of observed numbers of colonization events without genetic data for this dataset). The latter approach is implemented in the argument *method* = 1 of the *maxCol* function. The script to calculate the asymptotic estimator of colonization events for a species is available at Supplementary Material S4.



**FIGURE 3** Preliminary estimations of the number of colonization events of *Cistus monspeliensis* (after controlling field sampling in genetic estimation or genetic sampling in field estimation) using different numbers of replicates. Estimations were obtained using 100 replicates per variable (constant replicates) or three, five or 10 times the number of levels in each variable as number of replicates (estimates for 10-times replicates were impossible to calculate using individuals as field sampling unit because of computational limitations). Estimations were obtained using both individuals (ind) and populations (pop) as field sampling unit. Genetic estimation is shown in red and field estimation in blue. Rarefaction curves for these estimations are shown in Supplementary Material S9

## 2.5 | Comparison among 10 animal and plant species

We used suitable data from previous phylogeographic studies in oceanic archipelagos to infer the minimum number of colonization events and the asymptotic estimators of colonization events between islands for 10 species (Supplementary Material S5). These studies contain data from maternally inherited DNA (mitochondrial or plastid) and haplotype networks (but haplotype networks were not obtained in the same way in the 10 studies). For every species, we inferred the number of observed colonization events (minimum number) using the *colonization* function of the PAICE package considering current emerged islands as geographic units to obtain comparable results with those of the previous study (see Vargas et al., 2015). However, if two or more extant islands formed one larger palaeoisland in the past, it is not possible to distinguish colonization between these extant islands from population diffusion within the palaeoisland followed by vicariance. In these cases, to avoid overestimating the number of colonization events, it would be advisable to consider the palaeoisland as a single unit (see Arjona et al., 2018). We estimated the number of colonization events using the parameters described above for *C. monspeliensis* and also using individuals as sample units to enable comparison among studies. Rarefaction curves for the field sample were replicated five times the number of levels of each variable (e.g. if a species was sampled in 10 populations, this number of populations corresponds to 10 levels of field sampling, and therefore 50 replicates are performed). This was shown to represent the best balance between estimation and computing time (see Results). As PAICE resamples variable positions when conducting genetic resampling, in particular studies in which there was no information about which connection between haplotypes corresponded to which variable position in the alignment, we assumed that every connection of the haplotype network represented an independent mutation step. Given that our method requires knowledge about ancestor–descendant

relationships among haplotypes, we repeated the inference of the asymptotic estimator of colonization events using all possible ancestral haplotypes when information about ancestry was not available (i.e. we repeated the inference using every haplotype as the ancestral haplotype in the archipelago). When loops were present in the genealogy, we repeated the inference of asymptotic estimators by breaking loops in all possible ways (i.e. we repeated the inference, every time deleting one of the connections responsible for the loop, as shown in Supplementary Material S6). The assumptions made for each species are collected in Supplementary Material S7.

For the sake of comparison with previous methods, we contrasted the performance of PAICE with a widely used method to estimate colonization events—DPA—(Lemey et al., 2009). This method models geographic locations of sampled individuals as character states and then reconstructs ancestral states across phylogenies to infer changes of location (i.e. colonization events). DPA incorporates information from branch lengths to estimate recurrent colonization events across the phylogeny, but does not consider the effect of sample sizes on reconstructions. Therefore, to compare PAICE with an alternative method to estimate inter-island colonization events, we applied DPA to study cases in which original sequence alignments were available for re-analysis and we calculated the number of inter-island colonization events by reconstructing the counts of state changes using Markov Jumps, as described by Minin and Suchard (2008) (Supplementary Material S8).

## 3 | RESULTS

### 3.1 | Inference of asymptotic estimators of colonization events in *Cistus monspeliensis*

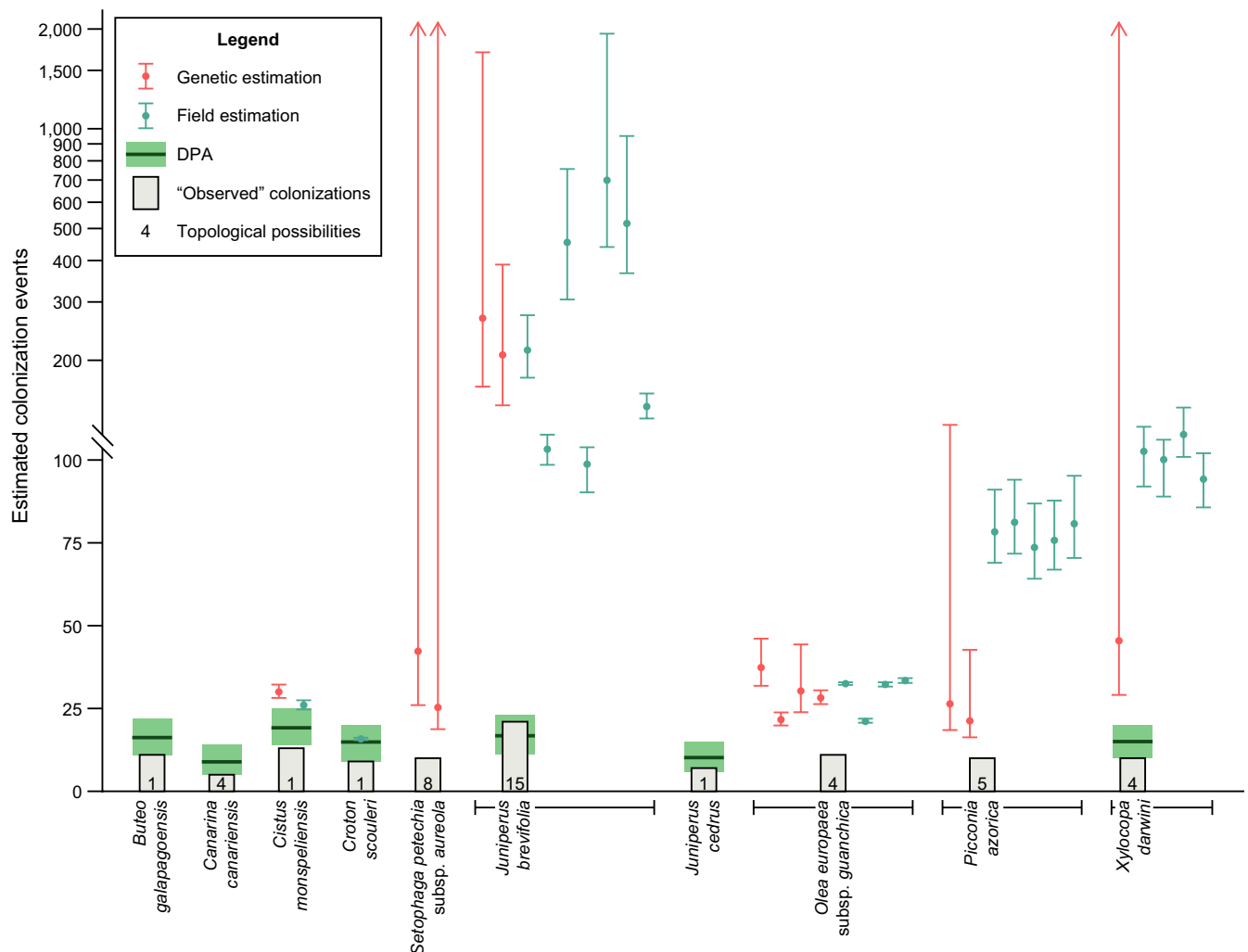
Both asymptotic estimators (genetic and field estimators) projected a number of colonization events between 25.9 and 38.0 for *C. monspeliensis* across the Canary Islands (Figure 3; Supplementary Material

S9). In general, more replicates resulted in smaller confidence intervals. Estimates based on genetic sampling effort had a better fit to the Michaelis–Menten equation than estimates based on field sampling effort. Considering field sampling units, estimates based on populations were always slightly higher than those based on individuals, but differences were not significant (95% confidence intervals usually overlapped).

### 3.2 | Comparison among the 10 species

The minimum number of colonization events and the asymptotic estimators for the number of inter-island colonization events for 10 animal and plant species are summarized in Figure 4 (see also Supplementary Materials S10 and S11). For *Cistus monspeliensis*

and *Olea europaea* subsp. *guanchica*, it was possible to obtain asymptotic estimators for both genetic and field procedures. These two species displayed the best fit of rarefaction curves with the smallest differences between both genetic and field estimators. For five species (*Croton scouleri*, *Setophaga petechia aureola*, *Juniperus brevifolia*, *Picconia azorica* and *Xylocopa darwini*), it was possible to infer at least one asymptotic estimator (also considering cases in which the upper limit of the 95% confidence interval was infinite). However, other estimators could not be calculated because some curves could not be fitted. For the remaining three species, sampling effort was insufficient to infer any asymptotic estimator (genetic or field): *Buteo galapagoensis*, *Canarina canariensis* and *Juniperus cedrus*. In addition, six species required multiple analyses because their genealogical relationships had network



**FIGURE 4** Estimation of the number of colonization events using PAICE and datasets from 10 animal and plant species taken from the literature (see text). Bars indicate the number of 'observed' colonization events (i.e. minimum number of colonization events needed to explain the geographical distribution of haplotype diversity). Numbers within grey bars indicate the number of topological possibilities considered (by breaking loops or by considering different ancestral haplotypes) in the estimation of colonization events. Red dots indicate average 'genetic estimators' of colonization events and blue dots indicate average 'field estimators', both with 95% confidence intervals. Dark green lines indicate the estimates of the number of colonization events based on discrete phylogeographic analysis (DPA), with 95% highest posterior density intervals in light green. When the upper limit of a 95% confidence interval is infinite, it is indicated as an upward arrow. To facilitate the visualization of results, the y-axis scale is linear from 0 to 100 colonization events and logarithmic from 100 colonization events



loops (*C. canariensis*, *J. brevifolia*, *O. europaea* subsp. *guanchica*, *X. darwini*; for these species all possible topologies breaking the loops in their networks were analysed independently) or due to the absence of any outgroup to infer haplotype ancestry (*S. petechia aureola* and *P. azorica*; as outgroup species were not connected to the network, every species was analysed multiple times considering every haplotype as a possible ancestral haplotype). Rarefaction curves of colonization events for all species are available in Supplementary Material S10.

In comparison to DPA, our approach estimated a higher number of colonization events, but DPA was able to estimate colonization events for all species with available DNA sequences, including cases in which sampling effort was insufficient for rarefaction curves to approach an asymptote (Figure 4, Supplementary Materials S10 and S11). However, it was problematic to apply DPA when original data were not available (e.g. microsatellites were used in *P. azorica*) while PAICE can be applied even when only the haplotype genealogy is available. Furthermore, DPA results were more similar to the minimum number of colonization events (i.e. without sampling effort correction) than PAICE estimations considering sampling effort.

## 4 | DISCUSSION

In this study, a new approach is proposed and implemented into a new R package (PAICE) to estimate the number of colonization events accounting for species and haplotype distributions in any island or island-like system, and regardless of the direction of such colonization events. Therefore, PAICE can be used as a complementary tool for studying the phylogeographic history of species by estimating the number of inter-island colonization events. Our approach accounts for sampling bias emerging from suboptimal sampling effort at both the genetic level (number of variable positions that determine genealogical relationships among haplotypes) and the field level (number of populations or individuals sampled). As a consequence, PAICE estimates a greater number of colonization events that cannot be derived directly (i.e. they are intrinsically 'hidden') from direct counting based on the current distributions of species and haplotype sharing. In doing so, PAICE standardizes the comparison of colonization events across species, studies and archipelagos.

### 4.1 | Performance of colonization event inference using PAICE

Our inference of a minimum number of colonization events extends the previous approach of Vargas et al. (2015), which only considered haplotype sharing, that is  $c_1$  (Figure 1a). Considering the three types of colonization together (Figure 1) allows the inference of additional colonization events that were previously hidden (e.g. *Buteo galapagoensis* and *Juniperus brevifolia*, Table S11.1).

Because the number of inferred colonization events increases with sampling effort (Coello et al., 2021), PAICE is the first method that incorporates asymptotic estimators to control for sampling effort when estimating the number of colonization events. This correction is of paramount importance when comparing the colonization success of multiple species that have been sampled with different protocols, such as the 10 species compared in this study. Without the explicit consideration of sampling bias, it is not possible to compare among species as each one has different sampling sizes (field and genetic), and therefore inference of a minimum number of colonization events without a sampling correction would be significantly biased.

The estimation of colonization events for *C. monspeliensis* using different numbers of replicates when building rarefaction curves showed similar estimates (Figure 3). However, the number of colonization events obtained from asymptotic field estimators was always lower than that obtained from genetic estimators (Supplementary Material S9). We consider that field estimators for *C. monspeliensis* lead to underestimations due to the rapid initial increase in the inferred number of colonization events with few individuals (first 20%), compared to ideal rarefaction curves (Gotelli & Colwell, 2001, 2011) (Supplementary Material S9). We interpret that the little hump around the first 20% of values (Figure 2) tends to deteriorate the fit of the curve, and therefore makes the field estimator less reliable. This effect is more marked when using individuals as field sampling unit, which makes sense considering that *C. monspeliensis* was sampled using a population strategy (Coello et al., 2021; Fernández-Mazuecos & Vargas, 2011). In future analyses, we recommend that the resampling unit used in PAICE be congruent with the field sampling unit of the study (individuals or populations).

### 4.2 | PAICE versus DPA

PAICE brings the opportunity to infer a number of colonization events using a fine-scale approach that analyses haplotype sharing and genealogies. In addition, a wide variety of techniques that do not necessarily produce DNA sequences can also be used, such as plastid and mitochondrial microsatellites (Ferreira et al., 2011). In contrast to methods based on coalescent modelling (primarily DPA; Lemey et al., 2009), PAICE (a haplotype network-based method) can estimate the number of colonization events corrected by sampling effort regardless of colonization routes and their direction, which are not reconstructed. This correction increases the power to estimate colonization events masked by poor sampling, yielding an estimate of a larger number of colonization events than DPA, that is up to one order of magnitude more colonization events estimated with PAICE in extreme cases such as that of *J. brevifolia* (Figure 4). The number of such 'hidden' colonization events decreases with sampling effort, being more pronounced in clearly undersampled studies (see Supplementary Materials S10 and S11), and less relevant in studies with a considerable sampling effort (e.g. *C. monspeliensis*;

Figure S10.6, Table S11.1). Therefore, the ability of PAICE to correct sampling effort bias helps evaluate island biogeographic hypotheses requiring an estimation of the number of inter-islands colonization events without the necessity of reconstructing ancestral areas over phylogenetic trees, for example a hypothesis dealing with differences in colonization ability between taxa with and without dispersal syndromes.

The downside of PAICE is that it generally requires larger sample sizes (both field and genetic sampling) than DPA in order to provide estimators. This is the case of some study species like *B. galapagensis* (Bollmer et al., 2006), *C. canariensis* (Mairal et al., 2015) and *J. cedrus* (Rumeu et al., 2014), in which their data could not be used to obtain any PAICE estimator. However, the minimum number of colonization events (i.e. those required to explain haplotype distributions and lineage sharing) can always be calculated by using the *colonization* function of the PAICE package and it is usually close to the result obtained by DPA (Figure 4). This suggests that, although PAICE does not consider phylogenetic tree topology and branch lengths, this method is still powerful enough to infer a minimum number of colonization events similar to that obtained by DPA (Lemey et al., 2009). It is important to mention that, although DPA can estimate the number of colonization events (i.e. changes in the 'location' character; Minin & Suchard, 2008), this method is mainly aimed at estimating migration rates using genetic markers with high mutation rate. This probably leads to the different results obtained by DPA and PAICE (Figure 4).

Additionally, PAICE requires an outgroup to determine ancestor-descendant relationships. The lack of connection to an outgroup outside the island system can be compensated by repetition of the algorithm multiple times to test all possible ancestral haplotypes, although results were usually similar across the different possibilities (see *Setophaga petechia aureola*, Chaves et al., 2012; *P. azorica*, Ferreira et al., 2011; Figure 4). Importantly, concomitant extinction processes inevitably lead to an underestimation of colonization events (as in DPA) as extinct haplotypes erode the signal of past colonizations (García-Verdugo et al., 2019). However, extinct populations may be considered part of the unsampled data if extinction happened at a low rate and it was uniformly distributed along the species phylogeography. Therefore, in these cases, colonization events hidden as a result of extinction may also be included in PAICE estimates.

In summary, PAICE is a powerful tool for studies of colonization success across island and island-like systems, most notably because it enables controlling for sample size at both field and genetic levels. In doing so, PAICE provides estimators that are comparable not only among species of the same island system but also among species across island-like systems, such as archipelagos, mountains, lakes or landscape fragments. In particular, the estimation of the number of inter-island colonization events corrected by sampling effort allows the comparison between species (from the same or from different archipelagos), as shown in Figure 4. Recent studies tested the relationship between dispersal syndromes and the number of islands within an archipelago occupied by species (Arjona et al., 2018; Heleno &

Vargas, 2015; Vargas et al., 2014). However, the number of colonization events between islands is more relevant as a potential correlate of dispersal syndromes than the number of occupied islands because a species can colonize the same island multiple times (Coello et al., 2021; Vargas et al., 2015). PAICE allows us to compare species more objectively based on the number of colonization events, and therefore will enable future analyses to test if dispersal syndromes are related to higher numbers of colonization events, as traditionally assumed (Carlquist, 1966). Furthermore, PAICE estimations will also enable comparisons of the contribution to colonization processes of different factors such as dispersal capabilities (Thomson et al., 2018) or habitat availability (Coello et al., 2021). Additionally, model-based approaches are continuously being developed, in combination with massive sequencing (providing highly variable genetic information) (Edwards et al., 2022; Xue & Hickerson, 2020). The synergy between PAICE and these other advanced techniques in phylogeography will improve our understanding of island biogeography.

### 4.3 | Insular species comparison

Our analysis reveals that the two estimations with the best fitted curves were those from *C. monspeliensis* (Coello et al., 2021) and the westernmost lineage of *Olea europaea* subsp. *guanchica* (García-Verdugo et al., 2010), thanks to their large sample sizes. Coincidentally, these two plants show a similar minimum number of colonization events (13 inter-island colonizations for *C. monspeliensis* and 11 for *O. europaea* subsp. *guanchica*) and a similar estimated number of colonization events corrected by sampling effort (average of 30.0 and 26.0 colonization events for genetic and field estimators, respectively, in *C. monspeliensis*; 29.4 and 29.8, respectively, in *O. europaea* subsp. *guanchica*) despite displaying different dispersal syndromes (unspecialized capsules vs zoochorous drupes) (Figure 4). Furthermore, both species first colonized the archipelago in a similar time frame (<1 Ma for *C. monspeliensis* and c. 1.5 Ma for *O. europaea* subsp. *guanchica*) (Besnard et al., 2009; Fernández-Mazuecos & Vargas, 2011).

Our estimation of the number of colonization events reveals a high colonization potential for species in which previous studies suggested a limited colonization success (Figure 4). This is the case of *Xylocopa darwini* of the Galápagos, in which observed colonization events (10 colonization events) and DPA (15.0 colonization events, with a 90% confidence interval from 10 to 20 colonization events) suggested a relatively small number of colonization events. The dominance of one haplotype on several islands (Vargas et al., 2015) increases the possibility of detecting additional colonization events as genetic sampling increases and new haplotypes are obtained, which is clearly shown by our estimations (increasing estimated colonization events up to an order of magnitude; Figure 4). Very similar results are found for *J. brevifolia* (21 observed colonization events versus more than c. 150 estimated colonization events) and *Picconia azorica* (10 observed colonization events versus 20–95 estimated colonization events), for which one (Rumeu et al., 2011) and two



(Ferreira et al., 2011) haplotypes are widely distributed on several Azorean islands respectively. This pattern is extreme in *Setophaga petechia aureola*, whose data were appropriate to describe the overall distribution of genetic diversity (Chaves et al., 2012), but not enough to estimate colonization events. It is worth noting that a minimum sampling effort is always needed, as sample size correction may not be possible for any of the two sampling variables when data are too scarce. An example is *Croton scouleri* in the Galápagos Islands, in which resampling of variable DNA positions does not yield an increase in the number of inferred colonization events (Figure S10.7, Table S11.1) because the genealogy of this species shows a clear dominance of one haplotype on several islands of the archipelago (Rumeu et al., 2016).

In some cases, our approach could not solve the problem of insufficient sampling size (*Buteo galapagoensis*, Bollmer et al., 2006; *Canarina canariensis*, Mairal et al., 2015; *Juniperus cedrus*, Rumeu et al., 2014). In these species, an increase in sampling effort is clearly necessary at both sampling levels to allow an estimation of the number of colonization events. Nevertheless, PAICE results were congruent with those provided by other techniques, such as the BPEC approach of Manolopoulou et al. (2011) in the case of *C. canariensis*. Indeed, the cluster observed by Mairal et al. (2015) formed by West Tenerife and La Gomera populations supports the numerous colonization events suggested in our PAICE analysis by the increasing number of colonization events as sampling effort increases (Figure S10.2–S10.5). In general, we recommend large sample sizes guided by species distribution modelling in future studies particularly when the inference of colonization events is one of the objectives (Coello et al., 2021).

Despite the advance represented by PAICE, it is important to mention that the limitations of this method should not be overlooked. For example, although *J. brevifolia* could be analysed and the estimation of the number of colonization events was considerably higher than the previous ones based exclusively on haplotype sharing (Vargas et al., 2015), more work is needed to confirm these results. In particular, uniparental inheritance of cpDNA is uncertain in this case, given that a great proportion of gymnosperms display biparental inheritance of the plastid genome (Adams, 2019). Furthermore, there is growing evidence of infrequent biparental inheritance of mtDNA in animals (e.g. Wolff et al., 2013).

In a nutshell, PAICE paves the road for investigation of plant and animal colonization success in island systems, which is a fundamental topic in island biogeography (Patiño et al., 2017). Although PAICE is focused on inter-island colonization events, its results can provide a first indicator of species potential to colonize across broader distances—such as continent-archipelago colonization—which tend to be more stochastic, less frequent and harder to quantify (Heleno & Vargas, 2015). Colonization success is the result of several factors including dispersal capacity (e.g. Thomson et al., 2018), habitat availability (e.g. Coello et al., 2021), mating system (e.g. Grossenbacher et al., 2017) and time since initial arrival in the island system (Arjona et al., 2018), among others. PAICE provides a new and more informative approach to estimate colonization success from haplotype

networks obtained using low-variability genetic markers, allowing the detection of colonization events masked by poor sampling and thus permitting cross-taxa and cross-archipelago comparisons. The complementarity between PAICE and other methods to reconstruct migration history (e.g. DPA; Lemey et al., 2009) and to estimate population parameters of insular species (e.g. DAISIE; Valente et al., 2015) holds a great potential to understanding the drivers of colonization success.

## ACKNOWLEDGEMENTS

The authors thank Leticia Fernández-Prado for her assistance with mathematical notation; Daniela Jordão, Sara Carvalho and José Augusto Ferreira for their suggestions to improve asymptotic estimators; Beatriz Rumeu, Mario Mairal, Carlos García-Verdugo and Jaime Chaves for their help with collection of data from previously published studies. This study is part of projects CGL2015-67865-P and PGC2018-101650-B-I00, funded by the Spanish Ministry of Economy, Industry and Competitiveness. A.J.C. was supported by the Spanish Ministry of Education, Culture and Sport through an FPU fellowship (FPU16/05681) and a FPU mobility fellowship (EST18/00383). M.F.-M. was supported by a Juan de la Cierva fellowship (Spanish Ministry of Economy and Competitiveness; reference IJCI-2015-23459) and by a Special Intramural Project of the Spanish National Research Council (CSIC; reference 201930E078). R.H.H. was funded by the Portuguese Foundation for Science and Technology (grant: UID/BIA/04004/2021). No permits were needed for this study.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

Alberto J. Coello <https://orcid.org/0000-0002-2665-3726>

Mario Fernández-Mazuecos <https://orcid.org/0000-0003-4027-6477>

Ruben H. Heleno <https://orcid.org/0000-0002-4808-4907>

Pablo Vargas <https://orcid.org/0000-0003-4502-0382>

## REFERENCES

- Adams, R. P. (2019). Inheritance of chloroplasts and mitochondria in conifers: A review of paternal, maternal, leakage and facultative inheritance. *Phytologia*, 101(2), 134–138.
- Arjona, Y., Nogales, M., Heleno, R., & Vargas, P. (2018). Long-distance dispersal syndromes matter: Diaspore-trait effect on shaping plant distribution across the Canary Islands. *Ecography*, 41(5), 805–814. <https://doi.org/10.1111/ecog.02624>
- Avise, J. C. (2000). *Phylogeography: The history and formation of species* (3. print). Harvard University Press.
- Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., & Saunders, N. C. (1987). Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, 18, 489–522. <https://doi.org/10.1146/annurev.es.18.110187.002421>
- Besnard, G., Rubio de Casas, R., Christin, P.-A., & Vargas, P. (2009). Phylogenetics of *Olea* (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: Tertiary climatic shifts and lineage

- differentiation times. *Annals of Botany*, 104(1), 143–160. <https://doi.org/10.1093/aob/mcp105>
- Björklund, M., & Bergek, S. (2009). On the relationship between population differentiation and sampling effort: Is more always better? *Oikos*, 118(8), 1127–1129. <https://doi.org/10.1111/j.1600-0706.2009.17555.x>
- Bloomquist, E. W., Lemey, P., & Suchard, M. A. (2010). Three roads diverged? Routes to phylogeographic inference. *Trends in Ecology & Evolution*, 25(11), 626–632. <https://doi.org/10.1016/j.tree.2010.08.010>
- Bollmer, J. L., Kimball, R. T., Whiteman, N. K., Sarasola, J. H., & Parker, P. G. (2006). Phylogeography of the Galápagos hawk (*Buteo galapagoensis*): A recent arrival to the Galápagos Islands. *Molecular Phylogenetics and Evolution*, 39(1), 237–247. <https://doi.org/10.1016/j.ympev.2005.11.014>
- Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association*, 88(421), 364. <https://doi.org/10.2307/2290733>
- Carlquist, S. (1966). The biota of long-distance dispersal. I. Principles of dispersal and evolution. *The Quarterly Review of Biology*, 41(3), 247–270.
- Chaves, J. A., Parker, P. G., & Smith, T. B. (2012). Origin and population history of a recent colonizer, the yellow warbler in Galápagos and Cocos Islands. *Journal of Evolutionary Biology*, 25(3), 509–521. <https://doi.org/10.1111/j.1420-9101.2011.02447.x>
- Clement, M., Posada, D., & Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, 9(10), 1657–1659. <https://doi.org/10.1046/j.1365-294x.2000.01020.x>
- Coello, A. J., Fernández-Mazuecos, M., García-Verdugo, C., & Vargas, P. (2021). Phylogeographic sampling guided by species distribution modeling reveals the Quaternary history of the Mediterranean–Canarian *Cistus monspeliensis* (Cistaceae). *Journal of Systematics and Evolution*, 59(2), 262–277. <https://doi.org/10.1111/jse.12570>
- Corriveau, J. L., & Coleman, A. W. (1988). Rapid screening method to detect potential biparental inheritance of plastid dna and results for over 200 angiosperm species. *American Journal of Botany*, 75(10), 1443–1458. <https://doi.org/10.1002/j.1537-2197.1988.tb11219.x>
- Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.
- De Maio, N., Wu, C.-H., O'Reilly, K. M., & Wilson, D. (2015). New routes to phylogeography: A Bayesian structured coalescent approximation. *PLoS Genetics*, 11(8), e1005421. <https://doi.org/10.1371/journal.pgen.1005421>
- Duffy, S. (2018). Why are RNA virus mutation rates so damn high? *PLoS Biology*, 16(8), e3000003. <https://doi.org/10.1371/journal.pbio.3000003>
- Edwards, S. V., Robin, V., Ferrand, N., & Moritz, C. (2022). The evolution of comparative phylogeography: Putting the geography (and more) into comparative population genomics. *Genome Biology and Evolution*, 14(1), evab176. <https://doi.org/10.1093/gbe/evab176>
- Fernández-Mazuecos, M., & Vargas, P. (2010). Ecological rather than geographical isolation dominates Quaternary formation of Mediterranean *Cistus* species. *Molecular Ecology*, 19(7), 1381–1395. <https://doi.org/10.1111/j.1365-294X.2010.04549.x>
- Fernández-Mazuecos, M., & Vargas, P. (2011). Genetically depauperate in the continent but rich in oceanic islands: *Cistus monspeliensis* (Cistaceae) in the Canary Islands. *PLoS One*, 6(2), e17172–e17172. <https://doi.org/10.1371/journal.pone.0017172>
- Fernández-Mazuecos, M., Vargas, P., McCauley, R. A., Monjas, D., Otero, A., Chaves, J. A., Guevara Andino, J. E., & Rivas-Torres, G. (2020). The radiation of Darwin's giant daisies in the Galápagos Islands. *Current Biology*, 30(24), 4989–4998.e7. <https://doi.org/10.1016/j.cub.2020.09.019>
- Ferreira, R. C., Piredda, R., Bagnoli, F., Bellarosa, R., Attimonelli, M., Fineschi, S., Schirone, B., & Simeone, M. C. (2011). Phylogeography and conservation perspectives of an endangered macaronesian endemic: *Picconia azorica* (Tutin) Knobl. (Oleaceae). *European Journal of Forest Research*, 130(2), 181–195. <https://doi.org/10.1007/s10342-010-0420-1>
- García-Verdugo, C., Caujapé-Castells, J., & Sanmartín, I. (2019). Colonization time on Island settings: Lessons from the Hawaiian and Canary Island floras. *Botanical Journal of the Linnean Society*, 191(2), 155–163. <https://doi.org/10.1093/botlinnean/boz044>
- García-Verdugo, C., Forrest, A. D., Fay, M. F., & Vargas, P. (2010). The relevance of gene flow in metapopulation dynamics of an oceanic Island endemic, *Olea europaea* subsp. *Guanchica*. *Evolution*, 64(12), 3525–3536. <https://doi.org/10.1111/j.1558-5646.2010.01091.x>
- Gillespie, R., & Clague, D. (2009). *Encyclopedia of islands*. University of California Press. <http://www.ucpress.edu/book.php?isbn=9780520256491>
- Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4(4), 379–391. <https://doi.org/10.1046/j.1461-0248.2001.00230.x>
- Gotelli, N. J., & Colwell, R. K. (2011). Estimating species richness. In A. E. Magurran & B. J. McGill (Eds.), *Biological diversity: Frontiers in measurement and assessment* (pp. 39–54). Oxford University Press.
- Grant, P. R., & Grant, B. R. (1995). Predicting microevolutionary responses to directional selection on heritable variation. *Evolution*, 49(2), 241–251. <https://doi.org/10.1111/j.1558-5646.1995.tb02236.x>
- Grossenbacher, D. L., Brandvain, Y., Auld, J. R., Burd, M., Cheptou, P.-O., Conner, J. K., Grant, A. G., Hovick, S. M., Pannell, J. R., Pauw, A., Petanidou, T., Randle, A. M., Rubio de Casas, R., Vamasi, J., Winn, A., Igic, B., Busch, J. W., Kalisz, S., & Goldberg, E. E. (2017). Self-compatibility is over-represented on islands. *New Phytologist*, 215(1), 469–478. <https://doi.org/10.1111/nph.14534>
- Guzmán, B., & Vargas, P. (2009). Long-distance colonization of the Western Mediterranean by *Cistus ladanifer* (Cistaceae) despite the absence of special dispersal mechanisms. *Journal of Biogeography*, 36(5), 954–968. <https://doi.org/10.1111/j.1365-2699.2008.02040.x>
- Heleno, R., & Vargas, P. (2015). How do islands become green?: How do islands become green? *Global Ecology and Biogeography*, 24(5), 518–526. <https://doi.org/10.1111/geb.12273>
- Johnson, K. A., & Goody, R. S. (2011). The original Michaelis constant: Translation of the 1913 Michaelis-Menten paper. *Biochemistry*, 50(39), 8264–8269. <https://doi.org/10.1021/bi201284u>
- Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9), e1000520–e1000520. <https://doi.org/10.1371/journal.pcbi.1000520>
- Mairal, M., Sanmartín, I., Aldasoro, J. J., Culshaw, V., Manolopoulou, I., & Alarcón, M. (2015). Palaeo-islands as refugia and sources of genetic diversity within volcanic archipelagos: The case of the widespread endemic *Canarina canariensis* (Campanulaceae). *Molecular Ecology*, 24(15), 3944–3963. <https://doi.org/10.1111/mec.13282>
- Manolopoulou, I., Legarreta, L., Emerson, B. C., Brooks, S., & Tavaré, S. (2011). A Bayesian approach to phylogeographic clustering. *Interface Focus*, 1(6), 909–921. <https://doi.org/10.1098/rsfs.2011.0054>
- Minin, V. N., & Suchard, M. A. (2008). Fast, accurate and simulation-free stochastic mapping. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512), 3985–3995. <https://doi.org/10.1098/rstb.2008.0176>
- Müller, N. F., Rasmussen, D. A., & Stadler, T. (2017). The structured coalescent and its approximations. *Molecular Biology and Evolution*, 34(11), 2970–2981. <https://doi.org/10.1093/molbev/msx186>
- Patiño, J., Whittaker, R. J., Borges, P. A. V., Fernández-Palacios, J. M., Ah-Peng, C., Araújo, M. B., Ávila, S. P., Cardoso, P., Cornuault, J., de Boer, E. J., de Nascimento, L., Gil, A., González-Castro, A., Gruner, D. S., Heleno, R., Hortal, J., Illera, J. C., Kaiser-Bunbury, C. N., Matthews, T. J., ... Emerson, B. C. (2017). A roadmap for Island biology: 50 fundamental questions after 50 years of *The Theory*



- of Island Biogeography. *Journal of Biogeography*, 44(5), 963–983. <https://doi.org/10.1111/jbi.12986>
- Phillips, J. D., French, S. H., Hanner, R. H., & Gillis, D. J. (2020). HACSIm: An R package to estimate intraspecific sample sizes for genetic diversity assessment using haplotype accumulation curves. *PeerJ Computer Science*, 6, e243. <https://doi.org/10.7717/peerj-cs.243>
- Phillips, J. D., Gwiazdowski, R. A., Ashlock, D., & Hanner, R. (2015). An exploration of sufficient sampling effort to describe intraspecific DNA barcode haplotype diversity: Examples from the ray-finned fishes (Chordata: Actinopterygii). *DNA Barcodes*, 3(1), 66–73. <https://doi.org/10.1515/dna-2015-0008>
- Rumeu, B., Caujapé-Castells, J., Blanco-Pastor, J. L., Jaén-Molina, R., Nogales, M., Elias, R. B., & Vargas, P. (2011). The colonization history of *Juniperus brevifolia* (Cupressaceae) in the Azores Islands. *PLoS One*, 6(11), e27697. <https://doi.org/10.1371/journal.pone.0027697>
- Rumeu, B., Vargas, P., Jaén-Molina, R., Nogales, M., & Caujapé-Castells, J. (2014). Phylogeography and genetic structure of the threatened Canarian *Juniperus cedrus* (Cupressaceae). *Botanical Journal of the Linnean Society*, 175(3), 376–394. <https://doi.org/10.1111/boj.12172>
- Rumeu, B., Vargas, P., & Riina, R. (2016). Incipient radiation versus multiple origins of the Galápagos *Croton scouleri* (Euphorbiaceae). *Journal of Biogeography*, 43(9), 1717–1727. <https://doi.org/10.1111/jbi.12753>
- Sanmartín, I., van der Mark, P., & Ronquist, F. (2008). Inferring dispersal: A Bayesian approach to phylogeny-based Island biogeography, with special reference to the Canary Islands. *Journal of Biogeography*, 35(3), 428–449. <https://doi.org/10.1111/j.1365-2699.2008.01885.x>
- Templeton, A. R., Crandall, K. A., & Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, 132(2), 619–633.
- Thomson, F. J., Letten, A. D., Tamme, R., Edwards, W., & Moles, A. T. (2018). Can dispersal investment explain why tall plant species achieve longer dispersal distances than short plant species? *New Phytologist*, 217(1), 407–415. <https://doi.org/10.1111/nph.14735>
- Valente, L. M., Phillimore, A. B., & Etienne, R. S. (2015). Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands. *Ecology Letters*, 18(8), 844–852. <https://doi.org/10.1111/ele.12461>
- Vargas, P., Nogales, M., Jaramillo, P., Olesen, J. M., Traveset, A., & Heleno, R. (2014). Plant colonization across the Galápagos Islands: Success of the sea dispersal syndrome. *Botanical Journal of the Linnean Society*, 174(3), 349–358. <https://doi.org/10.1111/boj.12142>
- Vargas, P., Rumeu, B., Heleno, R. H., Traveset, A., & Nogales, M. (2015). Historical isolation of the Galápagos carpenter bee (*Xylocopa darwini*) despite strong flight capability and ecological amplitude. *PLoS One*, 10(3), 1–13. <https://doi.org/10.1371/journal.pone.0120597>
- Whittaker, R. J., & Fernández-Palacios, J. M. (2007). *Island biogeography: Ecology, evolution, and conservation*. Oxford University Press.
- Wolff, J. N., Nafisinia, M., Sutovsky, P., & Ballard, J. W. O. (2013). Paternal transmission of mitochondrial DNA as an integral part of mitochondrial inheritance in metapopulations of *Drosophila simulans*. *Heredity*, 110(1), 57–62. <https://doi.org/10.1038/hdy.2012.60>
- Xue, A. T., & Hickerson, M. J. (2020). Comparative phylogeographic inference with genome-wide data from aggregated population pairs. *Evolution*, 74(5), 808–830. <https://doi.org/10.1111/evo.13945>

## BIOSKETCH

**Alberto J. Coello** is interested in reconstructing biogeographic patterns of plant species in the Canary Island and the Mediterranean Floristic Region. To that end, he applies a multi-disciplinary approach based on phylogenetics, phylogeography, population genetics and species distribution modelling.

**Author contributions:** A.J.C., R.H.H., M.F.-M. and P.V. conceived and improved the PAICE method; A.J.C built the R package and analysed species data; A.J.C. and M.F.-M. wrote the paper and all authors critically revised it.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Coello, A. J., Fernández-Mazuecos, M., Heleno, R. H. & Vargas, P. (2022). PAICE: A new R package to estimate the number of inter-island colonizations considering haplotype data and sample size. *Journal of Biogeography*, 49, 577–589. <https://doi.org/10.1111/jbi.14341>