Species delimitation in the presence of strong incomplete lineage sorting and hybridization

Alexandra Anh-Thu Weber[12#], Sabine Stöhr[3] and Anne Chenuil[1]

[1] *Aix-Marseille Université, Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE) - CNRS - IRD – UAPV, Station Marine d'Endoume, Chemin de la Batterie des Lions, 13007 Marseille, France*

[2] *Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland*

[3] *Swedish Museum of Natural History, Department of Zoology, Box 50007, 10405 Stockholm, Sweden,* sabine.stohr@nrm.se

[#] Corresponding author: ale.weber@unibas.ch; Phone: +41 61 207 59 03; Fax: +41 61 207 03 01; ORCID: 0000-0002-7980-388X

Key words: Approximate Bayesian Computation, multi-species coalescent, Discriminant Analysis of Principal Components, cryptic species, Echinoderms, Ophiuroidea

Running head: Cryptic species delimitation

Abstract

Accurate species delimitation is essential to properly assess biodiversity, but also for management and conservation purposes. Yet, it is not always trivial to accurately define species boundaries in closely related species, as strong incomplete lineage sorting might still be present. Additional difficulties may be caused by hybridization, now evidenced as a frequent phenomenon. Here, we propose a three-step framework for species delimitation and divergence history inference: i) unsupervised species discovery based on multilocus genotypes; ii) species validation using the multi-species coalescent; iii) divergence scenario testing (including gene flow) using Approximate Bayesian Computation (ABC) methods. To test this framework, we used the brittle star cryptic species complex *Ophioderma longicauda* that encompasses six mitochondrial lineages, including broadcast spawners and internal brooders. 30 sequence markers (transcriptome-based, mitochondrial or non-coding) for 89 *O. longicauda* and outgroup individuals were used. First, multivariate analyses revealed six genetic clusters, which globally corresponded to the mitochondrial lineages, yet with many exceptions, suggesting ancient hybridization events and challenging traditional barcoding approaches. Second, multi-species coalescent-based analyses validated the six species and provided divergence time estimates, but the sole use of this method failed to accurately delimit species, highlighting the power of multilocus genotype clustering to delimit recently diverged species. Finally, Approximate Bayesian Computation showed that the most likely scenario involves hybridization between brooders and broadcasters. Our study shows that despite strong incomplete lineage sorting and complex speciation history, accurate species delimitation is possible using a three-step framework combining complementary methods.

Introduction

Accurate species delimitation and description is essential to properly assess biodiversity, but also for management and conservation purposes (Agapow et al., 2004; De Queiroz, 2007). Historically and still nowadays, the vast majority of species are delimited using morphological characters, based on descriptions of type specimens for each nominal species, uniquely identified by Latinized names in the binominal nomenclature codified first by Linnaeus in the 18th century (for zoology (Linnaeus, 1758)). However, genetically isolated groups of individuals were detected in many nominal species during the last decades, owing to the use of genetic markers in population genetic studies (Bickford et al., 2007; Knowlton, 1993; Pfenninger & Schwenk, 2007). Such groups are often called cryptic species (see Chenuil et al., submitted, for a rational classification of the different types of putative cryptic species), which are widely spread and homogeneously distributed across the metazoan biodiversity (Pfenninger & Schwenk, 2007). The high occurrence of cryptic species can be explained with three main factors: i) insufficient morphological characterization; ii) recent species divergence (i.e. morphological differences may not have evolved yet); iii) morphological stasis (Knowlton, 1993). Cryptic species were first identified by diagnostic codominant markers such as allozymes (e.g. Knowlton, 1993) or, more recently, by single mitochondrial markers.

Due to their lower effective size, genetic markers from haploid genomes are more affected by genetic drift and thus reach reciprocal monophyly (i.e. alleles of distinct species form separate monophyletic groups) more rapidly than markers from nuclear genomes after species divergence. This explains their power to detect recently diverged cryptic species and their wide use for biodiversity barcoding. However, absence of gene

flow among groups of individuals cannot be established on the basis of markers from single haploid genomes since past bottlenecks or selective sweeps may generate patterns of divergent groups of closely related haplotypes within a panmictic entity (i.e. a group of randomly mating individuals) (Chenuil et al., submitted). In addition, mitochondrial markers only reflect the history of maternal lineages, which can be significantly different from the species history if males and females display different dispersal behaviors. Finally, past hybridization events can be misleading for species identification based on mitochondrial barcodes, as mitochondrial lineages can be incongruent with the species history (Currat, Ruedi, Petit, & Excoffier, 2008; Melo-Ferreira et al., 2014).

Finding independent markers confirming mitochondrial divergence may be difficult for recently diverged species, especially in non-model organisms with scarce to non-existing available genomic data. The most intuitive and popular approach for cryptic species delimitation consists of finding independent markers displaying reciprocal monophyly that are congruently associated within individuals (in linkage disequilibrium) (De Queiroz, 2007; Mkare, Vuuren, & Teske, 2017). However, confirming absence of gene flow should not rely on finding reciprocal monophyly in independent markers (Zhang, Zhang, Zhu, & Yang, 2011) for the following reasons: i) Recently diverged species, especially when their reproductive isolation is conferred by a single locus or few loci, rarely display reciprocally monophyletic markers, as most loci are either monomorphic or display shared polymorphism among species. ii) Recently diverged species may display diagnostic markers but not reciprocal monophyly when diagnosticity resulted from genetic drift (allele loss) but mutation events were not sufficient to create a pattern of reciprocal monophyly. iii) There are other means to establish absence of gene flow using a set of independent genetic markers (e.g. the

diagnosticity of a single Mendelian marker is sufficient, and see approaches based on multilocus genotype developed below).

Various methods (reviewed in Carstens, Pelletier, Reid, & Satler, 2013) propose to delimit species using nucleotide sequences from several markers. Some of these methods are based on divergence levels or consider ratio of within-group and between-group diversity to delimit species (e.g. the automatic barcoding gap discovery ABGD, (Puillandre, Lambert, Brouillet, & Achaz, 2012). Although they provide powerful tools to propose primary species hypotheses for large datasets (Ratnasingham & Hebert, 2007), such approaches cannot prove the absence of gene flow and often depend on arbitrary thresholds or assumptions such as constant effective sizes among ancestor and daughter species. The multispecies coalescent theory (Fujita, Leaché, Burbrink, McGuire, & Moritz, 2012; Yang & Rannala, 2010) provides a statistical framework to model the coalescence of multiple markers during genetic isolation of groups of individuals. Bayesian methods applied to the multispecies coalescent allow establishing the probability of species partitions and phylogenies for a sample of allele sequences from various individuals (Rannala, 2015; Yang & Rannala, 2010) and the most recent development of this method, implemented in the software BPP (Yang, 2015), allows the joint inference of species delimitation and species phylogeny (Rannala & Yang, 2017; Yang & Rannala, 2014). However, despite these qualities and being able to handle some degree of Incomplete Lineage Sorting (ILS) (Carstens, Knowles, & Collins, 2007), the efficiency of these methods depend for a large part on the accumulation of mutations since species separation, and thus can only delimit entities isolated for long enough (Rannala, 2015). In addition, these methods do not integrate the possibility of gene flow after divergence (Ence & Carstens, 2011; Yang & Rannala, 2010). By contrast, methods based on multilocus genotypes (e.g. Structure (Falush, Stephens, & Pritchard, 2003), Structurama

(Huelsenbeck, Andolfatto, & Huelsenbeck, 2011), DAPC (Jombart, Devillard, & Balloux, 2010)), although they are rarely used for species delimitation, can potentially reveal absence of gene flow after a single generation as they do not rely on information on allele relationships, like sequences, but on multilocus genotypes for each individual. In addition, they provide a fast and unbiased way of finding genetically separated entities, as they do not rely on *a priori* knowledge on individual grouping.

Even though much progress has been made in species tree estimation methods, the use of species trees implies that speciation is represented as a dichotomic process. Yet, there is increasing evidence for the role of hybridization and reticulate evolution in or after speciation (R. Abbott et al., 2013; R. J. Abbott, Barton, & Good, 2016; Lamichhaney et al., 2017; Meier et al., 2017). Current species discovery and delimitation methods do not allow for testing such cases, but Approximate Bayesian Computation (ABC) provides powerful methods allowing to do so (Csilléry, Blum, Gaggiotti, & François, 2010; Lopes & Beaumont, 2010), with the simultaneous testing of several divergence scenarios (with or without hybridization) and estimation of demographic parameters (e.g. divergence times, effective population sizes, migration rates). These methods are computationally efficient, as they use simulated datasets for which several summary statistics are compared to the original dataset (instead of likelihood computations). Thus, information rich datasets such as sequence genotypes at tens of loci in a hundred individuals can be exploited using ABC.

In this study, we propose a three-step framework for cryptic species delimitation and divergence history inference, particularly well-suited for very recently diverged species displaying strong ILS: i) unsupervised species discovery using multilocus genotypes; ii) species validation and divergence time estimation using the multi-species

coalescent; iii) divergence scenario testing using ABC (Fig. 1). To test our framework, we used a brittle star cryptic species complex. Brittle stars (Ophiuroidea) encompass a large number of cryptic species (e.g. (Barboza, Mattos, & Paiva, 2015; Baric & Sturmbauer, 1999; Boissin, Féral, & Chenuil, 2008; Boissin, Hoareau, Paulay, & Bruggemann, 2017; Heimeier, Lavery, & Sewell, 2010; Hoareau, Boissin, Paulay, & Bruggemann, 2013; Hunter & Halanych, 2008; Muths, Davoult, Gentil, & Jollivet, 2006; Muths, Jollivet, Gentil, & Davoult, 2009; Naughton, O'Hara, Appleton, & Cisternas, 2014; Pérez-Portela, Almada, & Turon, 2013; Sponer & Roy, 2002; Stöhr & Muths, 2010; Taboada & Pérez-Portela, 2016). The *Ophioderma longicauda* (Bruzelius, 1805) species complex encompasses six mitochondrial lineages and at least two different biological species with contrasting reproductive strategies, namely the broadcast spawners C3 and the brooders C5 (Boissin, Stöhr, & Chenuil, 2011; Stöhr, Boissin, & Chenuil, 2009; Weber, Dupont, & Chenuil, 2013; Weber, Mérigot, Valière, & Chenuil, 2015; Weber, Stöhr, & Chenuil, 2014; Weber et al., 2017). We used 30 sequence markers to test our three-step framework for species delimitation and divergence history inference of this complex of cryptic species. We found that combining all three methods that use different properties of the data provides complementary information such as number of species, among species relationships, divergence time, effective size and gene flow estimations to best represent complex speciation history.

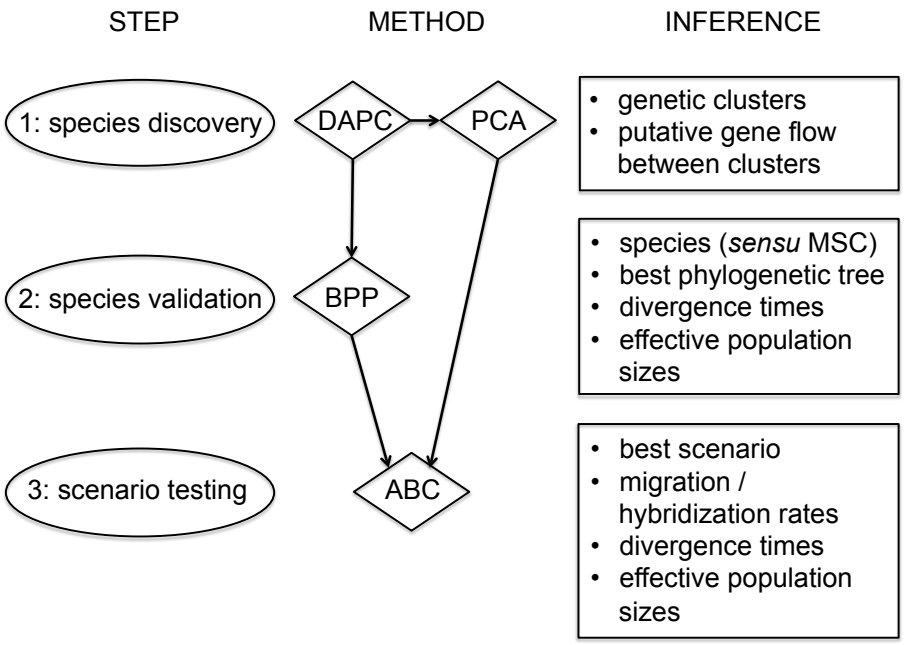

Figure 1: Proposed workflow for cryptic species delimitation including steps, methods used and inference of biologically relevant parameters.

Materials and Methods

Sampling, DNA extraction and marker development

89 individuals including the six *O. longicauda* mitochondrial lineages (L1-L6) and three outgroup species (*Ophioderma teres* (Lyman, 1860), *Ophioderma cinerea* (Müller & Troschel, 1842) and *Ophioderma phoenia* (H.L. Clark, 1918)) were used (Table S1). *Ophioderma* outgroups were used to estimate divergence times of the *O. longicauda* species complex, the latter occurring in the North-East Atlantic Ocean and in the Mediterranean Sea (Fig. 2). Indeed, the species *O. teres* (Eastern Pacific Ocean), *O. phoenia* and *O.cinerea* (West Atlantic Ocean) are geminate pairs that speciated after the closing of the Isthmus of Panama, so the divergence between these species pairs (*O. teres* - *O. phoenia* or *O. teres* - *O.cinerea*) is at least 2.8 Mya (Lessios, 2008).

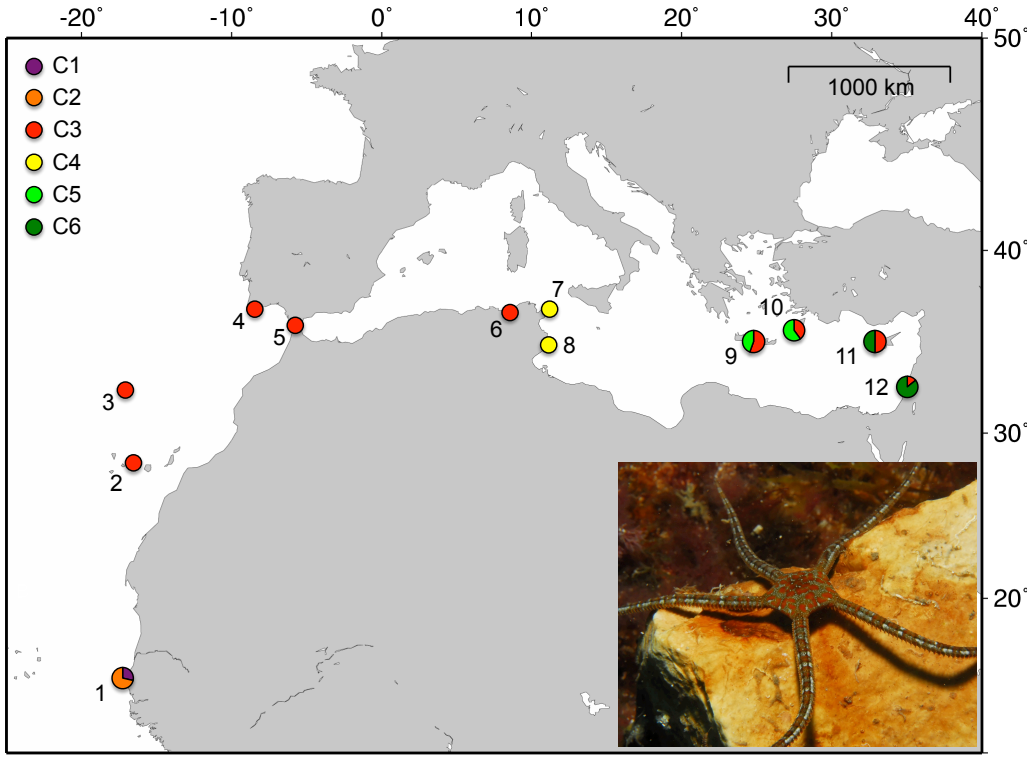

Figure 2: Map of sampling localities with colors corresponding to genetic clusters found with DAPC. 1: Dakar, Senegal. 2: Teneriffe, Canary Islands. 3: Madeira, Portugal. 4: Algarve, Portugal. 5: Ceuta, Spain. 6: Tabarka, Tunisia. 7: Kelibia, Tunisia. 8: Monastir, Tunisia. 9: Agios Pavlos, Crete. 10: Symi island, Greece. 11: Baths of Aphrodite, Cyprus. 12: Ramkine, Beirut and Raoucheh, Lebanon. A photograph of *O. longicauda* C3 from Marseilles, France, is displayed for illustration. Photo credit: Frédéric Zuberer.

DNA was extracted with MN-Tissue Kit (Macherey-Nagel) using an epimotion robot (Eppendorf) following the protocol of (Ribout & Carpentieri, 2013), and eluted in 200µl of sterile water. Extracted DNA was diluted 10x in sterile water prior to PCR. We used orthologous genes from transcriptomes of *O. longicauda* C3 and C5, corresponding to mitochondrial lineages L1 and L3, respectively (Weber et al., 2017, 2015) to develop 55 primer pairs to test. The criteria for marker development were: (i) the marker should be polymorphic; (ii) in half of the markers, at least one diagnostic SNP between C3 and C5 should be present; (iii) the length of the PCR product should be 300-400 bp (due to sequencing technology limitations). In addition, seven already existing markers were used, namely a mitochondrial marker (COI), ribosomal markers (ITS1, ITS2) and four

EPIC markers, introns i21, i36, i50 and i54b (Chenuil et al., 2010; Gérard et al., 2013; Penant, Aurelle, Feral, & Chenuil, 2013). Of the 55 exon-based markers tested, 16 amplified correctly in each lineage of *O. longicauda.* Furthermore, six out of seven existing markers amplified correctly. Finally, 22 markers were PCR amplified in the 89 specimens.

Amplicon sequencing and dataset processing

PCR products of different genes belonging to the same individual were pooled and 96 Illumina libraries were constructed. Paired-end (2x 250 pb) sequencing was performed on a MiSeq Sequencing System (Illumina) by the genomic platform Genotoul (www.genotoul.fr). About thirty millions raw reads were obtained after sequencing. Reads were cleaned, assembled and demultiplexed using the program MOTHUR v 1.31.2 (Schloss et al., 2009). On average, between 1000 and 10,000 sequences were obtained per marker and per individual. Then, identical sequences were clustered and the number of reads per sequence and per individual was counted for each marker.

As the number of reads greatly differed between markers (less than 100 reads to more than 1000 reads), applying a fixed threshold to keep final sequences was not possible. In addition, five markers displayed paralogous genes (e.g. more than two sequences with high and similar number of reads displayed). For this reason, selecting the sequence displaying the highest number of reads could lead to incorrectly selecting and clustering paralogous genes. Therefore, for each of the 22 markers, the number of reads obtained for 5-10 individuals was manually checked to determine a threshold to apply to each individual per marker. One to two sequences were kept per individual and per marker when paralogous genes were unambiguously absent, and up to ten sequences per individual and per marker were kept for genes displaying paralogs. Of the

22 markers, three could not be used due to a too low number of reads obtained after sequence cleaning. Finally, a total of 18 genes were obtained, and since five of them displayed paralogs, 30 markers were available for further analyses (Table S2).

Haplotype networks and concatenated phylogenetic analyses

For each marker, haplotype networks were generated using the median-joining algorithm of Network, version 4.6.1.1 (Bandelt, Forster, & Röhl, 1999). Kimura 2-parameter (K2P) pairwise distances (Kimura, 1980) among mitochondrial lineages (or among species when considering outgroups) were calculated using MEGA v7 (Kumar, Stecher, & Tamura, 2016). The within-group K2P distances were calculated in the same way. Then, the 30 markers were concatenated to create a supermatrix of 8,899 nucleotides for 87 individuals (two individuals were excluded due to a too low PCR success). Maximum Likelihood analyses were performed using RaxML 8.2.11 (Stamatakis, 2014), with 500 replicates (fast bootstrapping) and the GTR+G model of nucleotide evolution.

Species discovery: Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC)

In order to determine the number of existing genetic groups without prior knowledge (i.e. mitochondrial lineage or geographic origin), we performed a Discriminant Analysis of Principal Components (DAPC) (using genotype information, but not sequence information) using the *adegenet* 1.4-1 package from the R software (Jombart et al., 2010). The DAPC is a clustering method that maximizes the between-group variance while minimizing the within-groups variance. This analysis also uses genotypic information for each individual and each locus. First, the Bayesian

Information Criterion (BIC) was used to determine the optimal number of genetic clusters k. Then, DAPC was performed to define the clusters and visualize their relationships. It also provides membership probabilities, i.e. the probabilities for each individual to belong to a particular cluster. Analyses were performed including and excluding the mitochondrial marker COI to infer whether it significantly influenced the genetic clustering. Pairwise F_{ST} were calculated for the six genetic clusters found with the DAPC analysis (see Results) using Genetix (Belkhir, Borsa, Chikhi, Raufaste, & Bonhomme, 2004). We then performed a PCA on the multilocus diploid genotypes to explore the genetic relationships among individuals without constraint and without a priori knowledge on population membership. We visualized (using colors) the genetic proximity among individuals from the distinct clusters previously identified by the DAPC, but the PCA does not use this information and does not attempt to delimit divergent groups. For this reason, it can suggest incomplete separation or hybridizations between the clusters of individuals visualized by distinct colors.

Species validation: the multi-species coalescent

Since genetic clusters obtained after the first approach from multilocus genotypes appeared as separate genetic entities (see Results) we considered that their relationships could be described by tree-like topologies, possibly assuming gene flow events between some clusters. Based on the genetic clusters found with DAPC, we calculated the average distance between clusters using between group K2P distances (based on the 30 markers) implemented in MEGA 6.0.5. Then, we reconstructed a phylogenetic tree based on the distance between clusters, using the Neighbor-Joining method, to define a starting tree for multi-species coalescent based analyses (Fig. 3, scenario 1). Joint Bayesian species delimitation and species tree estimation was

conducted using the program BPP v3.3 (analysis A11; (Rannala & Yang, 2017; Yang, 2015)). The method uses the multispecies coalescent model to compare different models of species delimitation and species phylogeny in a Bayesian framework, accounting for incomplete lineage sorting due to ancestral polymorphism and gene tree-species tree conflicts (Rannala & Yang, 2013; Yang & Rannala, 2010, 2014). The population size parameters (θs) are assigned the gamma prior G(2, 1000), with mean 2/2000 = 0.001. The divergence time at the root of the species tree (τ0) is assigned the gamma prior G(2, 100), while the other divergence time parameters are assigned the Dirichlet prior (Yang & Rannala, 2010: equation 2). After 10,000 burnin iterations, 50,000 MCMC samples were recorded with a sample frequency of 2. Each analysis was run three times to confirm consistency between runs. Analyses were run using the full dataset and species were defined using the clusters found with DAPC.

As it was recently suggested that 'species discovery' methods may eventually not be necessary due to improving of algorithms and computational power (Rannala, 2015), we tested the accuracy of BPP alone to discover and delimit species, using a subset of our dataset for computational purposes. We used all C3 (9 individuals) & C5 (11 individuals) specimens from Greece, known to represent two biological species (brooding and broadcast spawning individuals; (Weber et al., 2015, 2014)). A DAPC was first performed on this sub-dataset. Then, each individual was set as a single species in BPP, and three replicate analyses A11 (joint species delimitation and species tree estimation) were performed using the same parameters as previously mentioned.

Divergence model testing using ABC

Species delimitation analyses perform well to determine species number and species phylogeny, but speciation history may be more complex than a simple

dichotomic process as is a species tree. We used the PCA results to identify possible cases of hybridization between groups of individuals. More specifically, the positioning of C4 (Tunisian) individuals in the PCA suggested possible hybridization event between C3 and C5 (see Results). In addition, the C4 individuals displayed incongruent genetic signals between mitochondrial and nuclear markers previously described (Weber et al., 2015, 2014), as their mitochondrial haplotypes (COI) were closely related to the brooding species C5, whereas their nuclear genotypes (intron i51) were shared with the broadcast spawning species C3. It is noteworthy that i51 was shown to be monomorphic in the brooding species C5 and C6 (Weber et al., 2015).

Then, we used an ABC framework to test eight different divergence scenarios for the *Ophioderma longicauda* species complex, including or excluding hybridization events (Fig. 3). The posterior probability of each scenario, as well as effective sizes, divergence times and admixture rate were estimated using ABC implemented in DIYABC v2.1.0 (Cornuet et al., 2014). Six summary statistics were used to estimate posterior probability of parameters: For the 'one sample summary statistics', the number of haplotypes, the number of segregating sites and the mean of pairwise differences were used. For the 'two-sample summary statistics', the number of haplotypes, the mean of pairwise differences (between groups) and the F_{ST} statistics (between groups) were used. For ABC analyses, data from the six *Ophioderma longicauda* clusters were used, excluding the outgroups. In addition, 9 markers were excluded due to their low amplification success in some clusters. Three sequence groups were defined, each one with a different mutation model. The first group included 19 transcriptome-based markers, the second group included the mitochondrial marker COI and the third group included two introns (Table S2). Default priors were used in preliminary analyses (800,000 simulated datasets) and were then adjusted using posterior distributions and

pre-evaluation verifications. When each posterior probability of parameters fell in the prior range in preliminary analyses, 8,000,000 simulated datasets were used to estimate posterior probabilities of parameters and each scenario. Model checking was performed using each available summary statistic, to verify that the parameter values of observed data belonged to posterior distributions.

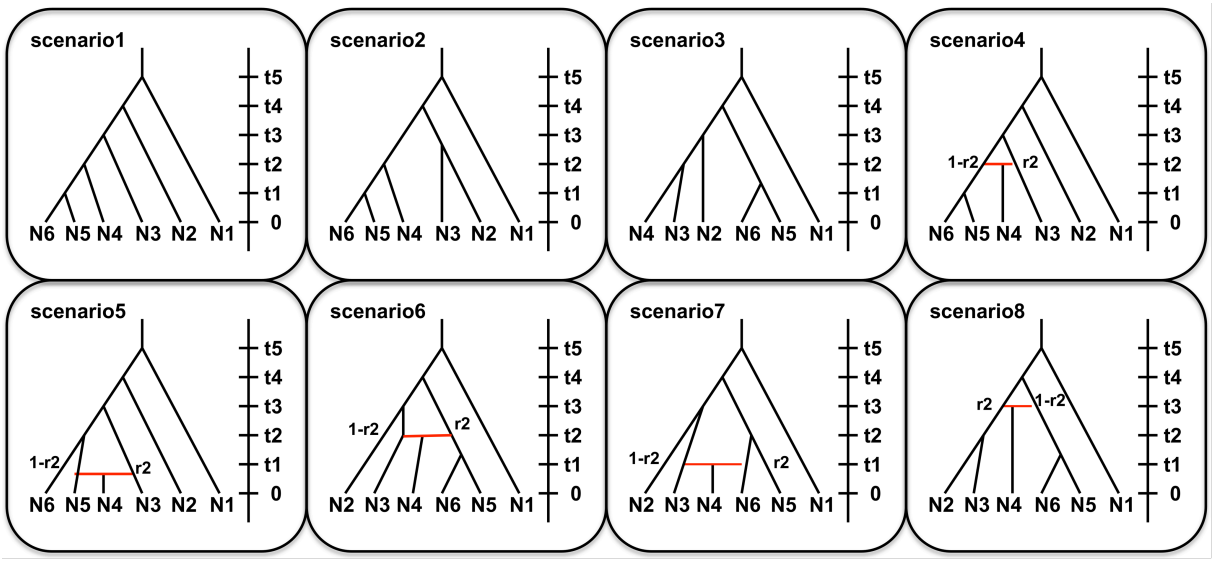

Figure 3: Eight divergence scenarios tested to infer evolutionary history of *O. longicauda* species complex divergence.

Results

Concatenated dataset analyses fail delimiting species

Using transcriptome based markers we successfully amplified, sequenced and sorted 30 informative genetic markers (Table S2). Network analyses showed that the majority of markers displayed incomplete lineage sorting, except the mitochondrial marker COI and, although partially, the markers 68241_I.I, i50_II and 98699 (Fig. S1). Not only reciprocal monophyly is not observed among previously-identified species (brooding C5 and broadcast spawning C3 in Crete, (Weber et al., 2017, 2015)) but the

large majority of alleles are shared among these species (Fig. S1). K2P distances among mitochondrial lineages ranged from 0.8% between L3 and L3b to 10.7% between L2 and L6 within the *O. longicauda* complex, whereas it ranged from 8.7% between L2 and *O. phoenia* to 11.8% between L6 and *O. phoenia* when considering the three outgroup species (Table S3). The Maximum Likelihood analysis of the concatenated dataset strongly supported the three *Ophioderma* outgroup species (85-100% Bootstrap Support) and provided moderate to good support for two groups of individuals from Dakar, namely all individuals from lineage L6 (89% BS) and individuals from lineage L5 (84% BS) (Fig. S2, purple and orange branches, respectively). There was no further supported group of individuals based on mitochondrial lineage or geography (Fig. S2), which was expected given that the species complex displays strong ILS in the majority of loci (Fig. S1).

Multilocus genotype analyses identify six genetic clusters

The DAPC showed that the optimal number of clusters (i.e. the minimal BIC value) was six (Fig. S3A). The six clusters were very distinct with nearly no overlapping in the 2D representation (Fig. 4A) and 100% probability of memberships for each individual (Figure S3B). The cluster C1, including all individuals of mitochondrial lineage L6 (from Dakar and Madeira) forms a well-defined group distant from the five other groups (Fig. 4A, Table 1). The cluster C2 includes all L5 individuals from Dakar, whereas the widely distributed broadcast-spawning cluster C3 encompasses all L1 and L5 individuals from Canary Island to Lebanon (Fig. 2, Table 1). The cluster C4 includes all L3b Tunisian individuals (Fig. 2). Finally, the brooding individuals are distributed in two different genetic clusters, incongruent with mitochondrial lineages but congruent with geography. Cluster C5 includes all L2 and L3 individuals sampled in Greece, whereas

cluster C6 includes all L2 and L4 individuals sampled in Cyprus and Lebanon (Table 1). The DAPC run without the mitochondrial COI marker provided the same clustering, showing that this marker did not bias the clustering process (Fig. S4). The PCA gave essentially the same results as the DAPC, except that the higher genetic diversity of the broadcast spawners was more visible (Fig. 4B-C). In addition, C4 was even closer to C3 and C5, highlighting the capacity of PCA to explore the natural distribution of individuals and its power to detect potential hybridization events. Furthermore, one individual from Madeira assigned to the cluster C1 was not clustering with the other C1 individuals from Dakar, but was rather at mid distance between the C1 and C3 individuals, suggesting the potential presence of a hybrid between C1 and C3. Yet, further sampling would be required to properly assess this hypothesis. Nevertheless, we excluded this individual from further analyses. Finally, pairwise F_{ST} among clusters were high, ranging from 0.19 between C2 and C3 to 0.47 between C1 and C5 (Table 2). In addition, all F_{ST} values were significant after a permutation test (Table 2).

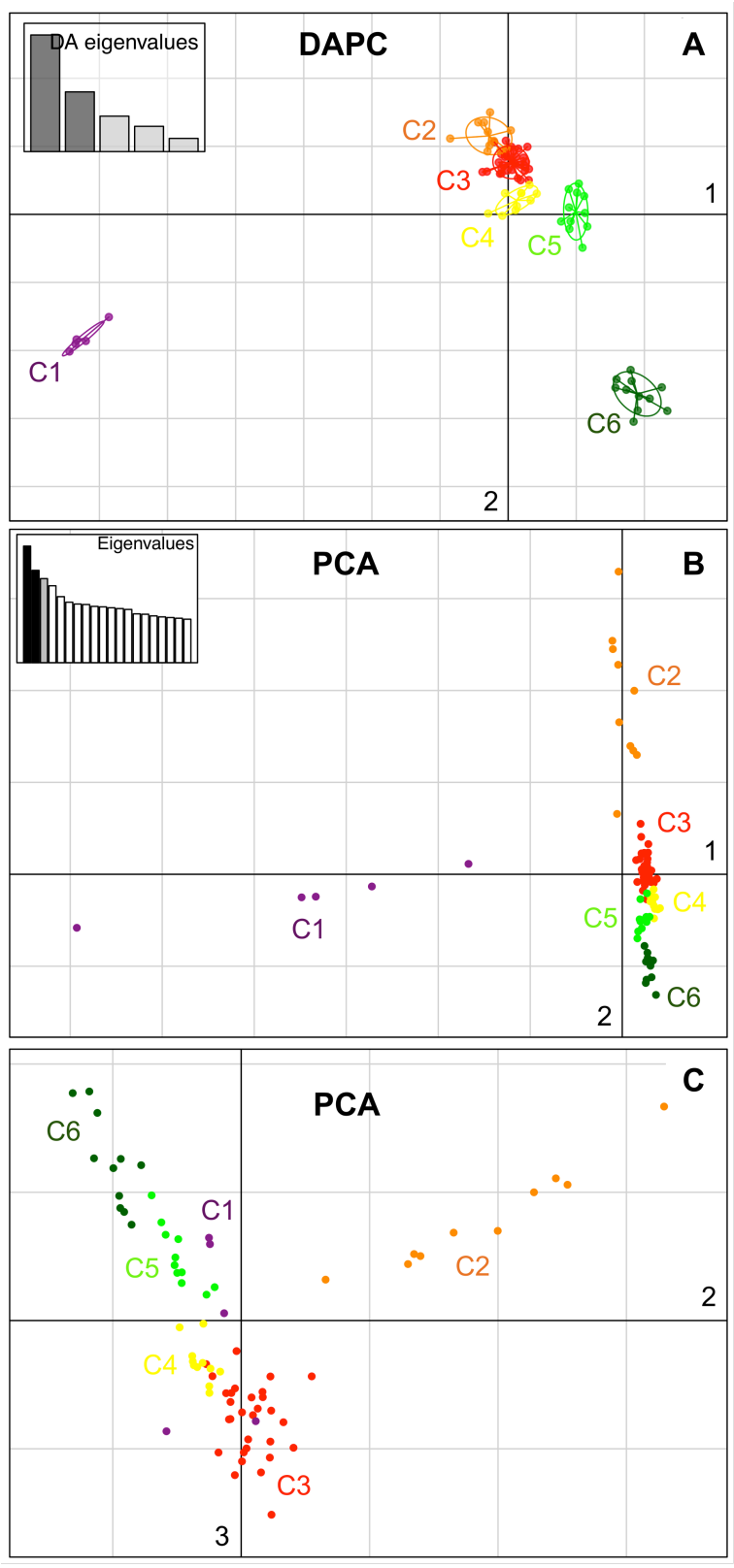

Figure 4: A: Results of the Discriminant Analysis of Principal Components (DAPC). The six different genetic clusters are displayed**. B:** Results of the Principal Components Analysis (PCA). Axes 1 and 2 are plotted. Individuals are colored according to their genetic cluster found in DAPC. **C:** Results of the Principal Components Analysis (PCA). Axes 1 and 3 are plotted. Individuals are colored according to their genetic cluster found in DAPC.

The multispecies coalescent validates the six species

BPP was used to jointly perform species delimitation and species tree estimation using the six genetic clusters and the three outgroup species. In the three replicate analyses, the six *Ophioderma* genetic clusters C1-C6 and the three outgroups species (*O. teres, O. cinerea* and *O. phoenia*) were fully supported (posterior probability of 9 species=1; Table S4). Furthermore, the species tree of the *O. longicauda* species complex was also highly supported (C1 most ancestral, C5 & C6 more recently diverged (Fig. S5); posterior probability=0.92-0.99), although the full topologies were different due to different placements of the three outgroups (Table S4). Using the divergence time of the geminate species *O. teres* and *O. cinerea / O. phoenia* (at least 2.8 mya; (Lessios, 2008)), the divergence times within the *Ophioderma longicauda* species complex were inferred to be at least 537,000 years ago [95% CI: 445,223-682,795] (Table 3; Fig. S5). It is noteworthy that the BPP analyses run on a sub–dataset of nine C3 individuals and eleven C5 individuals considering each individual as a candidate species in the starting tree gave unsupported results, with unstable numbers of estimated species and low posterior probabilities among replicate analyses (6, 3 and 2 species; Table S5). Therefore, BPP performed poorly to delimit species without a meaningful starting species tree. On the opposite, the DAPC succeeded in finding the true number of species and affecting individuals to them (Fig. S6).

A Divergence scenario supports past hybridization

After pre-evaluation of the priors for the eight scenarios (Fig. S7A), posterior probabilities of scenarios tested with ABC indicated that the most probable scenario was the scenario 5, including hybridization between C3 and C5 (PP=0.67; Table S6). The second most likely scenario was the scenario 7, also including a hybridization event

between C3 and C5 (PP=0.26; Table S6). The remaining scenarios were not supported. After model checking of scenario 5 (Fig. S7B), parameter estimation indicated that the widespread broadcast spawning cluster C3 displayed the largest effective population size, 3 to 10 times larger than the effective population sizes of the brooding species C4, C5 and C6 (Table 4; Fig. S8). The divergence time estimations indicated that C1 split from other *O. longicauda* clusters about 512,000 generations ago and that the broadcasters and the brooders split about 222,000 generations ago (Table 4). The hybridization event giving rise to C4 was estimated about 90,000 generations ago, with a high proportion of C4 genome originating from C3 (about 86.8%; Table 4). Overall, the divergence events of the *O. longicauda* species complex follow a pattern of West to East differentiation (Fig. 2). The divergence time estimations of C1 from the common ancestor of C2-C6 were similar between BPP [mode: 537,380; 95% CI: 445,223-682,795 years] and DIYABC [mode: 512,000; 95% CI: 331,000-928,000 generations] considering a generation time of one year. For the divergence times within *O. longicauda* species complex (C2-C6), we rather refer to the DIABC estimations as the divergence model is more accurate.

Discussion

Species limits and divergence history deciphered in the O. longicauda species complex

In this study, six *Ophioderma* species (C1-C6) were unambiguously delimited, one of them (C4) most likely originating from the hybridization of C3 and C5. So far, only two *Ophioderma* species have been described in the Eastern Atlantic; *Ophioderma longicauda* (Bruzelius, 1805), from Dakar to Spain in the Atlantic and in the Mediterranean, and *Ophioderma wahlbergii* Müller & Troschel, 1842 in South Africa, even though it was recently shown that the Mediterranean sympatric C3 & C5 and C3 & C6 are different

biological species (Weber et al., 2015, 2014). The emergence of the *Ophioderma* genus occurred most likely around the Caribbean Sea, before the closing of the Panama Isthmus. Indeed, most currently recognized extant species (26/28) of this genus thrive in this region (Stöhr et al., 2009) and the oldest confirmed *Ophioderma* fossil (about 10 million years old, Tortonian, early Late Miocene), is from South America (Martínez & Río, 2008). The most divergent species C1 occurring in West Africa split from the common ancestor of C2-C6 at least 537,000 years ago. The latter was the only clear species in the concatenated sequences phylogenetic analysis. Given that C1 and C2 were sampled in very close localities (11-17 km apart), this is further evidence that C1 and C2 are different biological species. Interestingly, Greef (1882) described a new species, *Ophioderma guineensis* Greef, 1882, from West Africa (Gulf of Guinea), which was later considered conspecific with O. *longicauda*, as its distinguishing morphological characters were assumed to fall within the variability of *O. longicauda* (Madsen, 1970). It is possible that this *O. longicauda* "variety" is actually the different biological species that we define here as C1. Yet, fresh samples from this locality are required to test this hypothesis.

Two other broadcast spawners were found, C2 in Senegal and the widespread C3 (from Canary Islands to Lebanon), whereas two species corresponded to brooders (C5 in Greece and C6 in Cyprus and Lebanon). The cluster C4, occurring in Tunisia, is most likely also a brooder, as it displays typical characteristics of brooders (e.g. mitochondrial lineage close to the brooding C5; small effective population size; ecological preference to low depth; (Weber et al., 2014)). Gonad examinations of C4 specimens were unsuccessful to determine their reproductive strategy as sampling was performed outside the reproductive season. Yet, it is known that brooders occur in this region as brooding specimens were previously sampled in Tunisia in 1849 and 1924 (Stöhr et al.,

2009). Unfortunately, molecular characterization of these samples failed due to poor DNA quality. Interestingly, the most likely origin of the cluster C4 is hybridization between C3 and C5, confirming our first hypothesis. A formal taxonomic revision of *Ophioderma longicauda* is in progress (Stöhr, Weber, Boissin, & Chenuil, in preparation). The ancestral strategy of the *Ophioderma* genus is broadcast spawning, as all *Ophioderma* from the Western Atlantic and *O. longicauda* C3 are broadcast spawners. Brooding evolved most likely about 222,000 generations ago in the common ancestor of C5 and C6. Interestingly, another independent evolution of brooding occurred in *O. wahlbergii*, which displays much larger and fewer young in its bursae than *O. longicauda* C5 (Landschoff & Griffiths, 2015).

COI: one marker is not enough

Mitochondrial barcodes such as COI have been widely used in species delimitation and species complex discovery (e.g. Hebert, Ratnasingham, & Waard, 2003) due to the numerous advantages of mitochondrial DNA such as its ubiquity, ease of amplification, high mutation rate and finally its reduced effective size compared to nuclear DNA which makes isolated populations diverge by genetic drift (and eventually reach reciprocal monophyly) more rapidly. In fact, it allowed in the first place the discovery of the *O. longicauda* species complex (Boissin et al., 2011; Stöhr et al., 2009), and it is still efficient to discover additional cryptic species, including brittle stars (Boissin et al., 2017). Nevertheless, mitochondrial lineages did not correspond to species (e.g. genetic clusters) in many cases. For instance, the cluster C3 encompasses individuals displaying the lineages L1 and L5. The same applied for the cluster C5 (lineages L2 and L3) and the cluster C6 (lineages L2 and L4). This is most likely the result of ancient introgression events. Mitochondrial DNA is particularly prone to both

selective and introgression sweeps (Currat et al., 2008; Galtier, Nabholz, Glémin, & Hurst, 2009; Pons, Sonsthagen, Dove, & Crochet, 2014; Toews & Brelsford, 2012), in contrast to nuclear DNA. Finally, selection events may be responsible for the retention of particular mitochondrial haplotypes. This study emphasizes the necessity of using nuclear markers to accurately delimit species.

DAPC, BPP and ABC: an efficient combination of methods for accurate species delimitation

Here, we propose a three-step approach to delimit recently diverged species and infer their divergence history. We used 30 genetic markers, of which 25 were transcriptome-based, to delimit species in the *Ophioderma longicauda* species complex. This is a high number of sequence markers, given that from the 28 studies presented in a review on species delimitation, only two used more than 10 genetic markers (Carstens et al., 2013). The first step ("species discovery") was performed using DAPC clustering, based on multi-locus genotypes (e.g. the sequence information was not used, only the allele frequencies) and revealed the presence of six distinct genetic clusters. We showed that this type of clustering approach is more powerful than the concatenation-based phylogenetic reconstructions, since diagnostic differences are not needed in order to find genetic clusters, only frequency differences. Therefore, clustering approaches are appropriate for recently diverged species displaying strong incomplete lineage sorting. Then, in a second step ("species validation"), we validated these six clusters as species using BPP that, in addition to providing a species tree and estimating the most likely number of species, allows the estimation of parameters such as effective population sizes and divergence times. Such a two-step approach has already been performed in other studies (e.g. Barrett & Freudenstein, 2011; Leaché & Fujita, 2010; Satler, Carstens, & Hedin, 2013) using Structurama (Huelsenbeck & Andolfatto, 2007; Huelsenbeck et al.,

2011) or Structure (Falush et al., 2003; Pritchard, Stephens, & Donnelly, 2000) to find genetic clusters based on Bayesian inference and BPP (Yang & Rannala, 2010) or spedeSTEM (Ence & Carstens, 2011) to delimit species based on those clusters, using an initial defined species tree. Besides being orders of magnitude faster than Bayesian clustering methods (Structure or Structurama), DAPC performed better under complex (i.e. departing from the island model) population genetic models (Jombart et al., 2010). In addition, we showed that the sole use of BPP failed to accurately delimit species, highlighting the need of the first "species discovery" step.

In a third step ("scenarios testing"), we propose to go one step further than discovering and delimiting species by inferring a more realistic divergence history, including hybridization, with model testing using ABC. Such methods allow the comparison of complex models including hybridization, reticulate evolution and demographic events (e.g. Roux, Tsagkogeorga, Bierne, & Galtier, 2013). We found that the most supported scenario included a hybridization event between the broadcast spawners C3 and the brooders C5. Some additional past hybridization events may also have occurred between the divergent C1 and the broadcast spawner C3. Indeed, an individual sampled in Madeira displayed many common alleles with C1, but also many common alleles with C3. Yet, due to the presence of a single potential C1-C3 hybrid, we were not able to test this hypothesis. Nevertheless, this suggests that hybridization and introgression may be common in *Ophioderma* species.

A previous study (Camargo, Morando, Avila, & Sites, 2012) tested the accuracy of BPP (Yang & Rannala, 2010), spedeSTEM (Ence & Carstens, 2011) and ABC methods (Csilléry, François, & Blum, 2012) for species delimitation. Based on simulations, the authors found that BPP was overall the most accurate, ABC displaying an intermediate

accuracy and spedeSTEM the lowest accuracy. All methods displayed lower accuracy when gene flow was incorporated, yet ABC displayed the lowest decrease in accuracy to delimit species. Rather than finding the overall best species delimitation method, we propose to use several consecutive methods to first find the number of distinct genetic entities, and then to estimate the divergence scenarios, therefore taking advantage of the best qualities of each method.

Albeit successful, our pipeline based on exonic amplified markers relies on pre-existing genomic resources to develop genetic markers, contrary to other methods such as RAD-sequencing and associated techniques (Davey et al., 2011). RAD-sequencing has been successfully used to delimit species (e.g. Pante et al., 2015), yet the efficiency of this method diminished drastically with genetic distance of compared species. Indeed, Pante et al. (2015) report that >70% of loci were lost when species displaying 0.028% of mitochondrial divergence were compared (1-2 myr divergence time) and 97% of loci were lost for species displaying 2.2% of mitochondrial divergence (9-16 myr divergence). This is expected given that the majority of RAD loci are found in non-coding fast evolving DNA. Here, we could successfully retrieve 84-100% of markers for C1-C6 (10.7% maximum mitochondrial divergence (Table S3); divergence at least 537,000 years ago (Table 3)) and 54% of the markers for the outgroup species (11.8% maximum mitochondrial divergence (Table S3); divergence at least 7.3 million years ago (Table 3)), highlighting that our exon-based method performs better than RAD sequencing for distantly related species. In addition, due to their longer sequences compared to RAD loci, our method allows the analysis of haplotype networks (Fig. S1). Therefore, coding sequence markers are useful to compare simultaneously closely and distantly related species. To circumvent the use of individual PCR amplification, one could use our analytic framework with exon-capture data, a method shown efficient to capture exons

displaying up to 12% of sequence divergence (Hancock-Hanser et al., 2013; Hugall, O'Hara, Hunjan, Nilsen, & Moussalli, 2015 for exon-capture specific to brittle stars). Until now, these data have mainly been used for phylogenomic purposes (e.g. O'Hara, Hugall, Thuy, & Moussalli, 2014; O'Hara, Hugall, Thuy, Stöhr, & Martynov, 2017) but they could as well be used for cryptic species delimitation with multilocus genotype approaches. To conclude, the use of this three-step approach with coding sequence markers allows comparisons at the within- and between-species levels, and bridging the gap between them. We emphasize the power of using these three steps consecutively and not individually, in order to uncover the most realistic divergence history of a species complex.

Acknowledgments

We are very grateful to the many people who contributed to sampling of *Ophioderma longicauda* specimens: Helmut Zibrowius, Christos Arvanitidis, Thanos Dailianis, Elena Sarropoulou, Magdalini Christodoulou, Zined Marzouk, Didier Weber, Thi Weber and Philipp Moser. Many thanks to Francisco Alonso Solis-Marin and Harilaos Lessios for providing *Ophioderma* outgroup samples. We also would like to thank Laurent Abi-Rached for his advice on phylogenetic analyses, Arnaud Estoup for his advice on DIYABC and the genomic sequencing platform Genotoul (INRA, Toulouse) for the Illumina sequencing. Finally, we thank the support team of sciCORE (center for scientific computing, University of Basel, http://scicore.unibas.ch/) for providing access to computational resources, especially Pablo Escobar Lopez.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., … Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*(2), 229–246. doi:10.1111/j.1420-9101.2012.02599.x

Abbott, R. J., Barton, N. H., & Good, J. M. (2016). Genomics of hybridization and its evolutionary consequences. *Molecular Ecology*, *25*(11), 2325–2332. doi:10.1111/mec.13685

Agapow, P., Bininda-Emonds, O. R. P., Crandall, K. A., Gittleman, J. L., Mace, G. M., Marshall, J. C., & Purvis, A. (2004). The Impact of Species Concept on Biodiversity Studies. *The Quarterly Review of Biology*, *79*(2), 161–179. doi:10.1086/383542

Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*(1), 37–48.

Barboza, C. A. de M., Mattos, G., & Paiva, P. C. (2015). Brittle stars from the Saint Peter and Saint Paul Archipelago: morphological and molecular data. *Marine Biodiversity Records*, *8*, e16. doi:10.1017/S1755267214001511

Baric, S., & Sturmbauer, C. (1999). Ecological parallelism and cryptic species in the genus *Ophiothrix* derived from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, *11*(1), 157–162. doi:10.1006/mpev.1998.0551

Barrett, C. F., & Freudenstein, J. V. (2011). An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. *Molecular Ecology*, *20*(13), 2771–2786. doi:10.1111/j.1365-294X.2011.05124.x

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., & Bonhomme, F. (2004). GENETIX 4.05, Population genetics software for Windows TM. *Université de Montpellier II. Montpellier*.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., … Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution*, *22*(3), 148–155. doi:10.1016/j.tree.2006.11.004

Boissin, E., Féral, J. P., & Chenuil, A. (2008). Defining reproductively isolated units in a cryptic and syntopic species complex using mitochondrial and nuclear markers: the brooding brittle star, *Amphipholis squamata* (Ophiuroidea). *Molecular Ecology*, *17*(7), 1732–1744. doi:10.1111/j.1365-294X.2007.03652.x

Boissin, E., Hoareau, T. B., Paulay, G., & Bruggemann, J. H. (2017). DNA barcoding of reef brittle stars (Ophiuroidea, Echinodermata) from the southwestern Indian Ocean evolutionary hot spot of biodiversity. *Ecology and Evolution*, n/a–n/a. doi:10.1002/ece3.3554

Boissin, E., Stöhr, S., & Chenuil, A. (2011). Did vicariance and adaptation drive cryptic speciation and evolution of brooding in *Ophioderma longicauda* (Echinodermata: Ophiuroidea), a common Atlanto-Mediterranean ophiuroid? *Molecular Ecology*, *20*(22), 4737–4755. doi:10.1111/j.1365-294X.2011.05309.x

Camargo, A., Morando, M., Avila, L. J., & Sites, J. W. (2012). Species Delimitation with Abc and Other Coalescent-Based Methods: A Test of Accuracy with Simulations and an Empirical Example with Lizards of the Liolaemus Darwinii Complex (squamata: Liolaemidae). *Evolution*, *66*(9), 2834–2849. doi:10.1111/j.1558-5646.2012.01640.x

Carstens, B. C., Knowles, L. L., & Collins, T. (2007). Estimating Species Phylogeny from Gene-Tree Probabilities Despite Incomplete Lineage Sorting: An Example from Melanoplus Grasshoppers. *Systematic Biology*, *56*(3), 400–411. doi:10.1080/10635150701405560

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. *Molecular Ecology*, *22*(17), 4369–4383. doi:10.1111/mec.12413

Chenuil, A., Hoareau, T. B., Egea, E., Penant, G., Rocher, C., Aurelle, D., … Mousset, S. (2010). An efficient method to find potentially universal population genetic markers, applied to metazoans. *BMC Evolutionary Biology*, *10*(1), 276. doi:10.1186/1471-2148-10-276

Chenuil, A., Cahill, A., Delemontey, N., Du Salliant du Luc, E. & Fanton, H. Problems and questions posed by cryptic species. A framework to guide future studies. Book chapter: From Assessing to Conserving Biodiversity - Beyond the Species Approach. *Submitted to Springer for the Springer series "History, Philosophy and Theory of the Life Sciences"*

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., … Estoup, A. (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, *30*(8), 1187–1189. doi:10.1093/bioinformatics/btt763

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. doi:10.1016/j.tree.2010.04.001

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. doi:10.1111/j.2041-210X.2011.00179.x

Currat, M., Ruedi, M., Petit, R. J., & Excoffier, L. (2008). The Hidden Side of Invasions: Massive Introgression by Local Genes. *Evolution*, *62*(8), 1908–1920. doi:10.1111/j.1558-5646.2008.00413.x

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510. doi:10.1038/nrg3012

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, *56*(6), 879–886.

Ence, D. D., & Carstens, B. C. (2011). SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources*, *11*(3), 473–480. doi:10.1111/j.1755-0998.2010.02947.x

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, *164*(4), 1567–1587.

Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., & Moritz, C. (2012). Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*, *27*(9), 480–488. doi:10.1016/j.tree.2012.04.012

Galtier, N., Nabholz, B., Glémin, S., & Hurst, G. D. D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology*, *18*(22), 4541–4550. doi:10.1111/j.1365-294X.2009.04380.x

Gérard, K., Guilloton, E., Arnaud-Haond, S., Aurelle, D., Bastrop, R., Chevaldonné, P., … Chenuil, A. (2013). PCR survey of 50 introns in animals: Cross-amplification of homologous EPIC loci in eight non-bilaterian, protostome and deuterostome phyla. *Marine Genomics*, *12*, 1–8. doi:10.1016/j.margen.2013.10.001

Greeff, R. (1882). Echinodermen beobachtet auf einer Reise nach der Guinea-Insel Sao-Thomé. *Zoologischer Anzeiger*, *107*, 156–159.

Hancock-Hanser, B. L., Frey, A., Leslie, M. S., Dutton, P. H., Archer, F. I., & Morin, P. A. (2013). Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, *13*(2), 254–268. doi:10.1111/1755-0998.12059

Hebert, P. D. N., Ratnasingham, S., & Waard, J. R. de. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(Suppl 1), S96–S99. doi:10.1098/rsbl.2003.0025

Heimeier, D., Lavery, S., & Sewell, M. A. (2010). Molecular species identification of *Astrotoma agassizii* from planktonic embryos: further evidence for a cryptic species complex. *Journal of Heredity*, *101*(6), 775–779. doi:10.1093/jhered/esq074

Hoareau, T. B., Boissin, E., Paulay, G., & Bruggemann, J. H. (2013). The Southwestern Indian Ocean as a potential marine evolutionary hotspot: perspectives from comparative phylogeography of reef brittle-stars. *Journal of Biogeography*, *40*(11), 2167–2179. doi:10.1111/jbi.12155

Huelsenbeck, J. P., & Andolfatto, P. (2007). Inference of Population Structure Under a Dirichlet Process Model. *Genetics*, *175*(4), 1787–1802. doi:10.1534/genetics.106.061317

Huelsenbeck, J. P., Andolfatto, P., & Huelsenbeck, E. T. (2011). Structurama: Bayesian Inference of Population Structure. *Evolutionary Bioinformatics Online*, *7*, 55–59. doi:10.4137/EBO.S6761

Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R., & Moussalli, A. (2015). An exon-capture system for the entire class Ophiuroidea. *Molecular Biology and Evolution*, *33*(1), 281–294.

Hunter, R. L., & Halanych, K. M. (2008). Evaluating connectivity in the brooding brittle star *Astrotoma agassizii* across the Drake passage in the Southern Ocean. *Journal of Heredity*, *99*(2), 137–148. doi:10.1093/jhered/esm119

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. doi:10.1186/1471-2156-11-94

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120. doi:10.1007/BF01731581

Knowlton, N. (1993). Sibling Species in the Sea. *Annual Review of Ecology and Systematics*, *24*, 189–216. doi:10.2307/2097177

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. doi:10.1093/molbev/msw054

Lamichhaney, S., Han, F., Webster, M. T., Andersson, L., Grant, B. R., & Grant, P. R. (2017). Rapid hybrid speciation in Darwin's finches. *Science*, eaao4593. doi:10.1126/science.aao4593

Landschoff, J., & Griffiths, C. L. (2015). Brooding behavior in the shallow-water brittle star Ophioderma wahlbergii. *Invertebrate Biology*, *134*(2), 168–179. doi:10.1111/ivb.12081

Leaché, A. D., & Fujita, M. K. (2010). Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). *Proceedings of the Royal Society B: Biological Sciences*, rspb20100662. doi:10.1098/rspb.2010.0662

Lessios, H. A. (2008). The great American schism: divergence of marine organisms after the rise of the Central American Isthmus. *Annual Review of Ecology, Evolution, and Systematics*, *39*, 63–91.

Linnaeus, C. (1758). *Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis* (10th ed., Vol. 1). Stockholm: Holmiae.

Lopes, J. S., & Beaumont, M. A. (2010). ABC: A useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, *10*(6), 825–832. doi:10.1016/j.meegid.2009.10.010

Madsen, F. J. (1970). West African Ophiuroids. *Atlantide Report*, *11*, 151–243.

Martínez, S., & Río, C. J. D. E. L. (2008). A new, first fossil species of. *Zootaxa*, *52*, 43 – 52.

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*, 14363. doi:10.1038/ncomms14363

Melo-Ferreira, J., Vilela, J., Fonseca, M. M., Fonseca, D., R, R., Boursot, P., & Alves, P. C. (2014). The Elusive Nature of Adaptive Mitochondrial DNA Evolution of an Arctic Lineage Prone to Frequent Introgression. *Genome Biology and Evolution*, *6*(4), 886–896. doi:10.1093/gbe/evu059

Mkare, T. K., Vuuren, B. J. van, & Teske, P. R. (2017). Conservation implications of significant population differentiation in an endangered estuarine seahorse. *Biodiversity and Conservation*, *26*(6), 1275–1293. doi:10.1007/s10531-017-1300-5

Muths, D., Davoult, D., Gentil, F., & Jollivet, D. (2006). Incomplete cryptic speciation between intertidal and subtidal morphs of *Acrocnida brachiata* (Echinodermata: Ophiuroidea) in the Northeast Atlantic. *Molecular Ecology*, *15*(11), 3303–3318. doi:10.1111/j.1365-294X.2006.03000.x

Muths, D., Jollivet, D., Gentil, F., & Davoult, D. (2009). Large-scale genetic patchiness among NE Atlantic populations of the brittle star *Ophiothrix fragilis*. *Aquatic Biology*, *5*, 117–132.

Naughton, K. M., O'Hara, T. D., Appleton, B., & Cisternas, P. A. (2014). Antitropical distributions and species delimitation in a group of ophiocomid brittle stars (Echinodermata: Ophiuroidea: Ophiocomidae). *Molecular Phylogenetics and Evolution*, *78*, 232–244. doi:10.1016/j.ympev.2014.05.020

O'Hara, T. D., Hugall, A. F., Thuy, B., & Moussalli, A. (2014). Phylogenomic resolution of the class Ophiuroidea unlocks a global microfossil record. *Current Biology*, *24*(16), 1874–1879.

O'Hara, T. D., Hugall, A. F., Thuy, B., Stöhr, S., & Martynov, A. V. (2017). Restructuring higher taxonomy using broad-scale phylogenomics: The living Ophiuroidea. *Molecular Phylogenetics and Evolution*, *107*, 415–430.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S. C., Boisselier, M. C., & Samadi, S. (2015). Use of RAD sequencing for delimiting species. *Heredity*, *114*(5), 450. doi:10.1038/hdy.2014.105

Penant, G., Aurelle, D., Feral, J., & Chenuil, A. (2013). Planktonic larvae do not ensure gene flow in the edible sea urchin Paracentrotus lividus. *Marine Ecology Progress Series*, *480*, 155–170. doi:10.3354/meps10194

Pérez-Portela, R., Almada, V., & Turon, X. (2013). Cryptic speciation and genetic structure of widely distributed brittle stars (Ophiuroidea) in Europe. *Zoologica Scripta*, *42*(2), 151–169. doi:10.1111/j.1463-6409.2012.00573.x

Pfenninger, M., & Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology*, *7*(1), 121. doi:10.1186/1471-2148-7-121

Pons, J.-M., Sonsthagen, S., Dove, C., & Crochet, P.-A. (2014). Extensive mitochondrial introgression in North American Great Black-backed Gulls (Larus marinus) from the American Herring Gull (Larus smithsonianus) with little nuclear DNA impact. *Heredity*, *112*(3), 226–239. doi:10.1038/hdy.2013.98

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959.

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, *21*(8), 1864–1877. doi:10.1111/j.1365-294X.2011.05239.x

Rannala, B. (2015). The art and science of species delimitation. *Current Zoology*, *61*(5), 846–853. doi:10.1093/czoolo/61.5.846

Rannala, B., & Yang, Z. (2013). Improved Reversible Jump Algorithms for Bayesian Species Delimitation. *Genetics*, *194*(1), 245–253. doi:10.1534/genetics.112.149039

Rannala, B., & Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*, syw119. doi:10.1093/sysbio/syw119

Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. doi:10.1111/j.1471-8286.2007.01678.x

Ribout, C., & Carpentieri, C. (2013). Automated genomic DNA purification of marine organisms on the epMotion® 5075 VAC from Eppendorf. *Eppendorf Application Note*, *281*, 1–6.

Roux, C., Tsagkogeorga, G., Bierne, N., & Galtier, N. (2013). Crossing the Species Barrier: Genomic Hotspots of Introgression between Two Highly Divergent Ciona

intestinalis Species. *Molecular Biology and Evolution*, *30*(7), 1574–1587. doi:10.1093/molbev/mst066

Satler, J. D., Carstens, B. C., & Hedin, M. (2013). Multilocus Species Delimitation in a Complex of Morphologically Conserved Trapdoor Spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). *Systematic Biology*, *62*(6), 805–823. doi:10.1093/sysbio/syt041

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. doi:10.1128/AEM.01541-09

Sponer, R., & Roy, M. S. (2002). Phylogeographic analysis of the brooding brittle star *Amphipholis Squamata* (Echinodermata) along the coast of New Zealand reveals high cryptic genetic variation and cryptic dispersal potential. *Evolution*, *56*(10), 1954–1967. doi:10.1111/j.0014-3820.2002.tb00121.x

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Stöhr, S., Boissin, E., & Chenuil, A. (2009). Potential cryptic speciation in Mediterranean populations of *Ophioderma* (Echinodermata: Ophiuroidea). *Zootaxa*, (2071), 1–20.

Stöhr, S., & Muths, D. (2010). Morphological diagnosis of the two genetic lineages of *Acrocnida brachiata* (Echinodermata: Ophiuroidea), with description of a new species. *Journal of the Marine Biological Association of the United Kingdom*, *90*(04), 831–843. doi:10.1017/S0025315409990749

Stöhr, S., Weber, A. A.-T., Boissin, E., & Chenuil, A. Resolving a cryptic species complex – the case of *Ophioderma longicauda* (Echinodermata: Ophiuroidea). *In Preparation*.

Taboada, S., & Pérez-Portela, R. (2016). Contrasted phylogeographic patterns on mitochondrial DNA of shallow and deep brittle stars across the Atlantic-Mediterranean area. *Scientific Reports*, *6*, 32425.

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907–3930. doi:10.1111/j.1365-294X.2012.05664.x

Weber, A. A.-T., Abi-Rached, L., Galtier, N., Bernard, A., Montoya-Burgos, J. I., & Chenuil, A. (2017). Positive selection on sperm ion channels in a brooding brittle star: consequence of life-history traits evolution. *Molecular Ecology*, *26*(14), 3744–3759.

Weber, A. A.-T., Dupont, S., & Chenuil, A. (2013). Thermotolerance and regeneration in the brittle star species complex *Ophioderma longicauda*: A preliminary study comparing lineages and Mediterranean basins. *Comptes Rendus Biologies*, *336*(11–12), 572–581. doi:10.1016/j.crvi.2013.10.004

Weber, A. A.-T., Mérigot, B., Valière, S., & Chenuil, A. (2015). Influence of the larval phase on connectivity: strong differences in the genetic structure of brooders and broadcasters in the *Ophioderma longicauda* species complex. *Molecular Ecology*, *24*(24), 6080–6094. doi:10.1111/mec.13456

Weber, A. A.-T., Stöhr, S., & Chenuil, A. (2014). Genetic data, reproduction season and reproductive strategy data support the existence of biological species in

Ophioderma longicauda. *Comptes Rendus Biologies*, *337*(10), 553–560. doi:10.1016/j.crvi.2014.07.007

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology*, *61*(5), 854–865. doi:10.1093/czoolo/61.5.854

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, *107*(20), 9264–9269. doi:10.1073/pnas.0913022107

Yang, Z., & Rannala, B. (2014). Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci. *Molecular Biology and Evolution*, *31*(12), 3125–3135. doi:10.1093/molbev/msu279

Zhang, C., Zhang, D.-X., Zhu, T., & Yang, Z. (2011). Evaluation of a Bayesian Coalescent Method of Species Delimitation. *Systematic Biology*, *60*(6), 747–761. doi:10.1093/sysbio/syr071

Data Accessibility

The raw Miseq reads were deposited on Dryad Digital Repository and are accessible on https://doi.org/10.5061/dryad.5ks03 (The raw data of the present study are from the same Miseq run as for the study: Weber et al., 2015).

Author Contributions

A.A.-T.W. and A.C. conceived the study; S.S. provided museum samples; A.A.-T.W. performed additional sampling, laboratory work and analyzed the data; A.A.-T.W., A.C. and S. S. wrote the manuscript.

Tables

Table 1: Correspondence between genetic clusters found in DAPC, mitochondrial lineages and sampling locations of individuals used in this study. The population numbers refer to the number indicated in Figure 1. C1-C6: genetic clusters or clusters found in DAPC analysis. L1-L6: *Ophioderma longicauda* mitochondrial lineages defined in Boissin et al., 2011. Congruent: congruence between mitochondrial lineage and nuclear data (genetic cluster).

Locality	Population number	Genetic cluster	Mitochondrial lineage	Congruent
Thiouriba / Cap Vert Peninsula, Dakar, Senegal	1a	C1	L6	Yes
Cap Manuel, Dakar, Senegal	1b	C2	L5	Yes
Madelene Island, Dakar, Senegal	1c	C2	L5	Yes
Gorée Island, Dakar, Senegal	1d	C2	L5	Yes
Teneriffe, Canary Islands	2	C3	L5	No
Madeira, Portugal	3	C3	L5	No
Algarve, Portugal	4	C3	L1	Yes
Ceuta, Spain	5	C3	L5	No
Tabarka, Tunisia	6	C3	L1	Yes
Kelibia, Tunisia	7	C4	L3b	Yes
Monastir, Tunisia	8	C4	L3b	Yes
Agios Pavlos, Crete	9	C3	L1	Yes
Agios Pavlos, Crete	9	C5	L3	Yes
Symi island, Greece	10	C3	L1	Yes
Symi island, Greece	10	C5	L2 & L3	No
Baths of Aphrodite, Cyprus	11	C3	L1	Yes
Baths of Aphrodite, Cyprus	11	C6	L4	Yes
Ramkine, Lebanon	12a	C3	L1	Yes
Ramkine, Lebanon	12a	C6	L2	No
Beirut & Raoucheh, Lebanon	12b	C6	L4	Yes

Table 2: F_{ST} values (W&C) among the six genetic clusters based on 30 genetic markers. Significant F_{ST} values after a permutation test (1000 permutations) are highlighted in bold. **: 0.001<P-value<0.01; ***: P-value < 0.001.

F_{ST}	C2	C3	C4	C5	C6
C1	**0.28742*** **	**0.22029*** **	**0.43463*** **	**0.47247*** **	**0.46308*** **
C2		**0.19154*** **	**0.35779*** **	**0.37697*** **	**0.38870*** **
C3			**0.20687*** **	**0.32060*** **	**0.30226*** **
C4				**0.39383*** **	**0.36829** **
C5					**0.36176*** **

Table 3: Posterior probabilities of parameters estimated with BPP. Theta = 4*Ne*μ. Tau = expected number of mutations per site. Minimum divergence times in years are calculated from the minimum divergence time of the geminate species pairs *O. teres* and *O. phoenia/O. cinerea* (2.8 mya, Lessios, 2008). Results for replicate 2 are displayed. See Fig. S5 for the full species tree and the positioning of branch lengths (Tau).

replicate 2	mean	median	mode	2.5% CI	97.5% CI	mode [years]	2.5% CI [years]	97.5% CI [years]
Theta_C1	0.005196151	0.005109	0.004971	0.003570775	0.007423225			
Theta_C2	0.007639499	0.007532	0.006841	0.005636	0.01023622			
Theta_C3	0.01682958	0.016739	0.014894	0.012518	0.02186435			
Theta_C4	0.002738336	0.002679	0.002408	0.0018	0.004092			
Theta_C5	0.001103252	0.001069	0.001277	0.000639775	0.001702			
Theta_C6	0.002360826	0.002304	0.002319	0.001494775	0.003635225			
Theta_Oteres	0.004313273	0.004187	0.004004	0.002433	0.006873225			
Theta_Ophoen	0.00736498	0.007178	0.006682	0.0046	0.01109122			
Theta_Ociner	0.01173528	0.011572	0.011885	0.008069	0.016437			
Tau_C1.C6	0.01596195	0.015895	0.014907	0.013522	0.018639	8920624	8091814	11153922
Tau_C2.C6	8.09527E-05	0.000054	0.00001	0.00001	0.000343	5984	5984	205258
Tau_C3_C4_C5_C6	0.000131744	0.000113	0.000072	0.000024	0.000325	43086	14362	194486
Tau_C4_C5_C6	0.000151832	0.000146	0.000156	0.000037	0.0003	93353	22141	179526
Tau_C5_C6	0.000152483	0.000142	0.000155	0.00003	0.000341	92755	17953	204061
Tau_C1	0.000899098	0.000883	0.000898	0.000744	0.001141	537380	445223	682795
Tau_C2	0.000818136	0.000811	0.000804	0.000675	0.000998	481128	403932	597222
Tau_C3	0.000686385	0.000688	0.000721	0.000533	0.000831	431460	318957	497286
Tau_C4	0.000534565	0.000535	0.000553	0.000399	0.000685	330925	238769	409917
Tau_C5	0.000382091	0.00038	0.00035	0.000247	0.000516	209446	147809	308784
Tau_C6	0.000382091	0.00038	0.00035	0.000247	0.000516	209446	147809	308784
tau_Oteres	0.004415418	0.004374	0.004679	0.002896	0.006119	2800000	1733020	3661723
Tau_Ophoen	0.004148281	0.004135	0.003983	0.00274	0.005535	2383501	1639667	3312246
Tau_Ociner	0.004148281	0.004135	0.003983	0.00274	0.005535	2383501	1639667	3312246
Tau_Ophoen_Ociner	0.000267135	0.000153	0.000003	0.000005	0.001263225	1795	2992	755937
Tau_Ot_Op_Oc	0.01244562	0.0124385	0.012251	0.0093731	0.01559	7331225	5609036	9329344

Table 4: Posterior probability values of estimated parameters for scenario 5 by ABC. N = effective population size; t = divergence times in number of generations, see details in Figure 3. r2 = admixture rate of C3 to C5. q025 and q975 indicate the range of 95% confidence interval.

Genetic Cluster	Parameter	mode	2.5% CI	97.5% CI
C1	N1	57,000	21,000	219,000
C2	N2	142,000	53,900	301,000
C3	N3	293,000	226,000	646,000
C4	N4	31,700	9,460	117,000
C5	N5	30,000	9,190	73,500
C6	N6	133,000	61,600	192,000
	r2	0.868	0.469	0.967
	t1	38,700	12,900	69,100
	t2	89,100	46,000	116,000
	t3	222,000	119,000	319,000
	t4	368,000	230,000	475,000
	t5	512,000	331,000	928,000

Supplementary figure captions

Figure S1: Haplotype networks for the 30 genetic markers used in this study. The colors correspond to the six genetic clusters found in DAPC.

Figure S2: RAxML phylogenetic tree of 30 concatenated markers. Numbers correspond to bootstrap support (500 bootstrap replicates). Details on individuals in Table S1.

Figure S3: DAPC results. **A:** The minimal value of Bayesian Information Criterion (BIC) indicates the optimal number of genetic clusters. **B:** Membership probability of each individual to belong to a particular genetic cluster.

Figure S4: Results of the Discriminant Analysis of Principal Components (DAPC) excluding the mitochondrial COI marker. The six genetic clusters are also retrieved in the absence of COI.

Figure S5: Species tree with the highest posterior probability inferred from BPP. Grey branch labels correspond to divergence times (Tau) found on Table 3.

Figure S6: DAPC results on the subsetted dataset (9 C3 individuals and 11 C5 individuals from Crete). **A:** The minimal value of Bayesian Information Criterion (BIC) indicates the optimal number of genetic clusters. **B:** The two species (C3 and C5) are retrieved in the DAPC analysis.

Figure S7: DIYABC results. **A:** Pre-evaluation for the prior values for the 8 tested scenarios. **B:** Model checking for scenario 5, to assess if the observed data fell within the range of estimated datasets. Scenario 5 was the scenario with the highest posterior probability (see Table S6).

Figure S8: Posterior probabilities of parameters estimated for scenario 5 using DIYABC (See Fig.3 for a drawing of scenario 5). **A:** N1 = effective population size of C1. **B:** N2 = effective population size of C2. **C:** N3 = effective population size of C3. **D:** N4 = effective population size of C4. **E:** N5 = effective population size of C5. **F:** N6 = effective population size of C6. **G**: r2 = proportion of C3 genome contributing to C4 genome. 1-r2 = proportion of C5 genome contributing to C4 genome. **H:** t1= Time of hybridization between C3 and C5. **I:** t2 = divergence time between C5 and C6**. J:** t3 = divergence time between C3 and C5/C6. **K:** t4 = divergence time between C2 and C3/C5/C6. **L:** t5 = divergence time between C1 and C2/C3/C5/C6.