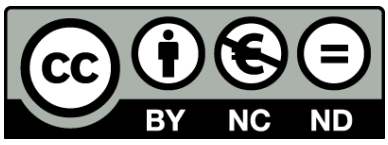
UNIVERSITAT DE BARCELONA

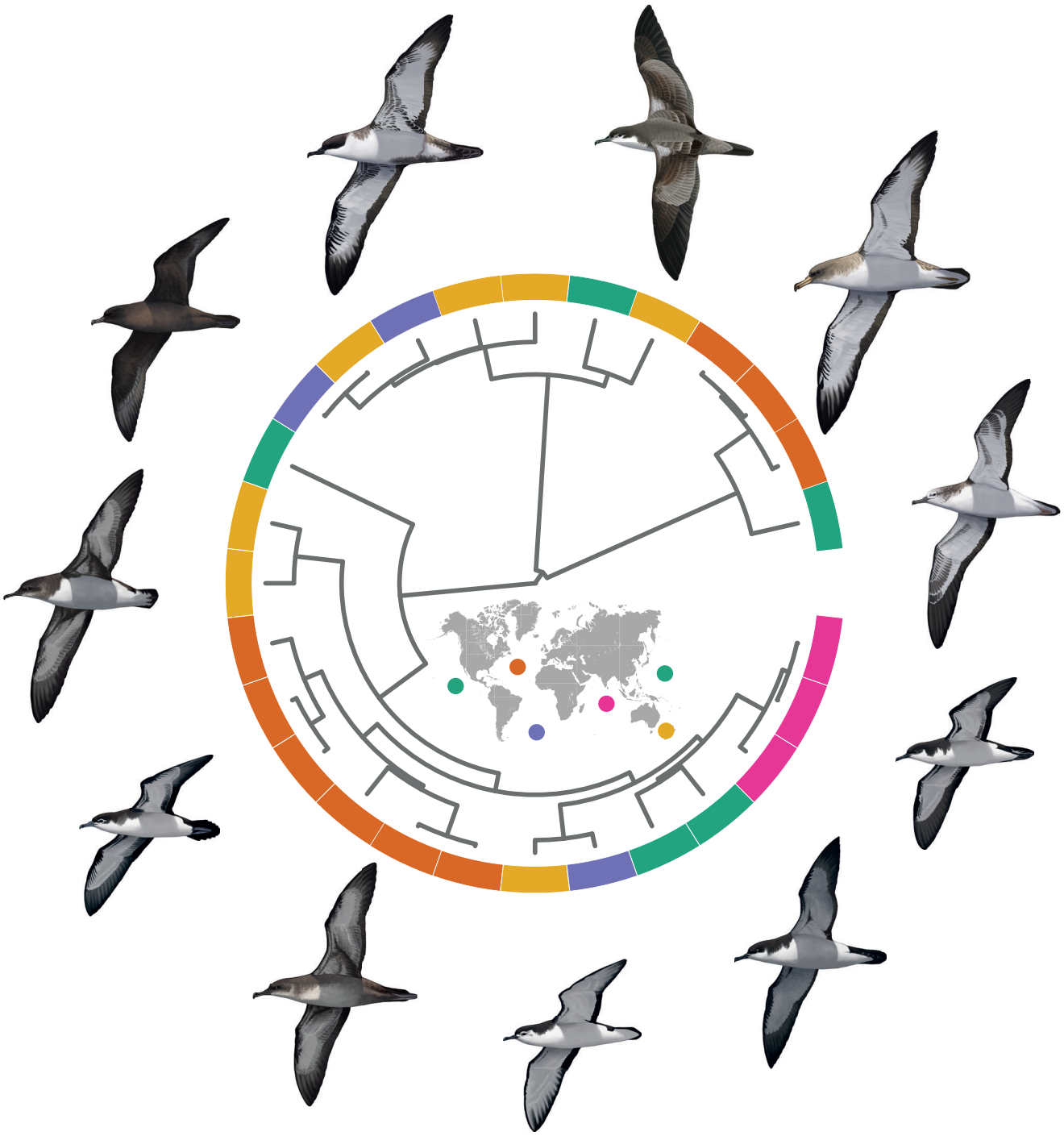
The evolutionary history of shearwaters: genomic analyses to resolve a radiation of pelagic seabrids

Joan Ferrer Obiol



Ph.D Thesis

The evolutionary history of shearwaters: genomic analyses to resolve a radiation of pelagic seabirds



Joan Ferrer Obiol

2020

Facultat de Biologia

Departament de Genètica, Microbiologia i Estadística

Programa de Doctorat en Genètica

The evolutionary history of shearwaters:

genomic analyses to resolve a radiation of pelagic seabirds

La història evolutiva de les baldrigues:

anàlisis genòmiques per a resoldre una radiació d'ocells marins pelàgics

Memòria presentada per

Joan Ferrer Obiol

per optar al grau de Doctor per la Universitat de Barcelona,

Barcelona, Desembre de 2020

Joan Ferrer Obiol	**Dr. Marta Riutort León**	**Dr. Julio Rozas Liras**
El doctorand	La directora i tutora	El director

Cover and chapter illustrations: Martí Franch

Glancing a wave with his wingtip, like
an arrow dashing cloud ward. It is only the
proud Shearwater who soars ever bold and
freely over the sea grey with sea foam.

"The song of the Stormy Petrel"

Maxim Gorky, 1901

Acknowledgements

Segons tinc entès aquesta secció està feta per agrair el suport a tots aquells que han permès que aquesta tesi sigui una realitat. I per tant, vull dedicar les primeres línies d'agraïment a uns individus molt especials que des del dia que els vaig conèixer m'han sigut font d'admiració i m'han ajudat a mirar la vida amb passió, moltes gràcies ocells i en especial ocells marins!

Un grand merci aussi à l'Équipe Prédateurs marins du CEBC de Chizé pour m'avoir donné l'opportunité de passer une année aux Kerguelen entouré d'albatros, pétrels et manchots qui me changeait la vie. Merci Henri, Yves, Nounours, Charlie, Christophe et Karine pour donner ailes à ma passion.

Marta, moltes gràcies per haver accedit a dirigir una tesi sobre ocells marins malgrat el teu amor incondicional a les planàries. Gràcies per no tenir por a començar una nova línia de recerca, per la teva confiança cega en mi i per treure temps d'on no en tenies per poder-me donar un cop de mà. I sobretot, gràcies per ser una gran persona. Julio, moltes gràcies per les oportunitats que m'has brindat (València, Buenos Aires) i gràcies pel teu suport i l'aportació del teu "know-how" durant la fase final d'aquesta tesi.

Companys de Fimol, abans de venir al laboratori era un biòleg de bota a qui la bata no interessava massa. Laia, Lisi, Ona, Edu, Dani, Cristian, Raquel, gràcies per ensenyar-me a posar-me la bata! Lisi gracias por los años juntos en Lluís el Piadós y por todos los ánimos que nos hemos dado el uno al otro. El meu inici al laboratori no hagués estat el mateix sense els cafès i les peces de fruita amb el tito Arnau. Tet, gràcies per aquests moments i per una gran amistat.

Aquesta tesi no hagués estat el mateix sense tots els estius a Veneguera. Jacob, gràcies per deixar-me treure el mono de camp cada any i a tots els jacobinos per la vostra ajuda i bons moments passats junts. Vir, ets la millor companya de campanya i una superstar de les bases de dades. Leia, gràcies pels nostres cafès. Teresa y Jose, gracias por iniciarme

en el mundo de las campañas. Y gracias a todos los compañeros de campaña y en especial a Naya por nuestras conversaciones existenciales.

Andreanna and Guojie, thank you for hosting me during my secondments in Durham and Copenhagen. I have been tremendously lucky to have the opportunity to work with you for a few months and start fruitful collaborations. Thanks to everyone I met during the secondments for making these months unforgettable: Andrea, Nilo, Federica, Alejandro, Erandi, Mónica, Menno, Jeroen, Josefin, Bitao, Jens, Justin, Manuel, Luigi, Lucie. Merci Claire Marie pour avoir partagé l'expérience danoise avec moi.

I would also like to thank the people in Exeter. Without you, the last period of my PhD would not have been the same. Thank you Jamie for giving me an office and Andy to be the best office mate. Thank you, Bonnie and Jim for your scientific advice and thank you Elliott, Mijke, Andrew, Phil, Guy, Molly, Shelly for bringing warmth to my life.

À mes copains français pour toujours pour m'avoir donnée une deuxième maison et avoir partagé des grands moments ensemble et à venir. Merci mon Pierrot, Dédé, Baptiste, Alizée, Alice et tous les autres copains chizéens. Merci copains kergueleniens pour avoir fait de la 63 la meilleure des missions et pour votre bonne humeur constante. Merci Thibaud, Christophe, Lolo, Piche, Thomas, Pachou, Mojito, Sophie, Cricri, mon frère ornitho Alex et Elsa pour toutes les rencontres annuelles à Tremouillat. Entre apéro et apéro on a quand même construit une maison! Merci Tim pour nos premiers moments partagés à Crozet, je me souviendrai toujours de notre première fois (et seule pour moi) à Pointe Basse. Un gros merci pour Gaia pour avoir été ma grande sœur en science et pour tous les moments partagés à Barcelone.

Companys del món dels ocells, gràcies per tot el que m'heu ensenyat i hem compartit junts. Viatges, sortides al Delta, Aiguamolls, gavines al port, sortides en barca,... Albert, Pere, Camil, Matxalen gràcies per ser un equip de xoc del Camp de Tarragona. Sergi, David, gràcies per tot el que m'heu ensenyat en els meus primers anys al Delta. Marcel, Martí, Dani, Guille gràcies per compartir generació d'ornitòlegs i per haver-nos iniciat en aquest món junts. Martí, moltíssimes gràcies pels teus dibuixos de baldrigues, ja saps que soc un gran fan.

No em puc oblidar tampoc dels amics castellers! Bandarres, gràcies per donar-me l'oportunitat de sentir els castells des de dins. Gràcies Mau, DCM, Lali i Cesc per ser unes persones excepcionals. Qui som nosaltres? Som poblesequins!!

Amics de la UB, els meus amics, biòlegs de món. Ens hem fet grans junts i m'heu fet millor persona. Gon, David, Martí, Miquel, Ferran, mes amis gràcies per ser els millors amics del món i per continuar sent-ho. Sou font d'inspiració, espatlles on agafar-se en els mals moments i la millor companyia per passar els bons moments plegats. Lucía, Júlia, Alba, Amaranta, Marta, us admiro i us estimo!

Tothom qui em coneix una mica bé sap que soc una persona de família. Gràcies tia i tio, padrina i Fono i Marieta per ser un motor d'amor. Gràcies a la família de Tarragona que tot i que ens veiem menys us sento molt a prop. I qui diu que la família és només aquella amb qui hi estem emparentats? La tia Teresa m'ha fet croquetes des que era petit, el tio Jesús em va fer despertar l'amor per la natura, la Noemi i el Jesús m'han fet de germans grans i la Fina m'ha ensenyat a estimar la llengua i el territori. My London family, Philippa, Melissa, Billy, Darren, Richard, thank you for treating me like family since the moment we met.

Moltes gràcies pares per ser els millors pares del món. Gràcies per sempre donar ales a les meves passions, per tots els bons principis que m'heu ensenyat, pel vostre suport incondicional en tot moment, per ser les dues persones que més admiro i per donar-me tot l'amor del món.

I pel final, moltes gràcies Josie. Gràcies per impregnar-me amb la teva passió per la ciència i per haver-me obert les portes del món de la genòmica i de tants altres. Gràcies per recórrer aquest últim tram junts, al nostre despatx del 30 The Mint, amb vistes al jardí i enmig de la pandèmia. Gràcies per ser la millor companya de feina i "the best wife"!

Abstract

How populations differentiate and become new species is a foundational question to the field of evolutionary biology and has important implications for the generation of both local and global patterns of species-level biodiversity. Ernst Mayr emphasised the importance of geographical isolation as a driver of speciation: "populations in separate locations begin a process of differentiation, and once differentiation is sufficient the populations have become two species". In marine environments, the lack of obvious physical barriers would suggest that panmixia, especially in highly mobile species, or isolation-by-distance, in other cases, will prevail. However, there is some counterintuitive evidence of fine-scale differentiation among populations and species in a number of mobile marine organisms, a phenomenon that has been described as the "marine species paradox". Seabirds of the order Procellariiformes present some of the most extreme examples of this paradox. On the one hand, Procellariiformes are highly mobile pelagic seabirds with a high dispersal ability and perform some of the longest animal migrations on Earth. On the other hand, they show high philopatry to their breeding grounds, which is expected to limit gene flow and therefore reinforce genetic differentiation.

This thesis aims to gain insights into the patterns and processes that contribute to genetic and phenotypic diversification, speciation and dispersal across multiple evolutionary timescales. To this end, I focus on shearwaters (*Calonectris*, *Puffinus* and *Ardenna*), a globally distributed and threatened group of Procellariiformes. Through an integrative approach combining two types of phylogenomic markers, which evolve at different nucleotide substitution rates, and state-of-the-art phylogenetic and introgression analyses, I inferred a robust phylogeny. This approach allowed to discover that the majority of the phylogenetic conflict in shearwaters is generated by high levels of incomplete lineage sorting (ILS) due to rapid speciation events. Divergence time estimation analyses highlighted a severe impact of the Pliocene marine megafauna extinction on shearwaters, probably caused by a sudden reduction in the availability of coastal habitat. Subsequently, the late Pliocene-early Pleistocene was inferred as a period of high and rapid speciation and dispersal, probably promoted by Pleistocene

climatic shifts. Biogeographic analyses showed that surface ocean currents promote species dispersal and founder events are a main mode of speciation in shearwaters. Our analysis, combining genomic data with morphological and ecological evidence, did not support any of the current taxonomic classifications for the North Atlantic and Mediterranean *Puffinus* shearwaters, and so I propose a more accurate taxonomy for the group. Moreover, the detection of fine-scale genetic structure within *Puffinus* shearwater species, highlights the need for management of evolutionary significant units below the species level. Population genomics analyses identified genetic drift as the major process shaping the genomic landscapes of divergence. In conclusion, the marriage of these various investigations identifies a prevalence of ILS across different timescales, highlights the important role of paleoceanographic events in promoting diversification, and demonstrates the importance of neutral evolution at driving population differentiation in pelagic seabirds. Overall, this thesis showcases the use of multiple genomic approaches, leveraging phylogenetic and population genetic analyses across multiple timescales, to shed light on the evolutionary history of shearwaters.

Sinopsi

Una de les preguntes fundacionals del camp de la biologia evolutiva és com es diferencien les poblacions i esdevenen noves espècies, i té importants implicacions en l'establiment dels patrons locals i globals de biodiversitat específica. Ernst Mayr va subratllar la importància de l'aïllament geogràfic com a mecanisme promotor d'especiació: "les poblacions en localitats aïllades comencen un procés de diferenciació i, quan aquesta diferenciació és suficient, les poblacions esdevenen dues espècies". En l'ambient marí, la manca de barreres físiques evidents suggereix que la panmixi (en espècies amb gran capacitat de moviment) i l'aïllament per distància (en espècies de mobilitat reduïda) són els patrons prevalents. En canvi, s'ha detectat diferenciació a petita escala entre diferents poblacions i diferents espècies en un elevat nombre d'organismes marins mòbils, un fenomen que es coneix com la "paradoxa de les espècies marines". Els ocells marins de l'ordre dels Procellariiformes representen un dels exemples més extrems d'aquesta paradoxa. Per una banda, els Procellariiformes són ocells marins amb gran capacitat de moviment i alta capacitat dispersiva que duen a terme algunes de les migracions més llargues del planeta. Per altra banda, són espècies molt fidels a les seves zones de cria, fenomen que probablement limiti el flux gènic i reforci la diferenciació genètica.

Aquesta tesi té com a objectiu caracteritzar els patrons i processos que contribueixen a la diversificació genètica i fenotípica, a l'especiació i a la dispersió a diferents escales evolutives. Per assolir aquest objectiu, m'he focalitzat en les baldrigues (*Calonectris*, *Puffinus* i *Ardenna*), un grup amenaçat de Procellariiformes distribuït per tot el planeta. Mitjançant una aproximació que combina dos tipus de marcadors filogenòmics que evolucionen amb diferents taxes de substitució nucleotídica, i anàlisis filogenètiques i de detecció d'introgressió infereixo una filogènia ben resolta. Aquesta aproximació ha permès descobrir que la majoria del conflicte filogenètic en les baldrigues és producte dels alts nivells de sorteig incomplet de llinatges (incomplete lineage sorting) degut a esdeveniments ràpids d'especiació. L'anàlisi dels temps de divergència ha demostrat un gran impacte de l'extinció de megafauna marina del Pliocè en les baldrigues, probablement a causa de la sobtada reducció en la disponibilitat d'hàbitats costaners.

Posteriorment, el Pliocè tardà i el principi del Pleistocè van ser períodes de molta i ràpida especiació i dispersió, probablement a causa dels canvis climàtics del Pleistocè. Les anàlisis biogeogràfiques han demostrat que els corrents marins promouen la dispersió i que l'efecte fundador és un mecanisme important d'especiació en les baldrigues. Les nostres anàlisis, combinant dades genòmiques amb evidència morfològica i ecològica, no han donat suport a les classificacions taxonòmiques actuals del grup de les baldrigues del gènere *Puffinus* del nord de l'Atlàntic i del Mediterrani i, per tant, proposo una classificació taxonòmica més adequada pel grup. A més a més, la detecció d'estructura genètica a petita escala en les espècies d'aquest grup de baldrigues destaca la necessitat de gestió d'unitats evolutives significatives per sota del nivell d'espècie. Les anàlisis de genòmica de poblacions han identificat la deriva genètica com a principal promotor de divergència en el genoma. Per concloure, el conjunt de les investigacions d'aquesta tesi identifiquen la prevalença del sorteig incomplet de llinatges al llarg de diferents escales evolutives, destaquen l'important paper dels esdeveniments paleoceanogràfics com a promotors de diversificació i indiquen la importància de l'evolució neutra com a causa de diferenciació poblacional en ocells marins pelàgics. Globalment, aquesta tesi demostra la utilitat de diferents aproximacions genòmiques, mitjançant anàlisis filogenètiques i de genètica de poblacions en diferents escales evolutives per proporcionar nou coneixement sobre la història evolutiva de les baldrigues.

Table of Contents

General Introduction

General Introduction

1 | The Shearwaters

1.1 | Systematics

Procellariiformes, an order of globally distributed pelagic seabirds, are characterised by a hooked bill covered in horny plates and raised tubular nostrils, which gave them the former name Tubinares (Warham 1990). The current name for the order comes from the Latin word *procella*, which means violent winds or a storm (Gotch 1979). The order comprises four extant families: the albatrosses (Diomedeidae), the petrels and shearwaters (Procellariidae) and two families of storm-petrels (Hydrobatidae and Oceanitidae), and an extinct family (Diomedeoididae).

Together with Gaviiformes (loons), Ciconiiformes (storks), Suliformes (gannets and cormorants), Pelecaniformes (pelicans, ibises and herons) and Sphenisciformes (penguins), they form the core waterbirds (Aequornithes) radiation (Figure 1a) (Hackett et al. 2008; Jarvis et al. 2014). This radiation took place shortly after the original neoavian radiation during the Cretaceous-Paleogene (K-Pg) boundary after a mass extinction event (Jarvis et al. 2014). The phylogenetic relationships resulting from this radiation have been challenging to resolve (i.e. Ericson et al. 2006). However, the advent of next-generation sequencing technologies has led to phylogenomic datasets (comprising hundreds to thousands of markers) that have been able to resolve the phylogenetic relationships among the orders of this group (Hackett et al. 2008; Jarvis et al. 2014; Prum et al. 2015). Specifically, Sphenisciformes are the sister group to Procellariiformes and together form a sister clade to the rest of the core waterbirds (Figure 1b). In addition to resolving the phylogenetic relationships, analysis using phylogenomic data have identified that phylogenetic conflict in the core waterbirds radiation was due to high levels of incomplete lineage sorting (ILS) caused by rapid speciation, although levels of ILS were less pronounced than during the neoavian radiation (Suh et al. 2015).

Due to their characteristic synapomorphies, the monophyly of Procellariiformes has long been established. However, the phylogenetic relationships among the different families remain controversial. Based on phylogenies inferred using sequences of the mitochondrial Cytochrome b gene (*cytb*), the family Hydrobatidae was recovered as the sister group to all the other Procellariiformes (Nunn and Stanley 1998; Kennedy and Page 2002). More recently, phylogenomic studies recovered the albatrosses as the sister group to all the other Procellariiformes, as well as other differences in the phylogenetic relationships among the families (Prum et al. 2015; Estandia 2019). That being said, more recent studies do however find consistent relationships among families (Figure 1c).

The family Procellariidae is the most species-rich within the Procellariiformes and is a diverse group which consists of fulmars, gadfly petrels, diving-petrels, prions, petrels and shearwaters. The phylogenetic relationships among them are also still controversial. Particularly contentious has been the position of the diving-petrels (*Pelecanoides* sp.) which were long-believed to be the sister group to the Procellariidae (Nunn and Stanley 1998), but more recent evidence indicated that they are embedded within Procellariidae (Figure 1c) (Kennedy and Page 2002; Estandia 2019).

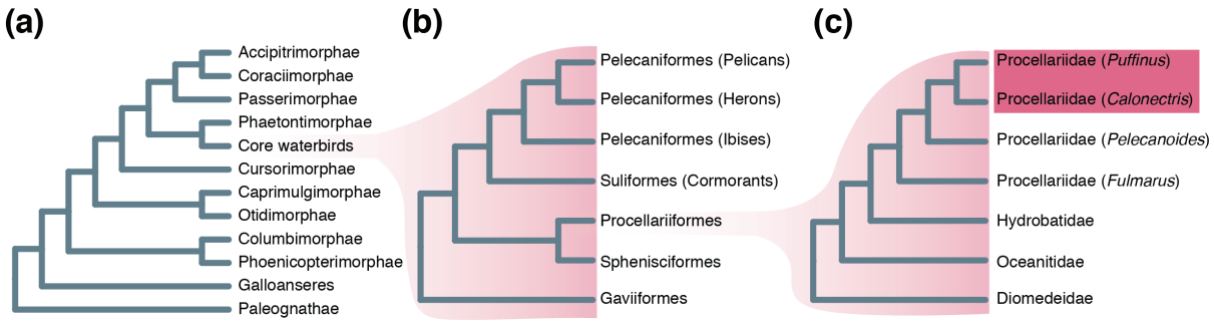


Figure 1 The phylogenetic position of shearwaters. (a) Phylogeny of birds showing the position of core waterbirds. (b) Phylogeny of core waterbirds showing the position of Procellariiformes. (c) Phylogeny of Procellariiformes showing the position of shearwaters (highlighted in magenta). Phylogenies based on Jarvis et al. (2014) and Estandia (2019).

1.2 | General Characteristics of the Three Genera of Shearwaters

Shearwaters form a monophyletic group of medium-sized seabirds that represent one of the two major radiations within Procellariidae. Some shearwater species are among the world's most abundant birds, with populations of >10 millions (Warham 1990), while others are amongst the world's most threatened birds. The group consists of three genera (Figure 2; *Calonectris*, *Ardenna* and *Puffinus*), which differ in morphology, flight and feeding habits (Carboneras and Bonan 1992).

The genus *Calonectris* comprises four large-sized species with grey-brown upperparts, white underparts and a large pale bill. They are aerial species that cover great distances and forage on pelagic fish and cephalopods by surface-seizing (Granadeiro et al. 1998). *Calonectris* species are distributed along the North Atlantic and North Pacific oceans and are long-distance migrants (González-Solís et al. 2007; Yamamoto et al. 2010).

The genus *Ardenna* is composed of seven species with polytypic coloration patterns. They are structurally similar to *Calonectris*, but their wings tend to be more pointed and their bills notably slimmer. *Ardenna* are also largely aerial species, although they exhibit better diving abilities than *Calonectris*, and some species can attain maximum depths of 70 m (Weimerskirch and Sagar 1996). Their main foraging strategies are surface-seizing and pursuit-plunging, and their diet is mainly dominated by fish and cephalopods, although some species also feed on crustaceans mainly during the breeding period (Brooke 2004). *Ardenna* species are distributed along subantarctic and temperate waters of the southern hemisphere and perform some of the longest animal migrations on Earth (covering > 70,000 km a year) (Shaffer et al. 2006).

The genus *Puffinus* is the most species rich, with 21 recognised species (Gill et al. 2020). Plumage colour is generally dark brown-to-black on the upperparts and brown or white on the underparts. There is considerable individual variation and intertaxon overlap, and some species show plumage polymorphism. *Puffinus* shearwaters are smaller than *Ardenna* and *Calonectris,* but size varies substantially within the group with the largest species weighing ~520 g and the smallest ~150 g. Contrary to the other two genera, *Puffinus* shearwaters are more aquatic, they cover shorter distances and

they routinely dive to access their prey (Shoji et al. 2016), which are crustaceans, cephalopods and fish. Most species in the genus are short distance migrants or disperse short distances around their breeding colonies, although one species, the Manx shearwater (*P. puffinus*) is a long-distance migrant flying from its colonies in Northern Europe to Patagonia (Guilford et al. 2009).

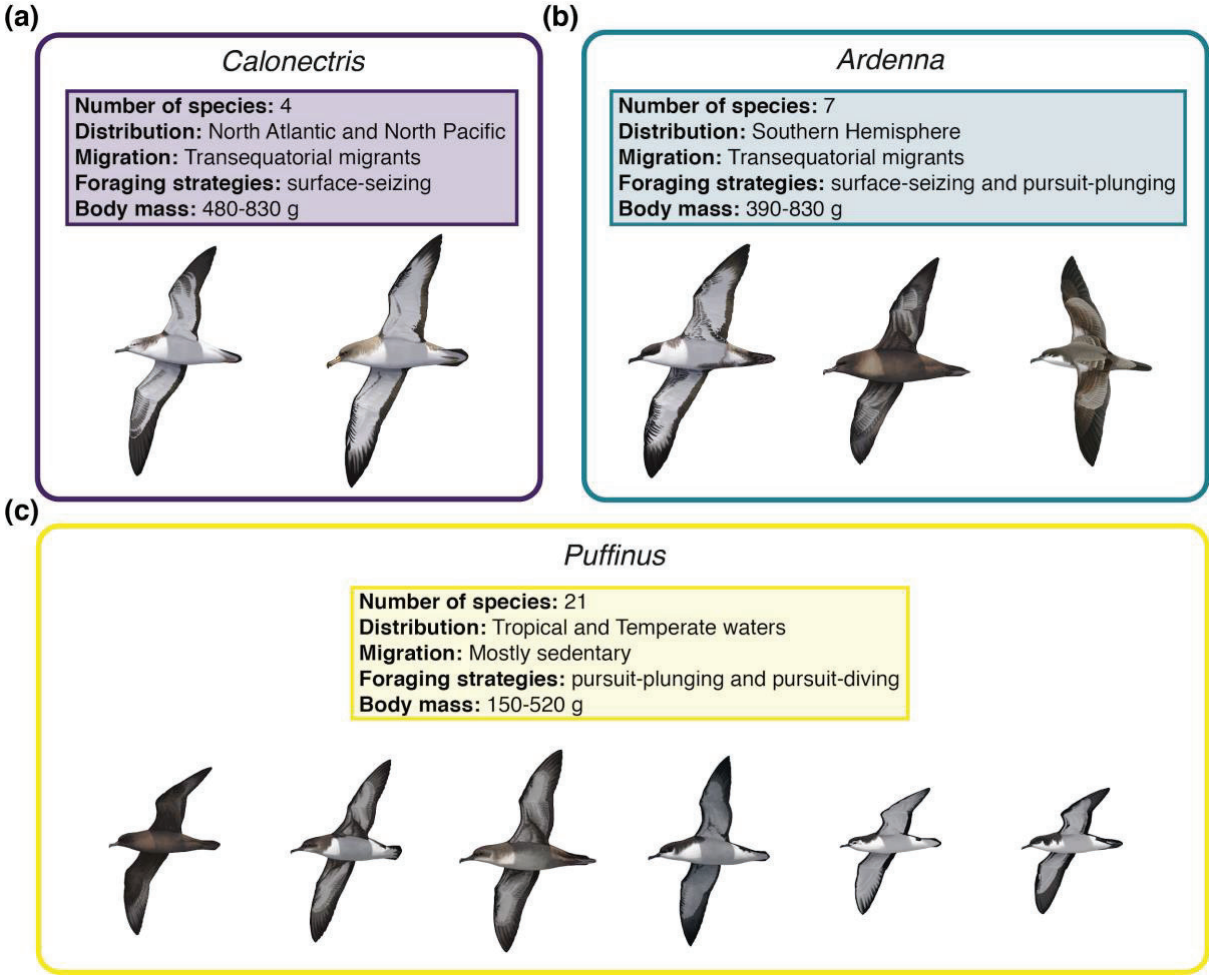


Figure 2 Shearwater genera showing the number of species per genus, their distribution, their migratory behaviour, their foraging strategies and their body mass. Shearwater illustrations by Martí Franch © are shown to scale and represent, from left to right: (a) *Calonectris leucomelas* and *C. diomedea*. (b) *Ardenna gravis* and *A. grisea*. (c) *Puffinus nativitatis*, *P. mauretanicus*, *P. puffinus*, *P. baroli* and *P. bailloni*. Note the differences in size, particularly within *Puffinus*.

1.3 | Breeding Cycle

Shearwaters are strictly pelagic seabirds that spend most of the year at sea and only approach land to visit their breeding colonies on islands during the breeding season. Breeding colonies are normally located in coastal habitats with direct access to and from the sea, although some species establish their colonies several kilometres inland, and up to 1500 m above sea level. They nest in burrows, which they excavate or enlarge themselves in soft soil, in crevices, in cavities on open ground or hidden under dense vegetation (Warham 1990). Shearwaters are monogamous and tend to create bonds for life with their partners. Long before egg-laying, they start visiting the colony in order to defend their burrows and maintain the pair-bond. There is little sexual dimorphism in shearwaters, which is mostly restricted to size (Ristow and Wink 1981). Sexual activity between mates takes place inside the burrow and not long after effective copulation, both members of the pair go back to the sea to regain weight (Brooke 2013). Shearwaters have bet-hedging life-histories typified by extended chick rearing periods. A single egg is laid, and incubation is shared between the sexes, over periods of several days, which alternate between incubation and foraging at sea. Arrival to the colonies takes place generally only during the night and after a few days of feeding at sea, adult shearwaters gather in the immediate vicinity of the colony in the evening, forming rafts to wait until night time. When the egg hatches, the duration of the turns of incubation decreases until the moment when the chick can regulate its own body temperature. At this time, the adults abandon the chick during the day and visit the colony only to feed the chick. Once the chick is fully fledged, the adults stop visiting the colony. A few days after the parents' desertion, the hungry chick develops an instinctive urge to learn to fly and therefore begins to leave the burrow during the night to flap its wings in order to develop its flight muscles, until the day that it flies out to sea on its own.

1.4 | Conservation Status and Threats

Human globalisation is changing the world’s ecosystems at an unprecedentedly fast pace, driving global declines and extinctions in a cornucopia of species (Jenkins 2003). Procellariiformes are particularly sensitive to anthropogenic changes, and as a

consequence, are one of the most endangered avian groups (Croxall et al. 2012). Within the Procellariiformes, shearwaters are a particularly sensitive group; 57% of the species are listed as threatened by the IUCN Red List of Threatened Species (Figure 3a). Shearwaters face several anthropogenic threats both at their breeding colonies and at sea (Dias et al. 2019; Rodríguez et al. 2019).

Inland, populations are severely affected by invasive alien species that predate on eggs, chicks and even adult birds (Spatz et al. 2017; Holmes et al. 2019). Indeed, predation by invasive mammals is the most harmful of all threats faced by shearwaters (Figure 3b) (Dias et al. 2019; Rodríguez et al. 2019). Rats and cats are the most widespread invasive species affecting shearwaters and have been described to cause colony extirpations, population declines, ultimately resulting in a higher risk of extinction (Jones et al. 2008; Bonnaud et al. 2011). Introduced herbivores such as rabbits can also cause damage to the breeding colonies, through erosion and alteration of the soil and damage to the burrows (Brodier et al. 2011).

Artificial lights can confuse and disorient shearwaters which are adapted to low light conditions due to their nocturnal activity inland. This can result in injuries or mortality due to collision with structures or the ground. Birds that fall to the ground without being injured are unlikely to be able to fly again and typically die from starvation, are predated by cats and dogs or are killed by human traffic (Rodríguez et al. 2012, 2017; Deppe et al. 2017).

Shearwaters have been, and are still, affected by direct exploitation by man. In the past, eggs, chicks and adults were harvested systematically from their burrows, a practice that has brought some species to extinction (Rando and Alcover 2008). Many species were also deliberately caught at sea for human consumption. These practices have fortunately significantly declined (Carboneras and Bonan 1992). However, harvesting of Short-tailed (*A. tenuirostris*) and Sooty (*A. grisea*) shearwaters still continues today in Tasmania and New Zealand, respectively (DPIPWE 2018; Newman et al. 2009). Regulations on quotas, however, have probably helped to reduce overall extinction risk.

At sea, fisheries bycatch is the main threat (Figure 3b) (Dias et al. 2019; Rodríguez et al. 2019), and one that has the potential to drive some species to extinction unless

conservation measures are efficiently implemented (Oro et al. 2004; Genovart et al. 2016). Longline fisheries are particularly dangerous for shearwaters and deep-diving species are the most vulnerable (Anderson et al. 2011; Cortés et al. 2017). Although with a lower severity, gillnets can also cause entanglements and mortality by drowning (Žydelis et al. 2013). Fortunately, seabird bycatch can be mitigated by applying operational measures such as avoiding discards during setting and hauling operations, or night setting (ACAP 2014; Cortés and González-Solís 2018).

Fisheries also have other impacts on shearwaters, such as direct competition for fish (Grémillet et al. 2018) and by providing an unpredictable source of food through discards. Discard volumes are decreasing globally (Zeller et al. 2018) and this reduction has the potential of impacting, at least in the short-term, several shearwater species that have developed a dependency on this foraging technique (Genovart et al. 2018).

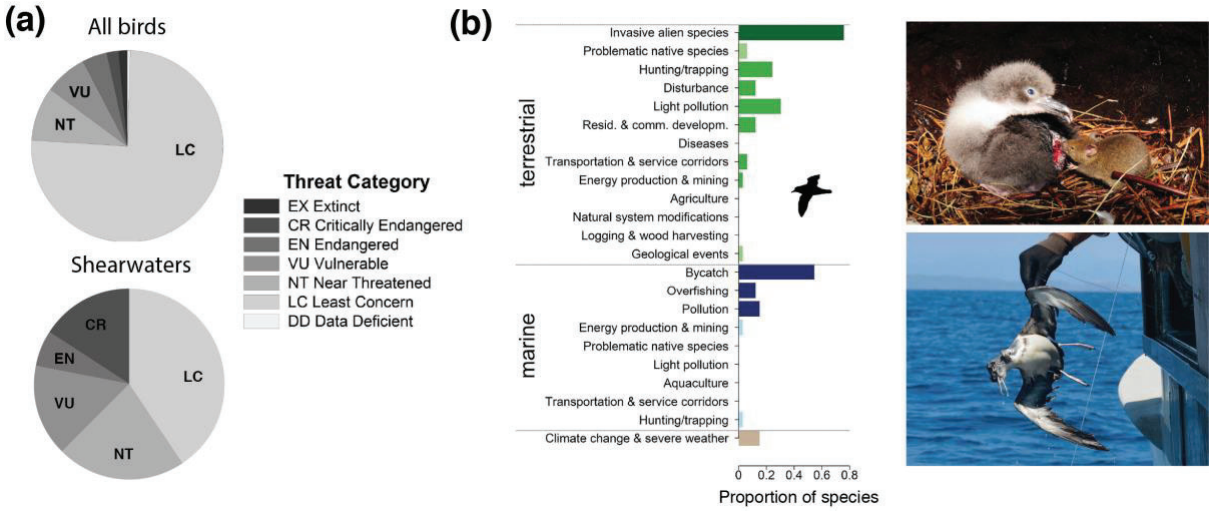


Figure 3 Conservation status and threats faced by shearwaters. (a) The percentage of species in each threat category from the IUCN Red List of Threatened Species for all bird species (above) and the shearwaters (below). Adapted from (Rodríguez et al. 2019). (b) Main threats (divided into marine and terrestrial) faced by shearwaters represented as the proportion of species affected by each threat. Reproduced from (Dias et al. 2019). Pictures showing a Great shearwater (*A. gravis*) chick being attacked by a mouse in Gough Island (photo Ben Dilley) and a Yelkouan shearwater (*P. yelkouan*) victim of longline fisheries bycatch.

2 | Mechanisms of Population Differentiation and Speciation in Seabirds

2.1 | Land Barriers

Seabirds, by nature of their name, are tightly associated with the sea. Given their strong marine ecology, they are not well-suited to life on land and actively avoid this habitat. Indeed, land masses present significant physical barriers to gene flow in seabirds (Friesen et al. 2007a). Most seabird species or groups of species that have breeding distributions fragmented by contemporary or historical land masses show genetic differentiation between the respective fragmented populations. For example, the American landmass plays a major role at promoting genetic differentiation between Atlantic and Pacific populations of Leach's storm petrels (*Hydrobates leucorhous*) (Bicknell et al. 2012) and between Atlantic and Indo-Pacific populations of Brown boobies (*Sula leucogaster*) (Morris-Pocock et al. 2011). This landmass has also promoted speciation in shearwaters and gadfly petrels, where Atlantic and Pacific species constitute sister monophyletic groups (Austin et al. 2004; Gómez-Díaz et al. 2006; Welch et al. 2014).

2.2 | Differences in Ocean Regimes

Land masses are not the only physical barriers to dispersal encountered by seabirds. Although not as obvious, contemporary and historical differences in ocean regimes can exert different selection pressures due to changes in water temperature and prey availability. Thus, ocean fronts can act as cryptic physical barriers to gene flow. In some cases, oceanographic conditions at either side of ocean fronts can be markedly different, limiting species dispersal and promoting genetic differentiation and speciation. For example, the Subtropical Convergence, which causes a 10 °C difference in sea-surface temperature north and south of the front, has driven population differentiation and speciation in albatrosses, petrels and penguins (Milot et al. 2008; Rexer-Huber et al. 2019; Vianna et al. 2020). Even fronts which cause less extreme oceanographic changes have been identified as barriers to gene flow, such as the Almeria-Oran front. This front

has been shown to constitute a barrier to gene flow between Cory's and Scopoli's shearwaters (*C. borealis* and *C. diomedea*) (Gómez-Díaz et al. 2009).

2.3 | Isolation-By-Distance

The aforementioned barriers are clearly important for population differentiation and speciation in seabirds. However, several species show genetic differentiation in the absence of physical barriers. This phenomenon requires the consideration of non-physical or intrinsic barriers to gene flow. Isolation-by-distance (IBD) is a well-known evolutionary consequence resulting from an increase in genetic divergence associated with geographical distance due to limited dispersal across space (Wright 1943). Pelagic seabirds tend to be highly mobile species and thus, this process is not predicted to be a major driver of population differentiation. Indeed, little evidence for IBD and little genetic differentiation, across long distances of a circumpolar distribution, has been found in a wide array of seabird species, including the Wandering albatross (*Diomedea exulans*) (Milot et al. 2008), the Emperor penguin (*Aptenodytes forsteri*) (Cristofari et al. 2016) and the White-chinnned petrel (*Procellaria aequinoctalis*) (Rexer-Huber et al. 2019). However, in less mobile seabird species, such as gulls and cormorants, IBD may play a role in driving genetic differentiation (Sternkopf et al. 2010; Thanou et al. 2017).

2.4 | Founder Events

The foundation of colonies is believed to be a rare event in most seabird species despite their great potential for long-range dispersal (Milot et al. 2008). However, in several species, including gulls and shearwaters, contemporary colony foundation events have been reported (A. Storey; J. Lien 1985; Oro and Ruxton 2001; Munilla et al. 2016). Founder events are generally characterised by strong initial bottlenecks due to a considerably reduced size of the founding population compared to the source population, which may promote genetic differentiation due to the effect of genetic drift (Slatkin 1996; Templeton 2008). Indeed, in some seabirds, founder events have been proposed as drivers of speciation. For example, the Armenian gull (*Larus armenicus*) is believed to be a relic of an ancient colonisation of Armenia from Atlantic populations (Liebers et al. 2001). Similarly, the Shy albatross (*Thalassarche cauta*) was potentially

the result of a range expansion of an ancestral population of White-capped albatrosses (*T. steadi*) (Abbott and Double 2003).

2.5 | Philopatry and Colony Dispersal

Ringing studies have indicated that seabirds are highly philopatric to their breeding grounds and intercolony dispersal is generally restricted to neighbouring colonies (Weimerskirch et al. 1985; Coulson 2002). Philopatry may have evolved due to the benefits of coloniality, such as facilitation of prey location, mate choice and defence against predation (Coulson 2002). In addition, breeding sites for seabirds are generally scarce, which may encourage young birds to return to their natal colonies in order to avoid losing time and energy at looking for alternative breeding sites. This behaviour is expected to limit gene flow and therefore reinforce genetic differentiation even across small spatial scales. A good example where philopatry may have been a main driver of genetic differentiation across a small range, are Shy albatrosses from different islands near Tasmania, which show differentiation at both mitochondrial and nuclear markers (Abbott and Double 2003).

2.6 | Non-Breeding and Foraging Distributions

As pelagic seabirds spend the majority of their lives at sea, non-breeding and foraging areas play an important role in social interactions. In species with population-specific non-breeding or foraging areas, or in species where individuals remain around their breeding colonies all-year-round, the probability of encountering individuals from other colonies is much lower than in species that have common non-breeding or foraging areas among populations. As a result, genetic structure in the former species may be stronger than in the latter. Empirical evidence has confirmed this pattern; for example, Brünnich's guillemots (*Uria lomvia*) from throughout the North Atlantic tend to congregate in the northwest Atlantic during the non-breeding season and are genetically panmictic (Tigano et al. 2015). On the other hand, Black guillemots (*Cepphus grylle*) tend to spend all-year-round at their breeding colony sites and therefore show strong genetic structure (Kidd and Friesen 1998). Regarding foraging

distributions, Burg and Croxall (2001) found that genetic differentiation among albatross taxa corresponded to differences in their foraging areas.

2.7 | Allochrony

Geographical variation in breeding phenology is common in birds and, in seabirds, seasonal populations can co-occur within the same breeding colonies (Brooke and Rowe 1996). Differences in breeding phenology (allochrony) can reduce or even completely eliminate gene flow, resulting in genetic differentiation and speciation (Hendry and Day 2005). The band-rumped storm-petrel complex provides an excellent example of sympatric speciation due to allochrony (Friesen et al. 2007b). In the complex, Cape Verde individuals, which were the first to diverge from the rest, breed all year-round. In the remaining populations breeding is seasonal and at least, six independent breeding season switches have occurred in different locations (Taylor et al. 2019). When allochronic populations coexist, they represent a continuum of speciation spanning from completely undifferentiated populations to full allochronic species.

2.8 | Interplay of Mechanisms of Population Differentiation and Speciation

The aforementioned mechanisms are not likely to act in isolation. Indeed, current seabird diversity is probably the result of an interplay between different contributions of these mechanisms, across both time and space. For instance, using mtDNA sequences in combination with tracking data, stable isotopes and breeding surveys, Rayner et al. (2011) showed that segregation during the non-breeding season may result in different arrival times to the breeding grounds, which results in breeding asynchrony, that in conjunction with philopatry, may restrict gene flow. Thus, studies combining different types of data have the potential to explore the interplay between these mechanisms and may provide important insights into the speciation and diversification processes in seabirds.

3 | Molecular Approaches in Evolutionary Biology

3.1 | Phylogenetics: Inferring Evolutionary Relationships

How extant and extinct species are related to one another is a question that underpins much of evolutionary biology. Reconstructing the Tree of Life has been a central endeavour of evolutionary biology that started in earnest with Darwinism. Phylogenies are trees containing nodes that are connected by branches. Each branch represents the persistence of a genetic lineage through time, and each internal node the birth of a new lineage. If the tree represents the relationships among a group of species, then internal nodes represent the speciation events.

Phylogenetic trees were originally used, almost exclusively, to infer the evolutionary relationships among species. However, their use extended to nearly every biological field, including the inference of histories of populations (Edwards 2009), the evolutionary and epidemiological dynamics of pathogens (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses 2020), the identification of genes and regulatory elements in newly sequenced genomes (Lindblad-Toh et al. 2011) and even the evolution of language (Gray et al. 2009).

'Tree thinking' has also transformed the field of population genetics with the development of the coalescent theory (Kingman 1982). In this case, the main interest is not in the trees themselves (the topology), but in the pattern of the included gene trees to infer relevant population genetics parameters and the underlying evolutionary processes (Rozas and Sánchez-Gracia 2014).

3.2 | Brief History of Phylogenetic Markers

Originally, phylogenies were almost exclusively intended to describe relationships among species in systematics and taxonomy and were reconstructed using morphological characters (Yang and Rannala 2012). In order to be used for phylogenetic reconstruction, morphological characters have to be homologous, that is, they need to have derived by divergence from a common ancestral structure. Inference of the phylogenetic relationships in this case is based on synapomorphies: derived character

states that distinguish a clade from other organisms. While morphological data provided the first reconstructions of the Tree of Life, this approach suffered from a lack of resolution due to the limited number of unambiguous characters (Scotland et al. 2003). With the development of DNA sequencing, the issues of morphological characters to reconstruct phylogenies were rapidly overcome by the use of molecular data (DNA and proteins), which provided more information and more power (Fitch and Margoliash 1967). However, morphology remains a powerful independent source of evidence for testing molecular clades, and the primary means for time-scaling phylogenies through fossil phenotypes (Lee and Palci 2015).

In the same way as morphological characters, molecular phylogenetics also relies on homology. Two genes (molecular markers) are homologous if they are derived from an ancestral gene. However, genes can diverge from a common ancestor by two different processes: duplication (termed 'paralogs') and speciation (termed 'orthologs'). Orthologous genes recapitulate the relationships among the species they derive from. Because paralogy does not reflect the relationships among species, determining orthologous genes is an essential step for reconstructing species phylogenies (Koonin 2005). The early years of molecular phylogenetics were dominated by studies using a small set of universal orthologous genes, including ribosomal RNAs (Woese and Fox 1977) and mitochondrial genes (Zardoya and Meyer 1996). These markers provided a unique opportunity to resolve phylogenetic relationships at an unprecedented resolution as they provided an array of different rates of evolution, making it possible for researchers to choose the right gene depending on the phylogenetic scale they were interrogating. Moreover, these markers were easy to amplify with universal primers, their orthology was well established, and databases of these marker sequences grew rapidly. However, these genes did not provide enough resolution to resolve the phylogenetic relationships of many areas of the Tree of Life, particularly in rapidly radiating clades (Edwards et al. 1991; Halanych et al. 1995).

The advances of high-throughput sequencing technologies during the last two decades have pushed phylogenetics into the era of genome-scale data sets. The increasing availability of phylogenomic data (genome-scale datasets) has certainly aided the resolution of many contentious relationships across the Tree of Life (Rokas et al.

2003; Jarvis et al. 2014; Hughes et al. 2018). Perhaps even more importantly, genome-wide data has provided an unprecedented power to detect patterns of phylogenetic discordance and to uncover the evolutionary processes responsible for this discordance (Arcila et al. 2017). The usefulness of different phylogenomic markers depends on the evolutionary timescale at play (Collins and Hrbek 2018). For shallow phylogenomics, Restriction site-Associated DNA sequencing (RAD-Seq; (Miller et al. 2007) and sequence capture of ultraconserved elements (UCE; (Faircloth et al. 2012), are two of the most widely-used techniques.

RAD-Seq and related genotyping-by-sequencing approaches (Figure 4) (e.g. ddRAD-Seq, (Peterson et al. 2012); GBS, (Elshire et al. 2011)) have been primarily used for studying polymorphism in population genomics analyses (Hohenlohe et al. 2010; Lescak et al. 2015). However, they have also been successfully used for phylogeographic and interspecific phylogenetic studies of a wide variety of organisms (Emerson et al. 2010; Rubin et al. 2012; Wagner et al. 2013; Cruaud et al. 2014; Díaz-Arce et al. 2016). The number of orthologous loci recovered using RAD-seq approaches depends on divergence times between lineages (Rubin et al. 2012; Jones et al. 2013; Leaché et al. 2015). Thus, RAD-seq markers are particularly suitable for exploring shallow phylogenetic scales, where it allows the recovery of thousands to tens of thousands of loci.

Conversely, UCEs (Figure 4) were developed for studying deep evolutionary timescales and have been successfully used for this purpose (Faircloth et al. 2012; Longo et al. 2017; Alfaro et al. 2018; Oliveros et al. 2019). However, their flanking regions are variable enough to also be informative at shallow evolutionary timescales (Smith et al. 2014), particularly when data are phased to retrieve polymorphism information (Andermann et al. 2018). Due to the sequence capture step in the UCE protocol, the same loci are usually captured across samples independently of the divergence times between them. The number of loci recovered depends on the probe set used yet it tends to be lower than in RAD-seq datasets.

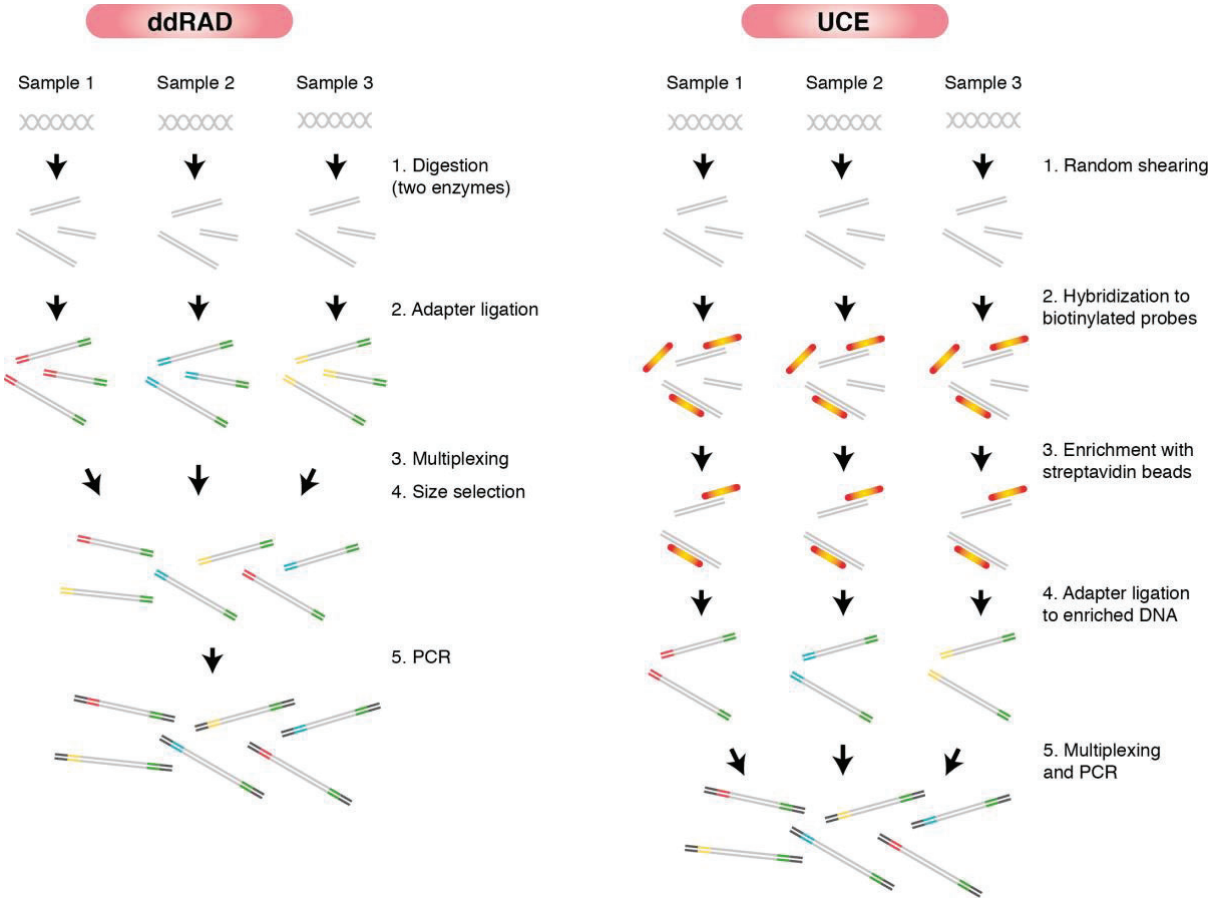


Figure 4 Workflow for ddRAD and UCE library preparation protocols used in this thesis. **ddRAD**: 1) Samples are digested using an uncommon and a common cutter in a single reaction. 2) After digestion, barcoded P1 Illumina adapters (represented in red, yellow and blue) and P2 Illumina adapters (represented in green) are ligated onto the fragments and 3) individually barcoded samples are pooled. 4) Pooled libraries are size selected and 5) amplified in a final PCR step. **UCE:** 1) genomic DNA is fragmented using a sonicator, and 2) UCE are captured using the Tetrapods UCE-5Kv1 probe set (available at ultraconserved.org) and 3) enriched using streptavidin beads. 4) Barcoded adapters (red, blue and yellow represent barcoded P1 Illumina adapters and green P2 Illumina adapters) are ligated to enriched DNA and 5) PCR is used to amplify the enriched library.

3.3 | Sources of Gene Tree Discordance

Large phylogenomic datasets have certainly helped in resolving many contentious relationships in the Tree of Life (Rokas et al. 2003; Jarvis et al. 2014; Hughes et al. 2018). But perhaps more importantly, they have also enabled the detection of gene tree discordance at an unprecedented scale, thereby allowing investigation into the causes of phylogenetic incongruence (Edwards 2009; Arcila et al. 2017). Indeed, individual gene trees may be in conflict with the species tree due to multiple biological processes such as incomplete lineage sorting (ILS), hybridisation and introgression, gene

duplication, GC-biased gene conversion (gBGC), and rate heterogeneity, *inter alia* (Maddison 1997; Nichols 2001).

ILS is considered to be the most widespread of the major biological causes of gene tree heterogeneity (Edwards 2009). ILS describes a phenomenon whereby orthologous genes from different species coalesce into a common ancestral copy before (or long before) the common ancestral species (Figure 5). As a result, the genes may not track the species phylogeny and may result in a different tree topology. Levels of ILS are especially high when the rate of genetic drift is low compared to the length of the internodes in the tree. Thus, ILS is more likely to occur when the ancestral species had large population sizes and especially when the interior branches of the species tree are short (Pamilo and Nei 1988; Rosenberg and Nordborg 2002), independently of the evolutionary timescale. When ILS is the main cause of phylogenetic discordance, in general, relationships represented by long internodes will be resolved with confidence. However, when radiations or other rapid speciation events occurred, ILS may seriously hinder species tree estimation (Edwards 2009).

Introgression (or horizontal gene transfer in bacteria) can also lead to phylogenetic incongruence between loci (Figure 5). Horizontal gene transfer is a common cause of discordance in the microbial world, and such a prevalent one that whether a coherent Tree of Life exists for microbes has been questioned (Doolittle and Bapteste 2007). Introgression is the incorporation of alleles from one species to the gene pool of a second species, usually via hybridization and backcrossing (Anderson and Others 1949). Introgression is a widespread process that plays an important role in the process of diversification (Abbott et al. 2016) and has been widely described to occur between sister or recently diverged species of a wide variety of taxa (Green et al. 2010; Kutschera et al. 2014; Schumer et al. 2016; Wen et al. 2016; Zarza et al. 2016). At deeper time scales, signals of introgression may erode or obscure the true species tree, which makes detection of old introgression challenging (Eaton et al. 2015; Schumer et al. 2016).

Other causes of gene tree discordance are less widespread but can be prevalent in certain taxa and can subvert phylogenetic analyses if not recognised. Gene duplication leads to the generation of paralogous genes. Because paralogs do not reflect the relationships among species, the unnoticed inclusion of paralogs instead of orthologs

can lead to misleading phylogenetic signal (Koonin 2005). Taxa that have experienced genome duplications, such as the salmonids, or polyploids, such as some plants, are more likely to suffer from orthology prediction errors.

gBGC is a process that takes place during meiotic recombination, at sites that are heterozygous for a 'weak' (AT) allele and a 'strong' (GC) allele, strong or weak defined by the number of hydrogen bonds between the two nucleotides within base pairs (i.e. three between G and C and two between A and T). In such cases, gene conversion tends to be biased towards the fixation of strong over weak alleles (Mancera et al. 2008; Lesecque et al. 2013) and results in higher GC content in areas of high recombination rates. This process is more effective in species with large population sizes (Romiguier et al. 2010; Lartillot 2013; Weber et al. 2014), and can result in convergence between non-sister lineages due to similar nucleotide composition, leading topological incongruence (Pease and Hahn 2013; Bossert et al. 2017).

Evolutionary rates can be affected by many factors such as generation time, population size or basal metabolic rate (Reis et al. 2015), which may cause rate heterogeneity among lineages. This may be an additional driver of phylogenetic discordance and, when rate variation among lineages is significant, can result in the artifact of 'long-edge attraction' (Felsenstein 1978).

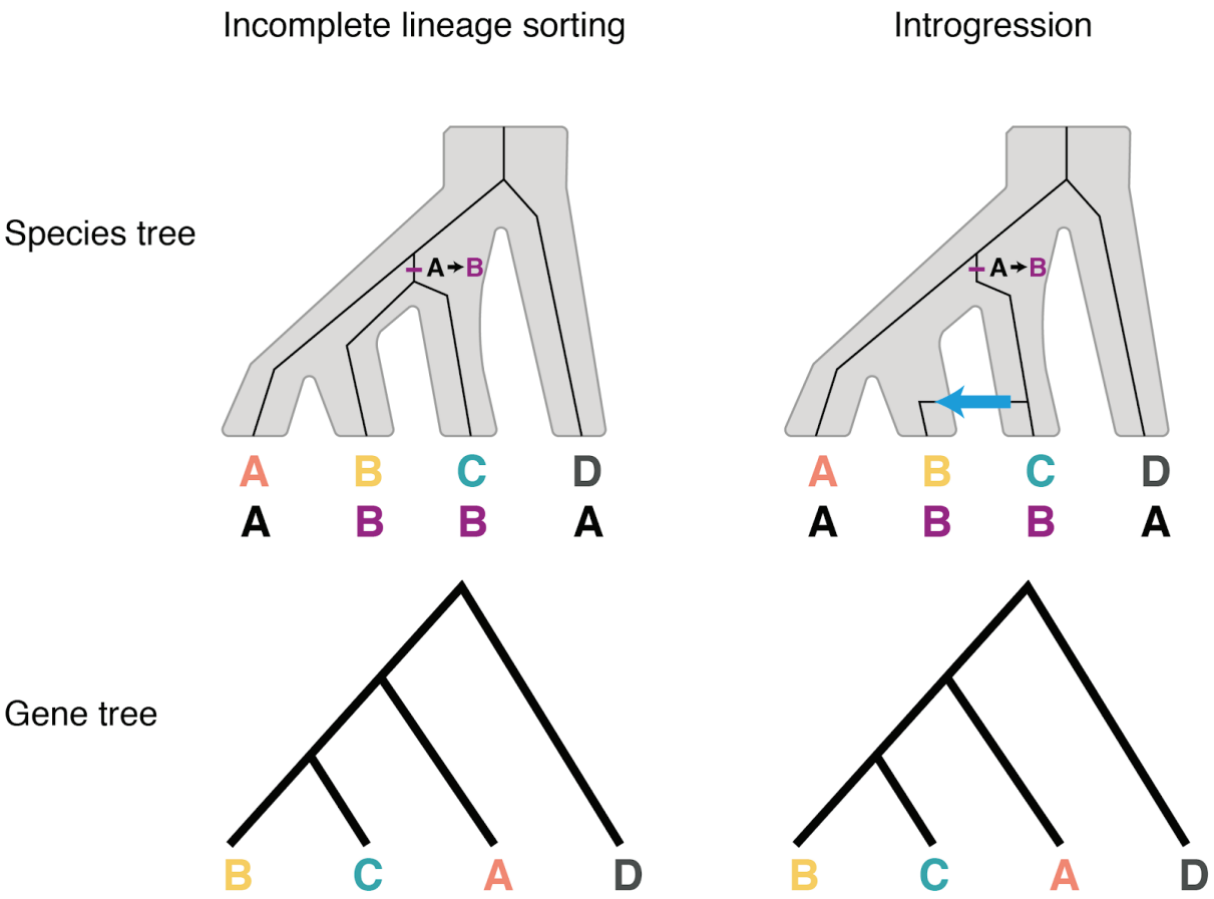


Figure 5 Schematic representing how ILS and introgression can result in phylogenetic conflict. Incomplete lineage sorting (ILS) and introgression are two of the major causes of gene tree heterogeneity. The diagrams on the top show a species tree with two sister species (A and B), a third external species (C) and an outgroup (D). The black lines represent the gene tree for a given locus and the purple dash represents a mutation from the ancestral allele A to a derived allele B. The diagrams on the bottom represent the topology of the gene tree, which in both cases is in conflict with the species tree topology. As exemplified here, both introgression and ILS can lead to the same conflicting topologies. If the incongruence is caused by ILS, the frequencies of ABBA and the alternative BABA are expected to be equal. Alternatively, if the cause of incongruence is introgression between C and either A or B, one of the patterns is expected to occur with a higher frequency.

3.4 | Phylogenetic Inference Methods

Phylogenetic inference methods have been paramount in resolving the evolutionary history of species and, ultimately, the Tree of Life. These methods have experienced two main waves of development coinciding with the discovery of the polymerase chain reaction (PCR) (Mullis 1990) and more recently with the emergence of genome-scale data (Liu et al. 2015).

With the purpose of inferring phylogenetic relationships using molecular data, distance-based methods were the first to be developed. In these methods, the genetic distance between every pair of sequences is calculated and the resulting distance matrix is used for building the tree. Distance methods have the advantage of being computationally efficient and, while their use to infer phylogenies is currently less widespread, they remain a useful approach to analyse large datasets from recently diverged taxa (Yang and Rannala 2012). However, they can perform poorly when analysing very divergent sequences because the variance of distance estimates becomes larger and distance methods, such as neighbour-joining (Saitou and Nei 1987), do not account for large variances. Phylogenetic networks methods based on distances have also been developed and remain as useful approaches for detecting areas of reticulation due to the aforementioned causes of discordance (Bryant and Moulton 2004; Huson and Bryant 2006; Suh et al. 2015).

Subsequently, the first character-based methods were developed during the 1970s and 1980s (Farris 1970; Felsenstein 1981). These methods simultaneously compare all sequences in the alignment, considering one site in the alignment at a time to calculate a score for each tree being considered. The calculation of this score relies on evolutionary models that consider the substitution process based on substitution rates and nucleotide frequencies (Jukes and Cantor 1969; Kimura 1980; Rodríguez et al. 1990) and can incorporate additional parameters such as invariant sites or a discrete gamma approximation that allows to model for variation in the rate of evolution across sites (Yang 1994). The two most widely used probabilistic approaches are maximum likelihood (ML) and bayesian inference (BI). ML approaches rely on the likelihood function which is defined as the probability of the data given the parameters of the model and a topology with branch lengths. In order to obtain the species tree, ML approaches search for the parameter values and topology that maximise the likelihood (maximum likelihood estimate) (Felsenstein 1981; Kozlov et al. 2019). BI differs from ML in considering that parameters in the model are random variables with statistical distributions instead of unknown fixed constants as assumed by ML (Rannala and Yang 1996; Ronquist et al. 2012; Aberer et al. 2014). Prior distributions are assigned to the parameters and combined with the data in order to generate the posterior distribution.

Markov chain Monte Carlo algorithms (MCMC) are then used to obtain a sample of the posterior distributions because these cannot be directly calculated.

To analyse multilocus data and especially phylogenomic data, two prevalent approaches are currently used: concatenation and coalescent methods. Despite intense ongoing debate about which of these methods is more appropriate to analyse multilocus sequence data (Liu et al. 2009; Gatesy and Springer 2014; Zhong et al. 2014; Springer and Gatesy 2016), in the majority of cases, both methods yield largely congruent results and areas where they differ generally provide important evolutionary insights (Tonini et al. 2015).

Concatenation methods rely on concatenating all genes into a supermatrix and using ML or BI to infer a single tree that is assumed to underlie all genes and to correspond to the species tree (Figure 6). Because the failure to account for substitution rate variation can seriously mislead phylogenetic analyses (Buckley et al. 2001; Telford and Copley 2011), datasets can be partitioned in order to ensure the appropriate evolutionary model for each partition (Lanfear et al. 2012).

By contrast, coalescent methods do not assume that a single tree underlies all genes. Instead, these methods treat gene trees as conditionally independent random variables given the species tree and parameters such as species divergence times and population sizes under the multispecies coalescent (MSC) model (Kingman 1982; Rannala and Yang 2003). There are two major approaches to species tree inference using the MSC. The summary or two-step methods use ML or BI to estimate gene trees for individual loci and then infer the species tree using summary statistics or a pseudolikelihood function (Figure 6) (Liu et al. 2010; Mirarab and Warnow 2015). These methods can easily analyse thousands of genes due to their computational efficiency. However, if errors in gene tree reconstruction are not taken into account, they can produce misleading results. On the other hand, full coalescent or single-step methods estimate both gene trees and species trees concurrently according to multilocus sequence data, priors, and the MSC model under a maximum likelihood framework (Figure 6) (Ogilvie et al. 2017; Flouri et al. 2018). These methods are more robust than summary methods but are computationally intensive, which still precludes the analysis of datasets containing thousands of genes. The MSC framework allows to strictly accommodate ILS (Rannala and Yang 2003;

Degnan and Rosenberg 2009) and recent developments have extended the MSC model in order to also incorporate introgression (Solís-Lemus and Ané 2016; Wen and Nakhleh 2018; Flouri et al. 2020).

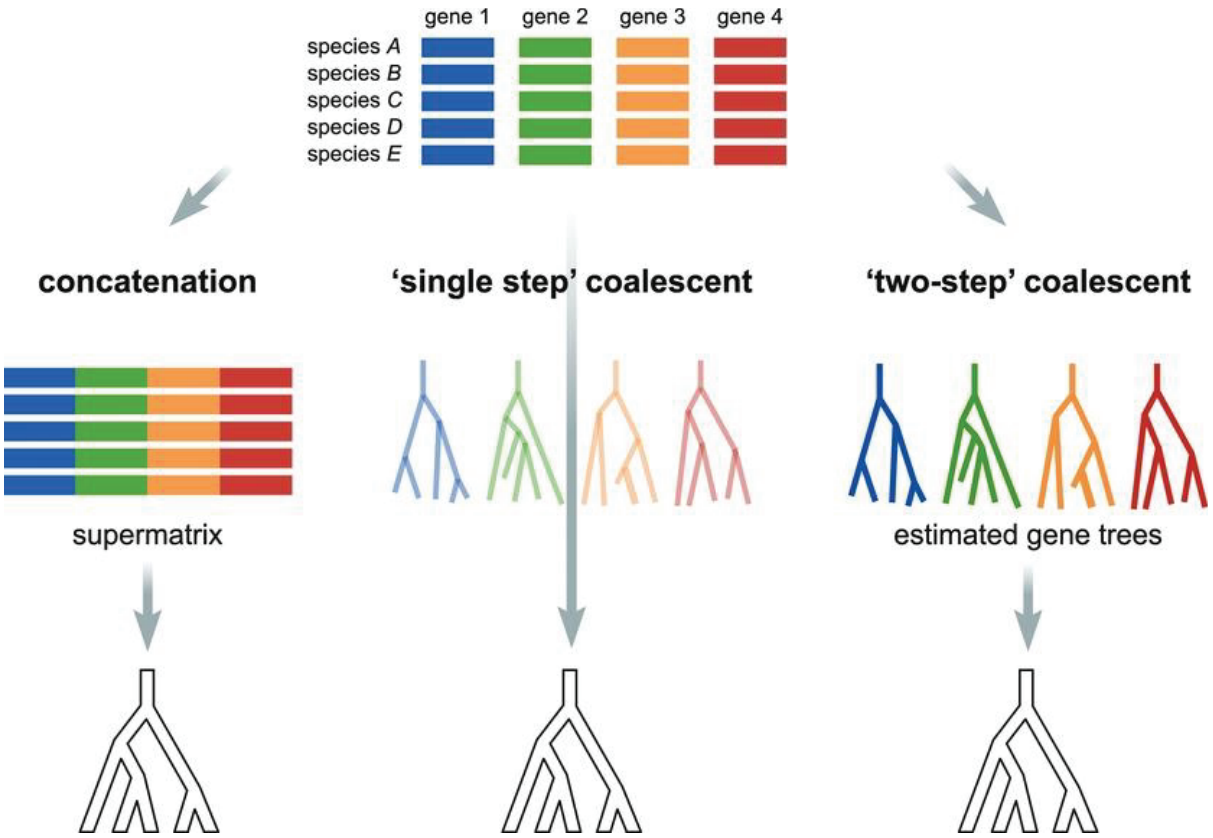


Figure 6 Schematic showing the three most widely used phylogenetic approaches for the analysis of phylogenomic datasets. On the left, the supermatrix approach (concatenation) relies on the concatenation of gene alignments and inference of a phylogeny using ML or BI methods. In the centre, the full coalescent approach simultaneously estimates gene trees and the species tree according to sequence data, priors and the MSC. On the right, summary methods estimate gene trees for individual loci using ML or BI approaches and then use the gene trees to estimate the species tree. Extracted from (Liu et al. 2015).

However, to detect signatures of introgression in a phylogenetic context most studies make use of the *D*-statistic (Green et al. 2010; Durand et al. 2011; Patterson et al. 2012), also known as ABBA-BABA tests. Given a pectinate tree topology (((A,B),C),D) (Figure 5), this test is based on the premise that the genome-wide frequencies at which two incongruent allele patterns appear across the tips (ABBA and BABA) can be used to infer introgression. These patterns, in which taxon C exhibits a derived allele relative to the outgroup O that is shared only by A or B (but not both), are incongruent with the

phylogeny. If the incongruence is caused by ILS, the frequencies of ABBA and BABA are expected to be equal. Alternatively, if the cause of incongruence is introgression between C and either A or B, one of the patterns is expected to occur with a higher frequency.

3.5 | Divergence Time Estimation

Phylogenetic methods allow the inference of the evolutionary relationships among organisms, but do not provide information about the times at which different clades diverged. In order to understand historical events that have shaped diversification in a group, it is thus necessary to estimate divergence times. The idea that divergence times could be estimated was first introduced in the 1960s by Zuckerkandl and Pauling. By comparing protein sequences, these authors highlighted that the rate of evolution at the molecular level is approximately constant through time and among species (known as the molecular clock hypothesis) (Zuckerkandl and Pauling 1965).

Despite many concerns about the accuracy and general applicability of this theory, it revolutionised the field of molecular evolution. Biologists adopted the use of the molecular clock to infer the divergence times of major species divergence events in the Tree of Life (Doolittle et al. 1996). The first implementations of the molecular clock assumed a constant rate (strict clock) and used fossil-age calibrations as point values, despite the fact that the fossil record can never provide a precise date estimate for a clade. Subsequently, evidence gathered that the molecular evolutionary rate is not constant (Langley and Fitch 1974; Felsenstein 1981). Indeed, as highlighted earlier, many factors can influence the variability in the evolutionary rate, such as generation time, population size or basal metabolic rate (Reis et al. 2015) and these factors remain a matter of debate (Ho 2014).

For closely related species, or in the analysis of population data, the strict clock can be a good approximation of reality. However, for calculating divergence times of more distantly related species, it has been necessary to develop methods that can accommodate variation in the evolutionary rate, as well as uncertainty in the fossil record. These methods, which use 'relaxed clocks', were developed by way of the advent

of Bayesian methods and a growing availability of genetic data (Kumar 2005; Heath and Moore 2014; Ho 2014).

Calibrations are a crucial step in divergence time estimation analyses. They are primarily based on fossil or geological evidence, and their choice can tremendously impact molecular estimates of evolutionary timescales (Ho and Phillips 2009; Jun Inoue 2010). In divergence time estimation analyses, it is thus very important to carefully choose the calibrations (Parham et al. 2012).

The aforementioned methods use a concatenation approach. However, analyses based on concatenation can lead to biases in branch lengths and misleading age estimates, particularly at recent timescales (McCormack et al. 2011; Angelis and Dos Reis 2015; Mendes and Hahn 2016). For such events, the multispecies coalescent model (MSC) offers a more accurate solution by incorporating the effects of incomplete lineage sorting (ILS) (Edwards et al. 2016). Recent methods have overcome this problem by allowing divergence-time estimation with genomic data using the MSC model (Stange et al. 2018).

3.6 | Historical Biogeography

Historical biogeography is a field which studies the patterns of species geographical ranges through geological time and focuses on the study of the historical processes (paleogeographic, paleoceanographic and paleoclimatic processes) that have shaped present and past species distributions (Crisci 2001). In recent years, the field has been dominated by phylogenetic biogeography (Barry Cox and Moore 2005; Maguire and Stigall 2008), wherein phylogenetic hypotheses and more recently, time-calibrated phylogenies are used to trace the evolution of geographic range.

A plethora of methods have been developed for inferring ancestral geographic ranges on phylogenies, which have different assumptions and consider different cladogenetic processes (Ronquist 1997; Ree and Smith 2008; Landis et al. 2013). Using the same datasets, these competing models can now be compared under a common statistical framework (Matzke 2013). This approach has the potential to answer long-standing questions in the field such as: 'What is the relative importance of founder-event

speciation in island clades versus continental clades?' (Matzke 2014). The possibility of testing different models under a common statistical framework has led to an increasing number of studies that have utilised this framework to gain insights into the biogeographical history of a wide diversity of groups (Rojas et al. 2016; Feng et al. 2017; Oliveros et al. 2019; Vianna et al. 2020; Wang et al. 2020).

3.7 | Species Delimitation

Species are a fundamental category of biological organisation and are used as basic units in several fields of research. One of the most influential species concepts, the biological species concept, was proposed by Ernst Mayr and has become a textbook standard: 'species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups' (Mayr 1999). Despite its wide acceptance, Mayr's definition also provoked critiques and the debate about the definition of species. This debate intensified particularly in the 1970s (and continues today), stimulating the proposal of a plethora of new species concepts (Mayden 1997) and compromising the status of the species as a basic category of biological organisation. These different concepts disagreed in adopting different properties acquired by lineages during the divergence continuum (e.g. intrinsic reproductive isolation, diagnosability, monophyly) (De Queiroz 2005; Hey 2006). More recently, De Queiroz (2005) identified the feature that unifies all alternative species concepts, which is the acceptance of 'existence as separately evolving metapopulation lineage' as the primary defining property of the species, and proposed a unified species concept that considered this common feature as the only necessary property of species (De Queiroz 2007) (known as the 'General lineage concept').

Due to the intense debate about the definition of species, species delimitation approaches have as a result, changed quite dramatically over the years. Traditionally, species were defined based on morphological differences (apomorphies), which often resulted in artificially broad delineations and the difficulty of identifying cryptic species. More recently, molecular data has emerged as one of the most useful sources of evidence to identify independently evolving lineages and particularly since the rise in the availability of genome-scale data. The availability of genome-scale data, together with

the unification of species concepts (De Queiroz 2007) have stimulated the advance in the development of MSC methods to detect lineage separation even in the presence of ILS (Knowles and Carstens 2007) and numerous methods are now available (Carstens et al. 2013; Leaché et al. 2014; Rannala 2015). These methods are increasingly being used to delimit species in a wide range of taxa (Abdelkrim et al. 2018; Tonzo et al. 2019; Ewart et al. 2020; Hosegood et al. 2020; Newton et al. 2020). However, the high resolution of genomic data makes it difficult to distinguish population structure within a metapopulation from species boundaries when using MSC methods (Sukumaran and Knowles 2017; Chambers and Hillis 2020). Therefore, to be able to delimit species as separately evolving metapopulation lineages as defined by De Queiroz (2005), the combination of MSC species delimitation approaches with other lines of evidence such as morphological, ecological or phenological data, provides a robust framework (Carstens et al. 2013; Sukumaran and Knowles 2017; Chambers and Hillis 2020).

3.8 | Evolutionary Significant Units

The implementation of efficient conservation policies relies on an understanding and characterisation of diversity within and among species to define evolutionary significant units (ESUs), originally defined as a group of organisms that have been isolated from other conspecific groups for a sufficient period of time to have undergone 'meaningful' genetic divergence (Moritz 1994). From a conservation perspective, ESUs are defined as populations that merit separate management and are the priority for conservation (Crandall et al. 2000). Similar to the species concept, the ESU concept is also under long-standing debate (Fraser and Bernatchez 2001) and has been applied from the species level to the population level. Within the context of this thesis, I wish to emphasise the adaptive evolutionary conservation view of ESUs (Fraser and Bernatchez 2001) under the species level. Specifically, addressing the questions: 1) 'Is the population genetically distinct from other conspecific populations?'; and 2) 'Does the population show any ecological or morphological signs of adaptation to its environment?'

3.9 | Speciation Genomics

Genetic differentiation accumulates across the genome as populations diverge. Along the speciation continuum, numerous evolutionary processes can promote, stall or even reverse the trajectory of the speciation process, including gene flow, mutation, recombination, genetic drift, and selection (Figure 7a) (Coyne and Orr 1989; Ravinet et al. 2017; Wolf and Ellegren 2017; Kearns et al. 2018). The effects of the speciation process vary along the genome and generally lead to a heterogeneous genomic landscape with peaks and troughs of differentiation and divergence.

The first genome scan approaches began with the idea that regions of high differentiation (outlier loci or 'islands of differentiation', usually defined as high F_{ST} regions) arose due to reproductive barriers, while the rest of the genome would show low differentiation due to the homogenising action of gene flow (Wu 2001; Turner et al. 2005). With the rapid development of high--throughput sequencing technologies during the past decade, the field of speciation genomics has expanded into wild populations allowing a rapid advancement of the field (Wolf et al. 2010; Seehausen et al. 2014; Ravinet et al. 2017; Wolf and Ellegren 2017). Despite the progress in documenting genomic landscapes of divergence in a myriad of organisms and different stages of the speciation continuum, interpreting these patterns has been anything but straightforward. This has been due to the numerous processes that interact to drive peaks and troughs of differentiation and divergence, which include external processes such as genetic drift, selection or gene flow but also genome features such as mutation, structural variants, recombination or gene density (Figure 7a). Such processes interact with one another and, at the same time, are also influenced by the demographic and evolutionary history of the species (Ravinet et al. 2017). Indeed, differentiation peaks cannot be simply attributed to have arisen due to reduced gene flow as originally thought, and instead they can be caused by reductions in diversity due to intrinsic genome features such as background selection due to reduced recombination or divergent selection (Figure 7b) (Nachman and Payseur 2012; Cruickshank and Hahn 2014).

Despite the complexity of the speciation process, applying a suite of summary statistics and exploring how F_{ST} relates with absolute measures of divergence (D_{XY}) and within-population nucleotide diversity (π) can provide insights into the processes involved in the formation of islands of differentiation (Figure 7c) (Cruickshank and Hahn 2014; Han et al. 2017; Irwin et al. 2018). However, in order to draw a more complete picture of the processes affecting differentiation, a solid understanding of the demographic history and the recombination landscapes are essential (Charlesworth 2009; Ferchaud and Hansen 2016; Burri 2017). The growing availability of genomic data together with the development of powerful simulation software that can integrate the multiple factors affecting the landscapes of divergence (i.e. Haller and Messer 2019) have the potential to broaden our understanding of how the speciation process shapes the genome and to help at understanding how genomic divergence relates to the speciation process across groups of organisms, a central question that remains elusive.

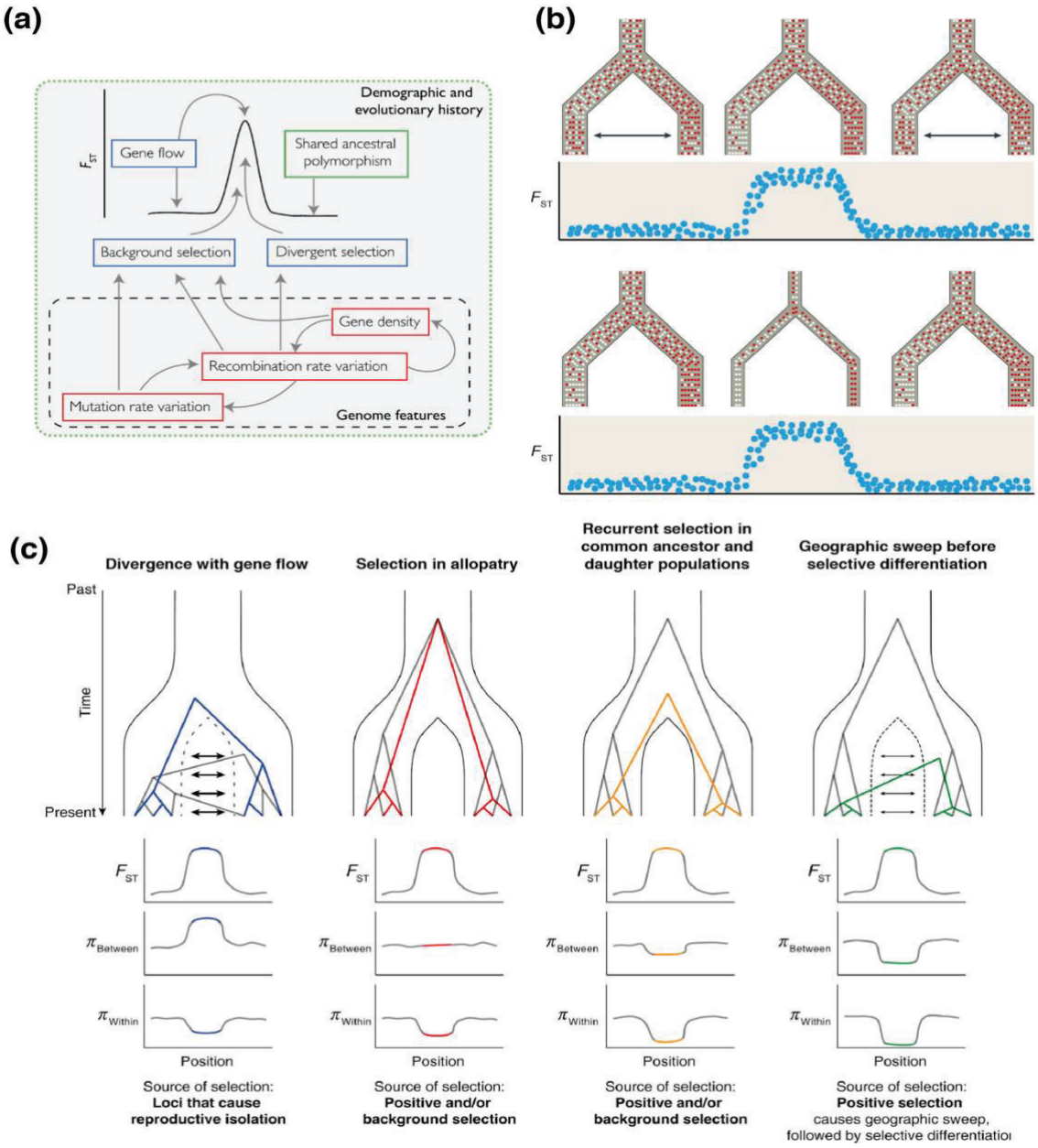


Figure 7 Processes shaping the genomic landscapes of divergence. (a) Schematic of the interplay between external processes (blue boxes) and genome features (red boxes) contributing and interacting to shape the genomic landscapes of divergence. At the same time, the effects and extent of these processes are shaped by the influence of demographic and evolutionary history. Extracted from (Ravinet et al. 2017). (b) Schematic of a region of a chromosome where two alternative processes generate islands of differentiation. On the top, a peak in F_{ST} is driven by reduced gene flow compared to a background of low differentiation due to strong gene flow. Conversely, on the bottom, a region with reduced diversity due to a reduced effective population size (N_e) produces a peak in F_{ST} due to an enhanced rate of lineage sorting compared to the background. Modified from (Wolf and Ellegren 2017). (c) Schematic of four alternative models for the formation of genomic islands of differentiation. Top panels illustrate a population splitting into two over time and gene genealogies for a neutral (in grey) and outlier loci (in colour) are shown. Bottom panels show variation in F_{ST}, absolute divergence ($\pi_{between}$) and nucleotide diversity within the populations (π_{within}) along a region subject to selection flanked by neutral regions. Sources of selection are listed for each model.

4 | Evolutionary History of Shearwaters

The importance of geographical isolation as a driver of population differentiation and speciation has long been established (Mayr 1963, 1999). Shearwaters represent an interesting case study to investigate the drivers of speciation and diversification in geographical isolation. On the one hand, they are highly mobile species that live in the marine environment, which has a lack of obvious physical barriers. On the other hand, they have global disjunct distributions (island-breeders) and are highly philopatric to their breeding grounds. These characteristics are expected to play opposite roles in the process of differentiation and speciation, and understanding how they interact to shape the evolutionary process is a challenging endeavour that remains unanswered. Below, I summarise what research has previously been undertaken to study the evolutionary history of shearwaters and what we know about their potential drivers of diversification.

4.1 | Phylogenetic Relationships of Shearwaters

Resolving the phylogenetic relationships among shearwaters has long been challenging. The first attempts to address this issue were undertaken by Kuroda (1954) who examined behaviour, osteology, external morphology, distribution and the fossil record, and Wragg (1985) who used an independent dataset of osteological characters. The phylogenies of Kuroda and Wragg were largely congruent, but unresolved, and the systematics of shearwaters remained controversial. This was likely due to slowly evolving osteological characters and remarkable similarities in plumage colouration (e.g. Figure 2c), which also challenges shearwater identification in the wild (Howell 2012; Gil-Velasco et al. 2015; Flood and Fisher 2020). In the late 1990s, the first molecular phylogenies based on partial mitochondrial *cytb* sequences (Austin 1996; Heidrich et al. 1998) revealed several conflicts with previous phylogenies based on morphology. However, these phylogenies were also unresolved and showed several polytomies among and within the three shearwater genera (particularly within major biogeographic groups in the genus *Puffinus*). Subsequent studies using complete *cytb* sequences were also unable to resolve these polytomies and suggested that hard polytomies may exist due to rapid diversification (Figure 8) (Austin et al. 2004; Pyle et

al. 2011). These studies also showed that the old genus *Puffinus* was polyphyletic, which was used as the reason to separate the species of this old genus in two different genera (*Ardenna* and *Puffinus*). From this point onwards in this thesis, the use of *Puffinus* will be restricted to the new genus *Puffinus* (not including *Ardenna*). More recently, a study using UCE loci to resolve the phylogenetic relationships among the major clades of Procellariiformes (Estandía 2019) recovered a generally well resolved topology, though maximum likelihood and species tree estimation methods yielded low support and conflicting topologies for the split among the three shearwater genera. Altogether, these results suggest that shearwater diversification has been shaped by periods of rapid speciation, when ILS and/or historical introgression may have occurred.

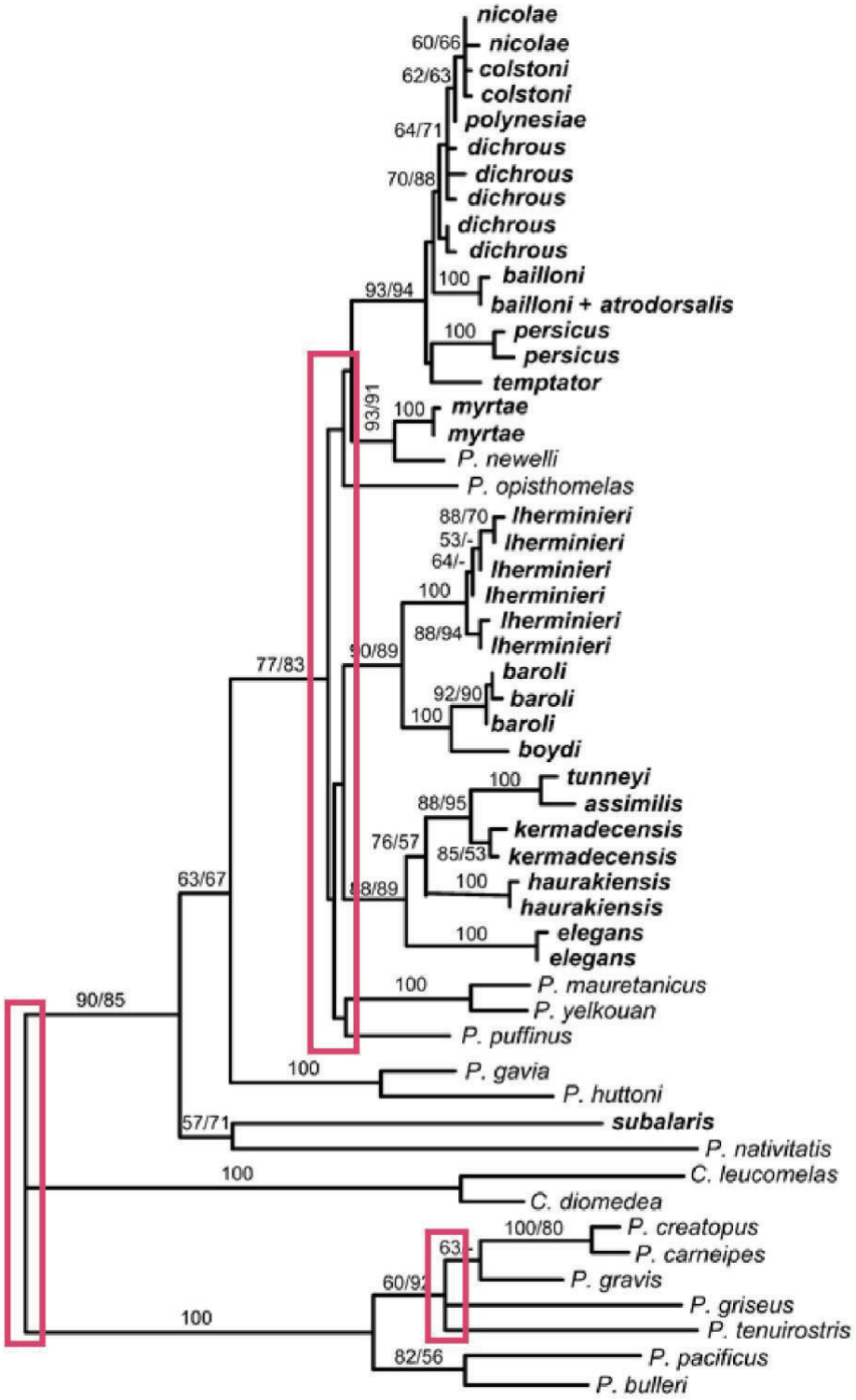


Figure 8 Austin et al. (2004) maximum-likelihood phylogeny of shearwaters based on 917 bp of *cytb* sequence alignments. Unresolved areas of the tree are highlighted with magenta boxes and correspond, from left to right, to the split among the three shearwater genera, the split among the major biogeographic groups within the genus *Puffinus*, and the basal split among most of *Ardenna* lineages.

4.2 | Biogeographic History and Drivers of Shearwater Diversification

To date, several biogeographical hypotheses have been proposed to explain the current distribution of *Puffinus* and *Ardenna* shearwater species (Kuroda 1954; Bourne et al. 1988; Austin 1996; Austin et al. 2004). As mentioned above, these two genera were originally classified under the same old genus *Puffinus* and thus were frequently analysed together. Based on his phylogeny, Kuroda (1954) suggested that the North Atlantic was the ancestral area where different *Ardenna* and *Puffinus* subgroups differentiated (Thyellodroma: *A. pacifica* and *A. bulleri*, Hemipuffinus: *A. carneipes* and *A. creatopus*, Ardenna: *A. gravis*, Neonectris: *A. grisea*, *A. tenuirostris* and *P. nativitatis* and Puffinus: including the rest of species of the genus *Puffinus*). Kuroda then proposed that the ancestors of the first four subgroups moved to the southern hemisphere via the flooded Panama land bridge and then differentiated and distributed throughout the Southern Ocean. For the Puffinus subgroup, Kuroda proposed that evolution of extant taxa in this group had occurred due to fragmentation of the original ancestral population and dispersal from the North Atlantic to the eastern Pacific and Indian Oceans. Subsequently, vicariant events in the late Tertiary would have isolated populations in the Pacific, the North Atlantic, the Mediterranean and the Indian Ocean. Finally, Kuroda proposed a secondary dispersal phase of the ancestors of the Little-Audubon's shearwater complex throughout the tropical and subtropical latitudes of the three major oceans.

Austin (1996) did not dispute the North Atlantic origin of *Ardenna* and *Puffinus*. However, instead of the previously proposed four different colonisations to the southern hemisphere by the ancestors of the different *Ardenna* lineages, Austin proposed a single colonisation, based on the monophyly of the *Ardenna* genus, and a subsequent radiation which gave rise to the different extant lineages. Austin agreed with Kuroda in attributing the diversification of *Puffinus* to the fragmentation of water bodies in the North Atlantic and central Europe region in the late Tertiary and through secondary dispersal into the Pacific and Indian Oceans. In order to explain the unexpected position of *P. gavia* and *P. huttoni* in his phylogeny, Austin proposed that these species would have evolved from a single ancestral population that dispersed to the west Pacific.

Focusing on the small *Puffinus* shearwaters, Austin et al. (2004) provided a more detailed picture, proposing that most of this complex had arisen via radiation in three geographically separate regions (Southern Ocean, Tropical Indian and Pacific Oceans and Tropical and Subtropical Atlantic Ocean). Austin et al. (2004) proposed that this event occurred around 2.5-5.2 million years ago (Mya), based on *cytb* evolutionary rate for the Procellariidae (Nunn and Stanley 1998). Austin et al. (2004) also proposed that the North Atlantic populations have been isolated from the Pacific ones since the formation of the Panama land bridge. Due to the separation of the Southern Ocean clade from the Tropical Indian and Pacific clades, they suggested that changes in sea-surface temperature and food availability across the boundary between tropical and subtropical oceanic zones may act as an ecological barrier to dispersal.

Historical biogeography of the genus *Calonectris* has received less attention. However, in a comprehensive phylogeographic study, Gómez-Díaz et al. (2006) suggested that a North Atlantic origin of the genus was most likely based on the fossil record and a greater diversity in that area. Based on *cytb* evolutionary rates, they matched the first speciation event within *Calonectris* with the Panama land bridge formation ~3 Mya, suggesting a vicariant scenario for the divergence of the Pacific and the Western Palearctic clades. For the separation between the Atlantic and Mediterranean clades, they suggested that this would have likely occurred by range contraction followed by local adaptation during the major biogeographic events of the Pleistocene.

In conclusion, the North Atlantic has been the only proposed ancestral area for shearwaters and several processes have been proposed to explain their biogeographic history: including dispersal to other ocean basins, allopatric speciation within and among ocean basins and ecological barriers to dispersal. However, none of the aforementioned studies used divergence time estimation analyses nor tested different biogeographical models under a statistical framework. Thus, a proper assessment of the historical processes that have shaped shearwater biogeographical history and a formal evaluation of the role of founder events and vicariance during their diversification is required.

4.3 | Species Limits in the North Atlantic and Mediterranean *Puffinus* Shearwaters

Species limits in shearwaters are controversial, mostly due to the high morphological stasis in the group, which is likely driven by similarities in their ecological patterns. North Atlantic and Mediterranean *Puffinus* shearwaters are under a particularly contentious ongoing taxonomic debate. Several sources of evidence have been used for assigning populations to species and subspecies in the group, including morphological, genetic, behavioural and ecological evidence (Heidrich et al. 1998; Sangster et al. 2005; Olson 2010; Genovart et al. 2012; Ramos et al. 2020; Rodríguez et al. 2020). However, there has been a lack of consensus on whether the evidence gathered for the taxa in the group supports species status or taxonomic treatment below the species status (for example, as ESUs). Given these uncertainties, in order to develop effective conservation measures there is an urgent need for a robust review of the current taxonomy.

Because birds are easy to observe in the field and have been widely studied, bird taxonomies have mostly relied on phenotypic, ecological and behavioural evidence (Price and Others 2008; Tobias et al. 2010; Del Hoyo et al. 2014). However, disagreement on what evidence supports species status in birds has led to the maintenance of four primary world bird lists, which differ in their primary goals and their taxonomic criteria (IOC v.10.2: Gill et al. 2020; Clements: Clements 2007; HBW & Birdlife International: del Hoyo et al. 2014; Howard & Moore v.4.1: Christidis et al. 2018). In bird taxonomy, there has been a general reluctance to incorporate genetic data to assist species delimitation, due in part to the patchiness of genetic data and the extent of disagreement about how they should be applied to delimit species (Edwards et al. 2005; Knowles and Carstens 2007; Price 2008). However, recent advances in MSC methods for species delimitation, and the increasing availability of genome-wide data are now providing a valuable source of evidence to assist species delimitation. Such approaches are increasingly being considered in order to update avian taxonomic classifications.

Objectives

The overall aim of this thesis is to evaluate the driving forces that shape the evolutionary history of shearwaters. Using genomic data derived from Restriction site-Associated DNA sequencing (RAD-seq) and Ultraconserved Elements (UCE), I will integrate phylogenetic and population genetic analyses across several evolutionary timescales to shed light on the patterns and processes that contribute to speciation, diversification, dispersal, and trait evolution in pelagic seabirds.

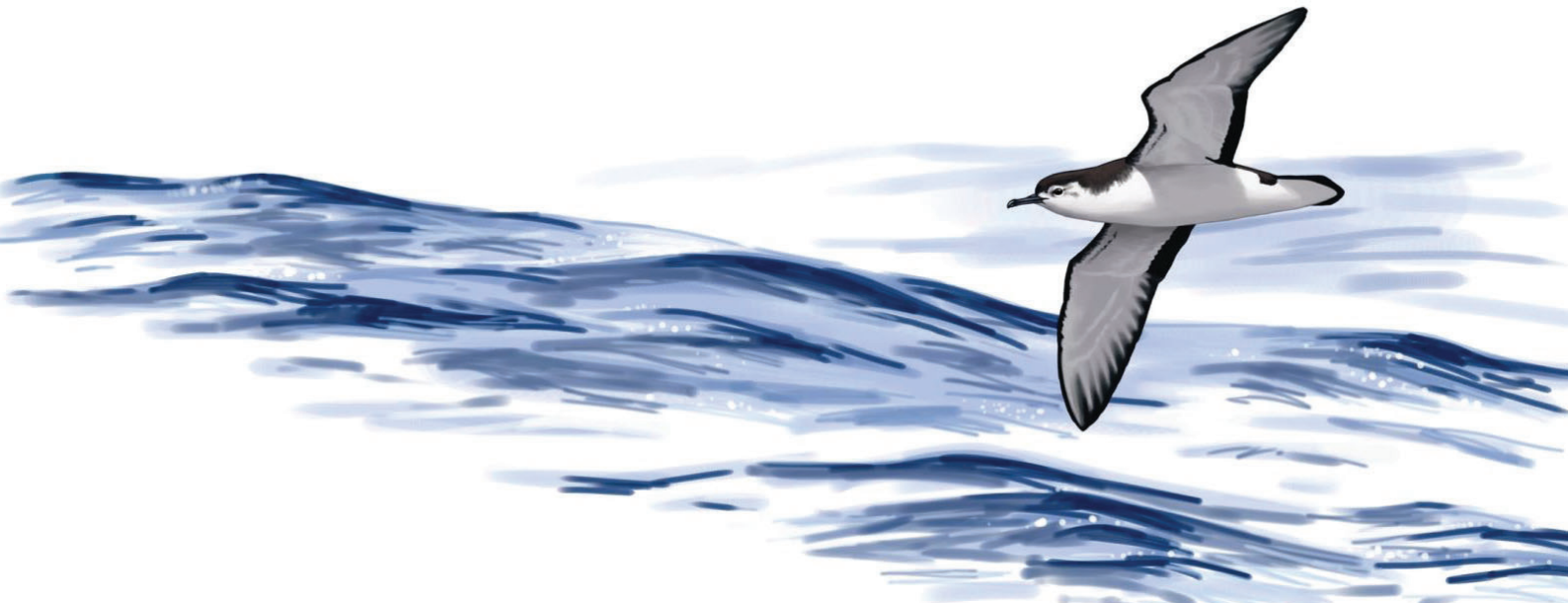
The specific objectives of this thesis are:

- Resolve the phylogenetic relationships of shearwaters.
- Disentangle the role of incomplete lineage sorting and introgression as causes of phylogenetic discordance during rapid diversification by applying concatenated and multispecies coalescent phylogenetic approaches on RAD-seq and UCE data.
- Explore the historical processes that have shaped their biogeographical history and evaluate the role of founder events, vicariance and ocean currents during their diversification.
- Investigate the ecological and geographical drivers of variability in a key phenotypic trait, the body size.
- Evaluate and update the taxonomy of the group combining genomic data analyses with consideration of morphological, behavioural and ecological evidence.
- Assess patterns of genomic variation among and within species of North Atlantic and Mediterranean *Puffinus* shearwaters.
- Characterise and investigate the main drivers shaping genomic landscapes of divergence across a speciation continuum.

Chapter I

Integrating Sequence Capture and Restriction Site-Associated DNA Sequencing to Resolve Radiations of Pelagic Seabirds

JOAN FERRER OBIOL, HELEN F. JAMES, R. TERRY CHESSER, VINCENT BRETAGNOLLE, JACOB GONZÁLEZ-SOLÍS, JULIO ROZAS, MARTA RIUTORT AND ANDREANNA J. WELCH



Integrating Sequence Capture and Restriction Site-Associated DNA Sequencing to Resolve Recent Radiations of Pelagic Seabirds

Joan Ferrer Obiol[1,2], Helen F. James[3], R. Terry Chesser[4,5], Vincent Bretagnolle[6], Jacob González-Solís[7,2], Julio Rozas[1,2], Marta Riutort[1,2], and Andreanna J. Welch[8]

[1]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

[2]Institut de Recerca de la Biodiversitat (IRBio), Barcelona, Catalonia, Spain

[3]Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

[4]U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, MD, USA

[5]National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

[6]Centre d'Études Biologiques de Chizé, CNRS & La Rochelle Université, 79360, Villiers en Bois, France

[7]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

[8]Department of Biosciences, Durham University, Durham, UK

Accepted in Systematic Biology

Abstract

The diversification of modern birds has been shaped by a number of radiations. Rapid diversification events make reconstructing the evolutionary relationships among taxa challenging due to the convoluted effects of incomplete lineage sorting (ILS) and introgression. Phylogenomic datasets have the potential to detect patterns of phylogenetic incongruence, and to address their causes. However, the footprints of ILS and introgression on sequence data can vary between different phylogenomic markers at different phylogenetic scales depending on factors such as their evolutionary rates or their selection pressures. We show that combining phylogenomic markers that evolve at different rates, such as paired-end double-digest restriction site-associated DNA (PE-ddRAD) and ultraconserved elements (UCEs), allows a comprehensive exploration of the causes of phylogenetic discordance associated with short internodes at different timescales. We used thousands of UCE and PE-ddRAD markers to produce the first well-resolved phylogeny of shearwaters, a group of medium-sized pelagic seabirds that are among the most phylogenetically controversial and endangered bird groups. We

found that phylogenomic conflict was mainly derived from high levels of ILS due to rapid speciation events. We also documented a case of introgression, despite the high philopatry of shearwaters to their breeding sites, which typically limits gene flow. We integrated state-of-the-art concatenated and coalescent-based approaches to expand on previous comparisons of UCE and RAD-Seq datasets for phylogenetics, divergence time estimation and inference of introgression, and we propose a strategy to optimise RAD-Seq data for phylogenetic analyses. Our results highlight the usefulness of combining phylogenomic markers evolving at different rates to understand the causes of phylogenetic discordance at different timescales.

Keywords: Aves, shearwaters, phylogenomics, radiations, incomplete lineage sorting, introgression, UCEs, PE-ddRAD-Seq

1 | Introduction

Understanding the phylogenetic relationships among species is paramount in biology and provides a framework for understanding evolutionary processes such as temporal and biogeographical patterns of diversification. Radiations are one of the major challenges in reconstructing evolutionary history and examples of recalcitrant clades are widespread across the Tree of Life (Song et al. 2012; Wagner et al. 2013; Jarvis et al. 2014; Pease et al. 2016). Such rapid diversification events commonly generate patterns of phylogenetic incongruence that hinder the understanding of major evolutionary processes.

Two prevalent processes contribute to incongruence. One of them is incomplete lineage sorting (ILS), which occurs when gene lineages coalesce into their common ancestor prior to the speciation events (Maddison 1997). Under ILS, retention and stochastic sorting of ancestral polymorphisms may result in misleading resolution of relationships among species. ILS is particularly prevalent at short internal branches in species trees and especially when effective population sizes (*Ne*) are large relative to the time between divergences (Pamilo and Nei 1988; Rosenberg and Nordborg 2002). The second process is introgression, the incorporation (usually via hybridisation and backcrossing) of alleles from one species into the gene pool of a second species

(Anderson 1949). Introgression plays an important role in the process of diversification (Abbott et al. 2016) and is especially important in adaptive radiations (The Heliconius Genome Consortium 2012; Malinsky et al. 2018). Distinguishing ILS from introgression remains a major challenge in phylogenomics. Recently, several methodological approaches have been developed that simultaneously account for both ILS and gene flow when reconstructing the evolutionary history of a clade (Solís-Lemus et al. 2017; Wen et al. 2018). The footprints of these processes on sequence data can, however, vary between different phylogenomic markers at different phylogenetic scales depending on factors such as their evolutionary rates or the type of selection that they experience (Martin and Jiggins 2017; Knowles et al. 2018).

The increasing availability of tractable phylogenomic data has certainly helped in resolving many contentious relationships in the Tree of Life (Rokas et al. 2003; Jarvis et al. 2014; Hughes et al. 2018). Genome-wide data can detect patterns of gene tree discordance and can be used to investigate the causes of phylogenetic incongruence in rapid diversification events (Arcila et al. 2017). Restriction site-associated DNA sequencing (RAD-Seq; Miller et al. 2007) and sequence capture of ultraconserved elements (UCEs; Faircloth et al. 2012), are two of the most widely used methods for shallow phylogenomics. RAD-Seq and related genotyping-by-sequencing approaches (e.g. ddRAD-Seq, Peterson et al. 2012; GBS, Elshire et al. 2011) have been primarily used for studying polymorphism in population genomics analyses, although they have also been successfully used for phylogeographic and interspecific phylogenetic studies of a wide variety of organisms (Emerson et al. 2010; Hohenlohe et al. 2010; Rubin et al. 2012). Conversely, UCEs were developed for studying deep evolutionary timescales but their flanking regions are variable enough to be informative at shallow evolutionary timescales (Faircloth et al. 2012; Smith et al. 2014). These two methods have recovered concordant phylogenetic relationships when using large enough datasets (Leaché et al. 2015; Harvey et al. 2016; Manthey et al. 2016; Collins and Hrbek 2018). RAD-Seq and UCEs have also been compared in terms of divergence time estimation using fossil-calibrated molecular-clock models and have been shown to accurately estimate divergence times at recent timescales (Collins and Hrbek 2018). Integrating these two approaches may provide a powerful tool for disentangling the roles of ILS and

introgression in rapid diversification events; however, empirical studies to test this assumption are generally lacking.

Modern birds provide many case studies, as their diversification has been characterised by a succession of rapid radiations, posing a significant challenge in resolving the phylogenetic relationships of several avian clades (Jarvis et al. 2014; Oliveros et al. 2019). Hybridisation can be common at the species level (Mallet 2005) and birds show relatively high levels, with 16.4% of the species having been documented to hybridise in nature (Ottenburghs et al. 2015).

Evolutionary relationships of tube-nosed seabirds (order Procellariiformes) are unresolved at many phylogenetic levels (Penhallurick and Wink 2004; Rheindt and Austin 2005). Many species of Procellariiformes are endangered (Croxall et al. 2012) and shearwaters are amongst the most vulnerable groups; 55% of shearwater species are listed as threatened by the IUCN Red List of Threatened Species. Shearwaters form a monophyletic group of medium-sized pelagic seabirds (family Procellariidae) consisting of three genera. Understanding shearwater diversification and biogeographic patterns can provide important insights into their biology, and ultimately assist in providing species delimitations or evolutionarily significant units vital for their conservation (Purvis et al. 2005).

Resolving the evolutionary relationships among shearwaters has long been challenging. Osteological and morphological analyses have generated many conflicting hypotheses (Kuroda 1954), likely due in part to their slowly evolving osteological characters and remarkable similarities in plumage colouration, which also make their identification in the wild challenging (Gil-Velasco et al. 2019). Mitochondrial DNA (mtDNA) analysis (Austin 1996; Heidrich et al. 1998; Austin et al. 2004; Pyle et al. 2011) revealed further conflicts and suggested polytomies may exist among and within the three shearwater genera (particularly within major biogeographic groups in the genus *Puffinus*). A recent analysis of the major clades of Procellariiformes using only UCE loci (Estandía 2019) recovered a generally well resolved topology, though maximum likelihood and species tree estimation yielded lower support and conflicting topologies for the split among the shearwaters, suggesting a rapid radiation where ILS and/or historical introgression may have occurred. Hybridisation has been documented in

several species of Procellariiformes, including shearwaters (Genovart et al. 2012; Booth Jones et al. 2017; Masello et al. 2019) despite strong philopatry to breeding colonies and, in most cases, a lack of overlap between breeding areas of closely related species. Thus, shearwaters demonstrate the challenges typical of many other taxonomic groups, where rapid diversification and introgression hinder attempts to confidently resolve their evolutionary history, and provide a good case study for how combining genomic datasets can aid in clarifying relationships at recalcitrant nodes.

Here, we generate the first phylogenomic datasets for shearwaters. We use paired-end ddRADSeq (PE-ddRAD) and UCE datasets to explore the role of ILS and introgression as causes of phylogenetic discordance during rapid diversification. We adopt a thorough and integrative approach, comparing and combining the two datasets, and applying state-of-the-art concatenated and coalescent-based approaches using shearwaters as a case study. We expand previous comparisons of UCE and RAD-Seq datasets for phylogenetics, divergence time estimation and inference of introgression, and we propose a strategy to optimise RAD-Seq data for phylogenetic analyses. Our approach allows us to completely resolve the phylogenetic relationships of shearwaters. We detect high levels of gene tree discordance associated with short internodes, mainly caused by ILS, and report a potential case of historical introgression. We show that our integrative approach provides a good framework for exploring the processes underlying phylogenetic incongruence during rapid diversification events.

2 | Materials and Methods

2.1 | Sampling and Sequence Data Generation

We obtained blood (n = 45), high-quality tissue (n = 19) or dry tissue (n = 3) samples for 30 taxa representing 26 of the 30 recognised species of shearwaters (Carboneras and Bonan 2019) (Table S1). We also sampled three species of Procellariiformes as outgroups: Fulmarus glacialis, a species from the same family as the shearwaters (Procellariidae); and two more distantly related outgroups, *Thalassarche chlororhynchos* (Diomedeidae) and *Oceanites oceanicus* (Oceanitidae). Species that

could not be included (*Puffinus heinrothi*, *P. bannermani*, *P. persicus* and *P. subalaris*) are mostly very localised or critically endangered.

We used the Qiagen DNeasy Blood and Tissue Kit to extract genomic DNA according to the manufacturer's instructions (Qiagen GmbH, Hilden, Germany). For dry tissue samples, we extracted DNA using the Dabney et al. (2013) ancient DNA extraction protocol. We used a Qubit Fluorometer (Life Technologies) to quantify and standardise DNA concentrations of all samples. PE-ddRAD data for the outgroups were retrieved from whole genome assemblies available from the Bird 10K project (Feng et al. 2020; see description of the *in silico* digestion protocol in the Data Assembly – PE-ddRAD-Seq dataset section below).

Library preparation, capture enrichment and sequencing of the UCEs was conducted by RAPiD Genomics, LLC (Gainesville, FL, USA). Briefly, genomic DNA was fragmented, and sequencing libraries were prepared. Indexed samples were subjected to PCR for 7 cycles prior to pooling. UCE loci were captured using the Tetrapods UCE-5Kv1 probe set (available at ultraconserved.org) as described by Faircloth et al. (2012), and PCR was conducted for 11 cycles to amplify the enriched library. Sequencing was performed on two lanes of an Illumina HiSeq 3000 platform using 100 bp paired-end (PE) sequencing.

Library preparation of PE-ddRAD loci was performed by the Genomic Sequencing and Analysis Facility, University of Texas at Austin, following the Peterson et al. (2012) protocol. Samples were digested using an uncommon cutter *EcoRI* and a common cutter *MspI* in a single reaction. After digestion, modified P1 Illumina adapters containing 5 bp unique barcodes and P2 Illumina adapters were ligated onto the fragments and individually barcoded samples were pooled. Barcodes differed by at least two base pairs (based on a Hamming distance metric) to reduce the chance of errors caused by inaccurate barcode assignment. Pooled libraries were size selected (between 150 and 300 bp after accounting for adapter length) using a Pippin Prep size fractionator (Sage Science, Beverly, Ma). Libraries were amplified in a final PCR step for 10 PCR cycles in six pools differing by their Illumina index, prior to sequencing on a single lane of an Illumina HiSeq4000 platform using 150 bp PE sequencing. The combination of unique barcodes and Illumina indexes allowed the multiplexing of all samples into one sequencing lane.

2.2 | Data Assembly

2.2.1 | UCE dataset

Raw reads were quality-filtered and cleaned of adapter contamination with TRIMMOMATIC v0.36 (Bolger et al. 2014), and were assembled into contigs using TRINITY v2.0.6 (Grabherr et al. 2011) as implemented in the PHYLUCE pipeline (Faircloth 2016). To identify contigs representing UCE loci, we mapped all assembled contigs to the probes' reference sequences (uce-5k-probes.fasta) and discarded those contigs not matching any probe, matching more than one probe or matching probes that matched multiple contigs, using the PHYLUCE match_contigs_to_probes.py script. The remaining UCE sequences were aligned using MAFFT v7.130B (Katoh and Standley 2013) and internally trimmed using GBLOCKS v0.91b (Castresana 2000). To retrieve polymorphism information lost when collapsing multiple reads into a single contig sequence, we used the phasing protocol described by Andermann et al. (2018) to obtain International Union of Pure and Applied Chemistry (IUPAC) consensus sequence alignments comparable to the PE-ddRAD alignments. Finally, we assembled datasets containing UCE loci that were present in at least 75% and 95% of the taxa using contig alignments (UCE 75 contig and UCE 95 contig) and using IUPAC consensus sequence alignments (UCE 75 IUPAC and UCE 95 IUPAC) (Figure 1).

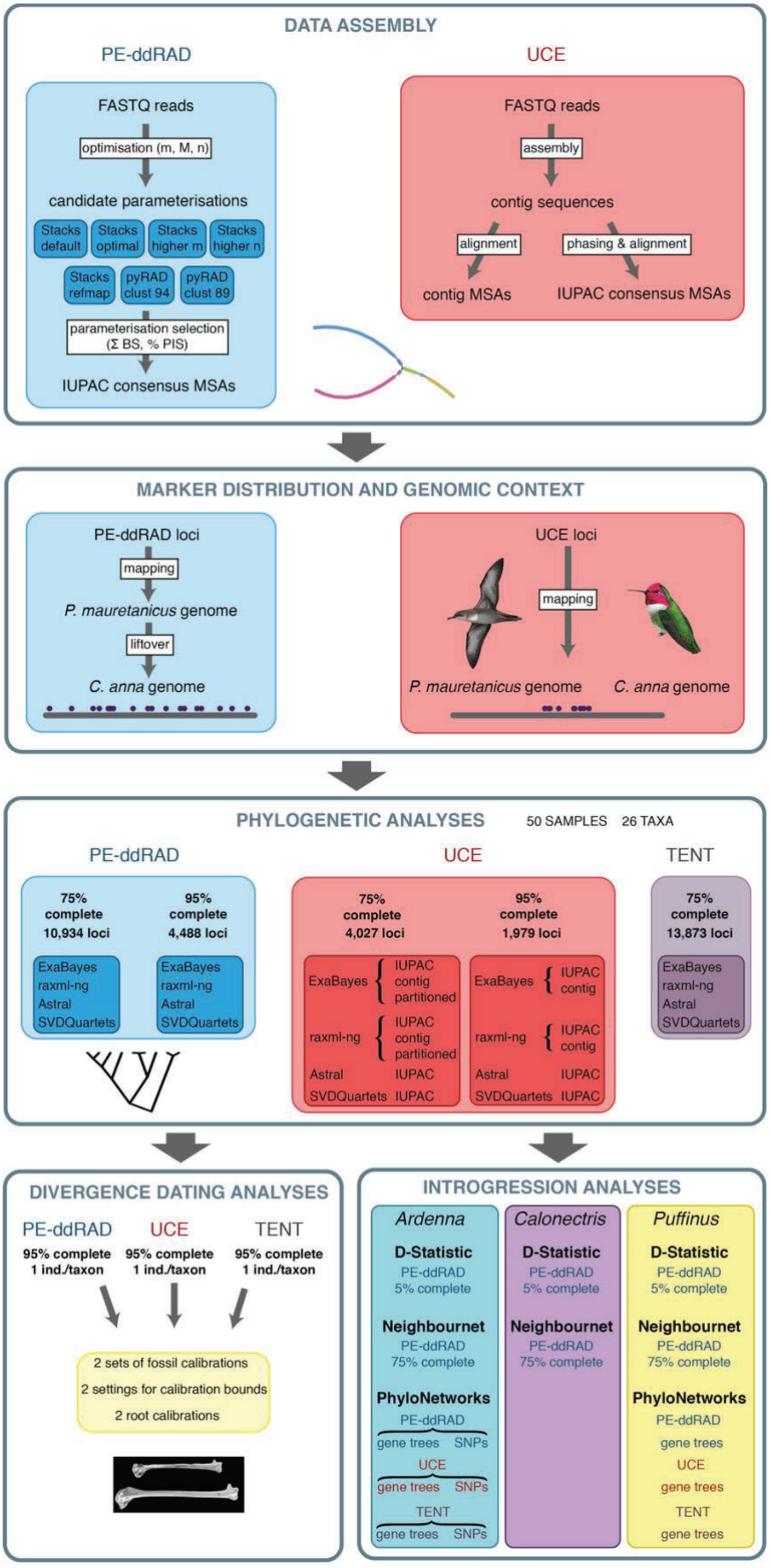


Figure 1 Graphic outline of the analyses carried out in this study. Balearic shearwater illustration by Martí Franch© and Anna's hummingbird illustration reproduced with permission from Lynx Edicions.

2.2.2 | PE-ddRAD-Seq dataset

We quality-filtered and demultiplexed reads using PROCESS_RADTAGS in STACKS v2.41 (Rochette et al. 2019). To obtain PE-ddRAD datasets optimised for phylogenomic analyses, we performed a two-step approach to assembling RAD loci (Figure 1). Here we provide an overview of this approach, which is described in detail in the Supplementary Information. First, we used a method for optimising *de novo* assembly of loci using STACKS (Paris et al. 2017; Rochette and Catchen 2017) which consists of varying each of the two key parameters (M: within-individual distance parameter, and n: between-individual distance parameter; Catchen et al. 2011) separately using DENOVO_MAP.PL, and selecting values of M and n under which the tendency of new polymorphic loci when increasing M or n becomes linear. We also assessed the frequency of putative sequencing errors in the data by selecting the value of m (stack-depth parameter) where the proportion of singletons in the dataset stabilised (Harvey et al. 2016). Secondly, we used two metrics to assess the effect of different pipelines and different parameterisations when assembling PE-ddRAD datasets: the number of parsimony informative sites (PIS) per locus relative to total locus length (pPIS); and the sum of bootstrap branch support (BS) obtained from maximum likelihood analyses (ML). For the second step, we used seven parameterisations from two assembly pipelines, STACKS and PYRAD v3.0.66 (Eaton 2014) (Figure 1 and Table S2). For all analyses, we extracted PE-ddRAD loci in silico from the outgroups using the python script Digital_RADs.py (DaCosta and Sorenson 2014).

We used PE-ddRAD IUPAC consensus sequence alignments from the selected parameterisation (STACKS higher m) to assemble two datasets containing loci present in at least 75% and 95% of the samples (PE-ddRAD 75 and PE-ddRAD 95) for downstream analyses.

2.2.3 | Total evidence dataset

To reduce the effect of data type on phylogenetic inference, we combined UCE and PE-ddRAD IUPAC consensus sequence alignments. We obtained a total evidence dataset with those UCE and PE-ddRAD loci present in at least 75% of the taxa (TENT 75). PE-ddRAD and UCE markers may overlap. To avoid including a sequence twice in our analyses, we used BLASTN (Altschul et al. 1997) to map representative PE-ddRAD loci

sequences to the UCE loci. We removed any PE-ddRAD loci that mapped to a UCE locus prior to concatenation.

2.3 | Marker Distribution and Genomic Context

Because the genomic distribution of phylogenomic markers can assist with understanding their informativeness across phylogenetic scales, we compared the distributions and genomic context of PE-ddRAD and UCE loci. We used BLASTN (Altschul et al. 1997) to map both sets of loci to the Balearic Shearwater (BaSh; *Puffinus mauretanicus*) draft genome assembly (Cuevas-Caballé et al. 2019) and to the most closely related chromosome-level genome assembly, the Anna's Hummingbird (AnHu; *Calypte anna*; (Korlach et al. 2017), which diverged between 62.7 to 71.1 Ma (Jarvis et al. 2015). Due to the large divergence time between shearwaters and AnHu, we also mapped the PE-ddRAD markers to the AnHu genome using a liftover approach (see Supplementary Information). To determine the level of clustering of UCE and PE-ddRAD loci, and to determine their degree of association with protein-coding genes, we applied a permutation procedure (Bioconductor package REGIONER; Gel et al. 2016), using 5000 permutations for each analysis (see Supplementary Information).

2.4 | Phylogenetic Analyses

We used unpartitioned UCE and PE-ddRAD concatenated datasets (UCE IUPAC 75, UCE IUPAC 95, PE-ddRAD 75, PE-ddRAD 95) and a total evidence concatenated dataset (TENT 75) partitioned by data type (UCE and PE-ddRAD) to estimate Bayesian and maximum-likelihood phylogenies using the MPI version of EXABAYES v.1.5 (Aberer et al. 2014) and RAXML-NG v.0.6.0 (Kozlov et al. 2019), respectively. Additionally, we estimated PE-ddRAD ML trees including the ingroup taxon *Puffinus assimilis haurakiensis*, for which no UCE data were available due to sampling constraints. For each dataset, we ran two independent EXABAYES runs with four coupled chains for 1,000,000 generations. We assessed runs for stationarity in TRACER v.1.7 (Rambaut et al. 2018) by checking for effective sample sizes > 300 for all model parameters. We created a consensus tree from the two independent runs using the CONSENSE programme from the EXABAYES package (burnin: 25%). We ran RAXML-NG with the GTR+G substitution

model to conduct 50 ML tree searches using 25 random and 25 parsimony-based starting trees. Following the best tree search, we generated 500 non-parametric bootstrap replicates. We checked for convergence post-hoc using the --bsconverge command in RAXML-NG with a cutoff value of 0.03; we computed branch support values and mapped the values onto the best-scoring ML tree using the RAXML-NG –support command.

We performed Bayesian and ML concatenated analyses using UCE unpartitioned contig alignments (UCE 75 contig and UCE 95 contig) to test the accuracy of the phylogenetic estimation of this approach (Figure 1). Concatenated unpartitioned analyses can converge to a tree other than the species tree (Roch and Steel 2015) and produce highly supported but incorrect nodes in the tree (Kainer and Lanfear 2015). To verify that our datasets were not affected by these issues, we also performed analyses of UCE partitioned IUPAC consensus alignments (75% complete). We used the Sliding-Window Site Characteristics (SWSC-EN) method described in Tagliacollo and Lanfear (2018) and PARTITIONFINDER2 (Lanfear et al. 2017) to partition the data, which yielded 131 partitions. These analyses are explained in detail in the Supplementary Information.

To account for coalescent stochasticity among individual loci, we inferred species trees with IUPAC UCE, PE-ddRAD and TENT datasets using two multispecies coalescent methods: the quartet-based method SVDQUARTETS (Chifman and Kubatko 2014), and the summary method ASTRAL-III (Zhang et al. 2018). SVDQUARTETS can handle both unlinked single nucleotide polymorphisms (SNPs) and multi-locus data. To allow a better comparison with analyses based on concatenation we ran SVDQUARTETS on the multi-locus sequence alignments. Analyses were run in PAUP* v.4 (Swofford 2002) evaluating all possible quartets. For each matrix, we conducted 100 bootstrap replicates, and results were summarized in a 50% majority-rule consensus tree. For ASTRAL-III, we used RAXML v.8 (Stamatakis 2014) to estimate gene trees for each PE-ddRAD and UCE locus in the IUPAC 75% complete datasets. We ran 500 rapid bootstrap replicates for each individual gene followed by a thorough ML search and we estimated species trees from the best-scoring ML gene trees and bootstrap replicates using ASTRAL-III. Two analyses were run for each dataset: one using the original gene trees and one using gene trees with very low support branches (BS < 10) contracted as this procedure can improve

tree accuracy (Zhang et al. 2018). Branch support values were inferred using local posterior probabilities (PP; Sayyari and Mirarab 2016). To avoid the negative impacts of fragmentary gene sequences on gene tree and species tree reconstruction (Sayyari et al. 2017), we removed three samples with mean missing data values per locus higher than 5% (*A. carneipes* 1, *A. grisea* 1 and *C. diomedea* 2) for ASTRAL-III analyses of UCE datasets. We annotated ASTRAL-III trees with local quartet supports for the main topology, and the first and second alternatives (ASTRAL-III option -t 2) to further investigate regions of the species tree that are potentially in the anomaly zone (Degnan and Rosenberg 2006). Finally, we performed a polytomy test (Sayyari and Mirarab 2018) to evaluate whether hard polytomies could be rejected at short internodes.

All phylogenetic analyses were conducted using only *F. glacialis* as an outgroup after checking that preliminary analyses using all outgroups yielded the same results (Figure S11). This decision was made to avoid long-branch attraction (Felsenstein 1978) and systematic error due to highly divergent outgroup taxa (Graham et al. 2002).

2.5 | Divergence Time Estimation

In divergence date estimation analyses, it has been common practice to reduce dataset size by selecting clock-like genes (Smith et al. 2018). Nonetheless, we decided to perform the analyses using the three 95% complete datasets (UCE, PE-ddRAD and TENT), because recent research has shown that divergence time analyses using complete phylogenomic datasets consistently show less variance in divergence times than estimates using subsets of clock-like genes (McGowen et al. 2019; Oliveros et al. 2019). We acknowledge that divergence time estimation methods based on concatenation may lead to branch-length bias and potentially misleading age estimates, particularly for younger divergence times (McCormack et al. 2011; Angelis and Dos Reis 2015). However, as our objective was not necessarily to calculate accurate estimates but rather to compare estimates based on different phylogenomic markers, we decided to use a concatenation-based method that allowed us to use complete phylogenomic datasets.

Divergence time analyses were performed using MCMCTREE v.4.9 from the PAML package (Yang 2007). MCMCTREE allows Bayesian divergence time inference of

phylogenomic datasets (dos Reis and Yang 2011) by implementing approximate likelihood calculation. We pruned the topology of our EXABAYES TENT 75% tree and the TENT 75 dataset so that they contained one individual per taxon to be used as input for divergence dating analyses, retaining the most complete individual. To decide the best-fitting clock model, we used the stepping-stones method (Xie et al. 2011) as implemented in the MCMC3R R package (dos Reis et al. 2018) to calculate marginal likelihoods for relaxed-clock models using the computationally expensive exact likelihood method because the approximate likelihood method cannot be used for marginal likelihood calculation (dos Reis et al. 2018). Thus, we carried out model selection on smaller subsets of the data suitable for exact likelihood calculation (two randomly selected subsets of 60 UCE loci and two randomly selected subsets of 120 PE-ddRAD loci, averaging 20,000 bp each). The marginal likelihoods were then used to calculate posterior probabilities for the strict, independent and autocorrelated rate models (ST, IR and AR, respectively).

We performed divergence dating analyses using: 1) two different fossil calibration strategies using 4 (A) and 3 (B) node calibrations (root calibration included in both strategies); 2) maximum and minimum soft bounds or only minimum bounds; and 3) two different maximum ages for the root calibration. For all analyses, maximum and minimum bounds were set for the root calibration. We provide detailed justifications for the four fossil calibrations used in the study (Marsh 1870; Miller 1961; Olson and Rasmussen 2001; Olson 2009) in the Supplementary Information.

We followed the two-step procedure outlined in dos Reis and Yang (2011) to infer divergence times using approximate likelihood calculation. For each analysis, we ran two independent Markov chain Monte Carlo (MCMC) chains, collecting 10,000 samples after a burn-in of 5,000 and a sample frequency of 500. We assessed likelihood convergence and parameters by examining trace plots in TRACER and checking that the estimated sample size (ESS) for each parameter was not smaller than 300, and by comparing results between independent runs. We also ran MCMCs with no data to generate joint prior distributions. Finally, we generated infinite-sites plots to assess how uncertainty in time estimates differed between analysis of the three datasets.

2.6 | GC-biased Gene Conversion

GC-biased gene conversion (gBGC) is known to strongly affect several features of avian genomes (Nabholz et al. 2011; Weber et al. 2014). To investigate potential signatures of GC-biased gene conversion (gBGC) in base composition and substitution rates in shearwaters, SNP data from the PE-ddRAD 75 dataset were output in VCF format using the populations program in STACKS. For biallelic variant sites, we computed reference and minor allele frequencies using VCFTOOLS v0.1.15 (Danecek et al. 2011). We assigned variant sites into one of the following mutation categories: strong-to-strong (S-to-S), strong-to-weak (S-to-W), weak-to-strong (W-to-S), and weak-to-weak (W-to-W), where C and G are strong bases (3 hydrogen bonds) and A and T are weak bases (2 hydrogen bonds). Although we recognize that the ancestral allele cannot be assigned with certainty (Keightley and Jackson 2018), due to the lack of outgroup sequences, we polarised variant sites based on their frequencies, with the minor allele being considered as derived. To investigate the role of gBGC on minor allele frequencies, we compared the distributions of minor allele frequencies between the different mutation categories and we explored the change in their prevalence depending on the local GC content.

To test expectations that gBGC is more effective in species with large population sizes and/or species with smaller body mass (Romiguier et al. 2010; Weber et al. 2014), we compared the number of breeding pairs and the average body mass to the overall proportion of W-to-S mutations per species. The number of breeding pairs per species and average body masses were retrieved from the Handbook of the Birds of the World (Carboneras and Bonan 2019).

2.7 | Introgression Analyses

2.7.1 | Split Networks

To better visualise patterns of genealogical discordance and potential areas of reticulate evolution, we computed phylogenetic networks for each genus using the Neighbour-Net approach (Bryant and Moulton 2004). Analyses were implemented in SPLITSTREE v.5 (Huson and Bryant 2006), using default parameters.

2.7.2 | Patterson's *D*-statistic (ABBA-BABA test)

To further explore whether tree discordances are due to past introgression or other forms of model misspecification, we quantified the Patterson's *D*-statistic (Green et al. 2010; Patterson et al. 2012) for all species quartets compatible with the time-calibrated topology. Calculations were performed using DSUITE DTRIOS (Malinsky et al. 2020) with a PE-ddRAD-derived SNP dataset that included loci with data in at least 5 taxa (ddRAD_min5, 295,779 SNPs). We performed analyses for the 3 genera separately. In each analysis, the sister species to the rest of the genus was used as the outgroup (i.e. *P. nativitatis*, *C. leucomelas*, and *A. bulleri* and *A. pacifica*). Block-jackknife resampling was used to evaluate significant deviations from zero in Patterson's *D*-statistic ($P < 0.001$). Significant results were validated using different outgroups and also using the PYRAD implementation of the statistic.

For those cases with a significant Patterson's *D*-statistic, we extracted SNPs with strong signatures of introgression (i.e. SNPs with ABBA configuration and fixed within species, hereafter 'ABBA SNPs') to evaluate alternative potential causes of these signatures, such as shared ancestral variation, mutational hotspots resulting in convergent mutations, gBGC, and levels of genetic variation (see Supplementary Information).

2.7.3 | Phylogenetic Network Analyses

We reconstructed phylogenetic networks using the maximum pseudolikelihood method implemented in SNAQ (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). This method accommodates ILS and gene flow under the multispecies network coalescent model (MSNC). We ran phylogenetic networks for each of the datasets (PE-ddRAD, UCE and TENT) and independently for *Puffinus* and *Ardenna*, to reduce computation time and to improve the accuracy of the inferred networks (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). In both cases, we used the time-calibrated topology, pruned to only contain taxa in the genus under study, as a starting topology. For each dataset, we conducted 10 independent runs with random seeds of SNAQ to infer the optimal coalescent tree with no hybridisation edges (h0). We then performed network searches from 1 to 4 (1 to 2 for *Ardenna* datasets) hybridisation edges providing, in each case, the optimal network with hmax-1 hybridisation edges. The preferred number of

hybridisations was selected based on the analysis of the slope of a plot of log-pseudolikelihood against the number of hybridisations (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). We expected a sharp improvement until the number of hybridisation edges reached the best value. We did not evaluate more than four (or two in the case of *Ardenna*) hybridisation edges because of the lack of change in the slope heuristic.

To assess whether the MSC adequately explained gene-tree discordance to our coalescent trees with no hybridisation edges (h0), we used the TICR test (Tree Incongruence Checking in R; Stenz et al. 2015), using the PHYLOLM R package (Tung Ho and Ané 2014). A chi-squared test was used to compare observed concordance factors (CF) with expected CF calculated from the h0 species trees under the MSC. We further looked for taxa that did not fit the tree model with ILS retrieving the outlier 4-taxon sets.

Because searches for an optimal network produced inconsistent results in independent runs and in different datasets, we focused on evaluating candidate reticulation events based on the results of the Patterson's *D*-statistics. We optimised the pseudo-deviance of candidate networks using the TOPOLOGYMAXQPSEUDOLIK! function and we visualized the estimated inheritance probabilities (γ).

3 | Results

3.1 | Data Assembly

We recovered a mean number of PE-ddRAD and UCE reads of 1,273,325 (SD = 876,938) and 1,851,330 (SD = 517,189) per sample, respectively (Table S1). We assembled an average of 25,716 PE-ddRAD tags per sample; the alignment lengths per locus ranged from 140 to 239 bp with a median of 198 bp (SD = 25.5). UCE TRINITY contigs ranged from 213 to 1,565 bp with a median length of 551 bp (SD = 126.6) and an average recovery of 83.6%.

Correlations between different UCE summary statistics showed that low sequencing yield not only resulted in low sequencing coverage but also in a lower number of

assembled loci, which were shorter on average, and in a lower percentage of sequencing on target (see Supplementary Information and Figure S1).

Our results for the *de novo* optimisation of PE-ddRAD are described in detail in the Supplementary Information. Briefly, the tendency of new polymorphic loci when increasing M or n parameters in STACKS, became linear at M=5 and n=8 (Figure S2) and the proportion of singletons in the dataset stabilised at m=7 (Figure S3). For the second optimisation step, different parameterisations had a minor effect on the pPIS with the exception of the default parameters in STACKS that yielded loci with a much lower number of PIS. With similar parameterisations, STACKS yielded approximately twice the number of loci than PYRAD for each level of missing data, but PYRAD loci had a slightly higher pPIS (Figure S4a and Table S2). Phylogenetic analyses using the different datasets and levels of missing data resulted in overall highly resolved and congruent phylogenies. However, the general trend was a slight increase in resolution when reducing missing data from a maximum of 35% to a maximum of 25% and thereafter, a slight decrease when reducing it to a maximum of 5% (Figure S4b). Phylogenetic analyses using datasets with a higher amount of missing data (35% and 25%) tended to yield higher bootstrap supports on recent splits, whereas analyses using datasets with a low level of missing data (5%) yielded higher bootstrap supports on more ancient splits (Figure S5). PYRAD datasets with a maximum of 25% missing data yielded the highest overall resolution but STACKS higher m datasets were the most consistent across different levels of missing data. Thus, we selected the latter parameterisation for downstream analyses.

For the same taxon coverage, PE-ddRAD concatenated alignments were both longer (i.e. 2,156,937 bp for the PE-ddRAD 75 matrix vs. 1,732,076 bp for the UCE 75 matrix) and contained more than double the number of loci than UCE alignments (Table 1). The number of PIS per locus was very similar between PE-ddRAD and UCE alignments despite the much shorter length of PE-ddRAD loci.

Table 1 Characteristics of assembled datasets used in phylogenetic analyses.

Method	Taxon coverage	Number of loci	Median locus length (SD)	Number of PIS	Median PIS per locus (SD)
PE-ddRAD	75%	10,934	198 (25.51)	85,070	7 (4.19)
PE-ddRAD	95%	4,488	205 (19.47)	34,505	7 (4.18)
UCE	75%	4,027	452 (121.06)	31,664	6 (7.13)
UCE	95%	1,979	484 (109.75)	13,900	6 (6.45)

3.2 | Marker Distribution and Genomic Context

Using BLASTN, 97.7% of the UCE and 95.4% of the PE-ddRAD loci successfully mapped to the *P. mauretanicus* draft genome assembly. When mapped to the more distant *C. anna* chromosome-level genome assembly, an even higher percentage of UCE loci successfully mapped (99.4%) which contrasted with the low percentage of successfully mapped PE-ddRAD loci (30.6%). Using the liftover approach, we managed to improve the percentage of successfully mapped PE-ddRAD loci to the *C. anna* genome assembly to 77.5%. UCE loci had a higher level of clustering than PE-ddRAD loci (median distance between the closest UCE loci = 16.2 kbp, median distance between the closest PE-ddRAD loci = 51.3 kbp; Figure 2, Figure S6, S7 and S8). PE-ddRAD loci were closer to protein-coding genes (51.4 ± 94.0 kbp) than UCEs (73.8 ± 104.8 kbp) (Figure S9 and S10).

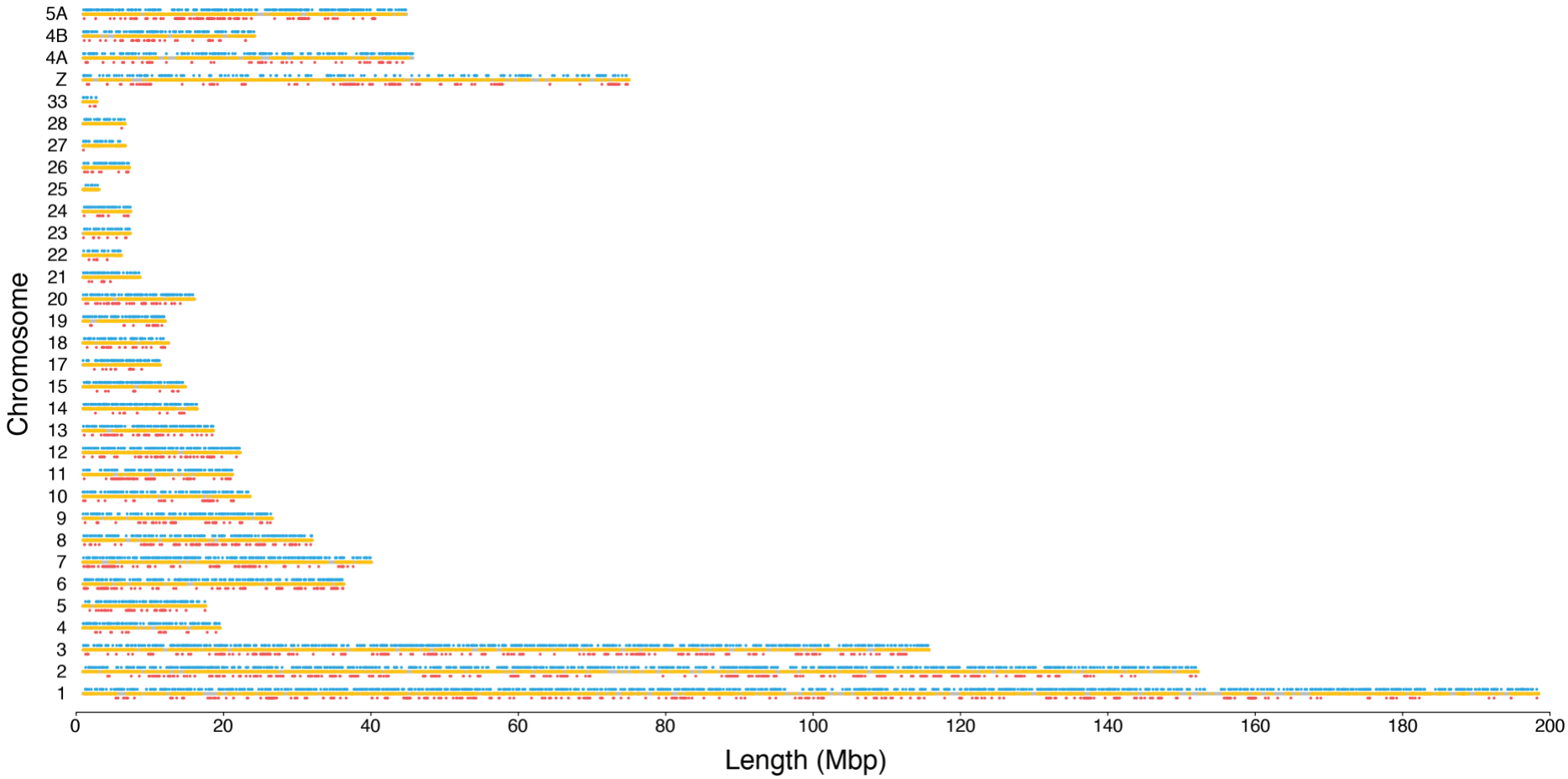


Figure 2 Genomic distribution of PE-ddRAD (blue) and ultraconserved elements (red) when mapped to the chromosome-level genome assembly for *Calypte anna*. Chromosomes are represented as grey lines and the yellow spots on the chromosomes represent the location of protein-coding genes.

3.3 | Phylogenomic Analyses

Phylogenetic analyses recovered largely the same well-resolved tree topology across different genomic markers, levels of missing data, and phylogenetic methods. The phylogenomic tree resulting from the TENT 75 EXABAYES analysis is shown in Figure 3a and results of all phylogenetic analyses performed in this study are detailed in Table S3. All analyses supported the monophyly of the three recognized genera of shearwaters: *Ardenna*, *Calonectris*, and *Puffinus*. *Ardenna* and *Calonectris* were sister genera and together were the sister lineage to the species-rich *Puffinus*. Within *Puffinus*, we recovered *P. nativitatis* as the sister taxon to the remaining *Puffinus* species, which formed five strongly supported and biogeographically defined clades: a clade from New Zealand and Australian waters (*P. gavia* and *P. huttoni*); a Subantarctic and New Zealand clade (*P. elegans* and *P. assimilis haurakiensis*); a North Pacific clade (*P. newelli* and *P. opisthomelas*); a Tropical Indian and South Pacific clade (*P. bailloni*); and a Caribbean, North Atlantic and Mediterranean clade (*P. puffinus*, *P. mauretanicus*, *P. yelkouan*, *P. lherminieri*, *P. boydi* and *P. baroli*) (Figure 3a and Figure S12).

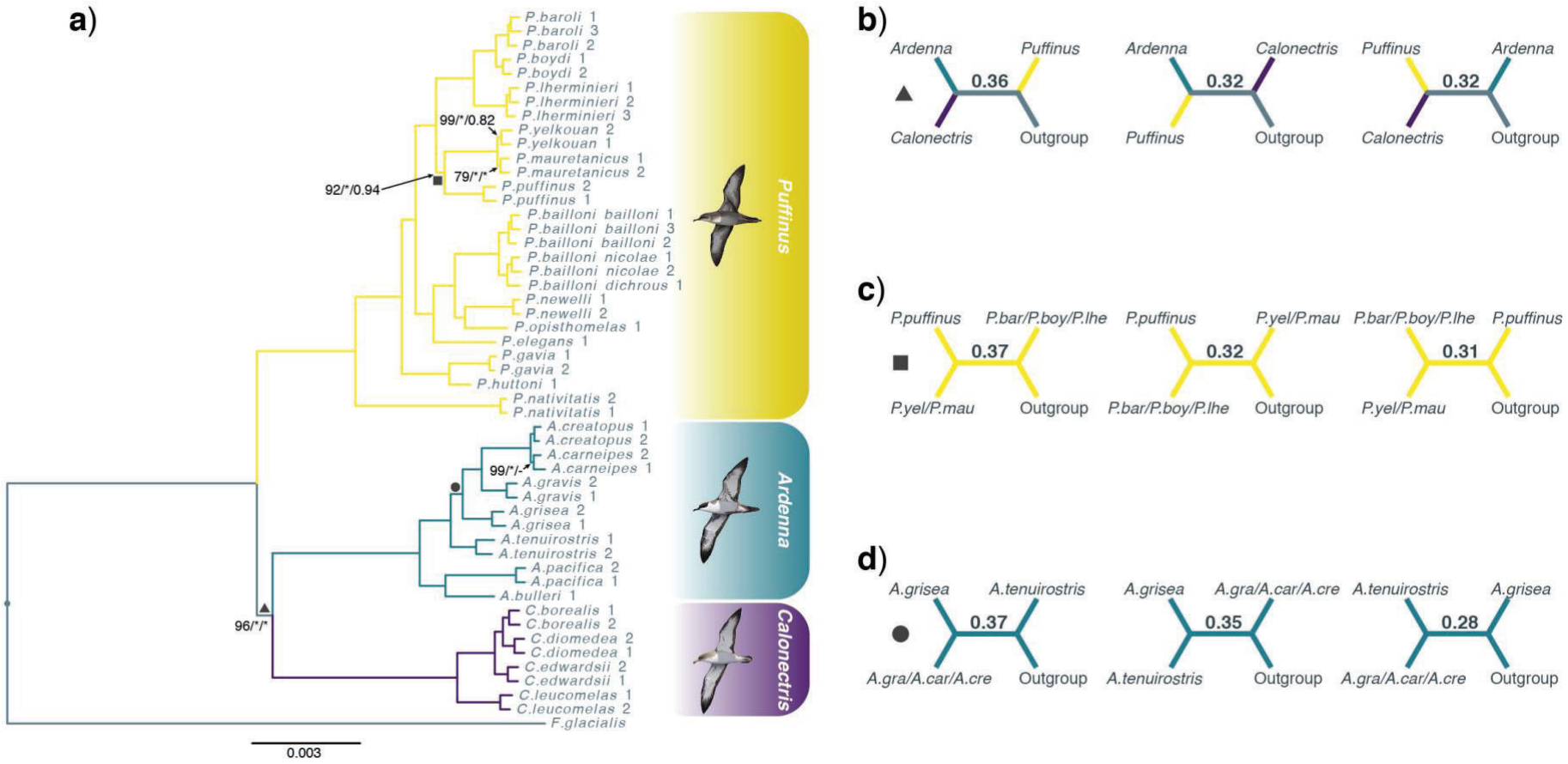


Figure 3 a) Phylogram of the concatenated bayesian tree inferred in STACKS using the total evidence (TENT) dataset: a 75% complete matrix of ultraconserved elements (UCE) and paired-end double-digest Restriction site-Associated DNA (PE-ddRAD) loci. All nodes have 100% bootstrap support values (BS; RAXML-NG) and 1.0 posterior probabilities (PP; EXABAYES and ASTRAL-III) unless labelled otherwise. Labels correspond to RAXML-NG BS / EXABAYES PP / ASTRAL-III PP with asterisks indicating full support and hyphens indicating nodes not recovered. The three nodes represented by different shapes resulted in incongruence in analyses using different genomic markers, levels of missing data and/or phylogenetic methods. Quartet supports for these nodes from the ASTRAL-III analysis using the same dataset are shown in b) to d) for the main and the two alternative quartet topologies. Illustrations by Martí Franch© represent the three shearwater genera.

Only three phylogenetic relationships, all localised to short internodes, were challenging to resolve due to discordances between analyses using different genomic markers, levels of missing data, or phylogenetic methods (Figure 3b-d). The first discordance was the relationship among the three shearwater genera (triangle). All analyses using the TENT dataset and most analyses using the UCE and the PE-ddRAD datasets recovered *Calonectris* as sister to *Ardenna* (Figure 4). However, RAXML-NG and ASTRAL-III trees based on the PE-ddRAD 75 dataset and the SVDQUARTETS tree based on the UCE 95 dataset recovered the alternative topology (*Ardenna* as sister to *Puffinus*). A polytomy test based on local quartet supports (Figure 3b) using the TENT dataset marginally ruled out that this branch should be replaced by a polytomy (P = 0.0324). Under a true polytomy we would expect local quartet supports for the three alternative topologies to be equal to 1/3. This test was also significant when using only UCE data but failed to discard the likelihood of a polytomy when using PE-ddRAD data (Table 2). The second discordance was the relationship between the three North Atlantic lineages of *Puffinus* (Figure 3c). In this case, all phylogenetic analyses recovered *P. puffinus* as the sister species to the Mediterranean *P. mauretanicus* and *P. yelkouan*, with the exception of the PE-ddRAD 95 EXABAYES tree, which recovered *P. puffinus* as the sister to *P. lherminieri*, *P. baroli* and *P. boydi*. Despite consistency in the recovered relationships for this case, only the polytomy tests based on the PE-ddRAD 75 and TENT 75 datasets were able to reject the null hypothesis that the branch should be replaced by a polytomy. The last case was the relationship between the two all dark species of *Ardenna* (*A. tenuirostris* and *A. grisea*) (Figure 3d). The only analyses that recovered *A. grisea* and *A. tenuirostris* as sister species used coalescent-based methods and UCE datasets (Figure 4), although it should be noted that ASTRAL-III analyses using UCEs were performed with only one *A. grisea* individual due to missing data filtering (See Phylogenomic Analyses in Materials and Methods). Interestingly, all tests rejected a polytomy although UCE datasets supported one topology and PE-ddRAD datasets another (Table 2), showing different phylogenetic signals between both types of markers. It is also noteworthy that the monophyly of the most recently diverged taxa *P. mauretanicus* - *P. yelkouan* and *A. creatopus* - *A. carneipes* was not supported in several phylogenetic analyses (Table S3).

Table 2 ASTRAL-III quartet supports for the main and the two alternative quartet topologies and polytomy test p-values for the challenging branches highlighted in Figure 2. The null hypothesis is polytomy and p-values < 0.05 reject a real polytomy. Polytomy test p-values are shown in parentheses when the alternative topology was recovered. Note that only the TENT dataset is able to reject real polytomies at the three challenging branches.

Branch	Value	UCE		ddRAD		TENT
		75%	95%	75%	95%	75%
▲	Polytomy test p-value	0.008	0.0103	- (0.2606)	0.2567	0.0324
	QS: (*Ardenna*,*Calonectris*)	0.39	0.4	0.34	0.35	0.36
	QS: (*Calonectris*,*Puffinus*)	0.32	0.32	0.31	0.31	0.32
	QS: (*Ardenna*,*Puffinus*)	0.29	0.28	0.35	0.34	0.32
■	Polytomy test p-value	0.1895	0.0335	0.0026	0.8022	0.001
	QS: (*P.puffinus*,(*P.mauretanicus*,*P.yelkouan*))	0.37	0.4	0.37	0.34	0.37
	QS: (*P.puffinus*,(*P.lherminieri*,*P.boydi*,*P.baroli*))	0.32	0.31	0.32	0.33	0.32
	QS: ((*P.mauretanicus*,*P.yelkouan*),(*P.lherminieri*,*P.boydi*,*P.baroli*))	0.31	0.29	0.31	0.33	0.31
●	Polytomy test p-value	- (0)	- (0.0016)	0	0	0
	QS: (*A.grisea*,(*A.gravis*,*A.creatopus*,*A.carneipes*))	0.38	0.37	0.38	0.38	0.37
	QS: (*A.grisea*,*A.tenuirostris*)	0.38	0.39	0.35	0.36	0.35
	QS: (*A.tenuirostris*,(*A.gravis*,*A.creatopus*,*A.carneipes*))	0.24	0.24	0.27	0.26	0.28

Figure 4 Heatmap showing support for the challenging branches highlighted in Figure 2 and the number of non-monophyletic species recovered in analyses using different genomic markers, levels of missing data and phylogenetic methods. Every row corresponds to a different analysis with a particular dataset and every column corresponds to one of the challenging branches highlighted in Figure 3, with the exception of the last column that corresponds to the number of non-monophyletic species recovered. Values are shown in each tile and tiles are coloured corresponding to the legend.

3.4 | Divergence Dating Analysis

For all sampled alignments except for one, the IR model had the highest posterior probability (Table S4). This model was interpreted as the best-fitting model and used in downstream MCMCTREE analyses. Mean posterior time estimates using PE-ddRAD, UCE or TENT datasets differed only marginally (Figure 5 and Figure S13). Analyses using the TENT dataset with two partitions produced posterior estimates with the narrowest credibility intervals while analyses using the UCE dataset produced the largest credibility intervals. The slope of the regression line in the infinite-sites plot was highest for UCE data (0.51), was slightly lower for PE-ddRAD data (0.49), and dropped to 0.43 when using the TENT dataset, meaning that 0.43 Ma of uncertainty was added to the

95% CI for every 1 Ma of divergence (Figure S14). Except for the root, the points of the infinite-sites plot from the combined dataset formed a straight line, indicating that the relatively high uncertainty in time estimates was mostly due to uncertainties in fossil calibrations (Rannala and Yang 2007).

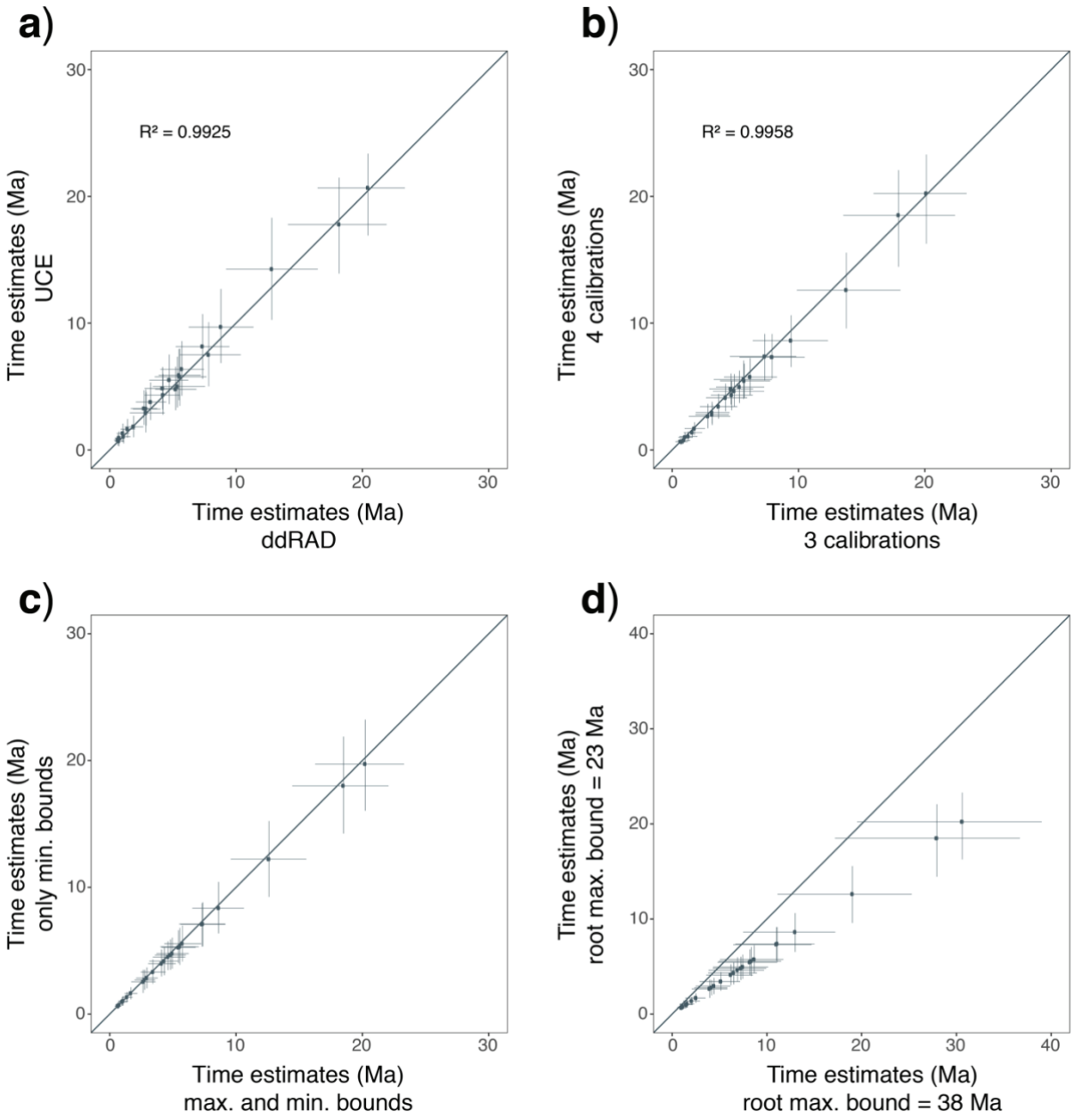


Figure 5 Effects of data type, calibration strategy, bound constraints and root maximum bounds. Posterior time estimates (points) and 95% credibility intervals (lines) using a) UCE versus PE-ddRAD data, b) four versus three calibration points, c) only minimum bounds versus minimum and maximum bounds and d) setting the root maximum bound at 38 Ma versus 23 Ma.

Setting the root maximum bound to 38 Ma had the strongest impact on the divergence time estimates and resulted in more ancient estimates, with larger differences observed near the root. When we used constraints on minimum bounds, posterior estimates tended to be more recent and credibility intervals larger. Despite the fact that using four calibrations resulted in a slight truncation of the prior density on the age of the *Calonectris - Ardenna* node, mean posterior time estimates were nearly identical and variances were higher when only using three calibrations (Figure 5 and Figure S13).

3.5 | GC-biased Gene Conversion

Analyses of the relative site-frequency spectrum (SFS) for each mutation class showed a shift in the relative proportion of putative W-to-S and S-to-W mutations (Figure S15a). Putative W-to-S mutations were skewed towards high frequencies, and S-to-W mutations towards low frequencies, consistent with a prevalent gBGC-driven fixation bias (Bolívar et al. 2016). In addition, we observed that putative W-to-S mutations tended to maintain higher frequencies of the minor allele than S-to-W mutations, particularly at GC-rich areas (Figure S15b). In shearwaters, in contrast with the general trend in birds, species with smaller body mass (genus *Puffinus*) have smaller census sizes and are expected to have smaller effective population sizes, which should increase both the number of meioses per unit time and the efficacy of gBGC (Romiguier et al. 2010; Weber et al. 2014). Concordant with these expectations, we observed strong positive correlations between the overall proportion of putative W-to-S mutations and both the number of breeding pairs (R^2 = 0.552 and P = 1.5 x 10-9) and the average body mass per taxon (R^2 = 0.731 and P = 1.1 x 10-14; Figure S16).

3.6 | Introgression Analyses

D-statistic tests found clear evidence for an excess of shared derived alleles consistent with introgression between *A. grisea* and *A. tenuirostris* (*D*-statistic = 0.1193) and even stronger evidence between *P. boydi* and *P. lherminieri* (*D*-statistic = 0.2538) (Figure 6a). For these two potential cases of introgression, shared ancestral variation accounted for 20-31% of shared derived alleles in SNPs with ABBA pattern (Table S5). Loci with ABBA

SNPs were not significantly more variable than average loci, although they were in the upper part of the distribution for the *P. boydi* - *P. lherminieri* case. In the *A. tenuirostris* - *A. grisea* case, we found a significantly higher proportion of ABBA patterns generated by putative W-to-S mutations than expected by chance (p-value=0.0257), which might indicate a role of gBGC in generating these patterns. Finally, we observed that potentially introgressed species had the highest individual heterozygosities (Figure S17).

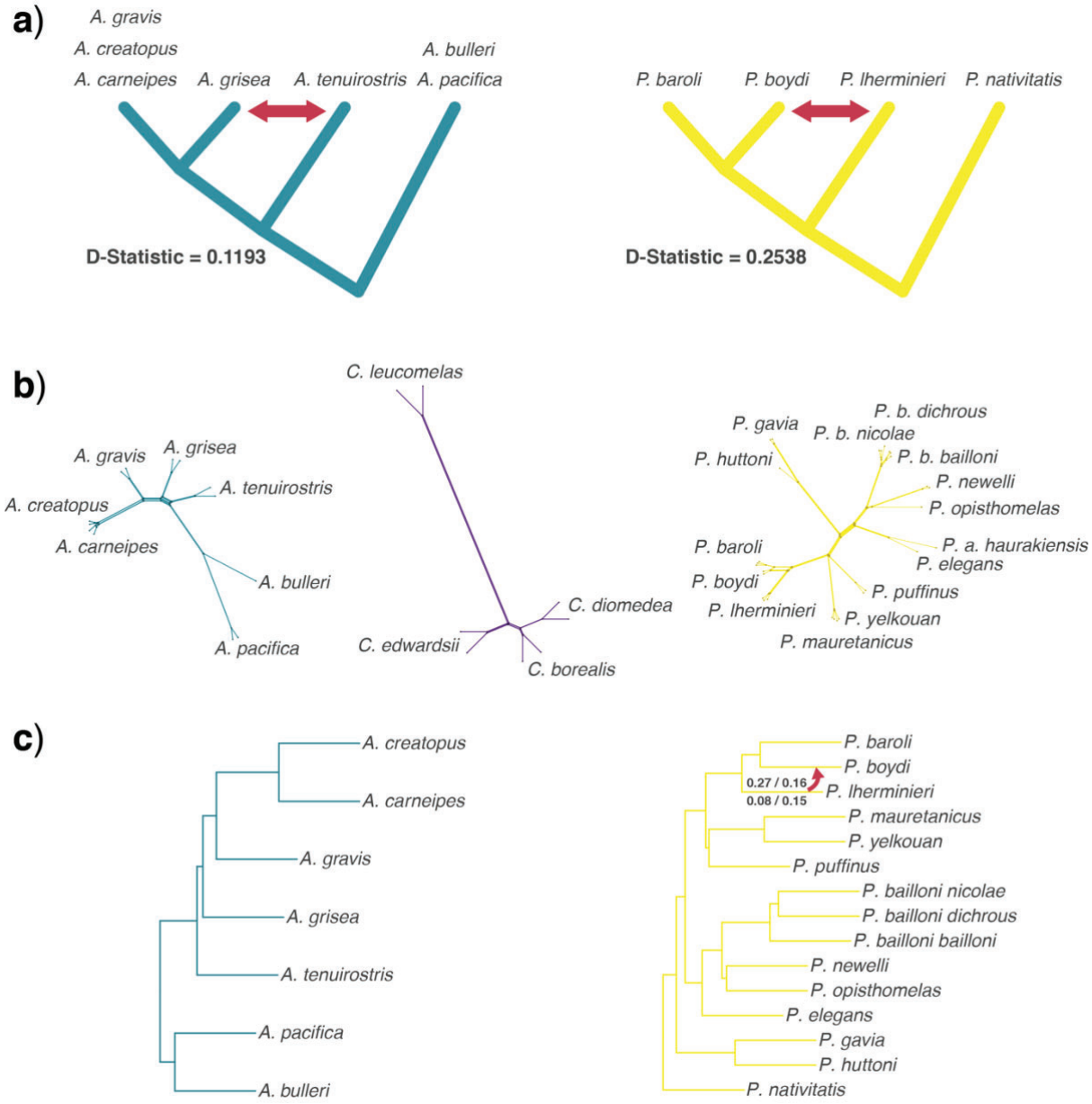


Figure 6 Introgression analyses in shearwaters. a) Gene-flow hypotheses obtained from *D*-statistic analyses (significant values) are shown with red arrows. Mean *D*-statistic values for the two cases are also shown. b) Neighbour-net networks for the three shearwater genera. c) Maximum pseudolikelihood SNAQ networks for *Ardenna* (h = 0) and *Puffinus* (h = 1). For the inferred hybridisation event in *Puffinus*, optimised inheritance probabilities for the minor hybrid edge (γ) using PE-ddRAD / UCE gene trees (above) and PE-ddRAD SNPs / UCE SNPs (below) are shown.

Neighbour-net networks for each genus showed low levels of reticulation and were consistent with concatenated and coalescent-based phylogenetic analyses (Figure 6b). However, we observed reticulation in areas where *D*-statistics showed evidence for an excess of shared derived alleles between non-sister taxa.

The TICR test detected a significant excess of outlier quartets in the data for *Puffinus*, when using PE-ddRAD or UCE gene tree data, suggesting that the coalescent tree inferred without introgression did not adequately fit the data. We recovered a deficit of high concordance factors (CF) and an excess of low CF (Figure S18). Nonetheless, the observed values at the extremes tended towards the expected values when increasing the BS threshold for collapsing branches in the gene trees. When branches with BS < 50 were collapsed, the TICR test was no longer significant, suggesting that the excess of outlier quartets was caused by including noise in form of inaccurate gene tree branches. Using a program like BUCKY (Larget et al. 2010) to calculate the CF from gene trees can provide an advantage, since it also considers uncertainty in gene tree estimation. It is noteworthy that in all PE-ddRAD datasets (but not in UCE datasets), *P. boydi* and *P. lherminieri* were found more frequently in 4-taxon sets that did not fit the tree model with ILS compared to any other taxon. The TICR test for *Ardenna* was not significant and all 4-taxon sets fitted the tree model with ILS.

Consistent with the TICR test results, we detected a lack of sharp improvement in the slope heuristic in both *Puffinus* and *Ardenna* datasets. Nonetheless, we detected a general sharper decrease up to h=2 in *Puffinus*, although inferred reticulation events had $\gamma < 5\%$. The only introgression event with $\gamma > 5\%$ was from *P. lherminieri* to *P. boydi*. For candidate networks assuming introgression between *P. lherminieri* and *P. boydi*, log pseudo-likelihood values obtained from PE-ddRAD data were either lower than or similar to those obtained in optimised networks with h=1, showing support for introgression between these taxa (Table S6). The directionality of introgression could not be determined reliably because optimisation of candidate networks showed very similar log pseudo-likelihood values for pairs of candidate networks with reversed direction of introgression. However, inheritance probabilities were much higher when assuming introgression from *P. lherminieri* to *P. boydi* than when assuming the opposite (Figure 6c and Table S6).

4 | Discussion

Our results demonstrate the power of integrating UCE and RAD markers for resolving the phylogenetic relationships of a group of pelagic seabirds characterised by rapid diversification events that have confounded previous phylogenetic studies. To our knowledge, our study is the first to compare and integrate UCE and paired-end ddRAD datasets in a phylogenomic context using comparably phased sequence alignments for both datasets. Here, we propose a strategy to optimise RAD-Seq data for phylogenetic analyses, we consider aspects of our methodological approach that may be of help to future studies, we discuss case-by-case how the integrative use of PE-ddRAD-Seq and UCE data with phylogenetic and introgression analyses allows identification of the causes of phylogenetic discordance, and we discuss the systematic implications of our phylogenetic results.

4.1 | RAD-Seq Dataset Optimisation for Phylogenetic Analyses

There are two main factors that affect the number of orthologous loci recovered in RAD-seq datasets: 1) divergence times between lineages and 2) filtering and assembly parameters applied to orthology inference. Considerable research attention has focussed on exploring how the number of orthologous loci decreases with increasing divergence times (Rubin et al. 2012; Cariou et al. 2013) and how filters based on taxon coverage affect dataset size and the ability to resolve phylogenetic relationships (Wagner et al. 2013; Leaché et al. 2015; Díaz-Arce et al. 2016; Tripp et al. 2017). However, phylogenomic studies have generally neglected the optimisation of assembly parameters in order to minimise the inclusion of paralogous or repetitive loci (but see Hosegood et al. 2020), a procedure that is common practice in population genomics analyses (Paris et al. 2017; Rochette and Catchen 2017). To fill this gap, we used a two-step optimisation process to assess the impact of common issues in RAD-Seq, such as sequencing error and paralog content, on phylogenetic reconstruction. However, our analyses revealed a minor effect of assembly parameters on phylogenetic reconstruction, compared to taxon coverage. Our study adds to previous evidence showing a major effect of taxon coverage on phylogenetic reconstruction when using

RAD-Seq datasets (Díaz-Arce et al. 2016; Tripp et al. 2017). We recommend that phylogenetic studies using RAD-Seq data should explore several taxon coverage filters in order to maximise the phylogenetic informativeness of their datasets.

4.2 | Considerations on Methodological Approaches for Phylogenetic Inference

As expected, differences in methodological approaches can severely affect the phylogenetic inference. First, in cases of extreme levels of ILS, concatenation analyses may result in an average topology that differs from the true species tree (Mendes and Hahn 2018). On the other hand, low phylogenetic information per locus in PE-ddRAD and UCE datasets might result in poorly resolved gene trees, and in such cases, concatenation methods can be more accurate than summary coalescent approaches (Mirarab et al. 2016; Springer and Gatesy 2016). In our case, concatenation and summary coalescent approaches produced largely congruent topologies that expressed a lower degree of incongruence than analyses using different datasets and different levels of missing data. Nevertheless, concatenation exacerbated systematic error in two cases, leading to high confidence in alternative relationships in areas of elevated ILS (Figure 4).

Second, Andermann et al. (2018) found that using IUPAC consensus sequences performed better than using contig sequences for estimating the tree topology of a recently diverged group of *Topaza* hummingbirds under the MSC model. Our concatenation analyses using UCE contig alignments produced topologies nearly identical to those from analyses using the IUPAC consensus sequence alignments. However, consistent with Andermann et al. (2018), we observed a tendency towards a poorer performance of contig sequences at recovering the monophyly of species and subspecies (Table S3). We evidenced that using IUPAC sequence alignments resulted in a reduction of mean locus length (~100 bp shorter). This was due to the inability of accurate phasing towards the extremes of UCEs because of lower read coverages (we required a minimum of 5 reads per haplotype to include a position). This approach is more conservative than using the full contig sequence and therefore results in a reduction of the number of PIS. However, it delivers a higher reliability in base calling

and a true representation of polymorphism that can be particularly useful at recent timescales. Researchers working with UCE data face a trade-off between longer alignments with higher amounts of PIS and more reliable base calling allowing the inclusion of polymorphism data.

Third, partitioning strategy can affect the outcome of phylogenetic analyses. For example, maximum likelihood, when used to analyse an unpartitioned concatenated alignment from different loci, can converge to a tree other than the species tree as the number of loci increases (Roch and Steel 2015). Our concatenation analyses using the partitioned UCE 75 dataset recovered exactly the same topology compared to unpartitioned analyses, with only slight changes in branch support, showing that the topology was relatively insensitive to the partitioning scheme used. This was likely due to a smaller effect of evolutionary rate heterogeneity among loci at shallow phylogenetic scales like the one we were working than at deep timescales.

Fourth, increasing the number of individuals strongly improves species tree estimation in ASTRAL-III when branch lengths are extremely short (Rabiee et al. 2019). We found evidence for this pattern in our UCE ASTRAL-III trees, where we found incongruences affecting nodes where individuals had been removed to avoid the inclusion of fragmentary sequences. We also observed an improvement in precision when analyses were performed after contracting very low support branches (BS < 10) in the gene trees, as previously suggested by (Zhang et al. 2017) (Table S3).

4.3 | Divergence Dating with UCE and PE-ddRAD

Collins and Hrbek (2018) showed consistent discrepancies in divergence time estimates using different phylogenomic markers under strict and relaxed clock models. The discrepancies were explained by temporal differences in phylogenetic informativeness (PI). However, their analyses showed that UCE and ddRAD datasets showed similar temporal patterns of PI and resulted in similar divergence time estimates. Our analyses using UCE and PE-ddRAD empirical datasets also showed very similar estimates. However, UCE estimates tended to be slightly older and have wider credibility intervals than PE-ddRAD estimates probably due to UCE datasets having lower PI at the timescales covered in this study.

Theoretically, for infinitely long alignments, an infinite-sites plot, which measures the uncertainty in the divergence time posterior (width of the credibility interval vs. posterior mean of node ages) should converge onto a straight line (Rannala and Yang 2007). The slope of this line represents the amount of uncertainty in time estimates per 1 million years (Myr) of divergence solely due to uncertainties in the fossil calibrations. The points of the infinite-sites plot from the UCE and the PE-ddRAD datasets did not form a straight line, suggesting that uncertainties in time estimates were due both to limited data as well as uncertainties in the fossil calibrations. On the other hand, the points from the TENT dataset formed a straight line (Figure S14), indicating that the relatively high uncertainty in time estimates was mostly due to uncertainties in fossil calibrations and that increasing the amount of data would only marginally reduce the credibility intervals (Rannala and Yang 2007). We thus show that combining both datasets resulted in lower uncertainties in divergence time estimates.

4.4 | Integrative Approach using UCE and PE-ddRAD to Disentangle Phylogenetic Discordance

The utility of RAD-seq and target capture approaches for phylogenetic estimation across different timescales has been widely demonstrated (Faircloth et al. 2012; Cruaud et al. 2014; Smith et al. 2014; McCluskey and Postlethwait 2015). However, only a handful of studies have explored the utility of both approaches in phylogenetic studies (Leaché et al. 2015; Harvey et al. 2016; Manthey et al. 2016; Collins and Hrbek 2018). Using a higher number of taxa and loci than these previous studies, we show the advantages of integrating PE-ddRAD-Seq and UCE data to infer the phylogenetic relationships of a challenging group, the shearwaters, across a range of timescales using concatenation and coalescent approaches.

Despite finding only minor data-type effects, datasets from different markers and levels of missing data tended to better resolve short internodes at different timescales (Figure 4 and Figure S5). For instance, the UCE 95 dataset, which contained the lowest number of PIS, showed the highest support across methods (with the exception of SVDQUARTETS) for the sister relationship between *Ardenna* and *Calonectris* (18.5 Ma), but showed the poorest performance at recovering the monophyly of recently diverged taxa

(Figure 4). This shows that the phylogenetic signal of this dataset is stronger near the root and weaker at shallow timescales. Conversely, the dataset with the highest number of PIS (PE-ddRAD 75) recovered alternative topologies (with low support) for the relationships between the three genera, but recovered the monophyly of all the species (and subspecies). Finally, phylogenetic analyses using the TENT 75 matrix performed well at both deep and shallow timescales, showing the advantages of combining different data types in a single analysis. Our study is consistent with previous findings that combining data types leads to higher resolution on short internodes (Jarvis et al. 2014), and it is noteworthy that, due to high rate heterogeneity between PE-ddRAD loci and UCEs, the observed data-type effects could also reflect poor model fit (Reddy et al. 2017).

Despite the strong support in most of our phylogeny, we found phylogenomic conflict associated with short internodes in three areas of the tree (Degnan and Rosenberg 2006). Short internode lengths are usually associated with topological conflict. This conflict arises because short speciation intervals 1) accumulate few substitutions, resulting in a low number of informative sites and, 2) they increase the probability of finding different gene histories due to ILS (Alda et al. 2019).

The first area of conflict was the short internode near the root separating *Ardenna* and *Calonectris* from *Puffinus* (triangle in Figure 3). When using the PE-ddRAD dataset with the highest number of PIS, a sister relationship between *Ardenna* and *Puffinus* received the highest quartet support in ASTRAL-III analyses. The remaining datasets provided the highest quartet support for a sister relationship between *Ardenna* and *Calonectris* and the support for this arrangement decreased when increasing the number of PIS in the dataset (Table 1 and 2). These results show how high levels of ILS resulted in different phylogenetic signals across datasets and could indicate that our most variable dataset is affected by homoplasy due to saturation at this timescale. In addition, the increased support for the *Ardenna* + *Puffinus* clade compared to the alternative topology in PE-ddRAD datasets could also indicate molecular convergence between *Ardenna* and *Puffinus*, which show similarities in morphology and diving behaviour and were historically placed in the same genus (Penhallurick and Wink 2004).

The split of the three North Atlantic lineages of *Puffinus* (1: *P. lherminieri*, *P. baroli* and *P. boydi*, 2: *P. puffinus*, and 3: *P. mauretanicus* and *P. yelkouan*) (Figure 3a) represents the second conflict associated with a short internode. In this case, our analyses largely supported a sister relationship between *P. puffinus* and the Mediterranean clade. Nonetheless, only the PE-ddRAD 75 and the TENT 75 datasets clearly rejected the polytomy test (Table 2). When two speciation events occur at the same time (hard polytomy) and gene tree heterogeneity is mostly caused by ILS, the multispecies coalescent model (MSC) predicts equal frequencies for each of the three topologically-informative unrooted quartet topologies defined around the short internode (Degnan and Rosenberg 2009). Our data show that these frequencies are nearly equal for most analyses (Table 2). In addition, our introgression analyses confirmed that no introgression had occurred between *P. puffinus* and the *P. boydi*, *P. baroli* and *P. lherminieri* lineage. These observations suggest a nearly simultaneous divergence of the three lineages that resulted in high levels of ILS.

The last short internode-associated phylogenetic conflict was the relationship between the two all-dark coloured *Ardenna* species (*A. tenuirostris* and *A. grisea*). Concatenation analyses showed strong support for the sister relationship between *A. grisea* and the *A. gravis*, *A. creatopus* and *A. carneipes* clade across data type and levels of missing data (Figure 4). On the other hand, the alternative topology of a sister relationship between the two all-dark species received ASTRAL-III quartet supports that were nearly as high as (PE-ddRAD data) or higher (UCE data) than the main topology, and at least 8% higher than the remaining alternative topology. These results, together with significant *D*-statistic values between *A. grisea* and *A. tenuirostris* (Figure 6), suggest that introgression could have generated this pattern. However, our PHYLONETWORKS analysis did not support introgression between these two species (Table S6). Moreover, our evaluation of SNPs with strong signals of introgression suggested a role of GC-biased gene conversion (gBGC) in generating these patterns (Table S5). Interestingly, *A. tenuirostris* and *A. grisea* have two of the highest population sizes amongst all shearwaters and thus a likely increased efficacy of gBGC (Weber et al. 2014). Mutation rate variation among different lineages can also lead *D*-statistics to incorrectly infer introgression (Blair and Ané 2019). Our clock model selection analysis

recovered the independent rates clock as the best-fit model, showing important rate heterogeneity among the shearwaters (Table S4). Branch lengths in phylogenetic analyses showed that rate heterogeneity is particularly high in the genus *Ardenna* (Figure 3a). In line with these observations, we recovered lower *D*-statistic values between *A. grisea* and *A. tenuirostris* when we used *A. creatopus* or *A. carneipes* (longer branches) as a sister group to *A. grisea* than when we used *A. gravis* (shorter branch). Furthermore, the shearwater ancestor was likely all dark in colouration, because most species of Procellaria, the sister genus to the shearwaters (Estandia 2019), are all dark. Consequently, *A. grisea* and *A. tenuirostris* may share ancestral variation that could also produce a signature of shared ancestry (Smith and Kronforst 2013). These two species also have large differences in range size which could result in *D*-statistics being misled by ancestral population structure. Taking all these observations into consideration, phylogenetic conflict in this case was likely driven by ILS in combination with rate heterogeneity, gBGC, shared ancestral variation and ancestral population structure.

Our results allow us to conclude that integrating different types of markers together with phylogenetic and introgression analyses provides a better understanding of the causes of phylogenetic incongruence and can be particularly useful for interpreting phylogenetic conflict at short internodes.

4.5 | Phylogeny of the Shearwaters

Previous phylogenies inferred using mtDNA provided poor support for the relationships among the major shearwater lineages (Austin 1996; Heidrich et al. 1998; Nunn and Stanley 1998; Austin et al. 2004). Historically, *Ardenna* species were included within *Puffinus* based on their morphology, osteology and behaviour. We obtained full support for the monophyly of the three shearwater genera. We also recovered *Calonectris* and *Ardenna* as sister lineages (Figure 3a), a novel arrangement, which differs from previous phylogenies. However, the internode of the clade which included *Ardenna* and *Calonectris* was very short, suggesting the succession of two rapid splits that gave rise to the three extant genera. The fossil record for the Procellariiformes is not rich in well-resolved older taxa, but it does document approximately simultaneous

primary records of the three shearwater genera stem lineages in the early to middle Miocene (~14-15.2 Ma) (Miller 1961; Olson 2009), supporting our results.

In the genus *Calonectris*, the short internode of the *C. borealis* - *C. diomedea* clade and its high levels of ILS suggest that the speciation events between the North Atlantic species occurred over a short period of time. Using mtDNA data, Gómez-Díaz et al. (2006) recovered *C. borealis* and *C. edwardsii* as sister species. *D*-statistic analyses showed that probably no introgression occurred between *C. edwardsii* and *C. borealis*, suggesting that phylogenetic discordance was likely caused by ILS alone.

Some lineages of *Ardenna* exhibit strong morphological stasis. For instance, fossils of the extinct *A. conradi* from the middle Miocene were very similar to the extant *A. gravis* (Wetmore 1926). Morphological stasis and similar diving adaptations may explain the resemblance of *A. grisea* and *A. tenuirostris* to *P. nativitatis*, and likely caused the previous placement of these species together under the polyphyletic group Neonectris (Kuroda 1954). Phylogenetic analyses based on mtDNA found the Neonectris group to be polyphyletic (Austin 1996; Nunn and Stanley 1998; Pyle et al. 2011). We also recovered this polyphyly and confirmed the phylogenetic relationships among *Ardenna* species recovered previously (Pyle et al. 2011).

We recovered *P. nativitatis* and the New Zealand species *P. gavia* and *P. huttoni* as the first two splits within *Puffinus*, in agreement with previous studies (Austin et al. 2004; Pyle et al. 2011). Relationships among the remaining species of *Puffinus* were previously unresolved. We recovered fully resolved relationships among these species, which have revealed consistent biogeographic patterns. One of the most relevant results is the polyphyly of the small-sized shearwaters of the *P. assimilis* - *lherminieri* complex, which were historically placed together based on their body size (Austin et al. 2004) (Figure 3a). Changes in body size are frequent in the evolutionary history of Procellariiformes (Nunn and Stanley 1998). Our results suggest that, on a smaller scale, changes in body size are also common along the shearwater phylogeny, and may represent an important trait in the diversification process of pelagic seabirds.

4.6 | Introgression between *P. boydi* and *P.lherminieri*

Despite strong philopatry to breeding colonies and, in most cases, a lack of overlap between breeding areas of closely related species, hybridisation has been documented between several sibling species of Procellariiformes, including shearwaters (Genovart et al. 2012; Booth Jones et al. 2017; Masello et al. 2019). However, the occurrence of ancestral introgression in Procellariiformes has not been studied. Our *D*-statistic tests found an excess of shared derived alleles between *P. boydi* and *P. lherminieri* (Figure 6). Despite full support for the sister relationship between *P. boydi* and *P. baroli* in all phylogenetic analyses, 28% of the gene trees supported a sister relationship between *P. boydi* and *P. lherminieri*, and 18% supported a sister relationship between *P. baroli* and *P. lherminieri*. Additionally, the NeighbourNet network showed smaller genetic distances between *P. boydi* and *P. lherminieri* than between *P. baroli* and *P. lherminieri*. Because *D*-statistic tests can be misled by factors such as ancestral population structure (Eriksson and Manica 2012) or low *Ne* (Martin et al. 2015), we used a phylogenetic network approach to simultaneously account for ILS and gene flow in *Puffinus*. We also evaluated the need to account for gene flow using the TICR test. Although the TICR test showed that a tree model with ILS could explain the observed concordance factors (Figure S18), candidate networks using PE-ddRAD data showed support for introgression between these taxa (Table S6).

The inference of the directionality of introgression using phylogenetic network approaches can be challenging. For instance, SNAQ assumes that no two edges can be part of the same cycle (Solís-Lemus and Ané 2016). This assumption precludes the possibility of inferring gene flow between two taxa in both directions in the same analysis. In addition, gene tree distributions from these alternative topologies may be indistinguishable (Pardi and Scornavacca 2015). However, the network with *P. boydi* as the recipient of genetic material from *P. lherminieri* had much higher inheritance probabilities than those with *P. lherminieri* as recipient. These facts, together with increased heterozygosity in *P. boydi* compared to *P. lherminieri* and *P. baroli*, point to *P. boydi* as the recipient of genetic material.

Optimised inheritance probabilities on the fixed network with gene flow from *P. lherminieri* to *P. boydi* were lower when using UCE data than when using PE-ddRAD data. Ultraconserved elements are likely under strong purifying selection (Bejerano et al. 2004; Harvey et al. 2016) and thus introgressed alleles may be more rapidly removed in these areas of the genome (Juric et al. 2016). In concordance with our results, although some introgressed alleles can be adaptive (Hedrick 2013), selection is primarily known to act against introgressed DNA (Sankararaman et al. 2014), particularly in regulatory regions (Petr et al. 2019) where UCEs are usually located (Bejerano et al. 2004).

The fossil record has documented the presence of both *P. boydi* and *P. lherminieri* in Bermuda during the Pleistocene, where *P. boydi* probably outcompeted *P. lherminieri* until it was extirpated, evidently due to human-introduced predators (Olson 2010). Thus, these two species may have hybridised during the Pleistocene. In addition, *P. boydi* and *P. lherminieri* show some contemporary overlap in their wintering areas (Ramos et al. 2020), which may be a relic of a previously higher overlap in distribution. Future studies should use population genomics data in order to confirm whether the observed patterns are due to ancestral introgression between these two species or to ancestral population structure.

4.7 | Conclusions

Here, we demonstrate the power of integrating UCE and RAD markers, and employing state-of-the-art phylogenetic and introgression analyses, for resolving phylogenetic discordances associated with short internodes. We show that using markers that evolve at different rates allows a detailed exploration of the causes of phylogenetic discordance at different timescales. Our approach provides power to fully resolve complex evolutionary scenarios, such as rapid radiations and introgression histories. Applied to the shearwater problem, our phylogenetic results identified novel relationships and resolved the rapid radiation in the genus *Puffinus*. We have demonstrated that most phylogenetic discordance in shearwaters is driven by high levels of ILS due to rapid speciation events. However, we found evidence for ancestral introgression between *P. boydi* and *P. lherminieri*.

Supplementary Material

Supplementary Material for this chapter may be found in Appendix I. Scripts used in this project are in the GitHub repository:

https://github.com/jferrerobiol/shearwater_phylogenomics.

Author Contributions

R.T.C., H.F.J., J.G., V.B., A.J.W. and J.F. contributed to data collection. J.F., A.J.W. and M.R. designed the study. J.F. and A.J.W. performed DNA extractions, J.F. processed and analysed the data and wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the many institutions that provided tissue loans for this research: Smithsonian National Museum of Natural History, University of Washington Burke Museum, American Museum of Natural History, Louisiana State University Museum of Natural Science, the University of Kansas Biodiversity Institute, and the Muséum National d'Histoire Naturelle. We are grateful to Gary Nunn, Jeremy J. Austin, Chris Gaskin, Kazuto Kawakami, Juan E. Martínez-Gómez, and Maite Louzao for collecting fresh or dry material from the field and/or providing tissue samples from museum skins for this study. We thank the pertinent authorities for issuing the permits needed for this work. We thank Cristian Cuevas for kindly providing the *Puffinus mauretanicus* reference genome. The *Calypte anna* illustration was reproduced with permission from Lynx Edicions. We thank Martí Franch for the shearwater illustrations. We would also like to thank Josephine R. Paris, Brant C. Faircloth and three anonymous reviewers for helpful comments on an earlier draft of the manuscript. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

5 | References

Abbott R.J., Barton N.H., Good J.M. 2016. Genomics of hybridization and its evolutionary consequences. Mol. Ecol. 25:2325–2332.

Aberer A.J., Kobert K., Stamatakis A. 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–2556.

Alda F., Tagliacollo V.A., Bernt M.J., Waltz B.T., Ludt W.B., Faircloth B.C., Alfaro M.E., Albert J.S., Chakrabarty P. 2019. Resolving Deep Nodes in an Ancient Radiation of Neotropical Fishes in the Presence of Conflicting Signals from Incomplete Lineage Sorting. Syst. Biol. 68:573–593.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Andermann T., Fernandes A.M., Olsson U., Töpel M., Pfeil B., Oxelman B., Aleixo A., Faircloth B.C., Antonelli A. 2018. Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements. Syst. Biol. 68:32–46.

Anderson E. 1949. Introgressive hybridization. New York: Wiley.

Angelis K., dos Reis M. 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. Curr. Zool. 61:874–885.

Arcila D., Ortí G., Vari R., Armbruster J.W., Stiassny M.L.J., Ko K.D., Sabaj M.H., Lundberg J., Revell L.J., Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:1-10.

Austin J.J. 1996. Molecular phylogenetics of *Puffinus* shearwaters: preliminary evidence from mitochondrial cytochrome b gene sequences. Mol. Phylogenet. Evol. 6:77–88.

Austin J.J., Bretagnolle V., Pasquet E. 2004. A global molecular phylogeny of the small *Puffinus* shearwaters and implications for systematics of the Little-Audubon's Shearwater complex. Auk 121:647–864.

Bejerano G., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D. 2004. Ultraconserved Elements in the Human Genome. Science 304:1321–1326.

Blair C., Ané C. 2019. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. Syst. Biol. 69:593-601.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bolívar P., Mugal C.F., Nater A., Ellegren H. 2016. Recombination Rate Variation Modulates Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson Interference, in an Avian System. Mol. Biol. Evol. 33:216–227.

Booth Jones K.A., Nicoll M.A.C., Raisin C., Dawson D.A., Hipperson H., Horsburgh G.J., Groombridge J.J., Ismar S.M.H., Sweet P., Jones C.G., Tatayah V., Ruhomaun K., Norris K. 2017. Widespread gene flow between oceans in a pelagic seabird species complex. Mol. Ecol. 26:5716–5728.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., Roychoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Bryant D., Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Carboneras C., Bonan, A. 2019. Petrels, Shearwaters (Procellariidae). In Handbook of the Birds of the World Alive (J. del Hoyo, A. Elliott, J. Sargatal, D. Christie, and E. de Juana, eds.). Lynx Edicions, Barcelona, Spain.

Cariou M., Duret L., Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. Ecol. Evol. 3:846–852.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H., De Koning D.-J. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. Genes|Genomes|Genetics 1:171–182.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Collins R.A., Hrbek T. 2018. An *In Silico* Comparison of Protocols for Dated Phylogenomics. Syst. Biol. 67:633-650.

Croxall J.P., Butchart S.H.M., Lascelles B., Stattersfield A.J., Sullivan B., Symes A., Taylor P. 2012. Seabird conservation status, threats and priority actions: a global assessment. Bird Conservation International 22:1–34.

Cruaud A., Gautier M., Galan M., Foucaud J., Sauné L., Genson G., Dubois E., Nidelet S., Deuve T., Rasplus J.Y. 2014. Empirical assessment of rad sequencing for interspecific phylogeny. Mol. Biol. Evol. 31:1272–1274.

Cuevas-Caballé C., Ferrer-Obiol, J., Genovart, M., Rozas, J., González-Solís, J., Riutort, M. 2019. Conservation genomics applied to the Balearic shearwater. G10K-VGP/EBP 2019. doi: 10.13140/RG.2.2.15751.21923.

Dabney J., Knapp M., Glocke I., Gansauge M.-T., Weihmann A., Nickel B., Valdiosera C., García N., Pääbo S., Arsuaga J.-L., Meyer M. 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. U. S. A. 110:15758–15763.

DaCosta J.M., Sorenson M.D. 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PLoS One 9:e106713.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R., 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Díaz-Arce N., Arrizabalaga, H., Murua, H., Irigoien, X., Rodríguez-Ezpeleta, N. 2016. RAD-seq derived genome-wide nuclear markers resolve the phylogeny of tunas. Mol. Phylogenet. Evol. 102:202-207.

dos Reis M., Gunnell G.F., Barba-Montoya J., Wilkins A., Yang Z., Yoder A.D. 2018. Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. Syst. Biol. 67:594–615.

dos Reis M., Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. 28:2161–2172.

Eaton D.A.R. 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 30:1844–1849.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379.

Emerson K.J., Merz C.R., Catchen J.M., Hohenlohe P. A., Cresko W. A., Bradshaw W.E., Holzapfel C.M. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. U. S. A. 107:16196–16200.

Eriksson A., Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. U. S. A. 109:13956–13960.

Estandia A. 2019. Genome-wide phylogenetic reconstruction for Procellariiform seabirds is robust to molecular rate variation. MSc Dissertation. Durham University.

Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32:786–788.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Syst. Biol. 27:401–410.

Feng S., Stiller, J., Deng, Y. et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. Nature 587:252–257.

Gel B., Díez-Villanueva A., Serra E., Buschbeck M., Peinado M.A., Malinverni R. 2016. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics 32:289–291.

Genovart M., Juste J., Contreras-Díaz H., Oro D. 2012. Genetic and phenotypic differentiation between the critically endangered balearic shearwater and neighboring colonies of its sibling species. J. Hered. 103:330–341.

Gil-Velasco M., Rodríguez, G., Menzie, S. and Arcos, J.M. 2015. Plumage variability and field identification of Manx, Yelkouan and Balearic Shearwaters. British Birds 108:514–539.

Gómez-Díaz E., González-Solís J., Peinado M.A., Page R.D.M. 2006. Phylogeography of the *Calonectris* shearwaters using molecular and morphometric data. Mol. Phylogenet. Evol. 41:322–332.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., Di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Graham S.W., Olmstead R.G., Barrett S.C.H. 2002. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. Mol. Biol. Evol. 19:1769–1781.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Harvey M.G., Smith B.T., Glenn T.C., Faircloth B.C., Brumfield R.T. 2016. Points of View Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Syst. Biol. 65:910–924.

Hedrick P.W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol. Ecol. 22:4606–4618.

Heidrich P., Amengual J., Wink M. 1998. Phylogenetic relationships in Mediterranean and North Atlantic shearwaters (Aves: Procellariidae) based on nucleotide sequences of mtDNA. Biochem. Syst. Ecol. 26:145–170.

Hohenlohe P.A., Bassham S., Etter P.D., Stiffler N., Johnson E.A., Cresko W.A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6:e1000862.

Hosegood J, Humble E, Ogden R, de Bruyn M., Creer S., Stevens G.M.W., Abudaya M., Bassos-Hull K., Bonfil R., Fernando D., Foote A.D., Hipperson H., Jabado R.W., Kaden J., Moazzam M., Peel L.R., Pollett S., Ponzo A., Poortvliet M., Salah J., Senn H., Stewart J.D., Wintner S., Carvalho G. 2020. Phylogenomics and species delimitation for effective conservation of manta and devil rays. Mol. Ecol. 00:1–14 .

Hughes L.C., Ortí G., Huang Y., Sun Y., Baldwin C.C., Thompson A.W., Arcila D., Betancur-R R., Li C., Becker L., Bellora N., Zhao X., Li X., Wang M., Fang C., Xie B., Zhou Z., Huang H., Chen S., Venkatesh B., Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. U. S. A. 115:6249–6254.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., Da Fonseca R.R., Li J.W., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X.J., Dixon A., Li S. B., Li N., Huang Y.H., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z.J., Zeng Y.L., Liu S.P., Li Z.Y., Liu B.H., Wu K., Xiao J., Yinqi X., Zheng Q.M., Zhang Y., Yang H.M., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jonsson K.A., Johnson W., Koepfli K.P., O'brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrom P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T. P., Zhang G.J. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Alfaro-Núñez A., Narula N., Liu L., Burt D., Ellegren H., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G., Avian Phylogenomics Consortium. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience. 4:4.

Juric I., Aeschbacher S., Coop G. 2016. The Strength of Selection against Neanderthal Introgression. PLoS Genet. 12:e1006340.

Kainer D., Lanfear, R. 2015. The effects of partitioning on phylogenetic inference. Mol. Biol. Evol. 32:1611-1627.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Keightley P.D., Jackson B.C. 2018. Inferring the Probability of the Derived vs. the Ancestral Allelic State at a Polymorphic Site. Genetics 209:897–906.

Knowles L.L., Huang H., Sukumaran J., Smith S.A. 2018. A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. Am. J. Bot. 105:376–384.

Korlach J., Gedman G., Kingan S.B., Chin C.-S., Howard J.T., Audet J.-N., Cantin L., Jarvis E.D. 2017. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience 6:1–16.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453-4455.

Kuroda N. 1954. On the classification and phylogeny of the order Tubinares, particularly the shearwaters (*Puffinus*), with special considerations on their osteology and habit differentiation (Aves). Tokyo: Herald Company.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2017. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Mol. Biol. Evol. 34:772–773.

Larget B.R., Kotha S.K., Dewey C.N., Ané C. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910–2911.

Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: Conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.

Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Malinsky M., Matschiner M., Svardal H. 2020. Dsuite - fast *D*-statistics and related admixture evidence from VCF files. bioRxiv 634477.

Malinsky M., Svardal H., Tyers A.M., Miska E.A., Genner M.J., Turner G.F., Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol. 2:1940–1955.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.

Manthey J.D., Campillo L.C., Burns K.J., Moyle R.G. 2016. Comparison of Target-Capture and Restriction-Site Associated DNA Sequencing for Phylogenomics: A Test in Cardinalid Tanagers (Aves, Genus: Piranga). Syst. Biol. 65:640–650.

Marsh C. 1870. ART. XXV.--Notice of some Fossil Birds, from the Cretaceous and Tertiary Formations of the United States. American Journal of Science and Arts (1820-1879) 49:205.

Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32:244–257.

Martin S.H., Jiggins C.D. 2017. Interpreting the genomic landscape of introgression. Curr. Opin. Genet. Dev. 47:69–74.

Masello J.F., Quillfeldt P., Sandoval-Castellanos E., Alderman R., Calderón L., Cherel Y., Cole T.L., Cuthbert R.J., Marin M., Massaro M., Navarro J., Phillips R.A., Ryan P.G., Shepherd L.D., Suazo C.G., Weimerskirch H., Moodley Y. 2019. Additive Traits Lead to Feeding Advantage and Reproductive Isolation, Promoting Homoploid Hybrid Speciation. Mol. Biol. Evol. 36:1671–1685.

McCluskey B.M., Postlethwait J.H. 2015. Phylogeny of Zebrafish, a "Model Species," within *Danio*, a "Model Genus." Mol. Biol. Evol. 32:635–652.

McCormack J.E., Heled J., Delaney K.S., Peterson A.T., Knowles L.L. 2011. Calibrating divergence times on species trees versus gene trees: implications for speciation history of *Aphelocoma* jays. Evolution. 65:184–202.

McGowen M.R., Tsagkogeorga G., Álvarez-Carretero S., Dos Reis M., Struebig M., Deaville R., Jepson P.D., Jarman S., Polanowski A., Morin P.A., Rossiter S.J. 2019. Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture. Syst. Biol. 69:479-501.

Mendes F.K., Hahn M.W. 2018. Why Concatenation Fails Near the Anomaly Zone. Syst. Biol. 67:158–169.

Miller L. 1961. Birds from the Miocene of Sharktooth Hill, California. Condor. 63:399–402.

Miller M., Dunham J., Amores A., Cresko W., Johnson E. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17:240–248.

Mirarab S., Bayzid M.S., Warnow T. 2016. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. Syst. Biol. 65:366–380.

Nabholz B., Künstner A., Wang R., Jarvis E.D., Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28:2197–2210.

Nunn G.B., Stanley S.E. 1998. Body size effects and rates of cytochrome b evolution in tube-nosed seabirds. Mol. Biol. Evol. 15:1360–1371.

Oliveros C.H., Field D.J., Ksepka D.T., Barker F.K., Aleixo A., Andersen M.J., Alström P., Benz B.W., Braun E.L., Braun M.J., Bravo G.A., Brumfield R.T., Chesser R.T., Claramunt S., Cracraft J., Cuervo A.M., Derryberry E.P., Glenn T.C., Harvey M.G., Hosner P.A., Joseph L., Kimball R.T., Mack A.L., Miskelly C.M., Peterson A.T., Robbins M.B., Sheldon F.H., Silveira L.F., Smith B.T., White N.D., Moyle R.G., Faircloth B.C. 2019. Earth history and the passerine superradiation. Proc. Natl. Acad. Sci. U. S. A. 116:7916–7925.

Olson S.L. 2009. A new diminutive species of shearwater of the genus *Calonectris* (Aves: Procellariidae) from the Middle Miocene Calvert Formation of Chesapeake Bay. Proceedings of the Biological Society of Washington 122:466–470.

Olson S.L. 2010. Stasis and turnover in small shearwaters on Bermuda over the last 400 000 years (Aves: Procellariidae: Puffinus lherminieri group). Biol. J. Linn. Soc. Lond. 99:699–707.

Olson S.L., Rasmussen P.C. 2001. Miocene and Pliocene birds from the Lee Creek Mine, North Carolina. Smithson. Contrib. Paleobiol. 90:233–365.

Ottenburghs J., Ydenberg R.C., Van Hooft P., Van Wieren S.E., Prins H.H.T. 2015. The Avian Hybrids Project: gathering the scientific literature on avian hybridization. Ibis 157:892–894.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Pardi F., Scornavacca C. 2015. Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable. PLoS Comput. Biol. 11:e1004135.

Paris J.R., Stevens J.R., Catchen J.M. 2017. Lost in parameter space: a road map for stacks. Methods in Ecology and Evolution. 8:1360–1373.

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. 2012. Ancient admixture in human history. Genetics 192:1065–1093.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14:e1002379.

Penhallurick J., Wink M. 2004. Analysis of the taxonomy and nomenclature of the Procellariiformes based on complete nucleotide sequences of the mitochondrial cytochrome b gene. Emu. 104:125–147.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7:e37135.

Petr M., Pääbo S., Kelso J., Vernot B. 2019. Limits of long-term selection against Neandertal introgression. Proc. Natl. Acad. Sci. U. S. A. 116:1639–1644.

Purvis A., Gittleman J.L., Brooks T. eds. 2005. Phylogeny and Conservation. New York. Cambridge University Press.

Pyle P., Welch A.J., Fleischer R.C. 2011. A New Species of Shearwater (*Puffinus*) Recorded from Midway Atoll, Northwestern Hawaiian Islands. Condor 113:518–527.

Rabiee M., Sayyari E., Mirarab S. 2019. Multi-allele species reconstruction using ASTRAL. Mol. Phylogenet. Evol. 130:286–296.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. 67:901–904.

Ramos R., Paiva V.H., Zajková Z., Precheur C., Fagundes A.I., Jodice P.G.R., Mackin W., Zino F., Bretagnolle V., González-Solís J. 2020. Spatial ecology of closely related taxa: the case of the little shearwater complex in the North Atlantic Ocean. Zool. J. Linn. Soc. zlaa045.

Rannala B., Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Syst. Biol. 56:453–466.

Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Hackett S.J., Han K.L., Harshman J., Huddleston C.J., Kingston S., Marks, B.D. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66:857-879.

Rheindt F.E., Austin J.J. 2005. Major analytical and conceptual shortcomings in a recent taxonomic revision of the Procellariiformes - A reply to Penhallurick and Wink (2004). Emu 105:181–186.

Rochette N.C., Catchen J.M. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. Nat. Protoc. 12:2640–2659.

Rochette N.C., Rivera-Colón A.G., Catchen J.M. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28:4737–4754.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100:56–62.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Romiguier J., Ranwez V., Douzery E.J.P., Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20:1001–1009.

Rosenberg N.A., Nordborg M. 2002. Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms. Nat. Rev. Genet. 3:380–390.

Rubin B.E.R., Ree R.H., Moreau C.S. 2012. Inferring phylogenies from RAD sequence data. PLoS One 7:e33394.

Sankararaman S., Mallick S., Dannemann M., Prüfer K., Kelso J., Pääbo S., Patterson N., Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354–357.

Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.

Sayyari E., Mirarab S. 2018. Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. Genes 9:132.

Sayyari E., Whitfield J.B., Mirarab S. 2017. Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. Mol. Biol. Evol. 34:3279–3291.

Smith B.T., Harvey M.G., Faircloth B.C., Glenn T.C., Brumfield R.T. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63:83–95.

Smith J., Kronforst M.R. 2013. Do *Heliconius* butterfly species exchange mimicry alleles? Biol. Lett. 9:20130503.

Smith S.A., Brown J.W., Walker J.F. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. PLoS One 13:e0197433.

Solís-Lemus C., Ané C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genet. 12:e1005896.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: A package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. U. S. A. 109:14942–14947.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stenz N.W.M., Larget B., Baum D.A., Ané C. 2015. Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species *Arabidopsis thaliana* (L.) heynh. Syst. Biol. 64:809–823.

Tagliacollo V.A., Lanfear R. 2018. Estimating Improved Partitioning Schemes for Ultraconserved Elements. Mol. Biol. Evol. 35:1798–1811.

The Heliconius Genome Consortium, Dasmahapatra K.K., Walters J.R., Briscoe A.D., Davey J.W., Whibley A., Nadeau N.J., Zimin A.V., Hughes D.S.T., Ferguson L.C., Martin S.H., Salazar C., Lewis J.J., Adler S., Ahn S.-J., Baker D. a., Baxter S.W., Chamberlain N.L., Chauhan R., Counterman B. A., Dalmay T., Gilbert L.E., Gordon K., Heckel D.G., Hines H.M., Hoff K.J., Holland P.W.H., Jacquin-Joly E., Jiggins F.M., Jones R.T., Kapan D.D., Kersey P., Lamas G., Lawson D., Mapleson D., Maroja L.S., Martin A., Moxon S., Palmer W.J., Papa R., Papanicolaou A., Pauchet Y., Ray D. a., Rosser N., Salzberg S.L., Supple M. a., Surridge A., Tenger-Trolander A., Vogel H., Wilkinson P. A., Wilson D., Yorke J. a., Yuan F., Balmuth A.L., Eland C., Gharbi K., Thomson M., Gibbs R. a., Han Y., Jayaseelan J.C., Kovar C., Mathew T., Muzny D.M., Ongeri F., Pu L.-L., Qu J., Thornton R.L., Worley K.C., Wu Y.-Q., Linares M., Blaxter M.L., Ffrench-Constant R.H., Joron M., Kronforst M.R., Mullen S.P., Reed R.D., Scherer S.E., Richards S., Mallet J., Owen McMillan W., Jiggins C.D. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

Tripp E. A., Tsai Y. H. E., Zhuang Y., Dexter K. G. 2017. RAD seq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceae) in the ultra-arid deserts of Namibia. Ecol. Evol. 7:7920-7936.

Tung Ho, L. S., Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst. Biol. 63:397–408.

Wagner C.E., Keller I., Wittwer S., Selz O.M., Mwaiko S., Greuter L., Sivasundar A., Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–798.

Weber C.C., Boussau B., Romiguier J., Jarvis E.D., Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol. 15:1-16.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet. Syst. Biol. 67:735–740.

Wetmore A. 1926. Observations on Fossil Birds Described from the Miocene of Maryland. Auk 43:462–468.

Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:15–30.

Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: RECOMB International Workshop on Comparative Genomics. Springer. p. 53–75

Chapter II

Paleoceanographic Changes in the Late Pliocene Promoted Rapid Diversification in Pelagic Seabirds

JOAN FERRER OBIOL, HELEN F. JAMES, R. TERRY CHESSER, VINCENT BRETAGNOLLE, JACOB GONZÁLEZ-SOLÍS, JULIO ROZAS, ANDREANNA J. WELCH AND MARTA RIUTORT

Paleoceanographic Changes in the Late Pliocene Promoted Rapid Diversification of Pelagic Seabirds

Joan Ferrer Obiol[1,2], Helen F. James[3], R. Terry Chesser[4,5], Vincent Bretagnolle[6], Jacob González-Solís[7,2], Julio Rozas[1,2], Andreanna J. Welch[8] and Marta Riutort[1,2]

[1]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain
[2]Institut de Recerca de la Biodiversitat (IRBio), Barcelona, Catalonia, Spain
[3]Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA
[4]U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, MD, USA
[5]National Museum of Natural History, Smithsonian Institution, Washington, DC, USA
[6]Centre d'Études Biologiques de Chizé, CNRS & La Rochelle Université, 79360, Villiers en Bois, France
[7]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain
[8]Department of Biosciences, Durham University, Durham, UK

In review in Journal of Biogeography

Abstract

Aim: Paleoceanographic changes can act as drivers of diversification and speciation, even in highly mobile marine organisms. Shearwaters are a globally distributed group of highly mobile pelagic seabirds with a recently well-resolved phylogeny and controversial species limits that show periods of both slow and rapid diversification. Here, we explore the role of paleoceanographic changes on the diversification and speciation in these highly mobile pelagic seabirds. We investigate shearwater biogeography and the evolution of a key phenotypic trait, body size, and we assess the validity of the current taxonomy of the group.

Location: Worldwide.

Taxa: Shearwaters (Order Procellariiformes, Family Procellariidae, Genera *Ardenna*, *Calonectris* and *Puffinus*).

Methods: We generated genomic data (double-digest restriction site-associated DNA) for almost all extant shearwater species to infer a time-calibrated species tree. We estimated ancestral ranges and evaluated the roles of founder events, vicariance and surface ocean currents in driving shearwater diversification. We performed

phylogenetic generalized least squares to explore the drivers of variability in body size along the phylogeny. To assess the validity of the current taxonomy of the group, we analysed genomic patterns of recent shared ancestry and differentiation among shearwater taxa.

Results: We identified a period of high dispersal and rapid speciation during the Plio-Pleistocene boundary. Species dispersal appears to be favoured by surface ocean currents, and founder events are a main process of diversification. Body mass shows significant associations with life strategies and local conditions. The current taxonomy shows some incongruences with the patterns of genomic divergence.

Main Conclusions: The Pliocene marine megafauna extinction had a severe effect on shearwaters, and the subsequent burst of speciation and dispersal was probably promoted by Pleistocene climatic shifts. Our findings extend our understanding on the drivers of speciation and dispersal of highly mobile pelagic seabirds and shed new light on the important role of paleoceanographic events.

Keywords: biogeography, diversification, molecular dating, seabirds, speciation, taxonomy

1 | Introduction

Speciation is a key evolutionary process that results from the independent evolution and adaptation of populations and ultimately acts as a major driver responsible for the generation of species-level biodiversity (Kopp 2010; Schluter and Pennell 2017). Species richness is unevenly distributed across the Tree of Life, and its current patterns of distribution result from biotic and abiotic processes that operate over space and time (Simpson 1953; Benton 2009; Vargas and Zardoya 2014). Evidence for the mechanisms that promote population differentiation and speciation are currently better understood in terrestrial than in marine environments (Coyne et al. 2004; Butlin et al. 2012; Nosil 2012), where the lack of obvious physical barriers would suggest that neutral processes of panmixia, or isolation-by-distance, will prevail, especially in highly mobile species (Moura et al. 2013). However, counterintuitive evidence of fine-scale differentiation among populations and species in a number of marine taxa has been described as the

'marine species paradox' (Palumbi 1994; Bierne et al. 2003). Thus, there is a need for explicit evaluations of the role of selective processes in driving patterns of differentiation in marine systems.

In species complexes that are geographically widespread, the gradual evolution of reproductive isolation in allopatry can make species delimitation challenging, especially in young radiations (Carstens et al. 2013; Cutter 2013). Many allospecies first tend to differ from their close relatives at traits subjected to sexual and other forms of social selection (Price 2008; Seddon et al. 2013). When this occurs, our ability to delimit species may be further hampered by morphological stasis, especially when changes in ecological niche in allopatry are minimal (Fišer et al. 2018). In cases of morphological stasis and limited behavioural information, genomic data can provide informed hypotheses on species limits of allopatric taxa and can be conclusive in parapatric or sympatric taxa. Despite the extent of disagreement about how genomic data should be applied to species delimitation (Sukumaran and Knowles 2017; Leaché et al. 2018), agreement exists that genomic data can provide additional perspective on species limits when used together with other data types such as phenotypic and ecological information.

Seabirds of the order Procellariiformes present some of the most extreme examples of the marine speciation paradox. Procellariiformes are highly mobile pelagic seabirds with a high dispersal ability and perform some of the longest animal migrations on Earth (covering more than 120,000 km a year) (Shaffer et al. 2006; Weimerskirch et al. 2015). However, Procellariiformes also show high philopatry to their breeding grounds (Weimerskirch et al. 1985), which is expected to limit gene flow and therefore reinforce genetic differentiation (Friesen et al. 2007a).

Shearwaters are a monophyletic group in the family Procellariidae, and they offer an excellent case study for examining the mechanisms of population differentiation and speciation in marine environments. First, shearwaters are globally distributed and breed mostly in allopatry. Second, the current taxonomy recognises three genera and 30 species with a recently well-resolved phylogeny showing clear periods of rapid diversification (Chapter I). Third, the three recognised genera exhibit different ecologies and degrees of species richness. Fourth, their high mobility makes them an ideal model

to evaluate the roles of founder events and vicariance using biogeographic analyses. Fifth, abiotic and biotic factors are known to promote speciation in the shearwaters and related Procellariiformes; for instance, paleoceanographic changes such as the Pleistocene climatic oscillations can act as historical drivers of speciation (Gómez-Díaz et al. 2006; Silva et al. 2015) and intrinsic biotic factors such as different foraging strategies and allochrony can also promote speciation (Friesen et al. 2007b; Rayner et al. 2011; Lombal et al. 2018). Sixth, species limits are controversial, mostly due to high morphological stasis (Austin 1996; Austin et al. 2004); indeed, only a few phenotypic traits, such as vocalisation characteristics, slight plumage colour differences and in particular, body size, may differ between closely related species. A comprehensive study using genomic data will assist in resolving species delimitation within the context of the factors that promote diversification and speciation.

The reconstruction of ancestral ranges and evaluation of alternative biogeographic models are critical to our understanding of shearwater diversification throughout the world in light of environmental and oceanographic events. Of particular interest is the importance of founder events during the evolution of shearwaters. The foundation of colonies is believed to be a rare event in most seabird species despite their great potential for long-range dispersal (Milot et al. 2008). However, in several shearwater species, contemporary colony foundation events have been reported (Munilla et al. 2016; Storey and Lien 1985). Understanding the relative importance of founder and vicariant events during the evolution of shearwaters can have important implications for the conservation of these endangered pelagic seabirds.

Here, we use paired-end double-digest restriction site-associated DNA sequencing (PE-ddRAD-Seq) for almost all extant shearwater taxa to explore the drivers of diversification and speciation in this group of pelagic seabirds. Specifically, we produce the first time-calibrated shearwater species tree using the multispecies coalescent approach (MSC) to account for the high levels of ILS affecting the shearwater phylogeny. We then infer the biogeographic history of the group by estimating ancestral ranges and evaluating the roles of founder events, vicariance and surface ocean currents in driving their diversification. Furthermore, we explore the ecological and geographical forces responsible for the variability in a key phenotypic trait, body size. Finally, we assess the

validity of the current taxonomy of the group by analysing genomic patterns of recent shared ancestry and differentiation among shearwater taxa.

2 | Materials and Methods

2.1 | Sampling and Sequence Data Generation

We collected 68 blood or tissue samples from 25 of the 32 recognised species of shearwaters (Gill et al. 2020) (Table S1) representing all the major lineages in the group (Austin 1996; 2004). Species that could not be included (*Puffinus heinrothi*, *P. bannermani*, *P. bryani*, *P. myrtae*, *P. auricularis*, *P. persicus* and *P. subalaris*) breed in remote islands, have very limited distributions and/or are categorised as critically endangered by the IUCN Red List of Threatened Species (http://www.iucnredlist.org/). Sampling was conducted under permits issued by the relevant authorities. Sequence data for 51 of these samples were generated previously in a recent phylogenomic study (Chapter I).

For the new samples generated here, we extracted genomic DNA using the Qiagen DNeasy Blood and Tissue Kit according to the manufacturer's instructions (Qiagen GmbH, Hilden, Germany). We used a Qubit Fluorometer (Life Technologies) to quantify and standardise DNA concentrations of all samples at 10 ng/ul. Approximately 250 ng of genomic DNA of each sample was sent to the Genomic Sequencing and Analysis Facility, University of Texas at Austin, to perform ddRAD library preparation following the Peterson et al. (2012) protocol. DNA was fragmented using an uncommon cutter *EcoRI* and a common cutter *MspI* in a single reaction. Illumina adaptors containing sample-specific barcodes and Illumina indexes were ligated onto the fragments and four pools were produced differing by their Illumina index. Barcodes differed by at least two base pairs to reduce the chance of inaccurate barcode assignment. Pooled libraries were size selected (between 150 and 300 bp after accounting for adapter length) using a Pippin Prep size fractionator (Sage Science, Beverly, Ma). Libraries were amplified in a final PCR step prior to sequencing on a single lane of an Illumina HiSeq4000 platform using a 150-bp paired-end (PE) metric.

2.2 | PE-ddRAD-Seq Data Filtering and Assembly

Raw reads were demultiplexed and cleaned using PROCESS_RADTAGS in STACKS v2.41 (Rochette et al. 2019). To maximise the amount of biological information, we built loci using the forward reads with parameters optimised for shearwater data (Chapter I) using the USTACKS-CSTACKS-SSTACKS core clustering algorithm. We used the TSV2BAM program to incorporate reverse reads by matching the set of forward read IDs in each locus. We then assembled a contig for each locus, called SNPs using the Bayesian genotype caller (Maruki and Lynch 2015, 2017) and phased haplotypes using GSTACKS. Subsequently, we mapped the GSTACKS catalog to the Balearic shearwater (*Puffinus mauretanicus*) genome assembly (Cuevas-Caballé et al. 2019) using BWA-MEM v. 0.7.17 (Li 2013). We sorted and indexed the mapped reads using SAMTOOLS v.1.4 (Li et al. 2009; Li 2011) and integrated alignment positions to the catalog using STACKS-INTEGRATE-ALIGNMENTS (Paris et al. 2017). Finally, we used the POPULATIONS program to filter SNP data requiring a minor allele frequency (MAF) above 5% and an observed heterozygosity below 50% to generate datasets for downstream analysis.

2.3 | Species Tree Inference

To estimate a time-calibrated species tree for shearwaters, we applied the SNP-based MSC approach of (Stange et al. 2018) implemented in the SNAPP v.1.3 (Bryant et al. 2012) package of the program BEAST2 v.2.5.0 (Bouckaert et al. 2019). To prepare a suitable dataset for this method, we selected a maximum of two individuals per subspecies (51 individuals in total) and we exported called variants to variant call format (VCF). Because SNAPP assumes a single nucleotide substitution rate, we performed the analyses including only transitions to reduce heterogeneity in the evolutionary rate. We further processed the VCF file with VCFTOOLS v.0.1.15 (Danecek et al. 2011) to include only biallelic SNPs without missing data, to mask genotypes if the per-sample read depth was below 5 or above 150, or if the genotype quality was below 30. Finally, we selected a single SNP per ddRAD locus to remove potentially linked SNPs (minimum distance between SNPs > 500 bp). After filtering, we retained a dataset of 1397 transitions.

We followed recommendations of Stange et al. (2018) by constraining the root of the species tree to follow a normal distribution with a mean of 20.23 Mya and a standard deviation (SD) of 2 (as reported in Chapter I) based on three fossil calibrations (see calibration strategy B there-in) and a relaxed clock. SD was calculated to fit the posterior distribution for the root as in Chapter I. This divergence time estimate for the root was further supported by a global study on birds using relaxed clocks (Jetz et al. 2012). As we were mainly interested in SNAPP's ability to estimate divergence times rather than the tree topology, we fixed the species tree topology to that inferred in Chapter I, using UCE and ddRAD data. We also tested the robustness of divergence-time estimates by performing two additional analyses. Firstly, we explored the effects of fixing the topology by also performing the analysis without the topology being fixed. We also evaluated the use of fossil calibrations using three different calibration points based on those described in strategy B of Chapter I. We used the ruby script snapp_prep.rb (https://github.com/mmatschiner/snapp_prep) to prepare the XML file for SNAPP analyses. For each analysis, we conducted three replicate runs, each with a run length of 500,000 Markov-chain Monte Carlo (MCMC) iterations. Convergence and stationarity were confirmed (effective sample sizes > 300) using TRACER v.1.7.1 (Rambaut et al. 2018). The first 10% of each MCMC was discarded as burn-in, and posterior distributions of run replicates were merged to generate maximum-clade-credibility (MCC) trees with node heights set to mean age estimates with TREEANNOTATOR (Heled and Bouckaert 2013). SNAPP trees were visualised in DENSITREE v.2.2.7 (Bouckaert 2010).

The finite-sites model implemented in SNAPP allows the estimation of both branch lengths (times) and population sizes (θ) (Bryant et al. 2012). Because the Stange et al. (2018) approach only estimates a single value of θ for all branches, we also constructed a SNAPP phylogeny without any age constraint, in order to estimate θ values for each branch.

2.4 | Ancestral Range Estimation

Biogeographic analyses were performed to estimate ancestral ranges and to examine patterns of shearwater dispersal across five broad areas. The five areas were chosen based on contemporary shearwater breeding ranges: Southern Australia and New

Zealand (A), Southern Ocean (B), North and Tropical Pacific Ocean (C), Tropical Indian Ocean (D), and North Atlantic Ocean and Mediterranean Sea (E). We set the limit between areas A and B at the Subtropical Front (Sutton 2001). The R package BIOGEOBEARS v. 1.1.2 (Matzke 2013) was used to estimate ancestral ranges using likelihood versions of three models: dispersal-extinction-cladogenesis (DEC; Ree and Smith (2008)), dispersal-vicariance (DIVA; Ronquist (1997)), and BayArea (Landis et al. (2013)), and the time-calibrated shearwater tree. We compared ancestral range estimates of these models with and without the founder-event speciation parameter (j) under two scenarios: one that allowed unrestricted dispersal between all areas and another that limited dispersal between areas connected by major surface ocean currents from the Pliocene to the present, when most of the shearwater diversification occurred (Figure 1). Corrected Akaike Information Criterion (AICc) and AICc weights were used to select the best-fit scenario for the models with and without the j parameter separately, because the DEC + j model has been criticised for not being statistically comparable to the DEC model (Ree and Sanmartín 2018). To infer the ancestral range of the shearwaters' most recent common ancestor (MRCA), we used the ranges of the two most closely related outgroup lineages (for which no genetic data are available): genus *Procellaria*, and genera *Pseudobulweria* and *Bulweria* (Estandía 2019). *Pseudobulweria rostrata*, *Bulweria bulwerii*, *Procellaria westlandica* and *Procellaria cinerea* were chosen because they represent the totality of ranges within their clades. Divergence times between the outgroups and shearwaters and among the outgroups were retrieved from the TIMETREE database (Kumar et al. 2017). Outgroups were incorporated into the time-calibrated shearwater tree using the BIND.TREE function from the APE package (Paradis and Schliep 2019) in R.

2.5 | Phylogenetic Comparative Analyses

To evaluate potential predictors of body size variation in shearwaters, we retrieved data for four body size measures: 1) mean body mass; 2) range of body mass (maximum body mass - minimum body mass); 3) wing length; and 4) total body length, and five predictors: 1) minimum, 2) mean and 3) maximum breeding latitudes (in absolute values), 4) latitudinal range occupied by a species year-round (maximum latitude -

minimum latitude where the species is present either during the breeding or the non-breeding period) and 5) migratory strategy (long-distance migrant, short distance migrant or dispersive/sedentary). Additionally, we retrieved wingspan measurements to obtain a mean body mass measure corrected by body surface (mean body mass / (body length x wingspan). Data were retrieved for all recognised species of shearwaters (Gill et al. 2020) with the exception of Heinroth's shearwater (*Puffinus heinrothi*), because no information about its phylogenetic placement is available. The majority of morphometric, distributional and behavioural data were retrieved from Birds of the World (Billerman et al. 2020) and additional morphometric data were extracted from (Onley and Scofield 2013). We performed phylogenetic generalised least squares regressions (PGLS) for all body size measures against each of the potential predictors using the R package CAPER (Orme et al. 2013) and we adjusted p-values by FDR correction for multiple testing. Due to the reduced number of species, we only performed univariate regressions to avoid overfitting (Mundry 2014). Following recommendations of Revell (2010), we simultaneously estimated the λ parameter (Pagel 1999) to account for deviations from a pure Brownian motion (BM). PGLS analyses were run using the time-calibrated tree with the species for which no sequencing data was available incorporated into the phylogeny using the BIND.TIP function from the R package PHYTOOLS (Revell 2012) according to the phylogenetic position and branch lengths from previous phylogenetic studies (Austin et al. 2004; Pyle et al. 2011; Martínez-Gómez et al. 2015; Kawakami et al. 2018). We estimated ancestral states for body size measures using the function FASTANC in the R package PHYTOOLS and visualised the reconstructions with phenograms using the R package GGTREE (Yu et al. 2017). We also reconstructed ancestral states for migratory behaviour using maximum likelihood (ML) with the function REROOTINGMETHOD in the R package PHYTOOLS.

To evaluate the effect of life-history traits (LHT) on the nucleotide substitution rate and the equilibrium of GC content (GC*), we modelled the correlation between these two parameters, and their correlations with mean body mass and the number of breeding pairs as a multivariate Brownian motion in COEVOL (Lartillot and Poujol 2011). We ran two independent Markov Chain Monte Carlo (MCMC) chains, and stopped the process after reaching convergence (effective sample size > 1,000 and discrepancy

between chains < 0.05 for all statistics; 5,000 generations) using TRACECOMP from the COEVOL package. The number of breeding pairs for each species were retrieved from Birds of the World (Billerman et al. 2020) and BirdLife International (2020).

The time-calibrated tree was also used to calculate evolutionary distinctness (ED) scores and EDGE scores (Isaac et al. 2007), based on IUCN Red List of Threatened Species threat-status (GE, as of June 2020; http://www.iucnredlist.org/), calculated in the R package CAPER. EDGE scores for each species were calculated as follows: EDGE = ln(1 + ED) + GE x ln(2).

2.6 | Patterns of Recent Coancestry and Sequence Divergence

To explore congruence between the current shearwater taxonomy and the genetic structure among species, we used FINERADSTRUCTURE v0.3.2 (Malinsky et al. 2018) to infer the shared ancestry among all individuals. FINERADSTRUCTURE uses haplotype linkage information to derive a co-ancestry matrix based on the most recent coalescent events. We exported haplotypes for loci present in at least 75% of the individuals to RADPAINTER format using POPULATIONS, resulting in a set of haplotypes for 8,049 PE-ddRAD loci containing a total of 63,492 SNPs. RADPAINTER was used to infer a coancestry matrix and the FINERADSTRUCTURE MCMC clustering algorithm was used to assign individuals into clusters, with a burn-in period of 100,000 generations and an extra 100,000 MCMC iterations sampled every 1,000 generations. To arrange the clusters based on their relationships within the coancestry matrix, we built a tree within FINERADSTRUCTURE using default parameters. To visualise the results, we used the R scripts fineradstructureplot.r and finestructurelibrary.r (available at http://cichlid.gurdon.cam.ac.uk/fineRADstructure.html).

As an additional approach to examining congruence between the current shearwater taxonomy and genomic divergence, we examined the distribution of pairwise genetic distances using loci present in at least 95% of the individuals (1,525 loci; 11,055 SNPs). Briefly, we exported variants into a VCF file using POPULATIONS in STACKS, we converted the VCF file into a DNABIN object using the R package VCFR (Knaus and Grünwald 2017), and we calculated pairwise distances using the DIST.DNA function from the APE package in R.

3 | Results

We recovered an average of 1,227,032 (SD = 815,798) PE-ddRAD reads per sample (Table S1) that were assembled to an average of 24,621 loci per sample, with a mean coverage per sample of 39x (SD = 19). Locus length ranged from 140 to 239 bp with a median of 198 bp (SD = 25.5).

3.1 | Bayesian Divergence Time Estimation with SNP Data

The SNAPP phylogeny revealed largely the same topology as a previous phylogenetic study based on the same data (Chapter I), except for the poorly supported relationship between *Ardenna* and *Puffinus* and the relationship between *A. grisea* and *A. tenuirostris* (Table S2; Figure S1). Both incongruences were already identified in the previous study using different methods and datasets, and were caused by high levels of ILS.

Using a constraint for the age of the root, we estimated the time-calibrated tree shown in Figure 1. The time to the most recent common ancestor (TMRCA) of *Puffinus* was the oldest among the three genera, estimated at 9.98 Mya (95% HPD: 12.26-7.65 Mya). The TMRCA of *Ardenna* was inferred to be 4.52 Mya (95% HPD: 5.64-3.53 Mya) and the TMRCA of *Calonectris* 3.54 Mya (95% HPD: 4.57-2.55 Mya). If the divergence times are accurate, then shearwater speciation increased during the Pliocene reaching a peak by the late Pliocene (~2.58 Mya; Figure 1), when most of the modern biogeographical groups of shearwaters were already present.

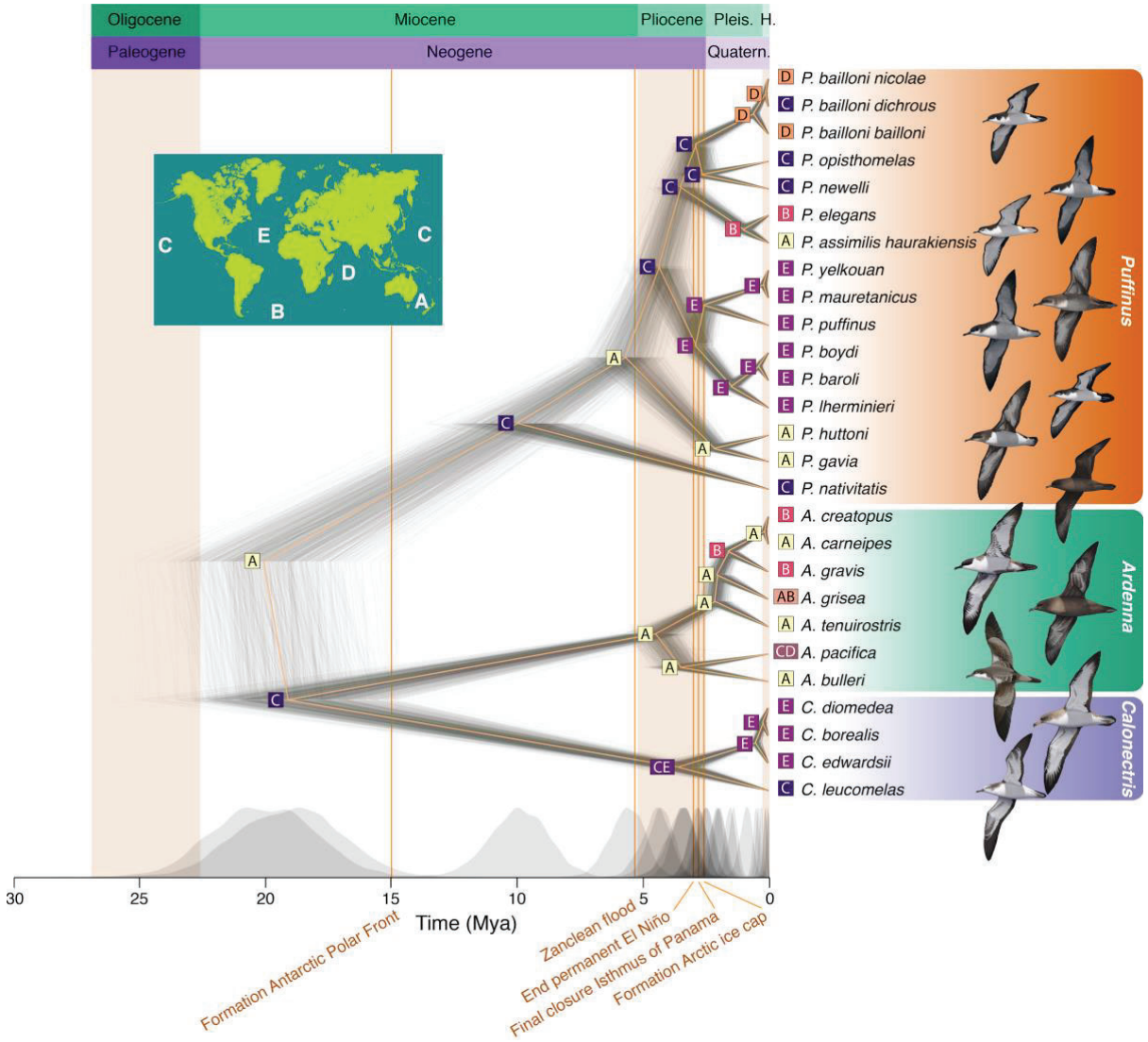


Figure 1 Time-calibrated species tree of the shearwaters using a constraint on the root age and a fixed topology. Geological periods and epochs are labelled above the tree. Posterior densities of divergence times are shown below the species tree. Note the speciation peak during the late Pliocene - early Pleistocene. Ancestral ranges were estimated under the DIVALIKE + j model using a dispersal matrix restricting dispersal between areas connected by main historical and present surface ocean currents in BIOGEOBEARS and are shown as boxes at nodes and tips coded according to the map (Inset). Posterior estimates of divergence times are summarized in Table S2. Illustrations by Martí Franch© are representative shearwater species depicted by their lineages.

Using the same three fossil calibrations (see Materials and Methods), shearwater divergence times inferred using the MSC were 28-94% younger than those estimated in Chapter I, using concatenation (Table S2). MSC analyses using these fossil calibrations resulted in slightly older estimates (8.9% older on average) compared to the same analyses using a single age constraint on the root (Figure S2, Table S2). Conversely, fixing the phylogeny had very little effect on age estimates (0.4% older on average).

The mean population size across all shearwater species estimated by SNAPP was N=63,555 individuals (95% HPD: 50,390–77,155) when assuming the lowest generation time estimated for a shearwater species (13 years; Genovart et al. 2016), and N=43,485 individuals (95% HPD: 34,477–52,790) when assuming the highest estimated value (19 years; Birdlife International 2020). However, SNAPP analysis without age constraints showed a notable variation in θ estimates even between sister species (Figure S3) suggesting frequent changes in population size in the evolutionary history of shearwaters.

3.2 | Biogeographic Analysis

Under all tested models, ancestral range estimation analyses, including a dispersal matrix restricting dispersal between areas connected by main historical and present surface ocean currents, had lower AICc than models with an unrestricted dispersal matrix (Table 1). DIVALIKE and DEC models had lower AICc than BAYAREALIKE models, especially when the founder event parameter (j) was not included, suggesting that vicariance is an important mode of speciation in shearwaters. The slightly lower AICc for DIVALIKE models compared to DEC models further supported the importance of widespread vicariance. However, in models including founder event speciation, the j parameter ranged from 0.0874 to 0.1733 and the rate of range expansion (d) was an order of magnitude smaller, showing that founder events have a higher probability of explaining most of the data than range expansion. Indeed, the likelihood ratio test (LRT) between the best DIVALIKE and DIVALIKE + j models showed that DIVALIKE + j was strongly favoured ($P = 1.9 \times 10^{-5}$).

Table 1 Comparison of models of ancestral range estimation for the shearwaters. Models with and without the founder event parameter (j) are shown separately and for each case the model with the highest AICc weight is shown in bold.

Model	Dispersal	LnL	Parameters	*d*	*e*	*j*	AICc	AICc weight (%)
DEC	Unrestricted	-61.35	2	0.0155	0.0027	0	127.14	4.0
DEC	Restricted to areas connected by currents	-60.17	2	0.0193	0.0009	0	124.78	13.0
DIVALIKE	Unrestricted	-60.38	2	0.0230	0.0045	0	125.20	10.5
DIVALIKE	Restricted to areas connected by currents	-58.45	2	0.0295	0.0035	0	121.34	**72.5**
BAYAREALIKE	Unrestricted	-92.31	2	0.0441	0.1288	0	189.06	0.0
BAYAREALIKE	Restricted to areas connected by currents	-90.71	2	0.0648	0.1309	0	185.86	0.0
DEC + J	Unrestricted	-52.77	3	0.0057	$1x10^{-12}$	0.1020	112.46	1.9
DEC + J	Restricted to areas connected by currents	-49.97	3	0.0071	$1x10^{-12}$	0.1644	106.86	31.8
DIVALIKE + J	Unrestricted	-52.16	3	0.0081	$1x10^{-12}$	0.0874	111.24	3.6
DIVALIKE + J	Restricted to areas connected by currents	-49.33	3	0.0100	$1x10^{-12}$	0.1410	105.58	**60.3**
BAYAREALIKE + J	Unrestricted	-54.24	3	0.0049	$1x10^{-7}$	0.1375	115.40	0.4
BAYAREALIKE + J	Unrestricted	-52.76	3	0.0055	$1x10^{-7}$	0.1733	112.44	1.95

Under the best DIVALIKE + j model, the South Australia - New Zealand area showed the highest support as the ancestral region of shearwaters (marginal ML probability = 0.44), followed by the North and Tropical Pacific (0.33) (Figure 1 and Figure S4). The origin of *Ardenna* was also traced to the South Australia - New Zealand area (0.54). On the other hand, *Calonectris* had an unequivocal origin in the Northern Hemisphere (North Atlantic and North and Tropical Pacific = 0.45, North and Tropical Pacific = 0.45), whereas the ancestral area of the MRCA of *Puffinus* was estimated as either the North and Tropical Pacific (0.37), the South Australia - New Zealand area (0.27) or both (0.16).

3.3 | Phylogenetic Generalized Least Squares of Body Size

The PGLS analyses recovered several significant correlations (FDR < 0.05) between body size measures and the predictors (Figure 2a; Table S3). Mean body mass showed significant correlations with all predictors, suggesting that this trait is strongly

influenced by ecological factors. Overall, migratory strategy and latitudinal range were the best predictors, suggesting that body size in shearwaters is associated with movement capacity. Indeed, migratory strategy explained 75% of the variance in mean body mass (Figure S5a; long-distance migrants were the heaviest and sedentary/dispersive species the lightest) and latitudinal range explained 67% of the variance in body mass range (Figure S5b shows the positive correlation between body mass range and latitudinal range occupied by a species year-round). Breeding latitude was also a good predictor of mean body mass, with the strongest correlation recovered for maximum breeding latitude (Figure S5c; adjusted R^2=0.212). As shown in the phenogram of ancestral state reconstructions for body mass in Figure 2b, striking differences in body mass between sister clades are common in shearwaters, showing that body mass changes may be important during speciation. The ancestral state reconstruction of migratory behaviour showed that the MRCAs of *Calonectris* and *Ardenna* were most likely long-distance migrants (Figure S6; marginal ML probability = 0.94 and 0.86, respectively). Conversely, the MRCA of *Puffinus* was most likely either a short-distance migrant or a sedentary species (marginal ML probability = 0.47 and 0.37, respectively).

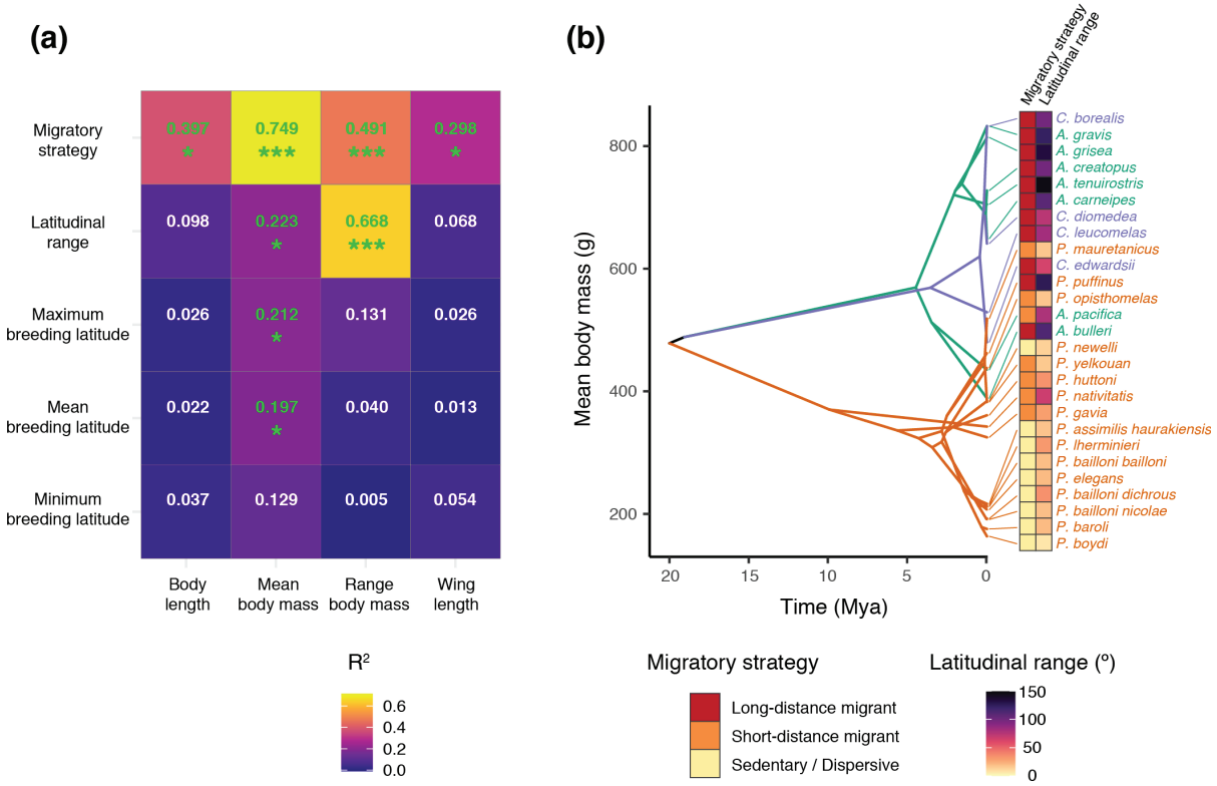


Figure 2 Migratory strategy and latitudinal range are the best predictors of body size. (a) Heatmap showing adjusted R^2 values for PGLS analyses of body size measures against the predictors. Positive correlations coefficients were recovered for each test. Numbers within each tile show the adjusted R^2 values and are coloured in green when significant after adjusting p-values by FDR correction for multiple testing (FDR: *** < 0.001 > ** < 0.01 > * < 0.05). (b) Phenogram of mean body mass constructed in PHYTOOLS showing abrupt differences in mean body mass between sister clades. Edge colours indicate the three genera: *Calonectris* (purple), *Ardenna* (green) and *Puffinus* (orange). Heatmaps next to the phenogram show the migratory strategy and the latitudinal range for each species.

3.4 | Effects of Life-history Traits on the Substitution Rate

Results from a previous study found that rates of mitochondrial DNA (mtDNA) evolution were slower for larger taxa in the Procellariiformes (Nunn and Stanley 1998), yet we did not find any significant correlations between the substitution rate of our PE-ddRAD-Seq dataset and the LHT (Table S4). This may have been influenced by the high variance in the substitution rates of our dataset. However, consistent with GC-biased gene conversion (gBGC): a recombination-associated mechanism that leads to the preferential fixation of G and C in AT/GC heterozygotes, we found that GC* had a positive significant correlation with the number of breeding pairs (correlation coefficient = 0.684, posterior probability = 0.94). These results are in agreement with

the hypothesis that the impact of gBGC is strongest in species with high population sizes (Weber et al. 2014).

3.5 | Genomic Divergence and Taxonomy

The FINERADSTRUCTURE analysis identified three major clusters corresponding to the three shearwater genera (Figure 3). Further subdivisions within each group largely supported the most recent shearwater phylogeny (Chapter I), and all the species and subspecies included in the study were recovered as unique clusters by the FINESTRUCTURE clustering algorithm (Lawson et al. 2012), except for *P. bailloni nicolae* and *P. bailloni dichrous*, *P. mauretanicus* and *P. yelkouan*, and *A. creatopus* and *A. carneipes*, that were in each case, clustered into a single population.

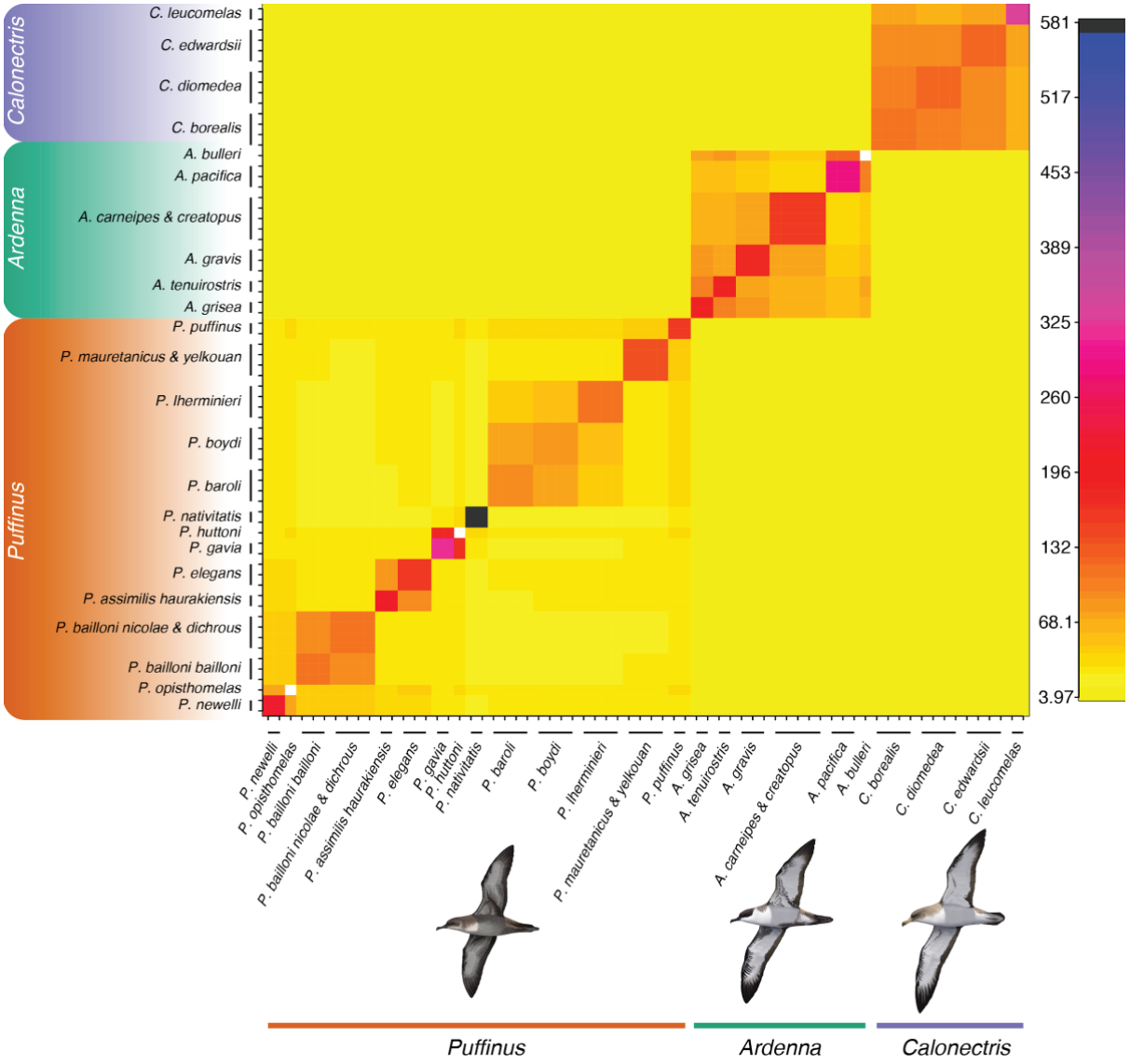


Figure 3 Clustered FINERADSTRUCTURE coancestry matrix based on 8,049 PE-ddRAD loci. Pairwise coancestry coefficients are colour coded from low (yellow) to high (black). Every name represents a discernible discrete cluster based on the pairwise matrix of coancestry coefficients, defined by a posterior probability > 0.9 in the FINESTRUCTURE tree. Note that the three major clusters represent the three genera and that most species and subspecies included in the study are recovered as unique clusters.

Overall, the distributions of genetic distances were consistent with the current taxonomy. However, the distributions of distances within and among species showed some overlap (Figure 4). The genetic distances between *A. creatopus* and *A. carneipes*, and between *P. mauretanicus* and *P. yelkouan*, were within the distribution of genetic distances within the same subspecies (first interval delimited in Figure 4b to ease qualitative comparison). In addition, the genetic distances between *P. boydi* and *P. baroli*, and between the different Atlantic *Calonectris* species were within the interval of genetic distances among different subspecies (second interval in Figure 4b).

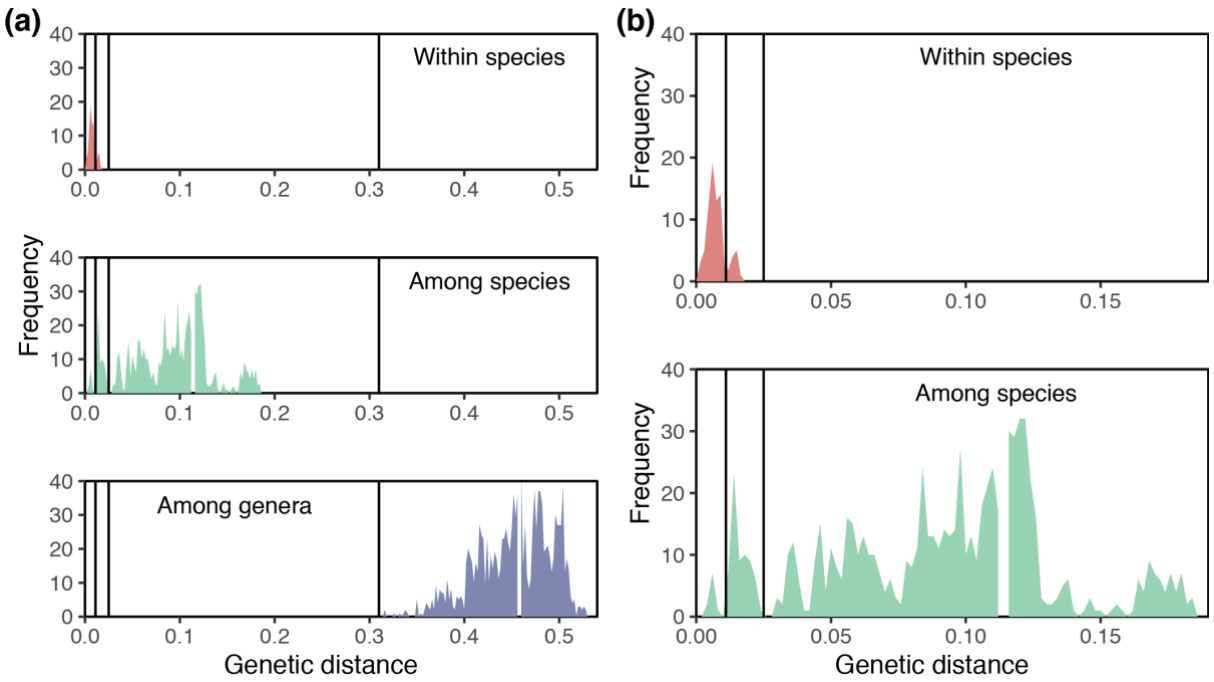


Figure 4 The distributions of genetic distances within and among species overlap. (a) Distribution of genetic distances at different taxonomic levels in the shearwaters (upper panel: within species, middle panel: among species and lower panel: among genera) according to current taxonomy. Vertical bars show proposed orientative limits to assist visualisation across panels for values within subspecies, among subspecies, among species and among genera from left to right. (b) Zoom-in of the distributions of genetic distances within and among species. Comparisons between *P. mauretanicus* and *P. yelkouan* and between *A. carneipes* and *A. creatopus* fall in the within subspecies range, and comparisons between *P. baroli* and *P. boydi* and between the three Atlantic *Calonectris* species fall in the among subspecies range.

4 | Discussion

This study presents a fundamental analysis of the drivers of diversification and speciation in a major group of seabirds, by constructing a MSC time-calibrated species tree and biogeographical analysis for shearwaters based on a fully resolved phylogeny.

This allowed us to explicitly explore the paleoceanographic events that may have driven the diversification of the group, as well as to infer their ancestral range using formal biogeographic analyses and evaluate the role of dispersal, vicariance and founder events in shearwater diversification. We also discuss the role of body size in shearwater diversification, and we consider potential ecological and evolutionary forces that may have shaped its evolution. Lastly, we used the evidence uncovered here to explain incongruences between the current taxonomy and the patterns of genomic divergence.

4.1 | Biogeographic History of Shearwaters

Our biogeographic analyses indicate that vicariance and founder events are probably the main mechanisms of speciation in shearwaters, as expected by their global distribution and high mobility. Unlike other Procellariiformes (Friesen et al. 2007b), sympatric speciation has not been described in shearwaters. Indeed, very few records of sister species inhabiting the same island exist in the wild and are limited to marginal overlaps between parapatric species (Navarro et al. 2009a). The biogeographic analyses suggest that shearwater dispersal is favoured by surface ocean currents; nevertheless, we cannot draw firm conclusions given the reduced differences in log-likelihood (< 3 units) between ancestral range estimation models with or without a dispersal matrix that restricted dispersal to areas connected by surface ocean currents (Table 1). Several studies have shown that winds are a major determinant of foraging ranges and migratory routes of seabirds, especially in the Procellariiformes (González-Solís et al. 2009; Weimerskirch et al. 2012). Winds are also a primary driver of surface ocean currents; hence, our study suggests that winds could also be an important determinant of species dispersal in the Procellariiformes.

Ancestral range estimation analyses inferred the South Australia - New Zealand area as the ancestral region of shearwaters with the highest support followed by the Northern and Tropical Pacific (Figure S4). The South Australia - New Zealand area is currently a hotspot of global seabird biodiversity (Croxall et al. 2012) and has the greatest number of shearwater species breeding in any single area (Dickson and Remsen 2013). On the other hand, the coast of California harbours the highest diversity of shearwater fossils from extinct species and some of the oldest ones (Miller 1961). These observations

suggest that the current biogeographic analyses represent a more probable hypothesis of the ancestral area of shearwaters than previous hypotheses, which suggested that the North Atlantic was the ancestral area based on the relatively rich shearwater fossil record in this area (Kuroda 1954; Austin 1996). The phylogenetic position of the oldest North Atlantic shearwater fossil species (*P. raemdonckii* and *P. arvernensis*) is still unclear (Olson 1985) and the age of *P. micraulax,* which was believed to be the oldest shearwater fossil species (lower Miocene, Hawthorne Formation, Florida) is uncertain (Chapter I). Thus, earlier suggestions of the North Atlantic as the ancestral area of shearwaters may have been misled by these uncertainties in the fossil record.

The MRCA of *Calonectris* had a North Pacific and North Atlantic distribution. Fossils of at least 5 species have been described from the North Atlantic dating back to ~14 Mya (Olson and Rasmussen 2001; Olson et al. 2008; Olson 2009), supporting this area as a speciation hotspot for the genus. However, considering the mobility of the genus and given that the MRCA of *Calonectris* was probably a long-distance migrant (Figure S6), we cannot eliminate the possibility that the regions where these fossils were found were not the breeding areas for the species. The estimated divergence time (~3.54 Mya) between the North Pacific and the North Atlantic clades is very similar to previous estimates based on mtDNA rates (~3.44 Mya; Gómez-Díaz et al. 2006) and suggests a vicariant event as the result of the gradual closure of the isthmus of Panama, as has been observed in other marine organisms (Lessios 2008).

Our analyses indicate that *Ardenna* had a South Australia - New Zealand origin and, thereafter, some lineages colonised the Southern Ocean (Figure 1), which disagrees with the North Atlantic origin of *Ardenna* proposed by (Austin 1996) based on the fossil record. Extant species are long-distance trans-equatorial migrants that can be locally common on North American and European coasts (Shaffer et al. 2006; Morrison 2009; Carey et al. 2014) and based on our ancestral state reconstruction, the MRCA of *Ardenna* was also most likely a long-distance migrant (Figure S6). We suggest that extinct taxa were also long-distance migrants breeding in the Southern Hemisphere, and that the fossils found in the North Atlantic likely represent birds that died during the non-breeding period.

The ancestors of *Puffinus* acquired the strongest diving adaptations of the three genera (Olson and Rasmussen 2001); these allow them to routinely dive to depths of 55 m (Shoji et al. 2016), providing advantages for reaching prey in the nutrient poor tropical and subtropical waters of the Pacific (inaccessible to most other tropical seabirds; Burger 2001), where the MRCA of *Puffinus* most probably originated based on the current ancestral range estimation analyses and the fossil record (Miller 1961). Although we could not obtain samples for *P. subalaris* from the Galapagos; in a previous study this species formed a clade with *P. nativitatis* (Austin et al. 2004), which further supports a Tropical Pacific origin of *Puffinus*. Most extant *Puffinus* species are short-distance migrants or dispersers that remain close to their breeding sites throughout the year (e.g. Ramos et al. 2020). Their lower dispersal compared to other shearwater genera may have reduced gene flow and promoted higher species richness. The population sizes of *Puffinus* species tend to be small and many had the highest EDGE scores (Table 2), which is a metric that identifies those threatened species that deserve particular attention because of their unique evolutionary history. Predation by invasive alien species is the main current threat for seabirds (Croxall et al. 2012) and is a principal cause of population declines among *Puffinus* species (Keitt et al. 2002; Sommer et al. 2009; Bonnaud et al. 2012). Enhanced by predation, intra- and inter-specific competition for nest sites plays an important role in limiting populations of small Procellariiformes, such as *Puffinus* shearwaters (Monteiro et al. 1996; Ramos et al. 1997). At sea, fisheries bycatch is also a main threat for *Puffinus* shearwaters (Bugoni et al. 2008; Cortés et al. 2017) and one that could drive some species to extinction unless conservation measures are put in place (Genovart et al. 2016). These are likely some of the main reasons why *Puffinus* shearwaters have the highest number of endangered species among the shearwaters.

Table 2 Number of breeding pairs, conservation status, evolutionary distinctness and EDGE scores for shearwater species and subspecies in the study.

Scientific name	Breeding pairs	IUCN Red List Status (GE score for EDGE calculation; IUCN 2019)	Evolutionary distinctness (ED)	EDGE Score
Ardenna bulleri	350,000	Vulnerable (2)	6.2	3.4
Ardenna carneipes	74,000	Near-threatened (1)	3.6	2.2
Ardenna creatopus	29,573	Vulnerable (2)	3.6	2.9
Ardenna gravis	6,800,000	Least concern (0)	4.4	1.7
Ardenna grisea	4,400,000	Near-threatened (1)	4.7	2.4
Ardenna pacifica	4,966,000	Least concern (0)	6.2	2.0
Ardenna tenuirostris	14,800,000	Least concern (0)	4.7	1.7
Calonectris borealis	252,500	Least concern (0)	5.3	1.8
Calonectris diomedea	182,000	Least concern (0)	5.3	1.8
Calonectris edwardsii	6,312	Near-threatened (1)	5.5	2.6
Calonectris leucomelas	1,000,000	Near-threatened (1)	7.5	2.8
Puffinus assimilis haurakiensis	10,000	Least concern (0)	3.3	1.5
Puffinus bailloni bailloni	4,080	Least concern (0)	2.6	1.3
Puffinus bailloni dichrous	60,500	Least concern (0)	2.3	1.2
Puffinus bailloni nicolae	120,000	Least concern (0)	2.3	1.2
Puffinus baroli	3,360	Vulnerable (2)	2.6	2.7
Puffinus boydi	5,000	Near-threatened (1)	2.6	2.0
Puffinus elegans	16,100	Least concern (0)	3.3	1.5
Puffinus gavia	100,000	Least concern (0)	4.8	1.8
Puffinus huttoni	114,000	Endangered (3)	4.8	3.8
Puffinus lherminieri	15,700	Near-threatened (1)	3.2	2.1
Puffinus mauretanicus	3,142	Critically endangered (4)	2.7	4.1
Puffinus nativitatis	50,000	Least concern (0)	10.6	2.4
Puffinus newelli	5,000	Critically endangered (4)	4.0	4.4
Puffinus opisthomelas	41,000	Near-threatened (1)	4.0	2.3
Puffinus puffinus	399,500	Least concern (0)	3.9	1.6
Puffinus yelkouan	22,928	Vulnerable (2)	2.7	2.7

Across the three genera, the Pliocene-Pleistocene boundary appeared as a period of high and rapid speciation and dispersal (Figure 1). For instance, *Puffinus* spread from the Pacific to the North Atlantic, the Southern Ocean, and the Indian Ocean during a rapid radiation. During the Cenozoic, the largest global sea-level changes and oscillations occurred in the Pliocene and Pleistocene (Zachos et al. 2001; Miller et al. 2005; Lisiecki and Raymo 2007). Neritic waters, which represent the main foraging grounds for medium and large shearwaters, especially during the breeding period, suffered a significant sudden reduction at the end of the Pliocene followed by extreme fluctuation and gradual reduction over the Pleistocene (De Boer et al. 2010). Global oceanographic changes, such as the end of permanent el Niño, the closure of the Isthmus of Panama and the formation of the Arctic ice cap (Fedorov et al. 2006; O'Dea

et al. 2016) may have been the cause of such reduction. This reduction has been hypothesised to be the cause of a three-fold increase in the extinction rate of megafauna associated with coastal habitats (O'Dea et al. 2007; Pimiento et al. 2017). In shearwaters, ~36% of the known extinct fossil species are from the Pliocene (Howard 1971; Olson 1985; Olson and Rasmussen 2001); together with the long stems in the three shearwater genera (Figure 1), this suggests that the late Pliocene extinction severely affected the group. The subsequent burst of speciation and dispersal was probably promoted by Pleistocene climatic shifts that probably promoted geographic splitting and bottlenecks (Avise and Walker 1998; Gómez-Díaz et al. 2006). An increase in diversification during this period has also been detected in other seabird groups such as penguins (Cole et al. 2019; Vianna et al. 2020) and even in deep sea species (Eilertsen and Malaquias 2015).

4.2 | Body Mass as a Key Phenotypic Trait

In the Procellariiformes, body mass is a trait closely related to fitness at the intraspecific level. For instance, body condition (body mass corrected by overall body size) of the progenitors affects breeding success in several species (Chastel et al. 1995; Tveraa et al. 1998; Barbraud and Chastel 1999). On the other hand, at the interspecific level, the drivers of body mass variation are poorly understood despite the high variation exhibited by the Procellariiformes (Nunn and Stanley 1998). Our results shed some new light on potential behavioural and distributional drivers that may be affecting body mass variation in the Procellariiformes, although caution must be taken at interpreting our findings that are merely correlational.

Migratory strategy was the best evaluated predictor for mean body mass (Figure 2). This correlation is likely twofold and additive. On the one hand, migratory species tend to be larger (i.e. longer wings in migratory species; Marchetti et al. 1995; Minias et al. 2015) as shown by the significant correlations of all the other body size measures with migratory strategy (Figure 2a). On the other hand, migratory behaviour allows the exploitation of additional resources leading to a higher accumulation of fat deposits. The weaker but significant correlation between migratory strategy and mean body mass when corrected by body surface (Table S3) supports the twofold and additive effect of this correlation.

Within an endothermic species or a group of closely related endothermic species, individuals inhabiting colder habitats and higher latitudes tend to be larger than those inhabiting warmer environments and lower latitudes (Bergmann 1848). This geographical pattern in body size holds for birds throughout the world at the intraspecific (Ashton 2002; Meiri and Dayan 2003) and interspecific level (Bergmann 1848) although the mechanisms responsible for the generation of this trend are subject to much debate (Ashton 2002; Meiri 2011). In shearwaters, this pattern has also been shown to apply to intraspecific body size variation in the Streaked Shearwater (*Calonectris leucomelas*; (Yamamoto et al. 2016). Among shearwater species, we also found a positive significant correlation between breeding latitude and mean body mass (Figure 2 and Figure S5C), despite previous studies that have shown that conformity to Bergmann's Rule tends to be weaker for migratory and enclosed nesting species (Meiri and Dayan 2003; Mainwaring and Street 2019). The correlation was strongest between maximum breeding latitude and mean body mass corrected by body surface ($R^2 = 0.387$; Table S3), suggesting that heavier bodies, independent of body size, might provide a better adaptation to thrive in higher and colder latitudes. However, these correlations could also be indirectly driven by a higher tendency of species living in higher latitudes to be migratory and/or by differences in diving behaviour, which could not be explored in this study.

The strong association between body mass range and latitudinal range is likely twofold. On the one hand, exploiting larger foraging areas may allow for ecological segregation between sexes and size dimorphism (Felipe et al. 2019). Indeed, ecological segregation has been shown to be the most likely cause of size dimorphism in other Procellariiformes (González-Solís 2004). On the other hand, larger body mass differences may arise between individuals that are more efficient at exploiting the available resources compared to those that are less efficient. This might provide the substrate for sexual selection to act on body mass. Higher body condition has been associated with higher breeding success in several species of Procellariiformes (Chastel et al. 1995; Barbraud and Chastel 1999; Barbraud and Weimerskirch 2005).

4.3 | Considerations of Shearwater Taxonomy

Species delimitation in shearwaters is a challenging and controversial topic, partly due to their remarkably similar morphology (Austin et al. 2004). Conflict has arisen amongst morphological studies, and analyses based on genetic data (i.e., mtDNA and microsatellites), and also between different genetic datasets (Kuroda 1954; Austin 1996; Gómez-Díaz et al. 2009; Genovart et al. 2013). In addition, despite being a promising trait for species delimitation, analyses of shearwater vocalizations are limited (Bretagnolle 1996). Genome-wide datasets have the potential to provide fine-scale population structure and genomic divergence estimates that can inform taxonomy. Despite the high resolution of our PE-ddRAD dataset, FINERADSTRUCTURE analysis showed no structure between two species pairs, *P. mauretanicus* and *P. yelkouan*, and *A. creatopus* and *A. carneipes* (Figure 3). Furthermore, although we do not consider there to be a genetic cutoff for species-level divergence, the genetic divergence between these recently diverged species were the lowest amongst any pair of species and overlapped with the genetic divergences observed between individuals of the same subspecies (Figure 4). *P. mauretanicus* and *P. yelkouan* were granted species status based on morphological, osteological and reciprocal monophyly using cytochrome b sequences (Heidrich et al. 1998; Sangster et al. 2002). However, more recently, a lack of correspondence at the individual level was found between phenotypic characters, stable isotope analyses, nuclear and mtDNA, and was attributed to admixture between the two species (Genovart et al. 2012; Militão et al. 2014). *A. creatopus* and *A. carneipes* are widely considered as two different species in taxonomic checklists (Carboneras and Bonan 2019; Gill et al. 2020), but some authors have argued that they should be considered conspecific based on the lack of uniform differentiation in colour and size (Bourne 1962) and on low mtDNA differentiation (Penhallurick and Wink 2004). These species pairs differ in plumage colouration and body size, which are known to be labile traits even within species of shearwaters. Dark and pale phases can be found within a single species (i.e., *A. pacifica*) and some species exhibit a continuum from pale to dark (i.e., *P. mauretanicus*). Body size covaries with migratory behaviour (see previous section), can be under selection (Navarro et al. 2009b), and thus could evolve rapidly under strong selection pressures. In addition to the aforementioned species pairs, other

shearwater species showed weak patterns of population structure and genetic distances within the interval among different subspecies: *P. boydi* and *P. baroli*, and the three Atlantic *Calonectris* species. These species complexes are the subject of ongoing taxonomic debate (Sangster et al. 2005; Gómez-Díaz et al. 2009; Olson 2010; Genovart et al. 2013). As a final consideration, our genomic data, together with ongoing taxonomic debate, suggest that these taxa should not be granted species status. Future studies should use species delimitation approaches combining genomic data with a thorough morphological revaluation including a detailed evaluation of vocalisations. Further research is also required to include the taxa that could not be sampled during this study, particularly taxa from the tropical Pacific that breed in remote islands and have very limited distributions and low population sizes.

Author Contributions

R.T.C., H.F.J., J.G., V.B., A.J.W. and J.F. contributed to data collection and all authors contributed to study design. J.F. performed DNA extractions, processed and analysed the data and wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Acknowledgements

We thank Gary Nunn, Jeremy J. Austin, Chris Gaskin, Juan E. Martínez-Gómez, and Maite Louzao for collecting tissue and blood samples from the field and/or providing tissue samples from museum skins for this study. We would also like to thank the many institutions that provided tissue loans for this research: Smithsonian National Museum of Natural History, Louisiana State University Museum of Natural Science, University of Washington Burke Museum, American Museum of Natural History, the University of Kansas Biodiversity Institute, and the Muséum National d'Histoire Naturelle. We thank the pertinent authorities for issuing the permits needed for this work. Finally, we would like to thank Josephine R. Paris for her comments on an early version of this manuscript and Martí Franch for his wonderful shearwater illustrations. Research funding was provided by the Fundación BBVA (program 'Ayudas a Equipos de Investigación Científica 2017', project code 062_17 to M.R.). Any use of trade, product, or firm names

is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Supplementary Material

Supplementary Material for this chapter may be found in Appendix II.

5 | References

Ashton K.G. 2002. Patterns of within-species body size variation of birds: strong evidence for Bergmann's rule. Glob. Ecol. Biogeogr. 11:505–523.

Austin J.J. 1996. Molecular phylogenetics of Puffinus shearwaters: preliminary evidence from mitochondrial cytochrome b gene sequences. Mol. Phylogenet. Evol. 6:77–88.

Austin J.J., Bretagnolle V., Pasquet E. 2004. A global molecular phylogeny of the small Puffinus shearwaters and implications for systematics of the Little-Audubon's Shearwater complex. Auk. 121:647–864.

Avise J.C., Walker D.E. 1998. Pleistocene phylogeographic effects on avian populations and the speciation process. Proceedings of the Royal Society of London. Series B: Biological Sciences. 265:457–463.

Barbraud C. 2000. Natural selection on body size traits in a long-lived bird, the snow petrel Pagodroma nivea. Journal of Evolutionary Biology 13: 81-88.

Barbraud C., Chastel O. 1999. Early body condition and hatching success in the snow petrel Pagodroma nivea. Polar Biol. 21:1–4.

Barbraud C., Weimerskirch H. 2005. Environmental conditions and breeding experience affect costs of reproduction in blue petrels. Ecology. 86:682–692.

Benton M.J. 2009. The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. Science. 323:728–732.

Bergmann C. 1848. Über die Verhältnisse der Wärmeökonomie der Thiere zu ihrer Größe.

Bierne N., Bonhomme F., David P. 2003. Habitat preference and the marine-speciation paradox. Proc. Biol. Sci. 270:1399–1406.

Billerman S. M., Keeney B. K., Rodewald P. G., Schulenberg T. S. 2020. Birds of the World. Cornell Laboratory of Ornithology, Ithaca, NY, USA.

BirdLife International. 2020. IUCN Red List for birds. Downloaded from http://www.birdlife.org on 17/08/2020.

Bonnaud E., Berger G., Bourgeois K., Legrand J., Vidal E. 2012. Predation by cats could lead to the extinction of the Mediterranean endemic Yelkouan Shearwater Puffinus yelkouan at a major breeding site: Shearwaters threatened by cats. Ibis . 154:566–577.

Bouckaert R.R. 2010. DensiTree: Making sense of sets of phylogenetic trees. Bioinformatics. 26:1372–1373.

Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., du Plessis L., Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu C.-H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15:e1006650.

Bourne W.R.P. 1962. Handbook of North American Birds.Yale University Press, New Haven.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Bugoni L., Mancini P.L., Monteiro D.S., Nascimento L., Neves T.S. 2008. Seabird bycatch in the Brazilian pelagic longline fishery and a review of capture rates in the southwestern Atlantic Ocean. Endanger. Species Res. 5:137–147.

Burger A.E. 2001. Diving depths of shearwaters. The Auk. 118: 755-759.

Butlin R., Debelle A., Kerth C., Snook R.R., Beukeboom L.W., Castillo R.F.C., Diao W., Maan M.E., Paolucci S., Weissing F.J., Others. 2012. What do we need to know about speciation? Trends Ecol. Evol. 27:27–39.

Carboneras C., Bonan, A. 2019. Petrels, Shearwaters (Procellariidae). In: del Hoyo, J., Elliott, A., Sargatal, J., Christie, D.A., de Juana, E. (eds.). Handbook of the Birds of the World Alive. Lynx Edicions, Barcelona.

Carey M.J., Phillips R.A., Silk J.R.D., Shaffer S.A. 2014. Trans-equatorial migration of Short-tailed Shearwaters revealed by geolocators. Emu - Austral Ornithology. 114:352–359.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. Mol. Ecol. 22:4369–4383.

Chastel O., Weimerskirch H., Jouventin P. 1995. Body Condition and Seabird Reproductive Performance: A Study of Three Petrel Species. Ecology. 76:2240–2246.

Cole T.L., Ksepka D.T., Mitchell K.J., Tennyson A.J.D., Thomas D.B., Pan H., Zhang G., Rawlence N.J., Wood J.R., Bover P., Bouzat J.L., Cooper A., Fiddaman S.R., Hart T., Miller G., Ryan P.G., Shepherd L.D., Wilmshurst J.M., Waters J.M. 2019. Mitogenomes Uncover Extinct Penguin Taxa and Reveal Island Formation as a Key Driver of Speciation. Mol. Biol. Evol. 36:784–797.

Cortés V., Arcos J.M., González-Solís J. 2017. Seabirds and demersal longliners in the northwestern Mediterranean: factors driving their interactions and bycatch rates. Mar. Ecol. Prog. Ser. 565:1–16.

Coulson J. 2002. Colonial breeding in seabirds. Biology of marine birds. 87-113.

Coyne J.A., Orr H.A.. 2004. Speciation (Vol. 37). Sinauer Associates Sunderland, MA.

Croxall J.P., Butchart S.H.M., Lascelles B., Stattersfield A.J., Sullivan B., Symes A., Taylor P. 2012. Seabird conservation status, threats and priority actions: a global assessment. Bird Conservation International. 22:1–34.

Cuevas-Caballé C., Ferrer-Obiol, J., Genovart, M., Rozas, J., González-Solís, J., Riutort, M. 2019. Conservation genomics applied to the Balearic shearwater. G10K-VGP/EBP 2019. doi: 10.13140/RG.2.2.15751.21923.

Cutter A.D. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Mol. Phylogenet. Evol. 69:1172–1185.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R., 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

De Boer B., van de Wal R.S.W., Bintanja R., Lourens L.J., Tuenter E. 2010. Cenozoic global ice-volume and temperature simulations with 1-D ice-sheet models forced by benthic δ18O records. Ann. Glaciol. 51:23–33.

De Felipe F., Reyes-González, J. M., Militão, T., Neves, V. C., Bried, J., Oro, D., Ramos R., González-Solís, J. 2019. Does sexual segregation occur during the non-breeding period? A comparative analysis in the spatial ecology of three Calonectris shearwaters. Ecology and Evolution, 9:10145-10162.

Dickson E.C., Remsen J.R. 2013. The Howard and Moore complete checklist of the birds of the World. Vol. 1. Non-passerines. Aves Press, Eastbourne, UK.

Eilertsen M.H., Malaquias M.A.E. 2015. Speciation in the dark: diversification and biogeography of the deep-sea gastropod genus Scaphander in the Atlantic Ocean. J. Biogeogr. 42:843–855.

Estandia A. 2019. Genome-wide phylogenetic reconstruction for Procellariiform seabirds is robust to molecular rate variation. Masters Thesis, Durham University.

Fedorov A.V., Dekens P.S., McCarthy M., Ravelo A.C., deMenocal P.B., Barreiro M., Pacanowski R.C., Philander S.G. 2006. The Pliocene paradox (mechanisms for a permanent El Niño). Science. 312:1485–1489.

Felipe F.D., De Felipe F., Reyes-González J.M., Militão T., Neves V.C., Bried J., Oro D., Ramos R., González-Solís J. 2019. Does sexual segregation occur during the nonbreeding period? A comparative analysis in spatial and feeding ecology of three Calonectris shearwaters. Ecology and Evolution. 9:10145–10162.

Fišer C., Robinson C.T., Malard F. 2018. Cryptic species as a window into the paradigm shift of the species concept. Mol. Ecol. 27:613–635.

Friesen V.L., Burg T.M., McCoy K.D. 2007a. Mechanisms of population differentiation in seabirds: Invited review. Mol. Ecol. 16:1765–1785.

Friesen V.L., Smith A.L., Gomez-Diaz E., Bolton M., Furness R.W., González-Solís J., Monteiro L.R. 2007b. Sympatric speciation by allochrony in a seabird. Proceedings of the National Academy of Sciences. 104:18589–18594.

Genovart M., Arcos J.M., Álvarez D., McMinn M., Meier R., Wynn R., Guilford T., Oro D. 2016. Demography of the critically endangered Balearic shearwater: the impact of fisheries and time to extinction. J. Appl. Ecol.

Genovart M., Juste J., Contreras-Díaz H., Oro D. 2012. Genetic and phenotypic differentiation between the critically endangered balearic shearwater and neighboring colonies of its sibling species. J. Hered. 103:330–341.

Genovart M., Thibault J.C., Igual J.M., Bauzà-Ribot M. del M., Rabouam C., Bretagnolle V. 2013. Population Structure and Dispersal Patterns within and between Atlantic and Mediterranean Populations of a Large-Range Pelagic Seabird. PLoS One. 8: e70711.

Gill F., Donsker, D., Rasmussen, P. 2020. IOC World Bird List (v10. 1). doi : 10.14344/IOC.ML.10.1.

Gómez-Díaz E., González-Solís J., Peinado M.A. 2009. Population structure in a highly pelagic seabird, the Cory's shearwater Calonectris diomedea: An examination of genetics, morphology and ecology. Mar. Ecol. Prog. Ser. 382:197–209.

Gómez-Díaz E., González-Solís J., Peinado M.A., Page R.D.M. 2006. Phylogeography of the Calonectris shearwaters using molecular and morphometric data. Mol. Phylogenet. Evol. 41:322–332.

González-Solís J. 2004. Sexual size dimorphism in northern giant petrels: ecological correlates and scaling. Oikos. 105: 247-254.

González-Solís J., Felicísimo A., Fox J.W., Afanasyev V., Kolbeinsson Y., Muñoz J. 2009. Influence of sea surface winds on shearwater migration detours. Mar. Ecol. Prog. Ser. 391:221–230.

González-Solís J., Croxall, J. P., Oro, D., Ruiz, X. 2007. Trans-equatorial migration and mixing in the wintering areas of a pelagic seabird. Frontiers in Ecology and the Environment, 5:297-301.

Heidrich P., Amengual J., Wink M. 1998. Phylogenetic relationships in Mediterranean and North Atlantic shearwaters (Aves: Procellariidae) based on nucleotide sequences of mtDNA. Biochem. Syst. Ecol. 26:145–170.

Heled J., Bouckaert R.R. 2013. Looking for trees in the forest: summary tree from posterior samples. BMC Evol. Biol. 13:221.

Howard H. 1971. Pliocene avian remains from Baja California. Los Angeles County Museum of Natural History.

Isaac N.J.B., Turvey S.T., Collen B., Waterman C., Baillie J.E.M. 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS One. 2:e296.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature. 491:444–448.

Kawakami K., Eda M., Izumi H., Horikoshi K., Suzuki H. 2018. Phylogenetic position of endangered Puffinus lherminieri bannermani. Ornithol. Sci. 17:11–18.

Keitt B.S., Wilcox C., Tershy B.R., Croll D.A., Donlan C.J. 2002. The effect of feral cats on the population viability of black-vented shearwaters (Puffinus opisthomelas) on Natividad Island, Mexico. Anim. Conserv. 5:217–223.

Knaus B.J., Grünwald N.J. 2017. vcfr: a package to manipulate and visualize variant call format data in R. Mol. Ecol. Resour. 17:44–53.

Kopp M. 2010. Speciation and the neutral theory of biodiversity. BioEssays. 32:564–570.

Kumar S., Stecher G., Suleski M., Hedges S.B. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34:1812–1819.

Kuroda N. 1954. On the Classification and Phylogeny of the Order Tubinares, Particularly the Shearwaters (Puffinus), with Special Considerations [ie Considerations] on Their Osteology and Habit Differentiation.

Landis M.J., Matzke N.J., Moore B.R., Huelsenbeck J.P. 2013. Bayesian analysis of biogeography when the number of areas is large. Syst. Biol. 62:789–804.

Lartillot N., Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. 28:729–744.

Lawson D.J., Hellenthal G., Myers S., Falush D. 2012. Inference of population structure using dense haplotype data. PLoS Genet. 8:e1002453.

Leaché A.D., Zhu T., Rannala B., Yang Z. 2018. The Spectre of Too Many Species. Syst. Biol. 68:168–181.

Lessios H.A. 2008. The Great American Schism: Divergence of Marine Organisms After the Rise of the Central American Isthmus. .

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 27:2987–2993.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Retrieved from http://arxiv.org/abs/1303.3997.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and samtools. Bioinformatics. 25:2078–2079.

Lisiecki L.E., Raymo M.E. 2007. Plio–Pleistocene climate evolution: trends and transitions in glacial cycle dynamics. Quat. Sci. Rev. 26:56–69.

Lombal A.J., Wenner T.J., Lavers J.L., Austin J.J. 2018. Genetic divergence between colonies of Flesh-footed Shearwater Ardenna carneipes exhibiting different foraging strategies. Conservation Genetics. 19:27-41.

Mainwaring M.C., Street S.E. 2019. Conformity to Bergmann's rule in birds depends on nest design and migration. bioRxiv, 686972, doi: 10.1101/686792.

Malinsky M., Trucchi E., Lawson D.J., Falush D. 2018. RADpainter and fineRADstructure: Population Inference from RADseq Data. Mol. Biol. Evol. 35:1284–1290.

Marchetti K., Price T., Richman A. 1995. Correlates of Wing Morphology with Foraging Behaviour and Migration Distance in the Genus *Phylloscopus*. J. Avian Biol. 26:177–181.

Martínez-Gómez J.E., Matías-Ferrer N., Sehgal R.N.M., Escalante P. 2015. Phylogenetic placement of the critically endangered Townsend's Shearwater (*Puffinus auricularis auricularis*): evidence for its

conspecific status with Newell's Shearwater (*Puffinus a. newelli*) and a mismatch between genetic and phenotypic differentiation. J. Ornithol. 156:1025–1034.

Maruki T., Lynch M. 2015. Genotype-Frequency Estimation from High-Throughput Sequencing Data. Genetics. 201:473–486.

Maruki T., Lynch M. 2017. Genotype Calling from Population-Genomic Sequencing Data. G3: Genes|Genomes|Genetics. 7:1393–1404.

Matzke N.J. 2013. BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. R package, version 0. 2. 1:2013.

Meiri S., Dayan T. 2003. On the validity of Bergmann's rule. J. Biogeogr. 30:331–351.

Meiri S. 2011. Bergmann's Rule–what's in a name?. Global Ecology and Biogeography 20:203-207.

Militão T., Gómez-Díaz E., Kaliontzopoulou A., González-Solís J. 2014. Comparing multiple criteria for species identification in two recently diverged seabirds. PLoS One. 9:e115650.

Miller K.G., Kominz M.A., Browning J.V., Wright J.D., Mountain G.S., Katz M.E., Sugarman P.J., Cramer B.S., Christie-Blick N., Pekar S.F. 2005. The Phanerozoic record of global sea-level change. Science. 310:1293–1298.

Miller L. 1961. Birds from the Miocene of Sharktooth Hill, California. Condor. 63:399–402.

Minias P., Meissner W., Włodarczyk R., Ożarowska A., Piasecka A., Kaczmarek K., Janiszewski T. 2015. Wing shape and migration in shorebirds: a comparative study. Ibis. 157:528–535.

Milot E., Weimerskirch, H., Bernatchez, L. 2008. The seabird paradox: dispersal, genetic structure and population dynamics in a highly mobile, but philopatric albatross species. Molecular Ecology. 17: 658-1673.

Monteiro L. R., Ramos, J. A., Furness, R. W. 1996. Past and present status and conservation of the seabirds breeding in the Azores Archipelago. Biological Conservation. 78: 319–328.

Morrison R.I.G. 2009. Migration and Winter Ranges of Birds in Greenland, by Peter Lyngs. ARCTIC. 59.

Moura A.E., Nielsen S.C.A., Vilstrup J.T., Moreno-Mayar J.V., Gilbert M.T.P., Gray H.W.I., Natoli A., Möller L., Hoelzel A.R. 2013. Recent diversification of a marine genus (Tursiops spp.) tracks habitat preference and environmental change. Syst. Biol. 62:865–877.

Mundry R. 2014. Statistical Issues and Assumptions of Phylogenetic Generalized Least Squares. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology. 131–153.

Munilla I., Genovart, M., Paiva, V. H., Velando, A. 2016. Colony foundation in an oceanic seabird. PloS one. 11: e0147222.

Navarro J., Forero M.G., González-Solís J., Igual J.M., Bécares J., Hobson K.A. 2009a. Foraging segregation between two closely related shearwaters breeding in sympatry. Biology Letters. 5:545–548.

Navarro J., Kaliontzopoulou A., González-Solís J. 2009b. Sexual dimorphism in bill morphology and feeding ecology in Cory's shearwater (Calonectris diomedea). Zoology. 112:128–138.

Nosil P. 2012. Ecological Speciation. Oxford University Press.

Nunn G.B., Stanley S.E. 1998. Body size effects and rates of cytochrome b evolution in tube-nosed seabirds. Mol. Biol. Evol. 15:1360–1371.

O'Dea A., Jackson J.B.C., Fortunato H., Smith J.T., D'Croz L., Johnson K.G., Todd J.A. 2007. Environmental change preceded Caribbean extinction by 2 million years. Proc. Natl. Acad. Sci. U. S. A. 104:5501–5506.

O'Dea A., Lessios H.A., Coates A.G., Eytan R.I., Restrepo-Moreno S.A., Cione A.L., Collins L.S., de Queiroz A., Farris D.W., Norris R.D., Stallard R.F., Woodburne M.O., Aguilera O., Aubry M.-P., Berggren W.A., Budd A.F., Cozzuol M.A., Coppard S.E., Duque-Caro H., Finnegan S., Gasparini G.M., Grossman E.L., Johnson K.G., Keigwin L.D., Knowlton N., Leigh E.G., Leonard-Pingel J.S., Marko

P.B., Pyenson N.D., Rachello-Dolmen P.G., Soibelzon E., Soibelzon L., Todd J.A., Vermeij G.J., Jackson J.B.C. 2016. Formation of the Isthmus of Panama. Sci Adv. 2:e1600883.

Olson S.L. 1985. The fossil record of birds. Avian Biology.

Olson SL. 2008. A new species of shearwater of the genus Calonectris (Aves: Procellariidae) from a middle Pleistocene deposit on Bermuda. Proceedings of the Biological Society of Washington. 121:398-409.

Olson S.L. 2009. A new diminutive species of shearwater of the genus Calonectris (Aves: Procellariidae) from the Middle Miocene Calvert Formation of Chesapeake Bay. Proceedings of the Biological Society of Washington. 122:466–470.

Olson S.L. 2010. Stasis and turnover in small shearwaters on Bermuda over the last 400 000 years (Aves: Procellariidae: Puffinus lherminieri group): Biol. J. Linn. Soc. Lond. 99:699–707.

Olson S.L., Rasmussen P.C. 2001. Miocene and Pliocene birds from the Lee Creek Mine, North Carolina. Smithson. Contrib. Paleobiol. 90:233–365.

Onley D., Scofield P. 2013. Albatrosses, Petrels and Shearwaters of the World. Bloomsbury Publishing.

Orme D., Freckleton R., Thomas G., Petzoldt T., Fritz S., Others. 2013. The caper package: comparative analysis of phylogenetics and evolution in R. R package version. 5:1–36.

Pagel M. 1999. The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies. Syst. Biol. 48:612–622.

Palumbi S.R. 1994. Genetic divergence, reproductive isolation, and marine speciation. Annu. Rev. Ecol. Syst. 25:547–572.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Paris J.R., Stevens J.R., Catchen J.M. 2017. Lost in parameter space: a road map for Stacks. Methods in Ecology and Evolution. 8:1360–1373.

Penhallurick J., Wink M. 2004. Analysis of the taxonomy and nomenclature of the Procellariiformes based on complete nucleotide sequences of the mitochondrial cytochrome b gene. Emu. 104:125–147.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 7: e37135.

Pimiento C., Griffin J.N., Clements C.F., Silvestro D., Varela S., Uhen M.D., Jaramillo C. 2017. The Pliocene marine megafauna extinction and its impact on functional diversity. Nat Ecol Evol. 1:1100–1106.

Price T. 2008. Speciation in birds. Roberts and Co.

Pyle P., Welch A.J., Fleischer R.C. 2011. A New Species of Shearwater (Puffinus) Recorded from Midway Atoll, Northwestern Hawaiian Islands. Condor. 113:518–527.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. 67:901–904.

Ramos J.A., Monteiro L.R., Sola E., Moniz Z. 1997. Characteristics and Competition for Nest Cavities in Burrowing Procellariiformes. Condor. 99:634–641.

Ramos R., Paiva V. H., Zajková Z., Precheur C., Fagundes A. I., Jodice P. G., Mackin W., Zino F., Bretagnolle, González-Solís, J. 2020. Spatial ecology of closely related taxa: the case of the little shearwater complex in the North Atlantic Ocean. Zoological Journal of the Linnean Society. zlaa045.

Rayner M.J., Hauber M.E., Steeves T.E., Lawrence H. a., Thompson D.R., Sagar P.M., Bury S.J., Landers T.J., Phillips R. A., Ranjard L., Shaffer S. a. 2011. Contemporary and historical separation of transequatorial migration between genetically distinct seabird populations. Nat. Commun. 2:332.

Ree R.H., Sanmartín I. 2018. Conceptual and statistical problems with the DEC+ J model of founder-event speciation and its comparison with DEC via model selection. J. Biogeogr. 45:741–749.

Ree R.H., Smith S.A. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Syst. Biol. 57:4–14.

Revell L.J. 2010. Phylogenetic signal and linear regression on species data: Phylogenetic regression. Methods Ecol. Evol. 1:319–329.

Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.

Rochette N.C., Rivera-Colón A.G., Catchen J.M. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28:4737–4754.

Provencher J., Raine A.F., Ramírez F., Rodríguez B., Ronconi R.A., Taylor R.S., Bonnaud E., Borrelle S.B., Cortés V., Descamps S., Friesen V.L., Genovart M., Hedd A., Hodum P., Humphries G.R.W., Le Corre M., Lebarbenchon C., Martin R., Melvin E.F., Montevecchi W.A., Pinet P., Pollet I.L., Ramos R., Russell J.C., Ryan P.G., Sanz-Aguilar A., Spatz D.R., Travers M., Votier S.C., Wanless R.M., Woehler E., Chiaradia A. 2019. Future directions in conservation research on petrels and shearwaters. Frontiers in Marine Science, 6, 94.

Ronquist F. 1997. Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography. Syst. Biol. 46:195–203.

Sangster G., Collinson J.M., Helbig A.J., Knox A.G., Parkin D.T. 2005. Taxonomic recommendations for British birds: third report. Ibis . 147:821–826.

Sangster G., Knox A.G., Helbig A.J., Parkin D.T. 2002. Taxonomic recommendations for European birds. Ibis . 144:153–159.

Schluter D., Pennell M.W. 2017. Speciation gradients and the distribution of biodiversity. Nature. 546:48–55.

Seddon N., Botero C.A., Tobias J.A., Dunn P.O., Macgregor H.E.A., Rubenstein D.R., Uy J.A.C., Weir J.T., Whittingham L.A., Safran R.J. 2013. Sexual selection accelerates signal evolution during speciation in birds. Proc. Biol. Sci. 280:20131065.

Shaffer S.A., Tremblay Y., Weimerskirch H., Scott D., Thompson D.R., Sagar P.M., Moller H., Taylor G.A., Foley D.G., Block B.A., Costa D.P. 2006. Migratory shearwaters integrate oceanic resources across the Pacific Ocean in an endless summer. Proc. Natl. Acad. Sci. U. S. A. 103:12799–12802.

Shoji A., Dean B., Kirk H., Freeman R., Perrins C.M., Guilford T. 2016. The diving behaviour of the Manx Shearwater Puffinus puffinus. Ibis . 158:598–606.

Silva M.C., Matias R., Wanless R.M., Ryan P.G., Stephenson B.M., Bolton M., Ferrand N., Coelho M.M. 2015. Understanding the mechanisms of antitropical divergence in the seabird White-faced Storm-petrel (Procellariiformes: Pelagodroma marina) using a multilocus approach. Mol. Ecol. 24:3122–3137.

Simpson G.G. 1953. The major features of evolution. No. 575 S55.

Sommer E., Bell M., Bradfield P., Dunlop K., Gaze P., Harrow G., McGahan P., Morrisey M., Walford D., Cuthbert R. 2009. Population trends, breeding success and predation rates of Hutton's shearwater (Puffinus huttoni): a 20 year assessment. Notornis. 56:144–153.

Stange M., Sánchez-Villagra M.R., Salzburger W., Matschiner M. 2018. Bayesian Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports Miocene Closure of the Panamanian Isthmus. Syst. Biol. 67:681–699.

Storey A. S., Lien, J. 1985. Development of the first North American colony of Manx Shearwaters. The Auk, 395-401.

Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. U. S. A. 114:1607–1612.

Sutton P. 2001. Detailed structure of the Subtropical Front over Chatham Rise, east of New Zealand. J. Geophys. Res. 106:31045–31056.

Tveraa T., Sether B., Aanes R., Erikstad K.E. 1998. Regulation of food provisioning in the Antarctic petrel; the importance of parental body condition and chick body mass. Journal of Animal Ecology. 67:699–704.

Vargas P., Zardoya R. 2014. The tree of life. Sunderland, MA.

Vianna J.A., Fernandes F.A.N., Frugone M.J., Figueiró H.V., Pertierra L.R., Noll D., Bi K., Wang-Claypool C.Y., Lowther A., Parker P., Le Bohec C., Bonadonna F., Wienecke B., Pistorius P., Steinfurth A., Burridge C.P., Dantas G.P.M., Poulin E., Simison W.B., Henderson J., Eizirik E., Nery M.F., Bowie R.C.K. 2020. Genome-wide analyses reveal drivers of penguin diversification. Proc. Natl. Acad. Sci. U. S. A.

Weber C.C., Boussau B., Romiguier J., Jarvis E.D., Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol. 15:549.

Weimerskirch H., Delord K., Guitteaud A., Phillips R.A., Pinet P. 2015. Extreme variation in migration strategies between and within wandering albatross populations during their sabbatical year, and their fitness consequences. Sci. Rep. 5:1–7.

Weimerskirch H., Jouventin P., Mougin J.L., Stahl J.C., Van B.M. 1985. Banding recoveries and the dispersal of seabirds breeding in French Austral and Antarctic Territories. Emu. 85:22–33.

Weimerskirch H., Louzao M., de Grissac S., Delord K. 2012. Changes in wind pattern alter albatross distribution and life-history traits. Science. 335:211–214.

Yamamoto T., Kohno H., Mizutani A., Yoda K., Matsumoto S., Kawabe R., Watanabe S., Oka N., Sato K., Yamamoto M., Sugawa H., Karino K., Shiomi K., Yonehara Y., Takahashi A. 2016. Geographical variation in body size of a pelagic seabird, the streaked shearwater Calonectris leucomelas. J. Biogeogr. 43:801–808.

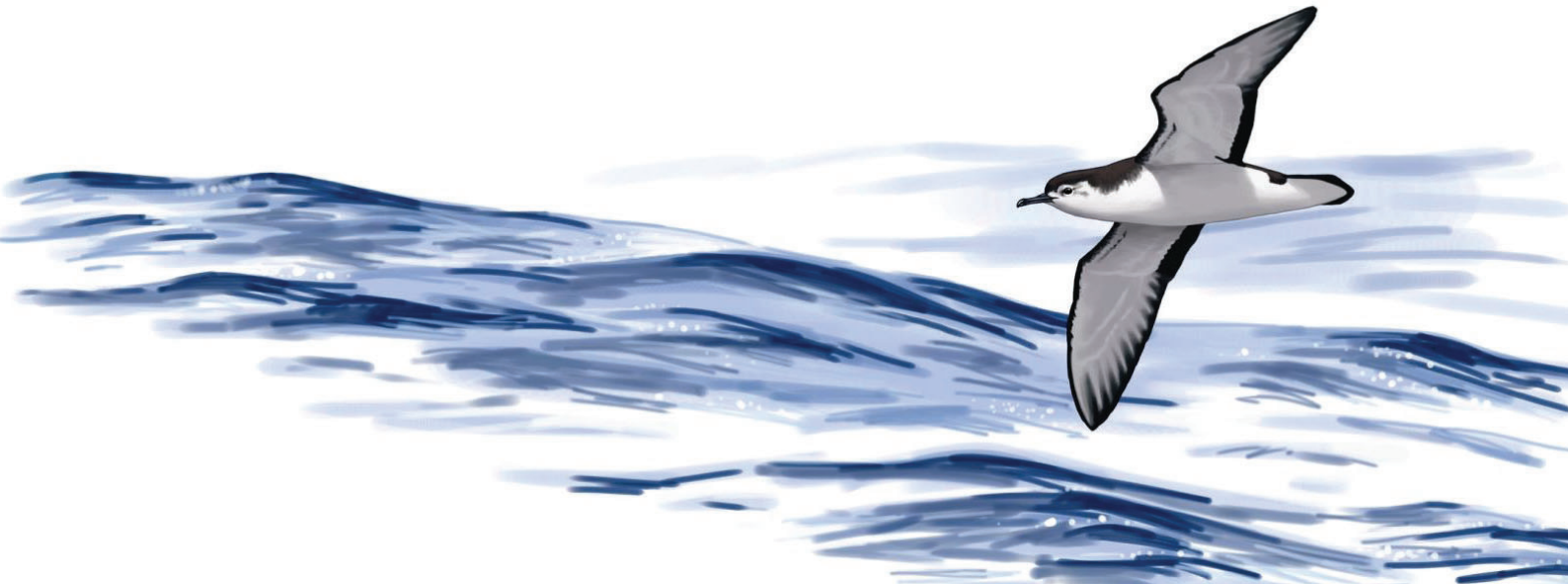
Yu G., Smith D.K., Zhu H., Guan Y., Lam T.T.Y. 2017. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol. Evol. 8:28–36.

Zachos J., Pagani M., Sloan L., Thomas E., Billups K. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science. 292:686–693

Chapter III

Neutral Processes Shape Landscapes of Divergence in a Speciation Continuum of Pelagic Seabirds

JOAN FERRER OBIOL, JOSE M. HERRANZ, JOSEPHINE R. PARIS, JAMES R. WHITING, JACOB GONZÁLEZ-SOLÍS, JULIO ROZAS, AND MARTA RIUTORT



Neutral Processes Shape Landscapes of Divergence in a Speciation Continuum of Pelagic Seabirds

Joan Ferrer Obiol[1,2], Jose M. Herranz[3,4], Josephine R. Paris[5], James R. Whiting[5], Jacob González-Solís[2,6], Julio Rozas[1,2], and Marta Riutort[1,2]

[1]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain
[2]Institut de Recerca de la Biodiversitat (IRBio), Barcelona, Catalonia, Spain
[3]National Institute for the Study of Liver and Gastrointestinal Diseases, CIBERehd, Carlos III Health Institute, Madrid, Spain
[4]Program of Hepatology, Center for Applied Medical Research (CIMA), University of Navarra, Pamplona, Spain
[5]Department of Biosciences, University of Exeter, Exeter, United Kingdom
[6]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

In preparation for Molecular Ecology

Abstract

Speciation is a continuous and complex process shaped by the interaction of numerous evolutionary processes. Despite the continuous nature of the speciation process, the implementation of conservation policies relies on the delimitation of species and evolutionary significant units (ESUs). *Puffinus* shearwaters are globally distributed and threatened pelagic seabirds that, as a result of remarkable morphological stasis, are under intense taxonomic debate. Here, we use double digest Restriction-Site Associated DNA sequencing (ddRAD-Seq) to provide dense genotyping of species and subspecies of North Atlantic and Mediterranean *Puffinus* shearwaters across their entire geographical range. We assess the phylogenetic relationships and population structure among and within the group, evaluate species boundaries and shed light on the processes that have shaped the genomic landscapes of divergence across a speciation continuum. We highlight that none of the current taxonomies are supported by genomic data and propose a more accurate taxonomy for the group integrating genomic information with other sources of evidence. Our data suggests that the genomic landscapes of divergence across the speciation continuum have been shaped primarily by neutral processes. Finally, combining different methods to detect introgression, we

provide empirical evidence of how differences in the effect of genetic drift among species or populations can lead to an incorrect inference of introgression. Our study illustrates a major role of neutral evolution at shaping the speciation process in highly-mobile pelagic seabirds and highlights the potential of genomic data to inform the management of threatened taxa.

Keywords: dd-RAD-Seq, shearwaters, speciation, genomic differentiation, species delimitation, conservation genomics

1 | Introduction

How populations diverge and become new species is one of the most fundamental questions in evolutionary biology (Darwin and Wallace 1858). The process of speciation is usually continuous (Mallet 2008; Nosil 2012) and involves the evolution of reproductive barriers (Coyne et al. 2004). With the increasing availability of genome-wide data, we can now characterise genome-wide patterns of divergence, and investigate the genetic architecture of reproductive isolation across the speciation continuum (Seehausen et al. 2014; Ravinet et al. 2017; Wolf and Ellegren 2017) and the evolutionary processes, including gene flow, mutation, recombination, drift, and selection that shape this genomic landscape. Understanding how these processes interact to shape the speciation process remains challenging (Nosil and Feder 2012; Ravinet et al. 2017). Speciation is therefore a complex process that can leave a myriad of different footprints on the genome depending on these evolutionary drivers, and the interplay between them.

Despite the continuous nature of the speciation process, the implementation of efficient conservation policies relies on the delimitation of species and evolutionary significant units (ESUs, Crandall et al. 2000; Moritz 2002). Under the general lineage concept (GLC), species represent separately evolving metapopulation lineages (De Queiroz 2007). Within the GLC framework, the combination of high-resolution genome-wide data with the development of multispecies coalescent (MSC) delimitation approaches has emerged as a powerful approach to test different hypotheses of lineage divergence (Knowles and Carstens 2007; Yang and Rannala 2010) and alternative

hypotheses of species delimitation (Leaché et al. 2014). Such methods are being used in a growing number of studies to delimit species in a wide range of taxa (Abdelkrim et al. 2018; Tonzo et al. 2019; Ewart et al. 2020; Hosegood et al. 2020; Newton et al. 2020). However, the high resolution of genomic data makes it difficult to distinguish population structure from species boundaries when using MSC methods (Sukumaran and Knowles 2017; Chambers and Hillis 2020) and introgression can further hinder species delimitation, especially in cases of limited geographical sampling (Chambers and Hillis 2020; Chan et al. 2020). Appropriate geographical sampling, including contact zones among putative species, can overcome the issue of over-splitting caused by sampling limitations and combined with other lines of evidence such as morphological, ecological or phenological data, provides a robust framework for species delimitation (Carstens et al. 2013; Chambers and Hillis 2020).

One group with an urgent need of well-defined species and ESUs is shearwaters, a globally distributed group of medium-sized pelagic seabirds. Over 50% of shearwater species are listed as threatened by the IUCN (http://www.iucnredlist.org). Shearwaters face several anthropogenic threats, both at their breeding colonies and at sea (Croxall et al. 2012; Dias et al. 2019; Rodríguez et al. 2019). Inland, shearwater populations are severely affected by the introduction of invasive alien species such as cats and rats, which predate on eggs, chicks and even adult birds (Spatz et al. 2017; Holmes et al. 2019). At sea, fisheries bycatch is the main threat (Bugoni et al. 2008; Oppel et al. 2011; Cortés et al. 2017), one that could drive some of the species to extinction unless conservation measures are promptly implemented (Oro et al. 2004; Genovart et al. 2016). However, resolving the evolutionary relationships among shearwaters has long posed a challenge (Austin 1996; Austin et al. 2004), and species limits are controversial, mostly due to high morphological stasis in the group (Austin et al. 2004). A recent phylogenomic study showed that *Puffinus* shearwaters from the North Atlantic and Mediterranean constitute a monophyletic group that is divided into a clade of medium-sized taxa (*P. puffinus*, *P. mauretanicus*, *P. yelkouan*) and a clade of small-sized taxa (*P. lherminieri*, *P. baroli*, *P. boydi*) (Chapter I). However, the group is still under contentious ongoing taxonomic debate (Sangster et al. 2005; Olson 2010; Genovart et al. 2012; Ramos et al. 2020; Rodríguez et al. 2020).

Taxa in the medium-sized group were originally considered to be conspecific and were placed together under *P. puffinus* (Mathews 1934) until the end of the 1980s, when morphology and vocalisation data (Bourne et al. 1988; Bretagnolle 1992), resulted in a split of the Mediterranean and North Atlantic lineages into two different species (*P. puffinus* and *P. yelkouan*). *P. yelkouan* included two subspecies (*mauretanicus* and *yelkouan*) that more recently, were elevated to species status based on morphological characters and reciprocal monophyly of cytochrome b sequences (Heidrich et al. 1998; Sangster et al. 2002). However, this split has not been unanimously integrated in bird taxonomies (i.e. Christidis et al. 2014). On the other hand, although after the split from its Mediterranean counterparts, *P. puffinus* was considered a monotypic species, most recently, based on multiple lines of evidence, the Canary Islands populations have been described as a new subspecies (*P. p. canariensis*) (Rodríguez et al. 2020). In addition, there is some uncertainty as to the taxonomic affinities of the Madeiran population of *P. puffinus* (Gil-Velasco et al. 2015; Rodríguez et al. 2020).

Small-sized species are under even more contentious taxonomic debate. Since Austin (1996) identified *lherminieri*, *baroli* and *boydi* to be a monophyletic group, the three taxa have been considered as one, two or three different species (del Hoyo et al. 2014; Sangster et al. 2005; Olson 2010). Given these uncertainties, in order to develop effective conservation measures there is an urgent need for a robust review of the current taxonomy of North Atlantic and Mediterranean *Puffinus* shearwaters to establish the real status of the currently recognized six species and their subspecies (*P. puffinus puffinus, P. p. canariensis, P. mauretanicus, P. yelkouan, P. baroli, P. boydi, P. lherminieri lherminieri, P. l. loyemilleri*).

Here, we use paired-end double digest Restriction-Site Associated DNA sequencing (PE-ddRAD-Seq) to (a) quantify the genomic levels of variation among and within every species and subspecies of North Atlantic and Mediterranean *Puffinus* shearwaters and reconstruct their phylogenetic relationships, (b) explore the number of independently evolving lineages by applying multiple coalescent-based species delimitation approaches and, integrating morphological, behavioural and ecological evidence, evaluate and update the taxonomy of the group, (c) explore a potential case of introgression, and (d) determine the main evolutionary drivers shaping genomic

divergence landscapes across a speciation continuum. Through examination of our results, we discuss the validity of current species designations and highlight the important role of neutral evolution in shaping the genomic landscapes of divergence in island-breeding pelagic seabirds.

2 | Materials and Methods

2.1 | Sampling, DNA Extraction and ddRAD-Seq Sequence Data Generation

We collected blood or tissue samples from 42 individuals of the eight recognised North Atlantic and Mediterranean *Puffinus* shearwater taxa across their geographical ranges (Figure 1a, Table S1). We also included *Puffinus nativitatis* and *Calonectris borealis* as outgroups (Table S1). Data for 14 individuals of the ingroup taxa and the outgroups were previously generated in Chapter I. Genomic DNA extraction and ddRAD-Seq library construction for the rest of individuals were performed as described in Chapter I.

2.2 | PE-ddRAD Data Processing

PE-ddRAD data were processed using STACKS v2.41 (Rochette et al. 2019). Raw reads were quality-filtered and demultiplexed using PROCESS_RADTAGS. Loci were built *de novo* using the forward reads with the USTACKS-CSTACKS-SSTACKS core clustering algorithm with optimised parameters for shearwater data (as per Chapter I). Reverse reads were incorporated using TSV2BAM and GSTACKS was used to assemble a contig for each locus, calling SNPs using the Bayesian genotype caller (BGC; Maruki and Lynch 2015, 2017) and phase haplotypes. We mapped GSTACKS catalog loci to the Balearic shearwater genome assembly (Cuevas-Caballé et al. 2019) using BWA-MEM 0.7.17 (Li 2013), sorted using SAMTOOLS v.0.1.19 (Li et al. 2009) with alignment positions integrated to the catalog using STACKS-INTEGRATE-ALIGNMENTS (Paris et al. 2017). The POPULATIONS module was used to export data in various formats for downstream analyses, requiring a minimum allele frequency (MAF) above 5% and an observed heterozygosity below 50%

to process a SNP. For SNP-based analyses, we further filtered VCF files using VCFTOOLS v.0.1.15 (Danecek et al. 2011) to include only biallelic SNPs and to mask genotypes if the per-sample read depth was < 5 or > 150, or if the genotype quality was < 30. Table S2 includes the subsets of the total dataset that were used for each downstream analysis, which include up to 16,339 loci and 141,767 SNPs.

2.3 | Analysis of Genomic Variation Among and Within Taxa

We studied the genomic variation among and within taxa using principal component analysis (PCA) and a discriminant analysis of principal components (DAPC, Jombart et al. 2010) using the R package ADEGENET (Jombart and Ahmed 2011).

We also performed a maximum-likelihood (ML) model-based clustering analysis to calculate individual ancestries using ADMIXTURE v.1.3.0 (Alexander and Lange 2011). We tested $K = 1$ to $K = 10$ and the optimal K was determined using the lowest cross-validation errors estimates (Evanno et al. 2005; Alexander et al. 2009). Additional values of K were also examined. To examine finer levels of structure, hierarchical analyses were performed, individually on the clusters identified using the optimal K.

FINERADSTRUCTURE v0.3.2 (Malinsky et al. 2018) was used to infer shared ancestry among all individuals. RADPAINTER was used to infer a coancestry matrix and the FINESTRUCTURE Monte Carlo Markov Chains (MCMC) clustering algorithm was used to assign individuals into clusters, running 100,000 MCMC iterations (following a burn-in period of 100,000 iterations) sampled every 1,000 generations. A tree of relationships based on the coancestry matrix was built in FINESTRUCTURE using default parameters. We used available R scripts (http://cichlid.gurdon.cam.ac.uk/fineRADstructure.html) to visualise the results. To detect finer-scale genetic structuring, we also performed FINERADSTRUCTURE analyses for each of the three main groups detected by our phylogenetic and populations structure analyses (*P. puffinus*, *P. mauretanicus* - *P. yelkouan* and *P. lherminieri* - *P. baroli* - *P. boydi*).

To further infer geographic structuring and visualise genealogical patterns, we computed Neighbor-net phylogenetic networks (Bryant and Moulton 2004), implemented in SPLITSTREE5 v.5.0.16 (Huson and Bryant 2006).

2.4 | Phylogenetic Analyses

To infer the phylogenetic relationships of the studied taxa and to evaluate the monophyly of clusters identified by FINERADSTRUCTURE, we estimated phylogenies based on concatenation and coalescent approaches using *C. borealis* and *P. nativitatis* as outgroups. For concatenation analyses, we used the MPI version of EXABAYES v.1.5 (Aberer et al. 2014) and RAXML-NG v.0.6.0 (Kozlov et al. 2019) to estimate unpartitioned Bayesian and maximum-likelihood (ML) phylogenies, respectively. For EXABAYES, two independent runs with four coupled chains for 1,000,000 generations were performed and assessed for stationarity (effective sample sizes (ESS) > 300 for all model parameters) in TRACER v.1.7 (Rambaut et al. 2018). For RAXML-NG, 50 ML tree searches were conducted with the GTR+G substitution model. Following the best tree search, we generated 500 non-parametric bootstrap replicates.

To directly model incomplete lineage sorting (ILS), we inferred species trees using two MSC methods: the Bayesian SNP-based SNAPP v.1.4.2 (Bryant et al. 2012) in BEAST v.2.5.0 (Bouckaert et al. 2019), and the summary method ASTRAL-III (Zhang et al. 2018). For SNAPP, we used uninformative priors as we do not assume strong *a priori* knowledge about the parameters. Two replicates were run for 100,000 burn-in iterations, followed by 1,000,000 MCMC iterations. Tree and parameter estimates were sampled every 1000 MCMC iterations. Convergence and stationarity were confirmed (ESS > 300) using TRACER. For ASTRAL-III, we used RAXML v.8 (Stamatakis 2014) to estimate gene trees for each PE-ddRAD locus running 100 rapid bootstrap replicates followed by a thorough ML search. We then used ASTRAL-III to estimate a species trees from the best-scoring ML gene trees and bootstrap replicates.

To estimate divergence times, we applied the MSC approach of Stange et al. (2018) implemented in SNAPP. To avoid the inclusion of potentially introgressed individuals, we performed the analysis on two individuals from the most geographically distant populations per taxon. We performed the analysis only with transitions to reduce rate heterogeneity. We followed Stange et al. (2018) in specifying an age constraint on the root as a normally distributed calibration density with a mean of 2.87 Mya and a standard deviation (SD) of 0.39 (based on Chapter II). We conducted three replicate

runs, each of 1,500,000 MCMC iterations after 100,000 burn-in iterations. More details of the phylogenetic analyses can be found in the Supplemental Information.

2.5 | Species Delimitation

To determine the number of independently evolving lineages, we applied two coalescent-based species delimitation approaches: BPP v.4.0 (Flouri et al. 2018) and BFD* (Leaché et al. 2014). BPP was run using option A11, which performs a joint comparison of species assignment and species tree models (Yang and Rannala 2014; Rannala and Yang 2017). To ensure computational tractability, we performed BPP analyses using two subsets of 500 loci, which has been shown to provide sufficient power for species delimitation (e.g. Tonzo et al. 2019). Subset 1 contained loci with at least four variable sites as such loci provide greater power in species delimitation (Huang 2018). We also selected a random subset of loci to evaluate the effects of including less informative loci on species delimitation. We followed the approach of Huang and Knowles (2016) to test for the impact of different evolutionary and demographic scenarios by using different inverse-gamma distributed diffuse priors ($\alpha = 3$) for the population sizes (θ) and root ages (τ_0) (Table 1). Each analysis was run for 100,000 generations, sampling every 10 generations after a pre-burnin of 100,000 generations.

We used BFD* (Leaché et al. 2014), using a matrix of 500 SNPs with no missing data, to rank ten competing species delimitation hypotheses (SDH) based on the five most popular world bird lists (IOC v.10.2: Gill et al. 2020; Clements v2019: Clements et al. 2019; HBW & Birdlife International: del Hoyo et al. 2014; Howard & Moore v.4.1: Christidis et al. 2014; Peters: Peters et al. 1931), and also using the results from the genetic clustering and phylogenetic analyses performed here (Table 2). For each SDH, we conducted species tree estimation and calculated marginal likelihoods estimates (MLE) using SNAPP. For MLE calculation, we performed path sampling analyses with 40 steps for 100,000 iterations after a pre-burnin of 12,000 iterations and setting alpha to 0.3. Every analysis was run twice using different seeds to assess consistency in marginal likelihood estimation. Because the number of SNPs included in the analysis has the potential to impact model ranks when using BFD* (Leaché et al. 2014), we also performed additional analyses using 2000 SNPs with no missing data. To ensure

computational tractability using this larger number of SNPs, we performed analyses separately for each of the three main groups detected by our phylogenetic and populations structure analyses. For both types of analyses, models were ranked by their MLE, and MLEs were compared using Bayes Factors (Kass and Raftery 1995).

2.6 | Genetic Diversity Within and Among Taxa

We further described genomic diversity within and among taxa using several summary statistics. Per taxon nucleotide diversity (π), inbreeding coefficient (F_{IS}), the ratio of polymorphic SNPs, and pairwise F_{ST} between taxa were calculated using the STACKS POPULATIONS program. Finally, we calculated the ratio of nonsynonymous to synonymous mutations following Perrier et al. (2017). Briefly, all loci were used in a BLAST query against *P. mauretanicus* annotated proteins (Cuevas-Caballé et al. 2019) using BLASTX. Hits with a similarity higher than 95% between the query and the annotated protein and spanning ≥ 25 amino acids were retained. We then identified nonsynonymous mutations across the significant hits, as described in Perrier et al. (2017) and report the ratio of nonsynonymous to synonymous mutations. Due to the low sample sizes for both subspecies of *P. lherminieri*, all summary statistics were calculated at the species level. To explore if patterns of genome-wide diversity relate to census size, we retrieved the number of breeding pairs from Birds of the World (Billerman et al. 2020) and BirdLife International (2020).

To assess the patterns of diversity and differentiation across the genome of three taxon pairs that showed low differentiation (see Results), we calculated per locus π, D_{XY} and Weir and Cockerham F_{ST} using the R package POPGENOME (Pfeifer et al. 2014). Because D_{XY} is associated with within-group diversity, we also calculated net divergence, D_a (Nei and Li 1979), to capture only the differences that have accumulated since the taxa split.

To better assess net divergence for each taxon pair and to explore potential divergence scenarios, we plotted D_{XY} against within-taxon π (π_{within}, Charlesworth 1998) and coloured the points by their F_{ST} values. To further explore the forces driving differentiation in outlier loci, we compared D_{XY}, D_a and $\Delta\pi$ between outlier loci and putatively neutral loci for each taxon pair. We defined outlier loci as those having values

above the 95th percentile of both F_{ST} and F_{ST}' (a haplotype measure of F_{ST} that is scaled to the theoretical maximum F_{ST} value at a particular locus (Meirmans 2006). To visualise relationships among taxa at outlier loci, we constructed haplotype networks for the most highly differentiated outliers in each taxon pair using the R package PEGAS (Paradis 2010).

To assign *Puffinus* loci to chromosomes, we used a liftover approach to map PE-ddRAD loci to the Anna's Hummingbird (*Calypte anna*) chromosome-level genome assembly (diverged from shearwaters between 62.7 to 71.1 Ma (Jarvis et al. 2015) and representing the most closely related chromosome-level genome assembly). Details of the liftover approach can be found in the Supplemental Information.

2.7 | Detecting Historical Introgression

A previous phylogenomic study reported potential historical introgression between *P. lherminieri* and *P. boydi* (Chapter I). To explore this further, we quantified Patterson's *D* statistic (Green et al. 2010; Patterson et al. 2012) for all taxon trios compatible with the species tree. *D*-statistics were calculated using DSUITE DTRIOS (Malinsky et al. 2020). Block-jackknife resampling was used to evaluate significant deviations from zero ($P < 0.001$).

We also used TREEMIX version 1.13 (Pickrell and Pritchard 2012) to test for gene flow in a phylogenetic context applying a likelihood ratio test (LRT) between tree models with and without gene flow. We calculated the total fraction of the variance explained by each migration edge using the GET_F() function in the OPTM R package. Next, we calculated the *f3*(C;A,B)-statistic (Patterson et al. 2012), which assesses whether taxon C is the result of admixture between ancestral taxa A and B using the THREEPOP program in TREEMIX designating *P. boydi* as taxon C, *P. baroli* as taxon A and *P. lherminieri* as taxon B.

To explore the putative heterogeneity in signatures of ancestral introgression/shared polymorphism between *P. lherminieri* and *P. boydi* across the genome, we calculated the f_D statistic (Martin et al. 2015) in sliding windows of 100 SNPs (25 SNP increments)

using DINVESTIGATE in DSUITE. We calculated mean f_D of all genomic windows for each chromosome with 10 or more windows.

3 | Results

3.1 | Population Structure and Phylogenetic Relationships

Genetic clustering analyses identified the majority of *Puffinus* shearwater species as distinct clusters with the exception of *P. mauretanicus* and *P. yelkouan*. PCA showed that the strongest population structure separated small-sized species from medium-sized species (Figure 1b), with PC1 (32.6% of the variance) showing a clear separation between these two groups and further subdividing each group into two (*P. puffinus* from *P. mauretanicus* and *P. yelkouan*, and *P. lherminieri* from *P. boydi* and *P. baroli*). PC2 (23.2%) further separated the medium-sized *P. puffinus* from *P. mauretanicus* and *P. yelkouan* highlighting higher differentiation among the medium-sized species compared to the small-sized species. PC3 (8.4%) further separated the three small-sized species into three different groups. Both DAPC and ADMIXTURE analyses identified $K = 4$ as optimal (Figure 1b,c), although cross-validation error in ADMIXTURE was lowest for K 3 - 5. Although increasing K to 5 did not provide additional interpretable resolution, analysing the small-sized species separately resulted in a complete discrimination of the three species (Figure 1c). Clustering analyses did not distinguish the Mediterranean species *P. mauretanicus* and *P. yelkouan* or the subspecies of *P. puffinus* and *P. lherminieri*. Taxa that were not found to be distinct using these analyses had low pairwise F_{ST} values ($F_{ST} < 0.12$, Figure 1d).

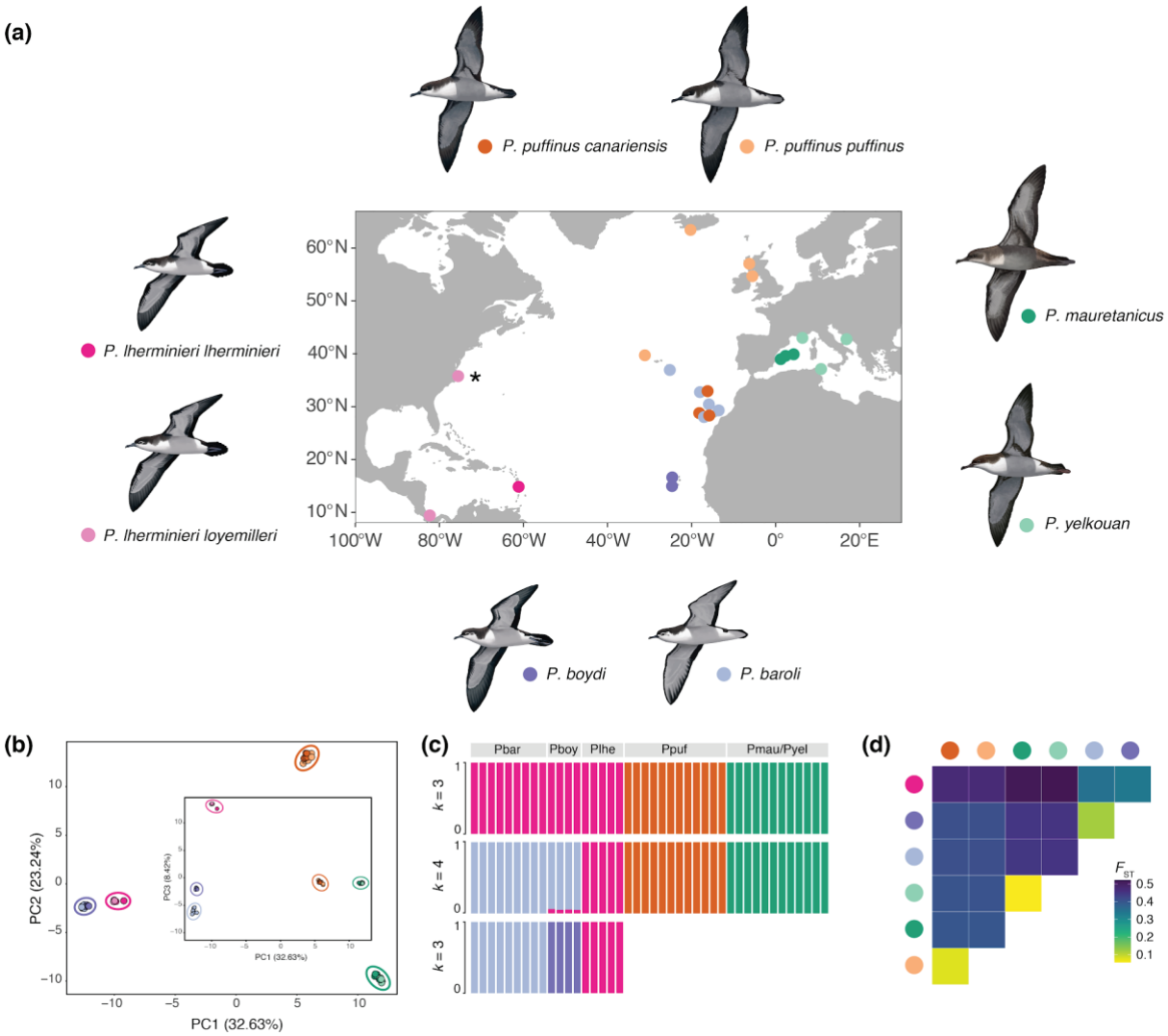


Figure 1 Sampling localities and population structure of *Puffinus* shearwaters included in this study. (a) Map with sampling sites coloured by taxon. Shearwaters were sampled at breeding colonies with the exception of the sampling site with an asterisk where stranded individuals were sampled. Illustrations by Martí Franch © represent shearwater species and subspecies included in this study. (b) PCA with taxa coloured following the same colour code used in (a) and PC1 versus PC3 presented in the insert. (c) ADMIXTURE results for $K = 3$ and $K = 4$ which had the lowest cross-validation error and results for $K = 3$ for the small-sized taxa only. Facet labels above the plots represent: *P. baroli* (Pbar), *P. boydi* (Pboy), *P. lherminieri* (Plhe), *P. puffinus* (Ppuf), *P. mauretanicus* (Pmau) and *P. yelkouan* (Pyel). (d) Heatmap of pairwise F_{ST} estimates between *Puffinus* shearwater taxa.

Overall, FINERADSTRUCTURE which emphasises recent coancestry produced results similar to the other genetic clustering methods, and clearly identified three main groups, one of them subdivided in three well-resolved groups, resulting in a total of five clusters that corresponded to four of the species, and a fifth cluster including *P. mauretanicus* and *P. yelkouan* (Figure 2a). However, FINERADSTRUCTURE detected finer-scale genetic structure within some groups, despite high levels of recent coancestry,

particularly in the analysis of each of the three main groups (Figure 2b,c,d). FINERADSTRUCTURE provided enough resolution to separate *P. mauretanicus* and *P. yelkouan*, and also showed that *P. mauretanicus* from Menorca share higher levels of recent coancestry with individuals of *P. yelkouan* (Figure 2b). Within *P. puffinus*, each sampling locality appeared as a distinct cluster, with the first division separating the individuals from the Canary Islands and Madeira from the North Atlantic populations and the Azores. Recent coancestry values in this species appeared to follow a pattern of isolation by distance (Figure 2c). Finally, the two subspecies of *P. lherminieri* either formed distinct groups or *P. l. lherminieri* was paraphyletic (Figure 2a and d).

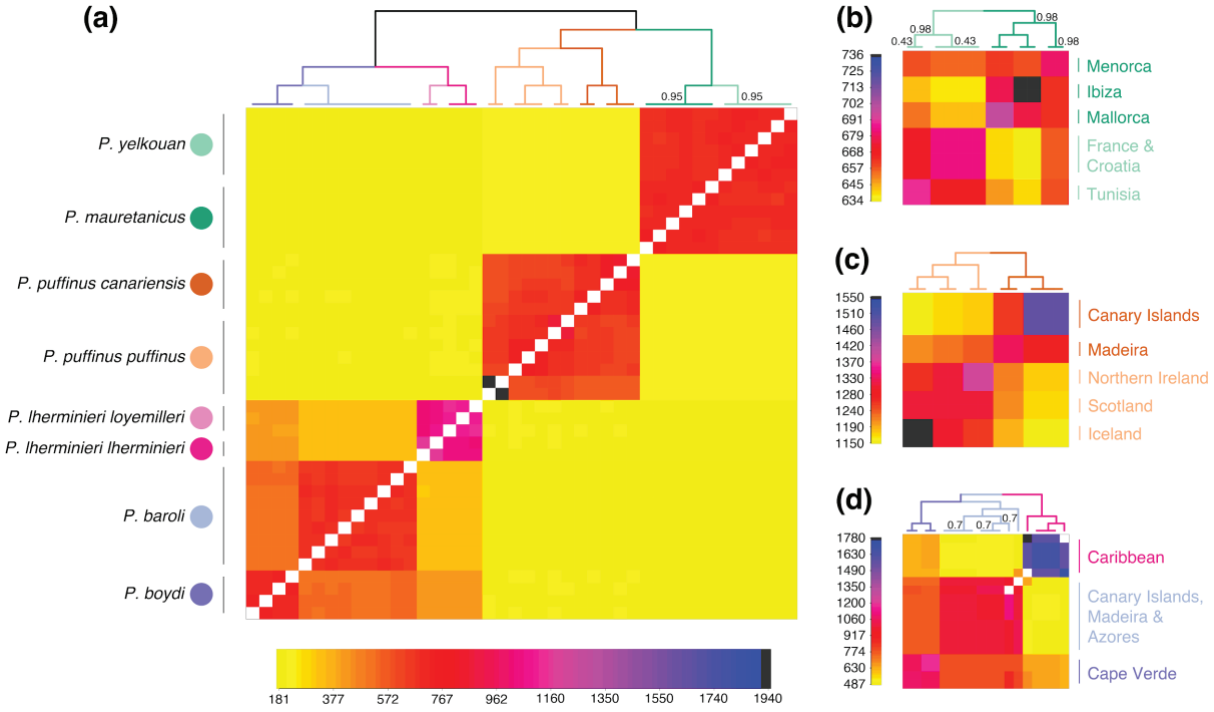


Figure 2 Patterns of shared coancestry inferred from FINERADSTRUCTURE. Each panel represents a heatmap showing coancestry coefficients between shearwater samples Coancestry coefficients are colour coded from low (yellow) to high (blue-black) corresponding to the values in the legend. Atop each heatmap is a FINERADSTRUCTURE clustering dendrogram based on the matrix of coancestry coefficients with branches coloured by taxon following the same colour code used next to the taxon labels on the left. Branch supports are shown for branches with posterior probabilities < 1. a) Coancestry coefficients among all samples. b) Average coancestry coefficients among all samples of *P. puffinus*, (c) *P. mauretanicus* and *P. yelkouan*, and (d) *P. lherminieri*, *P. boydi* and *P. baroli*.

Phylogenetic analyses recovered the five main clusters identified by FINERADSTRUCTURE as monophyletic groups. Phylogenetic trees recovered the same topology as in a previous shearwater phylogenomic study (Chapter I), and coalescent-based analyses confidently resolved the short internode separating the small-sized and the medium-sized species (Figure 3c). Neighbour-net networks and SNAPP analyses showed high levels of reticulation within and among *P. mauretanicus* and *P. yelkouan*, within *P. puffinus* and within *P. lherminieri* suggesting the presence of gene flow (Figure 3a and b). Accordingly, most of the phylogenetic analyses did not recover *P. mauretanicus* and *P. yelkouan*, and *P. puffinus* and *P. lherminieri* subspecies as monophyletic groups. Moreover, divergence time estimates between the two *P. lherminieri* subspecies and between *P. mauretanicus* and *P. yelkouan* included the present in the 95% HPD intervals (Figure 3d), suggesting that these taxa have not yet fully diverged. On the other hand, the split between *P. p canariensis* and *P. p. puffinus* was inferred during the upper-Pleistocene.

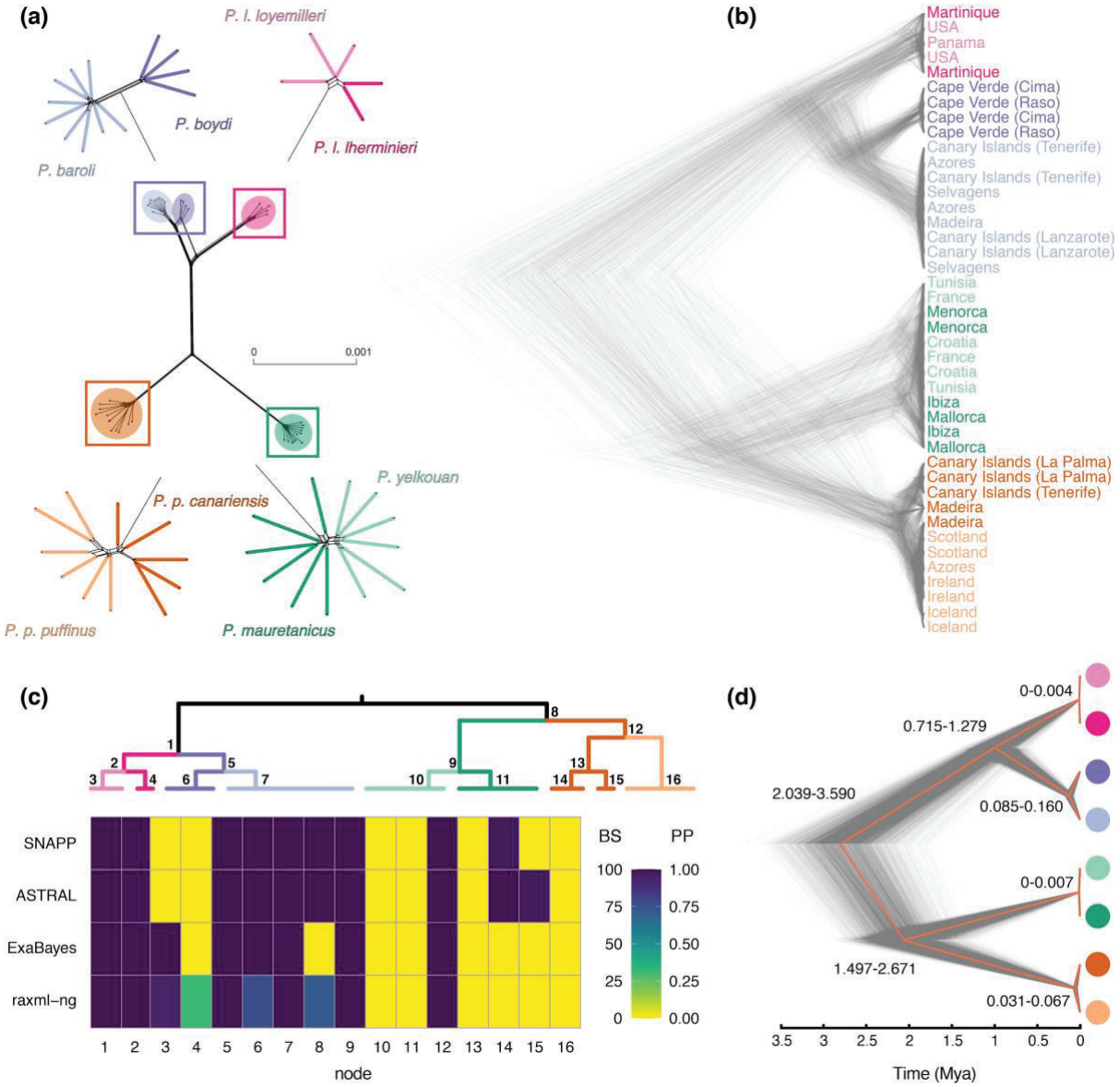


Figure 3 Phylogenetic analyses of *Puffinus* shearwaters. (a) Neighbour-net network inferred from 15,525 PE-ddRAD loci. Squares represent regions of the network that are examined in more detail adjacently. Note that reticulation denotes non-tree-like areas. Colours represent different taxa and are consistent across panels. (b) Cloudogram of SNAPP trees from the posterior tree distribution showing topological and branch length variation. Tip labels represent sampling localities. (c) Heatmap summarising phylogenetic analyses using different methods for selected nodes based on the FINERADSTRUCTURE dendrogram (shown above). Bootstrap support values or posterior probabilities are colour-coded as represented in the legend. (d) Time-calibrated SNAPP species tree (5403 transition sites). Individual trees shown in grey are samples from the posterior tree distribution and a maximum-clade-credibility summary tree is shown in orange.

3.2 | Species Delimitation

Species delimitation analyses using BPP consistently supported a SDH with five species (Table 1), considering each of the five FINERADSTRUCTURE main clusters (Figure 2) as a distinct species. The results of the analysis were largely robust to both the subset of loci and the prior combination used (Table 1).

Table 1 BPP species delimitation analysis results for each subset of loci (random: minimum 1 SNP per locus and informative: minimum 4 SNPs per locus) and different combinations of population size (θ) and root age (τ_0) priors. We report the species inferred by each analysis, the number of species, and the posterior probability of the number of species.

Min. num. of SNPs per locus	Population size prior (θ)	Root age prior (τ_0)	Number of Species	Species	Posterior probability
1	IG(3, 0.002)	IG(3, 0.003)	5	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus*	0.86
1	IG(3, 0.002)	IG(3, 0.03)	5	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus*	0.89
1	IG(3, 0.02)	IG(3, 0.003)	5	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus*	0.96
1	IG(3, 0.02)	IG(3, 0.03)	5	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus*	0.96
4	IG(3, 0.002)	IG(3, 0.003)	6	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus, puffinus* Ireland	0.47
4	IG(3, 0.002)	IG(3, 0.03)	6	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus, puffinus* Ireland	0.65
4	IG(3, 0.02)	IG(3, 0.003)	5	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus*	0.98
4	IG(3, 0.02)	IG(3, 0.03)	5	*baroli, boydi, lherminieri, mauretanicus/yelkouan, puffinus*	0.98

The 500 SNPs BFD* analyses tended to support SDHs with a lower number of species and showed strongest support for a four species model (H6), considering each of the genetic clusters identified with DAPC and ADMIXTURE with K = 4 as a distinct species (Table 2). The SDH based on a current taxonomy that received the highest support was the Howard & Moore World Bird List (v.4.1: Christidis et al. 2014) (H3, 2lnBF = 8.4). Increasing the number of SNPs and performing the analyses by group had a significant impact on model ranks with a tendency towards more splits (Table S3).

Table 2 BFD* analysis results for competing species delimitation hypothesis (SDH) based on five of the most popular world bird lists (H1 - H4), our genetic clustering and BPP analyses (H6 - H7) and other proposed taxonomic proposals (H5, H8 - H10). For each SDH, the number of species, marginal likelihood estimates (MLE), Bayes factors (2 x lnBF) and its rank are shown.

Species delimitation hypothesis (SDH)	Species	Num. species	Rank	MLE	2lnBF
H1: IOC 2020 & Clements 2020	*baroli, boydi, lherminieri, mauretanicus, yelkouan, puffinus*	6	8	-23,729.2	128.8
H2: HBW 2014-2016	*baroli-boydi-lherminieri, mauretanicus, yelkouan, puffinus*	4	5	-23,700.9	72.2
H3: Howard & Moore 2014	*baroli-boydi-lherminieri, mauretanicus-yelkouan, puffinus*	3	2	-23,669.0	8.4
H4: Peters 1931-1986	*baroli (assimilis), boydi-lherminieri, mauretanicus-yelkouan-puffinus*	3	10	-24,107.0	884.4
H5: All taxa	*baroli, boydi, l. lherminieri, l. loyemilleri mauretanicus, yelkouan, p. puffinus, p.* canariensis	8	9	-23,802.9	276.2
H6: ADMIXTURE & DAPC *K* = 4	*baroli-boydi, lherminieri, mauretanicus-yelkouan, puffinus*	4	1	-23,664.8	—
H7: BPP	*baroli, boydi, lherminieri, mauretanicus-yelkouan, puffinus*	5	3	-23,698.8	68.0
H8: Reassign Menorca	*baroli-boydi, lherminieri, mauretanicus, yelkouan* (incl. Menorca), *puffinus*	5	4	-23,700.7	71.8
H9: Split *P. puffinus*	*baroli-boydi, lherminieri, mauretanicus-yelkouan, p. puffinus, p. canariensis*	5	6	-23,701.4	73.2
H10: Split *P. puffinus* & reassign Madeira	*baroli-boydi, lherminieri, mauretanicus-yelkouan, p. puffinus* (incl. Madeira), *p. canariensis*	5	7	-23,703.3	77.0

3.3 | Patterns of Genome-wide Diversity

π ranged from 0.00147 (*P. lherminieri*) to 0.00214 (*P. boydi*) and the proportion of polymorphic SNPs varied markedly from 28.4% in *P. lherminieri* to 50.4% in *P. puffinus puffinus*. Inbreeding (F_{IS}) was relatively low for most taxa, ranging from 0.0808 (*P. baroli*) to 0.1506 (*P. puffinus puffinus*). Among an average of 5542 (3611-6397) polymorphic loci per taxon, an average of 115 (85-128) polymorphic loci had significant BLAST hits to *P. mauretanicus* proteins (Table S4). We identified 59-108 synonymous (average = 91) and 19-31 (average = 24) non-synonymous mutations per taxa. Per-locus π distributions only varied slightly across taxa (Figure S1), with the exception of *P. lherminieri*, which showed a much higher proportion of low π values. Accordingly, *P. lherminieri* had one of the highest F_{IS} values, the lowest ratio of polymorphic SNPs, and the highest ratio of non-synonymous to synonymous mutations amongst all taxa (Table 3), suggesting a reduction of diversity and a higher incidence of relaxed selection in this species despite the relatively high number of breeding pairs. Indeed, the number of breeding pairs did not appear to have a strong effect on the genome-wide levels of

genetic diversity. We found relatively high levels of global π in species with low census size (i.e. *P. boydi*). On the other hand, recently diverged sister taxa had very similar estimates for most of the statistics, suggesting that despite strong philopatry, high gene flow could be tempering potential loss of genetic diversity. However, for each taxon pair, the ratio of non-synonymous to synonymous mutations was always higher for the taxon with the lowest census sizes (Table 3), suggesting that these taxa are experiencing a stronger effect of relaxed selection.

Table 3 Genetic characteristics and number of breeding pairs for each taxon. Global π, inbreeding coefficient (F_{IS}) and the ratio of polymorphic SNPs are reported for each taxon. The ratio of non-synonymous to synonymous mutations was calculated based on significant hits from a BLAST query of polymorphic loci against the *P. mauretanicus* annotated proteins. Summary statistics for *P. lherminieri* were calculated at the species level, due to the low sample sizes for both subspecies of *P. lherminieri*.

Taxon	Number of breeding pairs	Global π	F_{IS}	Ratio of polymorphic SNPs	Ratio of non-synonymous mutations
P. baroli	3,360	0.00178	0.0808	0.483	0.326
P. boydi	5,000	0.00214	0.1105	0.446	0.255
P. lherminieri	15,700	0.00147	0.1308	0.284	0.441
P. mauretanicus	3,142	0.00186	0.1011	0.431	0.241
P. yelkouan	22,928	0.00184	0.0887	0.428	0.211
P. puffinus puffinus	399,500	0.00212	0.1506	0.504	0.176
P. puffinus canariensis	800	0.00210	0.1411	0.478	0.268

3.4 | Genome-wide Differentiation in Three Recently Diverged Taxon Pairs

Per-locus nucleotide diversity densities completely overlapped between each of the three taxon pairs, with the exception of the *P. baroli* and *P. boydi* pair where *P. baroli* had a higher proportion of loci with low π compared to *P. boydi* (Figure 4a). The variation in F_{ST} between the three recently diverged taxon pairs highlighted that the pairs represent three different initial stages of differentiation (Figure 4b). Differentiation between *P. mauretanicus* and *P. yelkouan* was the lowest (mean F_{ST} 0.04, 95th percentile 0.32), followed by *P. p. puffinus* and *P. p. canariensis* (mean F_{ST} 0.06, 95th

percentile CI: 0.44), and by *P. boydi* and *P. baroli* (mean F_{ST} 0.1, 95th percentile 0.65). Across the genome, pairwise F_{ST} showed only a few regions of high differentiation, particularly concentrated on the Z chromosome, in two of the three pairwise taxa comparisons (Figure S2). There was little overlap in differentiation peaks among different pairs (Figure S3) and the number of observed overlaps did not significantly differ from random expectations (Table S5). Variation in net divergence (D_a) showed a similar pattern to F_{ST} (Figure 4c), with mean values of 0.14% (*P. mauretanicus* versus *P. yelkouan*), 0.26% (*P. p. puffinus* versus *P. p. canariensis*) and 0.69% (*P. boydi* versus *P. baroli*). In agreement with the low levels of differentiation within each of the three taxon pairs, the majority of the genome showed that π_{within} was only slightly lower than D_{XY}, with most loci clustered along the 1:1 line (the expectation under a single panmictic population), despite marked heterogeneity in both π_{within} and D_{XY} (Figure 4d). However, the coefficient of determination (R^2) of the regression between π_{within} and D_{XY} also reflected three different levels of differentiation. As expected by their levels of differentiation, the comparison between *P. mauretanicus* and *P. yelkouan* had the highest coefficient of determination (R^2 = 0.94) and the comparison between *P. boydi* and *P. baroli* the lowest (R^2 = 0.74). In models of divergence with ongoing gene flow, after a sufficient amount of time since initial divergence we expect loci that are resistant to introgression to have higher D_{XY} and lower π compared to genome-wide values (Cruickshank and Hahn 2014). On the other hand, if post-speciation divergent selection is driving differentiation at outlier loci, these should show low π_{within} but not high D_{XY} (Cruickshank and Hahn 2014). Consistent with a prediction of a weak role of selection in driving differentiation, we did not detect an excess of loci with reduced π_{within} compared to D_{XY} (see Irwin et al. (2018)) in any of the taxon pairs (Figures 4d).

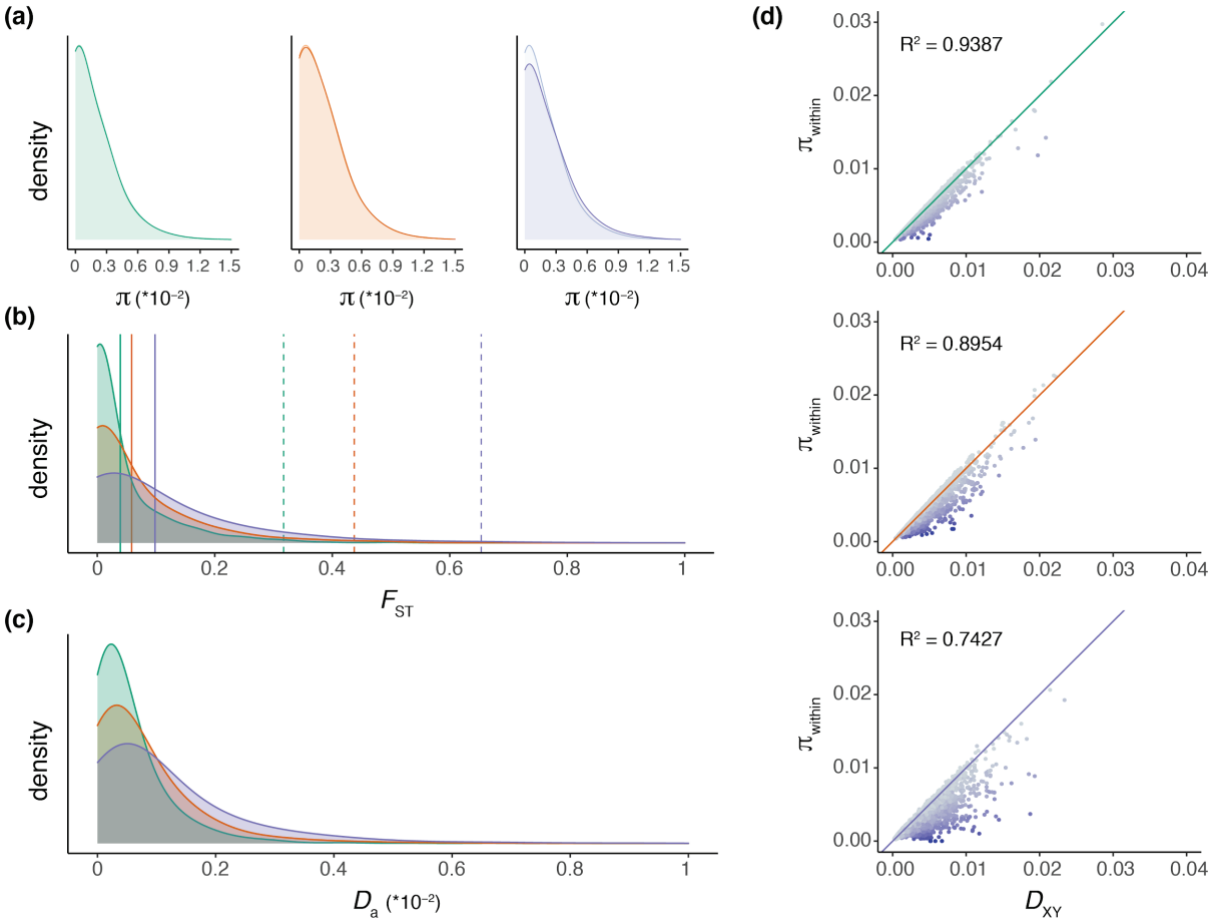


Figure 4 Genome-wide differentiation in three recently diverged taxon pairs. (a) Smoothed density distributions of per-locus π estimates for both taxa in each taxon pair. From left to right: *P. mauretanicus* (green) and *P. yelkouan* (light green), *P. p. puffinus* (light orange) and *P. p. canariensis* (orange), *P. boydi* (purple) and *P. baroli* (light purple). (b) Smoothed distributions of per-locus F_{ST} estimates for each taxon pair: *P. mauretanicus* versus *P. yelkouan* (green), *P. p. puffinus* versus *P. p. canariensis* (orange), *P. boydi* versus *P. baroli* (purple). Vertical continuous and dashed lines indicate mean and 95% percentiles, respectively. (c) Smoothed distributions of per-locus D_a estimates. (d) Relationship between D_{XY} and within-taxon nucleotide diversity (π_{within}). High F_{ST} loci (represented by increasing blue colour) are characterised by reduced π_{within} compared to D_{XY}. Taxon pair colour coding for panels (c) and (d) is identical as in panel (b).

Outlier loci were characterised by significantly higher relative and absolute divergence (F_{ST}, D_{XY} and D_a) than the rest of the genome in all taxon pairs (Figure 5), and reduced π in all taxa. Values of Δπ were not different between outlier loci and the rest of the genome showing no differential reduction in nucleotide diversity in outlier loci between the two members of each pair. Differences in F_{ST} and D_a between outlier loci and the rest of the genome were largest for the comparison between *P. boydi* and *P. baroli*, intermediate between *P. p. puffinus* and *P. p. canariensis,* and smallest between *P. mauretanicus* and *P. yelkouan* (Figure 5). In addition, haplotype networks of the four

highest outliers for the *P. boydi* and *P. baroli* comparison showed species diagnostic haplotypes, but haplotype networks of the highest outliers for the *P. mauretanicus* and *P. yelkouan,* and the *P. p. puffinus* and *P. p. canariensis* comparisons only showed allele frequency differences (Figure S4).

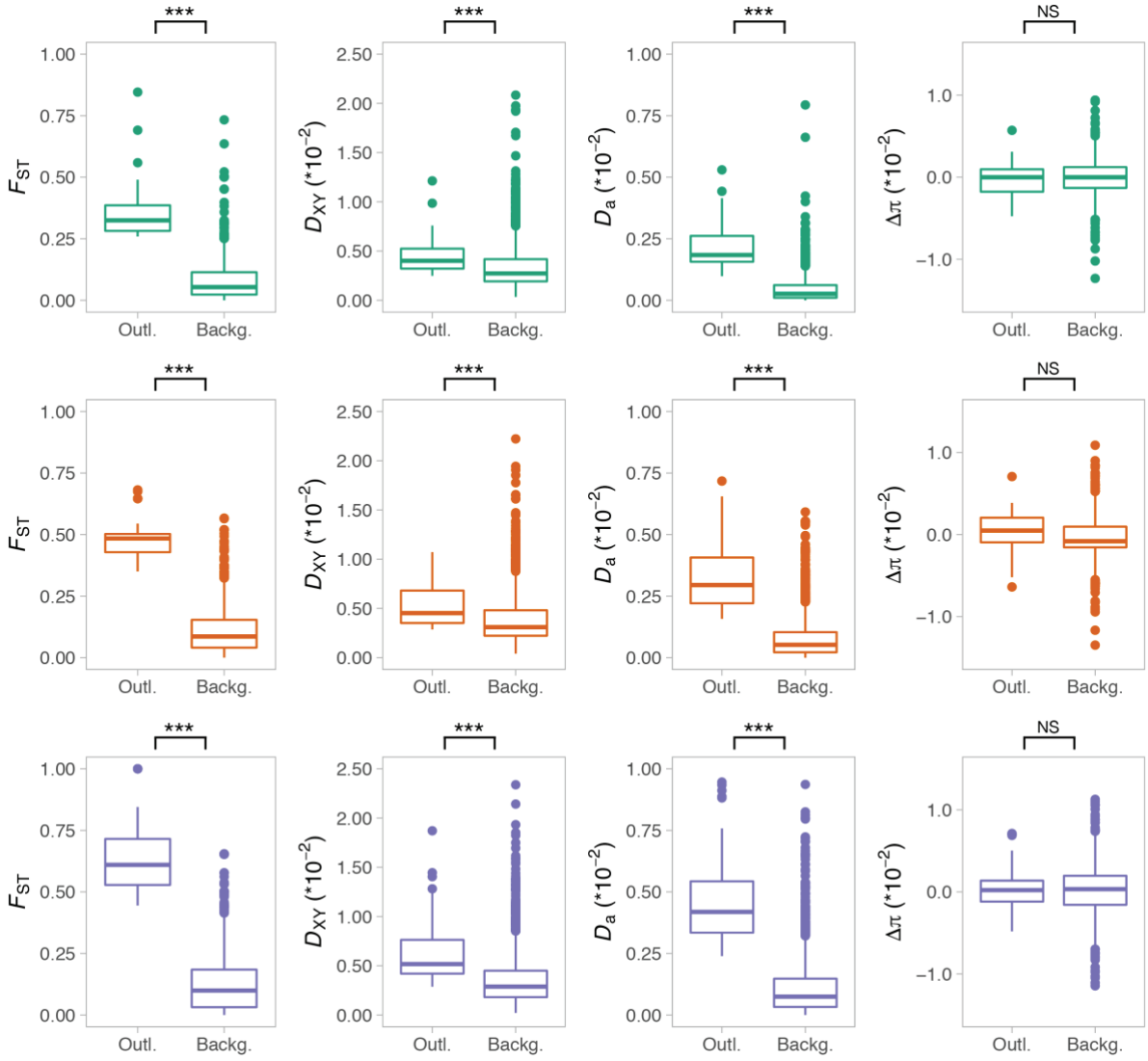


Figure 5 Comparison between population genetics statistics (F_{ST}, D_{XY}, D_a and Δπ) in outlier loci (Outl.) and the rest of the genome (Backg.) for the three taxon pairs: (a) *P. mauretanicus* versus *P. yelkouan*, (b) *P. p. canariensis* versus *P. p. puffinus*, and (c) *P. boydi* versus *P. baroli*. Outlier loci were defined as those having values above the 95th percentile of both F_{ST} and F_{ST}'. Mann–Whitney U tests p-values are shown for each comparison: p-value > 0.5 (NS), p-value < 0.05 (*), p-value < 0.01 (**), p-value < 0.001 (***).

3.5 | Historical Introgression or Different Rates of Neutral Evolution?

Analysis of Patterson's *D*-statistics found a significant excess of shared derived alleles between *P. lherminieri* and *P. boydi* (*D*-statistic = 0.2730). The excess of shared derived alleles was significant regardless of which *P. lherminieri* subspecies we used for the test (Figure 6a), suggesting historical introgression between *P. lherminieri* and *P. boydi* predating the split between *P. lherminieri* subspecies. TREEMIX analyses considering one migration event also inferred gene flow from *P. lherminieri* to *P. boydi*. However, adding gene flow to the model did not significantly improve the likelihood (LRT p-value = 0.85). Moreover, the tree topology inferred by TREEMIX accounted for the greatest majority of allele frequency variation (99.85%), and the migration event between *P. lherminieri* and *P. boydi* only resulted in a 0.13% increase in the variance explained. Consistent with the TREEMIX results, the *f*3(C;A,B)-statistic was not significant, suggesting that *P. boydi* was not the result of admixture between ancestral populations. The fraction of admixture (f_D) between *P. lherminieri* and *P. boydi* was very heterogeneous across chromosomes (Figure 6c) with the Z chromosome showing the lowest value.

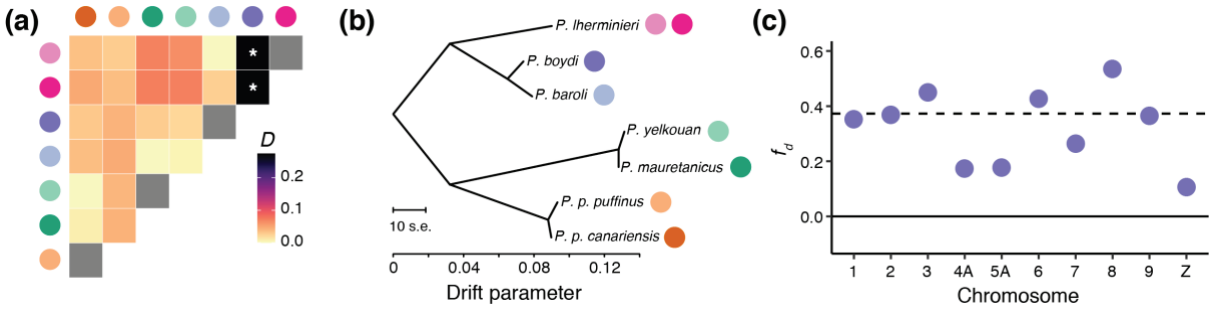


Figure 6 Potential introgression between *lherminieri* and *boydi* is confounded by genetic drift. (a) Heatmap indicating maximum pairwise Patterson's *D*-statistic among *Puffinus* shearwater taxa, showing significant *D*-statistics between *boydi* and both *lherminieri* subspecies. (b) TREEMIX maximum likelihood tree visualising the relationship among taxa. Horizontal branch lengths are proportional to the amount of genetic drift that has occurred along that branch. (c) Mean fraction of admixture (f_D) shows high heterogeneity of shared polymorphism between *boydi* and *lherminieri* across chromosomes.

4 | Discussion

4.1 | North Atlantic and Mediterranean *Puffinus* Shearwaters: How Many Species?

North Atlantic and Mediterranean *Puffinus* shearwaters have recently been identified as a monophyletic group using phylogenomic data (Chapter I). This gave us an opportunity to delve into the species delimitation and population structure on a finer scale, using high resolution genome-wide data for a group that is under highly contentious ongoing taxonomic debate (Sangster et al. 2005; Olson 2010; Genovart et al. 2012; Ramos et al. 2020; Rodríguez et al. 2020), exemplified by the fact that current world bird lists (Table 2) disagree about the number of North Atlantic and Mediterranean *Puffinus* shearwater species. Our species delimitation approaches found no support for any of the previously proposed taxonomies for the group. Taking our present results together with a recent phylogenetic study (Chapter II), and multiple additional lines of evidence (Genovart et al. 2012; Militão et al. 2014; Gil-Velasco et al. 2015; Flood and van der Vliet 2019; Ramos et al. 2020; Rodríguez et al. 2020) under an integrative taxonomic framework, we recommend a more accurate taxonomy for the group. We base our taxonomy on defining a species when all the species delimitation methods agree, and when additional evidence supports the species status. We define an ESU when we find no agreement between species delimitation methods, but when we do find evidence of genetic distinctiveness, and additional evidence for morphological or ecological distinctiveness. On this basis, we propose four species (*P. lherminieri*, *P. baroli*, *P. puffinus* and *P. yelkouan*) and six ESUs (*P. b. baroli*, *P. b. boydi*, *P. p. puffinus*, *P. p. canariensis* and *P. y. yelkouan*, *P. y. mauretanicus*). Below we discuss the consideration of each taxon case-by-case.

Integrating across different methods, we find no support for the split of *mauretanicus* and *yelkouan* into two different species. Genetic clustering analyses did not recover two distinct groups (Figure 1a and b), phylogenetic analyses failed to recover reciprocal monophyly between the two taxa (Figure 3c), and coalescent-based divergence time estimation included present time in the 95% HPD (Figure 3d), suggesting that

mauretanicus and *yelkouan* may have not yet split. Moreover, pairwise F_{ST} was extremely low (F_{ST} = 0.04) and we found no species-diagnostic SNPs. Our results, together with the continuous phenotypic diversity (Genovart et al. 2012; Militão et al. 2014) and non-breeding distributions (Austin et al. 2019), nearly indistinguishable vocalisations (Yésou et al. 1990), and a lack of correspondence at the individual level between phenotypic characters, stable isotope analyses, microsatellites and mtDNA (Genovart et al. 2012; Militão et al. 2014), allow us to propose that the two Mediterranean taxa should be considered as conspecific. However, fine-scale population structure analysis based on recent coancestry was able to separate *yelkouan* and *mauretanicus* into two distinct groups with finer-scale structure at the population level, especially in *mauretanicus* (Figure 2b). Our analyses suggest that *mauretanicus* and *yelkouan* are at a very initial stage of the speciation process, which is in contrast with a previous hypothesis which suggested a scenario of admixture between two well-differentiated species based on deeply divergent mitochondrial haplotypes (Genovart et al. 2005, 2012). Such deep mtDNA divergences are commonly found within species (Morales et al. 2015; Bernardo et al. 2019), and mito-nuclear discordance has been attributed to multiple mechanisms including adaptive introgression of mtDNA, demographic disparities, sex-biased asymmetries, or caused by differences in effective population size between mitochondrial and nuclear regions (Toews and Brelsford 2012). In some cases, epistatic interactions between the nuclear genome and mitochondrial haplotypes can form the basis of reproductive incompatibilities (Sloan et al. 2017). Exploring drivers of mito-nuclear discordance in this case represents an exciting avenue of future research. Taken together, these lead us to propose that these two taxa should be considered ESUs. We hypothesise that differentiation may be occurring due to different migratory strategies and associated changes in breeding phenology (Austin et al. 2019), which appears to be a common mode of population differentiation in Procellariiformes (Rayner et al. 2011).

In agreement with Rodríguez et al. (2020), our species delimitation analyses did not generally support the upgrade of the two *P. puffinus* subspecies (either including the Madeiran populations with the Canary Islands populations or with the northern populations) into separate species. However, FINERADSTRUCTURE analyses revealed that

recent coancestry was lowest between Canary Islands individuals and northern populations. Individuals from Madeira showed higher levels of coancestry with Canary Islands individuals than with northern populations suggesting that they should belong to *P. p. canariensis*. Our dating analyses showed that the Canary Islands populations diverged from its northern counterparts before the last glacial maximum (Figure 3d). These analyses together with morphological differences support the need to consider *P. p. canariensis* an independent ESU from *P. p. puffinus*.

In the small-sized species group, our phylogenetic and genetic clustering analyses were able to recover each of the three taxa as monophyletic/distinct groups. Divergence dating analyses placed the split between the West Atlantic clade (*lherminieri*) and the East Atlantic clade (*boydi* and *baroli*) at ~1 Mya (Figure 3d), and the divergence between *boydi* and *baroli* in the late Pliocene (~120,000 year ago), which is considerably more recent than has been previously proposed (Olson 2010). The relatively recent divergence time between *boydi* and *baroli*, together with low pairwise F_{ST} (mean = 0.10), a high overlap in morphological characters (Flood and van der Vliet 2019), and their shared ecological plasticity (Ramos et al. 2020), lead us to propose that these two taxa should be considered as two different ESUs under the same species (*P. baroli baroli* and *P. baroli boydi*), as previously suggested by Sangster et al. (2005). However, among the pairs analysed that represent different initial stages in the speciation continuum, these two constitute the case which is closest to separate species status. We also found evidence of fine-scale population structure within *P. lherminieri* suggesting that ESUs should be defined below the species level. However, our sampling was too reduced and too sparse (five individuals from two breeding colonies) to be able to draw strong conclusions here. Future phylogeographic studies are required in order to properly assess population structure in this species, which has suffered a 95% reduction in population size since the arrival of humans in the Caribbean (Mackin 2016).

4.2 | Conservation Implications

Conservation management relies on the classification of diversity into species and ESUs. Procellariiform species tend to function as metapopulations, with several populations representing independent ESUs (Friesen et al. 2007; Rexer-Huber et al.

2019; Taylor et al. 2019). The loss or significant population declines of ESUs within a species can significantly reduce the overall genetic diversity and the ability of species to cope with environmental perturbations (Friesen et al. 2007; Cristofari et al. 2019). In such cases, conservation and management of ESUs should be a priority (Palsbøll et al. 2007; Funk et al. 2012). This should be a major consideration in *P. puffinus*, for which the Canary Islands and Madeira populations could be in danger due to their low census sizes (Table 3). These populations harbour unique genetic diversity and a targeted conservation plan integrating evidence derived from these genetic data together with previous phenological, morphological, acoustic and mtDNA data should be developed to ensure preservation of these populations (Rodríguez et al. 2020). The proposed lump of *mauretanicus* and *yelkouan* based on low genetic differentiation does not preclude focussed conservation efforts on these two ESUs. Indeed, the detection of fine-scale population structure should be integrated as new evidence for future conservation plans. Currently, *mauretanicus* and *yelkouan* are catalogued as Critically Endangered and Vulnerable, respectively, and are severely affected by longline fisheries bycatch in the Mediterranean (Oppel et al. 2011; Genovart et al. 2016; Cortés et al. 2017). Identifying the origins of seabirds affected by bycatch is vital to identify the populations most severely affected by fishing practices. A previous integrative approach (Militão et al. 2014) was able to correctly identify 96% individuals to *mauretanicus* or *yelkouan* but lacked the resolution to assign individuals to populations. Our genomic dataset provided resolution at the population level (Figure 2b) and a selection of the most informative SNPs could be used to develop a management-relevant assay to determine the origin of shearwaters that die from fisheries bycatch. Such approaches have proved successful in the genetic assignment of other marine organisms (Nielsen et al. 2012; Meek et al. 2016; Jenkins et al. 2019).

4.3 | Divergence Landscapes Across a Speciation Continuum

The speciation continuum that we have found in the three studied taxon pairs highlights how divergence accumulates across the genome in the early stages of the speciation process. The F_{ST} density curve representing the earliest stage of divergence (*mauretanicus* vs. *yelkouan*) was the most skewed towards low values and there was a

consistent reduction of skew when comparing more diverged taxa (Figure 4b), following the same patterns found in previous studies comparing taxa along the speciation continuum (Martin et al. 2013; Burri et al. 2015; Stankowski et al. 2019; Sendell-Price et al. 2020).

To investigate the mechanisms driving the differentiation processes, we compared population genetic summary statistics between outlier loci and the rest of the genome for each taxon pair. For all comparisons, outlier loci were characterised by lower-than-average π and higher than average D_{XY}. Taken together, these patterns suggest that lineage-specific processes such as genetic drift and/or divergent selection have been the main drivers of differentiation in shearwaters. We are aware that working with individual PE-ddRAD loci may result in higher variance and lower precision of summary statistics estimates (Cruickshank and Hahn 2014), and RAD-Seq approaches can further bias these estimates due to allelic dropout and PCR duplicates (Arnold et al. 2013; Andrews et al. 2016). However, given the low genomic coverage (~0.14%) and the high spacing between PE-ddRAD loci (~80 kbp on average), we decided to use individual loci rather than using a sliding-window approach to be able to detect narrow genomic regions with high differentiation. Additionally, due to the low genomic coverage and high loci spacing, it is likely that other highly differentiated regions of the genome were undetected (Catchen et al. 2017; Lowry et al. 2017). However, as our interest was not in pinpointing loci under divergence, but rather in investigating the mechanisms driving differentiation, we only expect these issues to marginally affect our results and conclusions.

Our results are in contrast with previous genomic studies of closely related bird species where F_{ST} outliers were characterised by lower-than-average D_{XY} (Delmore et al. 2015; Burri 2017; Irwin et al. 2018). The latter pattern is usually attributed to conserved recombination landscapes that lead to regions with low recombination rates being more affected by genetic drift (Hill and Robertson 1966), thus showing elevated genetic differentiation due to a reduction of polymorphism but reduced absolute divergence (Cruickshank and Hahn 2014; Burri 2017; Ravinet et al. 2017). When recombination landscapes are conserved across species, as has been described for passerine birds (Singhal et al. 2015), outlier regions tend to be shared even across species from different

families (Vijay et al. 2016; Van Doren et al. 2017). However, the outlier loci were specific to each comparison and were not more shared across comparisons than expected by chance. This suggests that conserved recombination and causal factors common to all three pairs are not key factors shaping genomic divergence landscapes in shearwaters. Our results could indicate that recombination landscapes are not conserved in all bird groups and that exploring this is an exciting avenue of future research.

4.4 | Differences in the Effect of Genetic Drift Among Species Confound Introgression Analyses

Potential introgression between *lherminieri* and *boydi* has been recently proposed based on phylogenomic data (Chapter I). Here, we integrated different methods to reanalyse this potential introgression. We found support for differences in the effect of genetic drift among species driving patterns of shared alleles that confound several approaches to detect introgression. *D*-statistics yielded significant results suggesting introgression between these two taxa while ADMIXTURE analyses assigned hybrid ancestries to *boydi* individuals (Figure 1c), a pattern that is also known to arise in scenarios of recent bottlenecks (Lawson et al. 2018). However, the *f*3(C;A,B)-statistic was not significant, suggesting that the origin of *boydi* was not the admixture of ancestral *P. baroli* and *P. lherminieri* populations. Moreover, TREEMIX analyses, which strictly account for genetic drift, further demonstrated that the vast majority of allele frequency variation (99.85%) was explained by the North Atlantic and Mediterranean *Puffinus* shearwater tree topology, and the addition of migration between *lherminieri* and *boydi* only provided a marginal improvement (0.13%). Levels of genetic diversity were markedly different between *boydi*, *baroli* and *lherminieri*, with *baroli*, and particularly *lherminieri*, showing considerably reduced diversity (Table 3). On the other hand, the proportion of non-synonymous to synonymous mutations followed an opposite trend suggesting a higher effect of relaxed selection in the taxa with lower estimates of diversity. When we looked at shared variation (f_D) between *boydi* and *lherminieri* across different chromosomes, we found the lowest f_D levels in the Z chromosome (Figure 6c). This pattern could be consistent with a stronger effect of genetic drift in the Z chromosome (Mank et al. 2010) driving a faster reduction of shared

variation in this chromosome. Together, these analyses suggest that a higher effect of genetic drift in *baroli* and *lherminieri* compared to *boydi* results in the appearance of an artificial excess of shared variation between *boydi* and *lherminieri*. Our results show that *D*-statistics and population assignment analyses can be confounded by differences in the effect of genetic drift among species and add to the body of evidence that inferences of these analyses should be interpreted cautiously and confirmed using multiple methods (Eriksson and Manica 2012; Martin et al. 2015; Lawson et al. 2018).

4.5 | Conclusions

Our analysis using high resolution genome-wide data reveals the phylogenetic relationships and population structure of a group of globally threatened pelagic seabirds, the North Atlantic and Mediterranean *Puffinus* shearwaters. By integrating across multiple methods, we provide a robust framework for species delimitation of highly mobile pelagic organisms. We highlight that none of the current taxonomies provide an accurate delineation of shearwater species and propose a more accurate taxonomy for the group. By characterising fine-scale population level genetic structure, we further highlight the need for management of ESUs below the species level in shearwaters. Our findings have important implications for the conservation of these endangered seabirds as they provide detailed information of species limits and population connectivity and provide sufficient resolution for genetic assignment of shearwater bycatch in the North Atlantic Ocean and the Mediterranean Sea. By focusing on genetic differentiation between three recently diverged taxon pairs, we also provide insight into the process of genomic differentiation in island-breeding marine organisms across the speciation continuum. Specifically, we show that the divergence landscapes in *Puffinus* shearwaters appear to be shaped predominantly by neutral processes. Finally, combining different methods to detect introgression, we provide empirical evidence of how differences in the effect of genetic drift among species and populations can be a confounding factor causing ABBA-BABA tests and population assignment analyses to incorrectly infer introgression.

Author Contributions

J.F and M.R. conceived the study, J.G. contributed samples, J.F. and J.H. analysed the data and J.R.P., J.R.W and J.R. provided expert interpretation. J.F. wrote the article with input from all authors. All authors read and approved the final manuscript.

Acknowledgements

We thank Jeremy J. Austin, Yann Kolbeinsson, Meritxell Genovart, Maite Louzao, Karen Bourgeois, Vincent Bretagnolle and Andreanna J. Welch for collecting and/or providing shearwater blood samples from the field. We would also like to thank the Smithsonian National Museum of Natural History for providing tissue loans. We thank the pertinent authorities for issuing the permits needed for this work. Finally, we thank R. Andrew King for helpful comments on an earlier draft of the manuscript and Martí Franch for his wonderful shearwater illustrations. Research funding was provided by the Fundación BBVA (program 'Ayudas a Equipos de Investigación Científica 2017', project code 062_17 to M.R.).

Supplementary Material

Supplementary Material for this chapter may be found in Appendix III.

5 | References

Abdelkrim J., Aznar-Cormano L., Buge B., Fedosov A., Kantor Y., Zaharias P., Puillandre N. 2018. Delimiting species of marine gastropods (Turridae, Conoidea) using RAD sequencing in an integrative taxonomy framework. Mol. Ecol. 27:4591–4611.

Aberer A.J., Kobert K., Stamatakis A. 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–2556.

Alexander D.H., Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 12:246.

Alexander D.H., Novembre J., Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Andrews K.R., Good J.M., Miller M.R., Luikart G., Hohenlohe P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17:81–92.

Arnold B., Corbett-Detig R.B., Hartl D., Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22:3179–3190.

Austin J.J. 1996. Molecular phylogenetics of Puffinus shearwaters: preliminary evidence from mitochondrial cytochrome b gene sequences. Mol. Phylogenet. Evol. 6:77–88.

Austin J.J., Bretagnolle V., Pasquet E. 2004. A global molecular phylogeny of the small Puffinus shearwaters and implications for systematics of the Little-Audubon's Shearwater complex. Auk. 121:647–864.

Austin R.E., Wynn R.B., Votier S.C., Trueman C., McMinn M., Rodríguez A., Suberg L., Maurice L., Newton J., Genovart M., Péron C., Grémillet D., Guilford T. 2019. Patterns of at-sea behaviour at a hybrid zone between two threatened seabirds. Sci. Rep. 9:14720.

Bernardo P.H., Sánchez-Ramírez S., Sánchez-Pacheco S.J., Álvarez-Castañeda S.T., Aguilera-Miller E.F., Mendez-de la Cruz F.R., Murphy R.W. 2019. Extreme mito-nuclear discordance in a peninsular lizard: the role of drift, selection, and climate. Heredity 123:359–370.

Billerman S. M., Keeney B. K., Rodewald P. G., Schulenberg T. S. 2020. Birds of the World. Cornell Laboratory of Ornithology, Ithaca, NY, USA. https://birdsoftheworld.org/bow/home

BirdLife International. 2020. IUCN Red List for birds. Downloaded from http://www.birdlife.org on 17/08/2020.

Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., du Plessis L., Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu C.-H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15:e1006650.

Bourne W. R. P. Mackrill E. J. Paterson A. M. Yésou P. 1988. The Yelkouan Shearwater Puffinus (puffinus?) yelkouan. Br. Birds. 81:306—319.

Bretagnolle V. 1992. Variation géographique des vocalisations de pétrels ouest-paléarctiques et suggestions taxonomiques. Alauda. 60:251–252.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Bryant D., Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Bugoni L., Mancini P.L., Monteiro D.S., Nascimento L., Neves T.S. 2008. Seabird bycatch in the Brazilian pelagic longline fishery and a review of capture rates in the southwestern Atlantic Ocean. Endanger. Species Res. 5:137–147.

Burri R. 2017. Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. Mol. Ecol. 26:3853–3856.

Burri R., Nater A., Kawakami T., Mugal C.F., Olason P.I., Smeds L., Suh A., Dutoit L., Bureš S., Garamszegi L.Z., Hogner S., Moreno J., Qvarnström A., Ružić M., Sæther S.A., Sætre G.P., Török J., Ellegren H. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 25:1656–1665.

Carboneras, C. 1992. Family Procellariidae (petrels and shearwaters). In Handbook of the Birds of the World. Vol. 1 (J. del Hoyo, A. Elliott and J. Sargatal, Editors), Lynx Edicions, Barcelona, Spain. pp. 216–257.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. Mol. Ecol. 22:4369–4383.

Catchen J.M., Hohenlohe P.A., Bernatchez L., Funk W.C., Andrews K.R., Allendorf F.W. 2017. Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Mol. Ecol. Resour. 17:362–365.

Chambers E.A., Hillis D.M. 2020. The Multispecies Coalescent Over-Splits Species in the Case of Geographically Widespread Taxa. Syst. Biol. 69:184–193.

Chan K.O., Hutter C.R., Wood P.L. Jr, Grismer L.L., Das I., Brown R.M. 2020. Gene flow creates a mirage of cryptic species in a Southeast Asian spotted stream frog complex. Mol. Ecol. 29:3970–3987.

Charlesworth B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15:538–543.

Christidis L. 2014. The Howard and Moore Complete Checklist of the Birds of the World, version 4.1.

Clements J.F., Schulenberg T.S., Iliff M.J., Billerman S.M., Fredericks T.A., Sullivan B.L., Wood C.L. 2019. The eBird/Clements Checklist of Birds of the World: v2019.

Cortés V., Arcos J.M., González-Solís J. 2017. Seabirds and demersal longliners in the northwestern Mediterranean: factors driving their interactions and bycatch rates. Mar. Ecol. Prog. Ser. 565:1–16.

Coyne J.A., Orr H.A. 2004. Speciation. Sinauer Associates Sunderland, MA.

Crandall K.A., Bininda-Emonds O.R., Mace G.M., Wayne R.K. 2000. Considering evolutionary processes in conservation biology. Trends Ecol. Evol. 15:290–295.

Cristofari R., Plaza P., Fernández C.E., Trucchi E. 2019. Unexpected population fragmentation in an endangered seabird: the case of the Peruvian diving-petrel. Sci. Rep. 9:1–13

Croxall J.P., Butchart S.H.M., Lascelles B., Stattersfield A.J., Sullivan B., Symes A., Taylor P. 2012. Seabird conservation status, threats and priority actions: a global assessment. Bird Conservation International. 22:1–34.

Cruickshank T.E., Hahn M.W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23:3133–3157.

Cuevas-Caballé, C., Ferrer-Obiol, J., Genovart, M., Rozas, J., González-Solís, J., Riutort, M. 2019. Conservation genomics applied to the Balearic shearwater. G10K-VGP/EBP 2019. doi: 10.13140/RG.2.2.15751.21923

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R., 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

Darwin C., Wallace A. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. Journal of the proceedings of the Linnean Society of London. Zoology. 3:45–62.

del Hoyo J., Collar N. J., Christie D. A., Elliott A. , Fishpool L. D. C., Boesman P., Kirwan G. M. 2014. HBW and BirdLife International Illustrated Checklist of the Birds of the World, Volume 1, Lynx Edicions in association with BirdLife International, Barcelona, Spain and Cambridge, UK.

Delmore K.E., Hübner S., Kane N.C., Schuster R., Andrew R.L., Câmara F., Guigo R., Irwin D.E. 2015. Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. Mol. Ecol. 24:1873–1888.

Dias M.P., Martin R., Pearmain E.J., Burfield I.J., Small C., Phillips R.A., Yates O., Lascelles B., Borboroglu P.G., Croxall J.P. 2019. Threats to seabirds: A global assessment. Biol. Conserv. 237:525–537.

Eriksson A., Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. U. S. A. 109:13956–13960.

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

Ewart K.M., Lo N., Ogden R., Joseph L., Ho S.Y.W., Frankham G.J., Eldridge M.D.B., Schodde R., Johnson R.N. 2020. Correction: Phylogeography of the iconic Australian red-tailed black-cockatoo (Calyptorhynchus banksii) and implications for its conservation. Heredity. 125:167.

Flood R.L., van der Vliet R. 2019. Variation and identification of Barolo Shearwater and Boyd's Shearwater. Dutch Birding. 41:215–237.

Flouri T., Jiao X., Rannala B., Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35:2585–2593.

Friesen V.L., Burg T.M., McCoy K.D. 2007. Mechanisms of population differentiation in seabirds: Invited review. Mol. Ecol. 16:1765–1785.

Funk W.C., McKay J.K., Hohenlohe P.A., Allendorf F.W. 2012. Harnessing genomics for delineating conservation units. Trends Ecol. Evol. 27:489–496.

Genovart M., Arcos J.M., Álvarez D., McMinn M., Meier R., Wynn R., Guilford T., Oro D. 2016. Demography of the critically endangered Balearic shearwater: the impact of fisheries and time to extinction. J. Appl. Ecol. 53:1158–1168.

Genovart M., Juste J., Contreras-Díaz H., Oro D. 2012. Genetic and phenotypic differentiation between the critically endangered balearic shearwater and neighboring colonies of its sibling species. J. Hered. 103:330–341.

Genovart M., Juste J., Oro D. 2005. Two sibling species sympatrically breeding: A new conservation concern for the critically endangered Balearic shearwater. Conserv. Genet. 6:601–606.

Gil-Velasco M., Rodriguez G., Menzie S., Arcos J.M. 2015. Plumage variability and field identification of Manx, Yelkouan and Balearic Shearwaters. Br. Birds. 108:514–539.

Gill F., Donsker D., Rasmussen P. 2020. IOC World Bird List (v10. 1). doi : 10.14344/IOC.ML.10.1.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A draft sequence of the Neandertal genome. Science. 328:710–722.

Heidrich P., Amengual J., Wink M. 1998. Phylogenetic relationships in mediterranean and North Atlantic shearwaters (Aves: Procellariidae) based on nucleotide sequences of mtDNA. Biochem. Syst. Ecol. 26:145–170.

Hill W.G., Robertson A. 1966. The effect of linkage on limits to artificial selection. Genetical Research. 8:269–294.

Holmes N.D., Spatz D.R., Oppel S., Tershy B., Croll D.A., Keitt B., Genovesi P., Burfield I.J., Will D.J., Bond A.L., Wegmann A., Aguirre-Muñoz A., Raine A.F., Knapp C.R., Hung C.-H., Wingate D., Hagen E., Méndez-Sánchez F., Rocamora G., Yuan H.-W., Fric J., Millett J., Russell J., Liske-Clark J., Vidal E., Jourdan H., Campbell K., Springer K., Swinnerton K., Gibbons-Decherong L., Langrand O., Brooke M. de L., McMinn M., Bunbury N., Oliveira N., Sposimo P., Geraldes P., McClelland P., Hodum P., Ryan P.G., Borroto-Páez R., Pierce R., Griffiths R., Fisher R.N., Wanless R., Pasachnik S.A., Cranwell S., Micol T., Butchart S.H.M. 2019. Globally important islands where eradicating invasive mammals will benefit highly threatened vertebrates. PLoS One. 14:e0212128.

Hosegood J., Humble E., Ogden R., de Bruyn M., Creer S., Stevens G.M.W., Abudaya M., Bassos-Hull K., Bonfil R., Fernando D., Foote A.D., Hipperson H., Jabado R.W., Kaden J., Moazzam M., Peel L.R., Pollett S., Ponzo A., Poortvliet M., Salah J., Senn H., Stewart J.D., Wintner S., Carvalho G. 2020. Phylogenomics and species delimitation for effective conservation of manta and devil rays. Mol. Ecol.

Huang J.-P. 2018. What have been and what can be delimited as species using molecular data under the multi-species coalescent model? A case study using Hercules beetles (*Dynastes*; Dynastidae). Insect Systematics and Diversity. 2:3.

Huang J.-P., Knowles L.L. 2016. The Species versus Subspecies Conundrum: Quantitative Delimitation from Integrating Multiple Data Types within a Single Bayesian Approach in Hercules Beetles. Syst. Biol. 65:685–699.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Irwin D.E., Milá B., Toews D.P.L., Brelsford A., Kenyon H.L., Porter A.N., Grossen C., Delmore K.E., Alcaide M., Irwin J.H. 2018. A comparison of genomic islands of differentiation across three young avian species pairs. Mol. Ecol. 27:4839–4855.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Alfaro-Núñez A., Narula N., Liu L., Burt D., Ellegren H., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G., Avian Phylogenomics Consortium. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience. 4:4.

Jenkins T.L., Ellis C.D., Triantafyllidis A., Stevens J.R. 2019. Single nucleotide polymorphisms reveal a genetic cline across the north-east Atlantic and enable powerful population assignment in the European lobster. Evol. Appl. 12:1881–1899.

Jombart T., Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 27:3070–3071.

Jombart T., Devillard S., Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.

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

Knowles L.L., Carstens B.C. 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56:887–895.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35:4453–4455.

Lawson D.J., van Dorp L., Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat. Commun. 9:1–11.

Leaché A.D., Fujita M.K., Minin V.N., Bouckaert R.R. 2014. Species delimitation using genome-wide SNP Data. Syst. Biol. 63:534–542.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN].

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Lowry D.B., Hoban S., Kelley J.L., Lotterhos K.E., Reed L.K., Antolin M.F., Storfer A. 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol. Ecol. Resour. 17:142–152.

Mackin W.A. 2016. Current and former populations of Audubon's Shearwater (Puffinus lherminieri) in the Caribbean region. The Condor: Ornithological Applications. 118:655–673.

Malinsky M., Matschiner M., Svardal H. 2020. Dsuite - Fast D-statistics and related admixture evidence from VCF files. Mol. Ecol. Resour. 00:1–12.

Malinsky M., Trucchi E., Lawson D.J., Falush D. 2018. RADpainter and fineRADstructure: Population Inference from RADseq Data. Mol. Biol. Evol. 35:1284–1290.

Mallet J. 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363:2971–2986.

Mank J.E., Nam K., Ellegren H. 2010. Faster-Z evolution is predominantly due to genetic drift. Mol. Biol. Evol. 27:661–670.

Martin S.H., Dasmahapatra K.K., Nadeau N.J., Salazar C., Walters J.R., Simpson F., Blaxter M., Manica A., Mallet J., Jiggins C.D. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23:1817–1828.

Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32:244–257.

Maruki T., Lynch M. 2015. Genotype-Frequency Estimation from High-Throughput Sequencing Data. Genetics. 201:473–486.

Maruki T., Lynch M. 2017. Genotype Calling from Population-Genomic Sequencing Data. G3: Genes|Genomes|Genetics. 7:1393–1404.

Mathews G.M. 1934. A check-list of the order Procellariiformes.

Meek M.H., Baerwald M.R., Stephens M.R., Goodbla A., Miller M.R., Tomalty K.M.H., May B. 2016. Sequencing improves our ability to study threatened migratory species: Genetic population assignment in California's Central Valley Chinook salmon. Ecol. Evol. 6:7706–7716.

Meirmans P.G. 2006. Using the AMOVA framework to estimate a standardized genetic differentiation measure. Evolution. 60:2399–2402.

Militão T., Gómez-Díaz E., Kaliontzopoulou A., González-Solís J. 2014. Comparing multiple criteria for species identification in two recently diverged seabirds. PLoS One. 9:e115650.

Morales H.E., Pavlova A., Joseph L., Sunnucks P. 2015. Positive and purifying selection in mitochondrial genomes of a bird with mitonuclear discordance. Mol. Ecol. 24:2820–2837.

Moritz C. 2002. Strategies to protect biological diversity and the evolutionary processes that sustain it. Syst. Biol. 51:238–254.

Nei M., Li W.H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U. S. A. 76:5269–5273.

Newton L.G., Starrett J., Hendrixson B.E., Derkarabetian S., Bond J.E. 2020. Integrative species delimitation reveals cryptic diversity in the southern Appalachian Antrodiaetus unicolor (Araneae: Antrodiaetidae) species complex. Mol. Ecol. 29:2269–2287.

Nielsen E.E., Cariani A., Mac Aoidh E., Maes G.E., Milano I., Ogden R., Taylor M., Hemmer-Hansen J., Babbucci M., Bargelloni L., Bekkevold D., Diopere E., Grenfell L., Helyar S., Limborg M.T., Martinsohn J.T., McEwing R., Panitz F., Patarnello T., Tinti F., Van Houdt J.K.J., Volckaert F.A.M., Waples R.S., FishPopTrace consortium, Albin J.E.J., Vieites Baptista J.M., Barmintsev V., Bautista J.M., Bendixen C., Bergé J.-P., Blohm D., Cardazzo B., Diez A., Espiñeira M., Geffen A.J., Gonzalez E., González-Lavín N., Guarniero I., Jeráme M., Kochzius M., Krey G., Mouchel O., Negrisolo E., Piccinetti C., Puyet A., Rastorguev S., Smith J.P., Trentini M., Verrez-Bagnis V., Volkov A., Zanzi A., Carvalho G.R. 2012. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nat. Commun. 3:1–7.

Nosil P. 2012. Ecological Speciation. Oxford University Press.

Nosil P., Feder J.L. 2012. Genomic divergence during speciation: causes and consequences. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367:332–342.

Olson S.L. 2010. Stasis and turnover in small shearwaters on Bermuda over the last 400 000 years (Aves: Procellariidae: Puffinus lherminieri group): small Bermuda shearwaters. Biol. J. Linn. Soc. Lond. 99:699–707.

Oppel S., Raine A.F., Borg J.J., Raine H., Bonnaud E., Bourgeois K., Breton A.R. 2011. Is the Yelkouan shearwater Puffinus yelkouan threatened by low adult survival probabilities? Biol. Conserv. 144:2255–2263.

Oro D., Aguilar J.S., Igual J.M., Louzao M. 2004. Modelling demography and extinction risk in the endangered Balearic shearwater. Biol. Conserv. 116:93–102.

Palsbøll P.J., Bérubé M., Allendorf F.W. 2007. Identification of management units using population genetic data. Trends Ecol. Evol. 22:11–16.

Paradis E. 2010. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics. 26:419–420.

Paris J.R., Stevens J.R., Catchen J.M. 2017. Lost in parameter space: a road map for stacks. Methods in Ecology and Evolution. 8:1360–1373.

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. 2012. Ancient admixture in human history. Genetics. 192:1065–1093.

Perrier C., Ferchaud A.L., Sirois P., Thibault I., Bernatchez L. 2017. Do genetic drift and accumulation of deleterious mutations preclude adaptation? Empirical investigation using RADseq in a northern lacustrine fish. Mol. Ecol. 26:6317–6335.

Peters J.L. 1986. Check-list of Birds of the World, 1931-1986. Harvard University Press/Museum of Comparative Zoology.

Pfeifer B., Wittelsbürger U., Ramos-Onsins S.E., Lercher M.J. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31:1929–1936.

Pickrell J.K., Pritchard J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8:e1002967.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. 67:901–904.

Ramos R., Paiva V.H., Zajková Z., Precheur C., Fagundes A.I., Jodice P.G.R., Mackin W., Zino F., Bretagnolle V., González-Solís J. 2020. Spatial ecology of closely related taxa: the case of the little shearwater complex in the North Atlantic Ocean. Zool. J. Linn. Soc.

Rannala B., Yang Z. 2017. Efficient Bayesian Species Tree Inference under the Multispecies Coalescent. Syst. Biol. 66:823–842.

Ravinet M., Faria R., Butlin R.K., Galindo J., Bierne N., Rafajlović M., Noor M.A.F., Mehlig B., Westram A.M. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. 30:1450–1477.

Rayner M.J., Hauber M.E., Steeves T.E., Lawrence H.A., Thompson D.R., Sagar P.M., Bury S.J., Landers T.J., Phillips R. a., Ranjard L., Shaffer S.A. 2011. Contemporary and historical separation of transequatorial migration between genetically distinct seabird populations. Nat. Commun. 2:1–7.

Rexer-Huber K., Veale A.J., Catry P., Cherel Y. 2019. Genomics detects population structure within and between ocean basins in a circumpolar seabird: The white-chinned petrel. Mol. Ecol. 28:4552–4572.

Rochette N.C., Rivera-Colón A.G., Catchen J.M. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28:4737–4754.

Rodríguez A., Arcos J.M., Bretagnolle V., Dias M.P., Holmes N.D., Louzao M., Provencher J., Raine A.F., Ramírez F., Rodríguez B., Ronconi R.A., Taylor R.S., Bonnaud E., Borrelle S.B., Cortés V., Descamps S., Friesen V.L., Genovart M., Hedd A., Hodum P., Humphries G.R.W., Le Corre M., Lebarbenchon C., Martin R., Melvin E.F., Montevecchi W.A., Pinet P., Pollet I.L., Ramos R., Russell J.C., Ryan P.G., Sanz-Aguilar A., Spatz D.R., Travers M., Votier S.C., Wanless R.M., Woehler E., Chiaradia A. 2019. Future Directions in Conservation Research on Petrels and Shearwaters. Frontiers in Marine Science. 6:94.

Rodríguez A., Rodríguez B., Montelongo T., Garcia-Porta J., Pipa T., Carty M., Danielsen J., Nunes J., Silva C., Geraldes P., Medina F.M., Illera J.C. 2020. Cryptic differentiation in the Manx Shearwater hinders the identification of a new endemic subspecies. J. Avian Biol.

Sangster G., Collinson J.M., Helbig A.J., Knox A.G., Parkin D.T. 2005. Taxonomic recommendations for British birds: third report. Ibis . 147:821–826.

Sangster G., Knox A.G., Helbig A.J., Parkin D.T. 2002. Taxonomic recommendations for European birds. Ibis . 144:153–159.

Seehausen O., Butlin R.K., Keller I., Wagner C.E., Boughman J.W., Hohenlohe P.A., Peichel C.L., Saetre G.-P., Bank C., Brännström Å., Brelsford A., Clarkson C.S., Eroukhmanoff F., Feder J.L., Fischer M.C., Foote A.D., Franchini P., Jiggins C.D., Jones F.C., Lindholm A.K., Lucek K., Maan M.E., Marques D.A., Martin S.H., Matthews B., Meier J.I., Möst M., Nachman M.W., Nonaka E., Rennison D.J., Schwarzer J., Watson E.T., Westram A.M., Widmer A. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–192.

Sendell-Price A.T., Ruegg K.C., Anderson E.C., Quilodrán C.S., Van Doren B.M., Underwood V.L., Coulson T., Clegg S.M. 2020. The Genomic Landscape of Divergence Across the Speciation Continuum in Island-Colonising Silvereyes (Zosterops lateralis). G3. 10:3147–3163.

Singhal S., Leffler E.M., Sannareddy K., Turner I., Venn O., Hooper D.M., Strand A.I., Li Q., Raney B., Balakrishnan C.N., Griffith S.C., McVean G., Przeworski M. 2015. Stable recombination hotspots in birds. Science. 350:928–932.

Sloan D.B., Havird J.C., Sharbrough J. 2017. The on-again, off-again relationship between mitochondrial genomes and species boundaries. Mol. Ecol. 26:2212–2236.

Spatz D.R., Holmes N.D., Reguero B.G., Butchart S.H.M., Tershy B.R., Croll D.A. 2017. Managing Invasive Mammals to Conserve Globally Threatened Seabirds in a Changing Climate. Conservation Letters. 10:736–747.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Stange M., Sánchez-Villagra M.R., Salzburger W., Matschiner M. 2018. Bayesian Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports Miocene Closure of the Panamanian Isthmus. Syst. Biol. 67:681–699.

Stankowski S., Chase M.A., Fuiten A.M., Rodrigues M.F., Ralph P.L., Streisfeld M.A. 2019. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. PLoS Biol. 17:e3000391.

Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. U. S. A. 114:1607–1612.

Taylor R.S., Bolton M., Beard A., Birt T., Deane-Coe P., Raine A.F., González-Solís J., Lougheed S.C., Friesen V.L. 2019. Cryptic species and independent origins of allochronic populations within a seabird species complex (Hydrobates spp.). Mol. Phylogenet. Evol. 139:106552.

Toews D.P.L., Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. Mol. Ecol. 21:3907–3930.

Tonzo V., Papadopoulou A., Ortego J. 2019. Genomic data reveal deep genetic structure but no support for current taxonomic designation in a grasshopper species complex. Mol. Ecol. 28:3869–3886.

Van Doren B.M., Campagna L., Helm B., Illera J.C., Lovette I.J., Liedvogel M. 2017. Correlated patterns of genetic diversity and differentiation across an avian family. Mol. Ecol. 26:3982–3997.

Vijay N., Bossu C.M., Poelstra J.W., Weissensteiner M.H., Suh A., Kryukov A.P., Wolf J.B.W. 2016. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat. Commun. 7:1–10.

Wolf J.B.W., Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. Nat. Rev. Genet. 18:87–100.

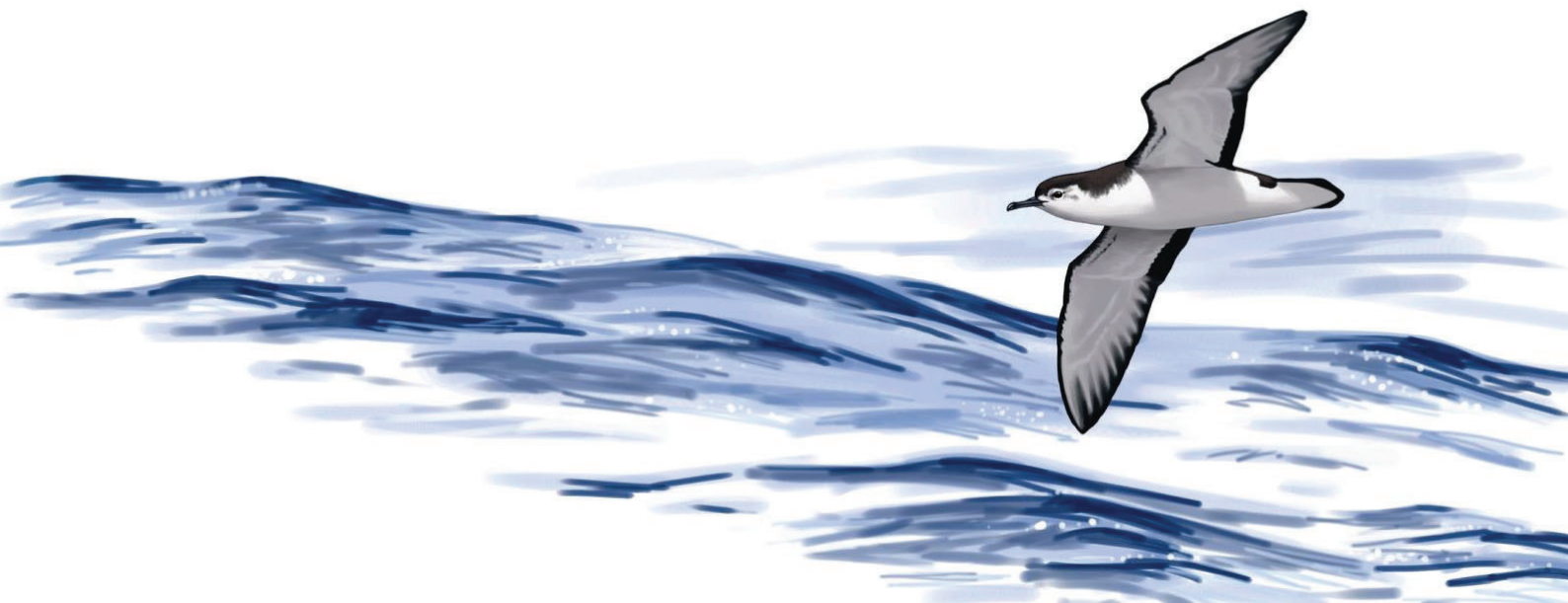
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U. S. A. 107:9264–9269.

Yang Z., Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple Loci. Mol. Biol. Evol. 31:3125–3135.

Yésou P., Paterson A.M., Mackrill E.J., Bourne W.R.P. 1990. Plumage variation and identification of the "Yelkouan Shearwater." Br. Birds. 83.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:15–30

General Discussion



General Discussion

In this thesis, I aimed to unveil the patterns and processes that contribute to genetic and phenotypic diversification, speciation and dispersal in shearwaters. To this end, I undertook a series of phylogenetic and population genetic analyses across several evolutionary timescales. Here, I will discuss a synthesis of the obtained results, their strengths and their shortcomings. Specifically, I will discuss an integrative approach to disentangle the role of incomplete lineage sorting (ILS) and introgression as causes of phylogenetic conflict; I will assess the role of founder events and paleoceanographic changes in the diversification of shearwaters; I will discuss neutral evolution as a main driver of differentiation between geographically isolated populations of pelagic seabirds; and I will highlight important implications from a conservation and management perspective.

1 | Disentangling the Role of ILS and Introgression as Causes of Phylogenetic Conflict

The advent of next-generation sequencing technologies has allowed researchers to move from phylogenetics to phylogenomics. With the availability of thousands of loci, we now have an extraordinary ability to detect phylogenetic conflict and investigate the causes of discordance. ILS is likely the most common cause of gene tree discordance (Edwards 2009), which can produce a high degree of gene tree incongruence in phylogenomic datasets (Pollard et al. 2006; Salichos and Rokas 2013). The radiation of neoaves presents an extreme example of ILS: Jarvis et al. (2014) reported that none of the gene trees inferred from different marker types (UCE, introns, exons) matched the species tree topology. However, ILS is not the only potential source of gene tree discordance. Hybridisation and introgression can also lead to phylogenetic incongruence between gene trees. Distinguishing ILS from introgression remains a major challenge in phylogenomics, and therefore in evolutionary biology. However, these processes can be discerned since their footprints at different genomic regions (that is, between different markers) vary across phylogenetic scales, depending on

factors such as evolutionary and/or recombination rates, or the type and strength of the underlying selection (Martin and Jiggins 2017; Knowles et al. 2018). However, empirical studies to test these assumptions are generally lacking.

In order to fill this gap, we adopted a comprehensive and integrative approach, comparing and combining UCE and ddRAD-seq datasets, and applying state-of-the-art concatenated and coalescent-based approaches using shearwaters as a case study (Chapter I). Consistent with the expectation that the phylogenetic signal of different markers varies across timescales, we found that datasets from different markers and levels of missing data tended to perform better at resolving conflictive internodes at different timescales. Phylogenetic analyses using UCE data, which have a slower evolutionary rate than ddRAD markers, tended to provide better support towards the root of the tree. On the other hand, phylogenetic analyses using ddRAD-seq data tended to provide higher resolution to resolve phylogenetic relationships towards the shallower nodes. However, these data type effects were only minor. We initially assumed that resolving the phylogenetic relationships among shearwaters was going to be a difficult task due to the lack of resolution of previous phylogenetic studies based on complete mitochondrial Cytochrome b (*cytb*) sequences. But unexpectedly, phylogenetic analyses based on genomic data provided a largely well-supported phylogeny with only three conflictive internodes. To resolve these conflictive internodes, our approach integrating data types proved very fruitful, and allowed us to investigate the causes of discordance. This demonstrates the power of our integrative approach at resolving and interrogating phylogenetic conflict and moreover, allows us to advocate the use of this approach in groups with higher levels of phylogenetic conflict than the shearwaters.

D-statistics have gained popularity for detecting introgression in a phylogenetic context because of their simplicity, but they suffer from drawbacks. Specifically, *D*-statistics require a high number of loci to hold enough power to reliably detect introgression (Zheng and Janke 2018) and can be misled by various factors such as ancestral population structure (Eriksson and Manica 2012), low effective population sizes (Martin et al. 2015) or rate heterogeneity (Blair and Ané 2019). Unfortunately, in our analyses, we could not compare *D*-statistic inferences between data types due to a lack of power in the more reduced UCE dataset. However, we detected two potential

cases of introgression using the ddRAD dataset (*A. grisea-A. tenuirostris* and *P. lherminieri-P. boydi*). In performing additional analyses to account for additional potential sources of bias driven by processes other than introgression in these cases, we detected that one of them (*A. grisea-A. tenuirostris*) was most likely the result of GC-biased gene conversion (gBGC) and rate heterogeneity rather than introgression. In addition, using Treemix, which infers introgression using population data based on allele frequencies, we also showed that in the other case (*P. lherminieri-P. boydi*), *D*-statistics had been confounded by genetic drift. Indeed, bottlenecked populations, due to their reduced diversity, tend to increase the likelihood of significant *D*-statistics between their sister populations (P2) and external populations (P3), simply because the loss of diversity in the bottlenecked populations, results in an excess of shared variation between P2 and P3. The incorporation of demographic information is thus essential in order to avoid confusion between reduced diversity and introgression when using *D*-statistics. This makes the *D*-statistics an unsuitable method for accurately testing for introgression in a phylogenomic context, where demographic information is generally lacking. Together, these two cases demonstrate the limitations of using *D*-statistics in isolation of other evidence, and therefore, a careful consideration of other factors is required in order to infer introgression with confidence.

In order to accurately infer introgression in a phylogenomic context, recent methodological developments make use of the multispecies coalescent framework (MSC) in order to simultaneously account for both ILS and gene flow when reconstructing the evolutionary history of a clade (Solís-Lemus et al. 2017; Wen et al. 2018; Flouri et al. 2020). By using the MSC, these methods (phylogenetic networks) do not suffer from the same drawbacks as *D*-statistics, but unfortunately, current implementations are computationally intensive which precludes their use in large phylogenies. In Chapter I, we applied one of these phylogenetic networks approaches (Solís-Lemus et al. 2017) in order to assess whether signatures of introgression were consistent across data types. We found that searches for an optimal network produced inconsistent results in independent runs and in different datasets. However, we did not find strong support for introgression in any of the analyses, suggesting that phylogenetic networks were correctly inferring a lack of introgression. Our comparison of data types

for phylogenetic network analyses provided a robust framework to disentangle the roles of ILS and introgression and future tests should focus on systems with a more important impact of introgression. To conclude, phylogenetic network approaches hold much promise for detecting introgression in a phylogenomic context, yet they are still in their infancy. More accurate inferences of introgression require population-level datasets that allow a proper evaluation of the effects of demography.

2 | Founder Events as a Common Mode of Speciation in Shearwaters

In this thesis, I evaluated the relative role of founder and vicariant events as modes of shearwater speciation. Vicariance due to physical barriers such as land masses or differences in oceanic regimes has been proposed as a major driver of speciation in seabirds (Friesen et al. 2007; Friesen 2015) and in shearwaters (Austin 1996; Austin et al. 2004). On the other hand, very few cases of speciation due to founder events have been reported in the seabird literature (Liebers et al. 2001; Abbott and Double 2003), although range expansions without being directly associated with speciation have been widely described. Indeed, although founder-event speciation has received extensive theoretical attention, due to the difficulty of directly associating founding bottlenecks with reproductive isolation, its role in speciation *per se* remains controversial (Coyne et al. 2004). Probably as a result of this controversy, biogeographic probabilistic models have, for a long time, incorporated anagenetic range-expansion events (Ronquist and Sanmartín 2011), but they have only recently incorporated founder events as an alternative type of cladogenesis event (Matzke 2014), and not without criticism (Ree and Sanmartín 2018). Despite controversy, the importance of founder-event speciation in oceanic island systems seems undeniable (Cowie and Holland 2006; Gillespie et al. 2012; Matzke 2014).

In Chapter II, we compared biogeographic models that include the j parameter, which accounts for founder-event speciation, to models without this parameter in order to evaluate this hypothesis in shearwaters. It is noteworthy that due to our designation of broad biogeographical areas, our analyses were limited in that we could only detect

founder events between different ocean basins and thus, the importance of this process at a finer scale could not be assessed. However, models including the *j* parameter consistently outperformed models without the parameter, and further, founder events had a higher probability of explaining most of the speciation events between taxa in separate biogeographic areas. These results should, however, be taken with caution due to recent criticism that suggests that models including the *j* parameter are poor models of founder-event speciation (Ree and Sanmartín 2018). In support for founder-event speciation, we detected notable variation in the effective population size (θ) estimates across branches in the shearwater phylogeny, even between sister species, suggesting frequent changes in population size that could be the result of founding bottlenecks. In Chapter III, we also detected contrasting levels of nucleotide diversity between sister taxa (*P. baroli boydi* and *P. b. baroli*), which could also indicate that this case of divergence could in fact, be the cause of a founder event.

Despite our results being strongly indicative that founder events are an important mode of shearwater speciation, there is a need for additional analyses to provide a more robust view of the real importance of founder-event speciation. Comparing the demographic histories of species that are expected to have arisen through founder events (such as *P. b. baroli*) with species that are expected to be the result of vicariant events would be a compelling approach to tackle this. Unfortunately, the data generated in this thesis was not suitable to perform accurate demographic inferences. Our sampling did not include enough individuals to produce accurate site frequency spectra (SFS), which form the backbone of several methods which infer historic population dynamics (Excoffier and Foll 2011; Kamm et al. 2019). Additionally, our RAD-Seq data lacked appropriate resolution to perform demographic analyses using sequentially markovian coalescent methods, which can produce accurate demographic inferences using a reduced number of individuals (Schiffels and Durbin 2014), or even a single individual (Li and Durbin 2011). A particularly promising avenue of future research would be to compare the demographic histories of species that have arisen through founder events with species that are expected to be the result of vicariant events. Such analyses would require the use of whole-genome sequencing (WGS) data and analyses which can simultaneously calculate population sizes across time and separation

histories, which are included in models in RELATE (Speidel et al. 2019). In order to provide a null hypothesis of the expectations for the demographic trajectories of the species under these two alternative scenarios, these results could be compared to simulations based on founder and vicariant events, using forward genetic simulations as provided in SLIM3 (Haller and Messer 2019).

3 | The Importance of Paleoceanographic Events

The divergence time estimation analyses carried out in Chapter II showed an extended period (>10 million years), during the Miocene and the beginning of the Pliocene, with barely any evidence of speciation. This was then followed by a burst of speciation at the Pliocene-Pleistocene boundary which gave rise to most of the extant lineages. Recently, Louca and Pennell (2020) showed that for any diversification scenario, there exists an infinite number of alternative diversification scenarios that are equally likely to have generated any given extant timetree. Such results imply that by using timetrees alone, these alternative diversification scenarios cannot be distinguished. Taking the shearwater timetree as an example, the long period with barely any speciation could actually be the result of a very low speciation rate during this time, or of a steadily high extinction rate, or even as a result of a single extinction event. On the other hand, the subsequent burst in speciation could be due to a sudden increase in the speciation rate, or alternatively to a sudden decrease in the extinction rate, among other possibilities.

In order to distinguish between alternative diversification scenarios, paleontological data can be a crucial source of information. Indeed, conducting a compilation of shearwater fossils, we found that ~82% of the known extinct shearwater fossil species lived during the Miocene (~46%) and the Pliocene (~36%) (Miller 1961; Howard 1971; Olson 1985; Olson et al. 2001). Taking these results, together with the estimated timetree, suggests that a scenario with high extinction rates (or sudden extinction events) is more probable than a scenario with low speciation rates during the Miocene and the Pliocene. Neritic areas represent the main foraging grounds for many shearwater species and, during the Pliocene, these areas suffered a significant sudden

reduction, which has been associated with a three-fold increase in the extinction rate of megafauna associated with coastal habitats (Pimiento et al. 2017). Our results provide new evidence for the Pliocene extinction event by revealing a severe effect of the Pliocene extinction on shearwaters. This corroborates the power of combining fossil data with timetrees in order to understand which of the infinite alternative scenarios are more likely to have generated a given extant timetree. A limitation to this interpretation is that despite the high proportion of shearwater species that lived during the Miocene and the Pliocene, we were unable to assess with certainty how many of these species represented extant versus extinct lineages. This was due to the low rate of evolution of shearwater osteological characters (Kuroda 1954), which precluded a proper evaluation of diversity loss during the Pliocene extinction. However, due to the diversity in size and some osteological characters of the extinct fossil species, some of these species may actually be representatives of extinct lineages.

After the sudden reduction of neritic areas in the late Pliocene, these areas suffered extreme oscillations over the Pleistocene. The shearwater timetree and the fossil record suggested that these oscillations did not seem to have such a severe effect on lineage extinction. Conversely, climatic oscillations during the Pleistocene seemed to have promoted speciation and dispersal, probably as a consequence of both geographic isolation during glacial periods and also by promoting dispersal between ocean basins aided by prevailing winds during interglacial periods. Increases in diversification during the Pleistocene have also been observed in a wide array of seabird species including penguins (Vianna et al. 2020), storm-petrels (Silva et al. 2015; Silva et al. 2016), boobies (Morris-Pocock et al. 2011) and gadfly petrels (Gangloff et al. 2013), providing additional evidence that oceanographic changes caused by climatic oscillations during this period played a major role in seabird diversification. An exciting avenue of ongoing research led by Max Levy and Andreanna J. Welch from Durham University (UK), aims to generate a timetree for all Procellariiformes using UCEs. Such research will allow an investigation into exploring how paleoceanographic events have affected speciation and extinction rates to shape species diversity across this group of pelagic seabirds.

4 | Speciation Driven by Neutral Evolution?

Speciation in shearwaters tends to occur due to geographic isolation after founder or vicariant events, as mentioned previously in this discussion. In most cases, recently diverged species show strong morphological and ecological similarities, and also occupy similar ecological niches (Ramos et al. 2020). We thus expected genetic drift to play a major role in driving genetic differentiation in shearwaters and conversely, for natural selection to play only a minor role. The results we obtained in Chapter III were in agreement with this expectation. Analysing three different pairs of taxa at different early stages in the speciation continuum we found no evidence of selective sweeps (positive selection). The analyses of outlier loci showed that higher differentiation in these loci was pair-specific and most likely due to higher-than-average rates of neutral evolution. Recently diverged species of shearwater generally show high morphological and ecological similarity. However, during the radiation that took place amid the Pliocene-Pleistocene boundary (Chapter II), several speciation events, especially within *Puffinus*, were associated with body size changes and behavioural specialisations (different migratory strategies and different preferences in foraging areas (oceanic vs. neritic)). In these speciation cases, we would expect a more prevalent role of selection in driving divergence, especially in regions of the genome affecting body size and the aforementioned behaviours.

Interestingly, two of the pairs analysed in Chapter III comprised sister species that differ in size, colouration and migratory behaviour. *P. puffinus puffinus* are bigger, and whiter on the underwing than its congeners from the Canary Islands (*P. p. canariensis*) (Rodríguez et al. 2020). *P. p. canariensis* might also use different migration routes and wintering areas although these are largely unknown. *P. yelkouan yelkouan* is also much smaller and much darker than *P. y. mauretanicus*. The latter migrates from the Mediterranean to Atlantic waters of western Iberia and northwest France whilst the former stays in the Mediterranean all year round or migrates to the Black Sea (Austin et al. 2019). Despite morphological and behavioural differences between the members of these two pairs, our ddRAD-Seq data did not detect an enrichment of loci showing

signatures of potential positive selection in these pairs compared to the *P. baroli baroli* - *P. b. boydi* pair.

This could be due to several reasons. Firstly, the ddRAD-Seq methodology does not provide enough statistical power; indeed, due to the low genomic coverage and large physical distances between neighbouring ddRAD loci, it is likely that some highly differentiated regions of the genome were undetected (Lowry et al. 2017). Secondly, the differentiated traits could have a polygenic architecture, or be the result of epistatic interactions and thus adaptive loci could go undetected as for such loci, we do not expect differentiation peaks (Rockman 2012). Finally, the low sample sizes used in our study could have precluded the detection of relevant differentiation peaks. Even if our ddRAD-Seq approach had provided enough resolution to detect loci potentially under positive selection, we may still have found it challenging to detect the relevant genes or regulatory regions under selection due to a lack of recombination and linkage information. For instance, if a locus potentially under selection was found in regions of low recombination, the actual genes or regulatory regions under selection could be physically distant from the ddRAD locus, and thus it would be difficult to pinpoint them using only ddRAD-Seq data. To overcome this limitation, future studies should aim to generate WGS population data across a higher sample size, and combine genome scans to detect positive and balancing selection with inference of genome-wide recombination. This would allow to discern the putative confounding effects of varying rates of recombination (Burri et al. 2015; Han et al. 2017; Hejase et al. 2020), and to identify regions of the genome potentially associated with phenotypic and behavioural changes between the diversifying taxa.

A particularly interesting research topic that emerges from this thesis is the discovery that within the genus *Puffinus*, reductions in body size seem to have occurred at least three times in a short period during the Pliocene-Pleistocene radiation (Chapter II). Small-sized shearwaters are similar morphologically and ecologically and previously had been considered to form a monophyletic group (Jouanin and Mougin 1979; Wragg 1985; Austin 1996). However, a more recent phylogenetic study based on *cytb* sequences showed that these species did not constitute a monophyletic group (Austin et al. 2004) and in Chapter I we confirmed these findings. This leads to the prediction that

morphological and ecological evolution in shearwaters seems to be constrained, and indeed such a hypothesis has previously been suggested by Austin et al. (2004).

The recurrence of the same phenotypic polymorphism in multiple species of a clade is a common pattern in a wide array of organisms including plants, insects, molluscs and vertebrates (reviewed in Jamie and Meier (2020)). Despite the ubiquity of these persistent phenotypic polymorphisms, relatively little is known about their evolutionary origins. However, we do now know that genetic polymorphisms can cross species boundaries in the form of ILS and standing genetic variation (Cortez et al. 2014) or through introgression (Jay et al. 2018), especially in cases of rapid speciation events, and that balancing selection can maintain these polymorphisms over time (Gray and McKinnon 2007). *Puffinus* shearwaters provide an excellent case-study to investigate the mechanisms that have allowed the recurrent evolution of small body size across a radiation in a group where introgression does not seem to play a major role (Chapter I). Specifically, two approaches would provide exciting avenues of future research. On the one hand, generating a genome for every species in the *Puffinus* radiation would allow us to evaluate if body size convergence is driven by genetic convergence at either the amino acid (or nucleotide), or the evolutionary rate level at either coding (McGowen et al. 2020) or non-coding regions (Sackton et al. 2019). On the other hand, generating WGS population data for several small-sized and medium-sized species in the radiation. This would allow us to better quantify the levels of shared ancestral polymorphism and to evaluate the importance of balancing selection and structural variants at maintaining phenotypically-relevant polymorphism across the radiation (Vijay et al. 2016; Weissensteiner et al. 2020).

5 | Species Delimitation and Conservation Implications

Procellariiformes face threats both at terrestrial breeding colonies and at sea, and are among the most endangered groups of birds in the world (Croxall et al. 2012). Shearwaters are a particularly sensitive group, with 57% of the species being listed as Threatened by the IUCN Red List of Threatened Species due to anthropogenic activity. A large majority of conservation actions focus specifically on species and evolutionary

significant units (ESUs) and thus an accurate delimitation of species and ESUs is required to ensure the effective management of species-level biodiversity. In this thesis, we have demonstrated the usefulness of genomic data as a unique source of evidence for species and ESUs delimitation. The combination of genomic data with other sources of evidence has allowed us to update the alpha taxonomy of a controversial group: the North Atlantic and Mediterranean *Puffinus* shearwaters (Chapter III). For example, we have provided unequivocal evidence for lumping *P. mauretanicus* and *P. yelkouan* together in a single species. Our data showed that these taxa have not yet fully diverged and that they are still part of the same evolutionary lineage. This has important implications for conservation as these taxa are both listed by the IUCN as Critically Endangered and Vulnerable, respectively. In addition, because these taxa are part of the same species, hybridisation between them is no longer a conservation concern as had been previously proposed (Genovart et al. 2005). However, this thesis also highlighted the importance of conservation below the species level through the delimitation of ESUs. Major population declines within an ESU risk a reduction in a species' overall genetic diversity and evolutionary potential, and slow the species' ability to recover from perturbations (Friesen et al. 2007). Hence, the definition of ESUs for biological conservation is particularly important in endangered species in order to assess the impact of declines in particular populations on overall genetic diversity. In this thesis, we showed that the application of RAD-Seq data provides potential to define ESUs and to detect fine-scale genetic structure. For example, we were able to assign *P. y. mauretanicus* and *P. y. yelkouan* individuals to the specific island where they were born (Chapter III). This opens up the possibility to develop management-relevant genetic-based assays to identify the population of origin of shearwaters that die from fisheries bycatch, by using a selection of the most informative SNPs. Such approaches have proved successful in the genetic assignment of other marine organisms (Nielsen et al. 2012; Meek et al. 2016; Jenkins et al. 2019) and applied to shearwaters would allow to pinpoint the specific populations more severely affected by fisheries bycatch. RAD-Seq approaches have proven useful to detect fine-scale population structure in several other seabird species (Younger et al. 2017; Vianna et al. 2017; Rexer-Huber et al. 2019; Taylor et al. 2019) and have the potential to become an excellent companion to wildlife management.

Conclusions

1) Integrating information from genomic markers that evolve at different rates for phylogenetic and introgression analyses allows a detailed exploration of the causes of phylogenetic incongruence at different timescales. This leads to improved resolution of phylogenetic relationships and divergence dating.

2) Phylogenomic analyses provided enough resolution to resolve the phylogenetic relationships among shearwaters, including the *Puffinus* genus radiation, and revealed that the majority of phylogenetic conflict is driven by high levels of incomplete lineage sorting (ILS) due to rapid speciation events. Yet conflict is also caused by processes such as GC-biased gene conversion, lineage-specific rate heterogeneity and ancestral population structure.

3) Divergence time estimation analyses and the fossil record pinpointed a severe impact of the Pliocene marine megafauna extinction on shearwaters, probably caused by a sudden reduction in the availability of coastal habitat. The late Pliocene-early Pleistocene was inferred as a period of high and rapid speciation and dispersal, probably promoted by Pleistocene climatic shifts.

4) Ancestral range estimation analyses suggest that shearwaters originated in the Pacific Ocean, most probably in the Southern Australia and New Zealand area. This is in contrast with previous studies, which suggested that the North Atlantic was the ancestral area.

5) Surface ocean currents have promoted shearwater species dispersal. Because winds are a major determinant of seabird movement and are a primary driver of surface ocean currents, this suggests that winds could be an important determinant of global seabird dispersal.

6) Founder-event speciation is a main mode of speciation in shearwaters, although widespread vicariance also seems to be important.

7) Shearwater body mass is highly correlated with their migratory strategy and the latitudinal range they occupy, suggesting that movement capacity is a major determinant of body mass in shearwaters.

8) Analyses of recent coancestry and genomic differentiation detected incongruences between genomic data and the current shearwater taxonomic classifications. These results highlighted that the species status of *P. mauretanicus, P. yelkouan, A. creatopus, A carneipes, P. boydi, P. baroli, C. edwardsii, C. diomedea* and *C. borealis* should be reconsidered.

9) The integration of genomic data with morphological and ecological evidence did not support the current hypotheses of species delimitation for the North Atlantic and Mediterranean *Puffinus* shearwaters, and allowed to propose a new and more robust delimitation.

10) The detection of fine-scale genetic structure within *Puffinus* species highlights the need for management of ESUs below the species level.

11) Genomic landscapes of divergence in North Atlantic and Mediterranean *Puffinus* shearwaters appear to be shaped by genetic drift and, to a minor extent, by divergent selection.

12) Different effects of genetic drift among populations and species can be a confounding factor for the detection of introgression using ABBA-BABA tests and population assignment programs.

References

Abbott C.L., Double M.C. 2003. Genetic structure, conservation genetics and evidence of speciation by range expansion in shy and white-capped albatrosses. Mol. Ecol. 12:2953–2962.

Abbott R.J., Barton N.H., Good J.M. 2016. Genomics of hybridization and its evolutionary consequences. Mol. Ecol. 25:2325–2332.

Abdelkrim J., Aznar-Cormano L., Buge B., Fedosov A., Kantor Y., Zaharias P., Puillandre N. 2018. Delimiting species of marine gastropods (Turridae, Conoidea) using RAD sequencing in an integrative taxonomy framework. Mol. Ecol. 27:4591–4611.

Aberer A.J., Kobert K., Stamatakis A. 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–2556.

ACAP. 2014. Best Practice Seabird Bycatch Mitigation Criteria And Definition. Available at: https://www.bmis-bycatch.org/references/s62anjgw [accessed October, 2020].

Alfaro M.E., Faircloth B.C., Harrington R.C., Sorenson L., Friedman M., Thacker C.E., Oliveros C.H., Černý D., Near T.J. 2018. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. Nature Ecology & Evolution.

Andermann T., Fernandes A.M., Olsson U., Töpel M., Pfeil B., Oxelman B., Aleixo A., Faircloth B.C., Antonelli A. 2018. Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements. Syst. Biol. 68:32–46.

Anderson E., Others. 1949. Introgressive hybridization. Introgressive hybridization.

Anderson O.R.J., Small C.J., Croxall J.P., Dunn E.K. 2011. Global seabird bycatch in longline fisheries. Endanger. Species Res.

Angelis K., Dos Reis M. 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. Curr. Zool. 61:874–885.

Arcila D., Ortí G., Vari R., Armbruster J.W., Stiassny M.L.J., Ko K.D., Sabaj M.H., Lundberg J., Revell L.J., Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat Ecol Evol. 1:20.

Austin J.J. 1996. Molecular phylogenetics of Puffinus shearwaters: preliminary evidence from mitochondrial cytochrome b gene sequences. Mol. Phylogenet. Evol. 6:77–88.

Austin J.J., Bretagnolle V., Pasquet E. 2004. A global molecular phylogeny of the small Puffinus shearwaters and implications for systematics of the Little-Audubon's Shearwater complex. Auk. 121:647–864.

Austin R.E., Wynn R.B., Votier S.C., Trueman C., McMinn M., Rodríguez A., Suberg L., Maurice L., Newton J., Genovart M., Péron C., Grémillet D., Guilford T. 2019. Patterns of at-sea behaviour at a hybrid zone between two threatened seabirds. Sci. Rep. 9:14720.

Barry Cox C., Moore P.D. 2005. Biogeography: An Ecological and Evolutionary Approach. Wiley.

Bicknell A.W.J., Knight M.E., Bilton D., Reid J.B., Burke T., Votier S.C. 2012. Population genetic structure and long-distance dispersal among seabird populations: Implications for colony persistence. Mol. Ecol. 21:2863–2876.

Blair C., Ané C. 2019. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. Syst. Biol.

Bonnaud E., Medina F.M., Vidal E., Nogales M. 2011. The diet of feral cats on islands: a review and a call for more studies. Biologicals.

Bossert S., Murray E.A., Blaimer B.B., Danforth B.N. 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol. Phylogenet. Evol. 111:149–157.

Bourne W. R. P. Mackrill E. J. Paterson A. M. Yésou P. 1988. The Yelkouan Shearwater Puffinus (puffinus?) yelkouan. Br. Birds. 81:306—319.

Brodier S., Pisanu B., Villers A., Pettex E. 2011. Responses of seabirds to the rabbit eradication on Ile Verte, sub-Antarctic Kerguelen Archipelago. Animal.

Brooke M. 2004. Albatrosses and petrels across the world. Oxford University Press.

Brooke M. 2013. The Manx Shearwater. A&C Black.

Brooke M. de L., Rowe G. 1996. Behavioural and molecular evidence for specific status of light and dark morphs of the Herald Petrel Pterodroma heraldica. Ibis . 138:420–432.

Bryant D., Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Buckley T.R., Simon C., Chambers G.K. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50:67–86.

Burg T.M., Croxall J.P. 2001. Global relationships amongst black-browed and grey-headed albatrosses: analysis of population structure using mitochondrial DNA and microsatellites. Mol. Ecol. 10:2647–2660.

Burri R. 2017. Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. Molecular Ecology. 26:3853–3856.

Burri R., Nater A., Kawakami T., Mugal C.F., Olason P.I., Smeds L., Suh A., Dutoit L., Bureš S., Garamszegi L.Z., Hogner S., Moreno J., Qvarnström A., Ružić M., Sæther S.A., Sætre G.P., Török J., Ellegren H. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 25:1656–1665.

Carboneras C., Bonan A. 1992. Petrels, Shearwaters (Procellariidae). in Handbook of the Birds of the World (J. del Hoyo, A. Elliott, J. Sargatal, D. Christie, and E. de Juana, eds.). Lynx Edicions, Barcelona, Spain.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. Mol. Ecol. 22:4369–4383.

Chambers E.A., Hillis D.M. 2020. The Multispecies Coalescent Over-Splits Species in the Case of Geographically Widespread Taxa. Syst. Biol. 69:184–193.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10:195–205.

Christidis L., Dickinson E.C., Remsen J.V., Cracraft J., Peters S., Kuziemko M., Lepage D. 2018. The Howard and Moore complete checklist of the birds of the world, version 4.1 (Downloadable checklist). .

Clements J.F. 2007. The Clements Checklist of Birds of the World 6th Edition. .

Collins R.A., Hrbek T. 2018. An In Silico Comparison of Protocols for Dated Phylogenomics. Syst. Biol. 0:1–11.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 5:536–544.

Cortés V., Arcos J.M., González-Solís J. 2017. Seabirds and demersal longliners in the northwestern Mediterranean: factors driving their interactions and bycatch rates. Mar. Ecol. Prog. Ser. 565:1–16.

Cortés V., González-Solís J. 2018. Seabird bycatch mitigation trials in artisanal demersal longliners of the Western Mediterranean. PLOS ONE. 13:e0196731.

Cortez D., Marin R., Toledo-Flores D., Froidevaux L., Liechti A., Waters P.D., Grützner F., Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. Nature. 508:488–493.

Coulson J. 2002. Colonial breeding in seabirds. Biology of marine birds.:87–113.

Cowie R.H., Holland B.S. 2006. Dispersal is fundamental to biogeography and the evolution of biodiversity on oceanic islands. Journal of Biogeography. 33:193–198.

Coyne J.A., Orr H.A. 1989. PATTERNS OF SPECIATION IN DROSOPHILA. Evolution. 43:362–381.

Coyne J.A., Orr H.A., Others. 2004. Speciation. Sinauer Associates Sunderland, MA.

Crandall K.A., Bininda-Emonds O.R., Mace G.M., Wayne R.K. 2000. Considering evolutionary processes in conservation biology. Trends Ecol. Evol. 15:290–295.

Crisci J.V. 2001. The voice of historical biogeography. Journal of Biogeography. 28:157–168.

Cristofari R., Bertorelle G., Ancel A., Benazzo A., Le Maho Y., Ponganis P.J., Stenseth N.C., Trathan P.N., Whittington J.D., Zanetti E., Zitterbart D.P., Le Bohec C., Trucchi E. 2016. Full circumpolar migration ensures evolutionary unity in the Emperor penguin. Nat. Commun. 7:11842.

Croxall J.P., Butchart S.H.M., Lascelles B., Stattersfield A.J., Sullivan B., Symes A., Taylor P. 2012. Seabird conservation status, threats and priority actions: a global assessment. Bird Conservation International. 22:1–34.

Cruaud A., Gautier M., Galan M., Foucaud J., Saun?? L., Genson G., Dubois E., Nidelet S., Deuve T., Rasplus J.Y. 2014. Empirical assessment of rad sequencing for interspecific phylogeny. Mol. Biol. Evol. 31:1272–1274.

Cruickshank T.E., Hahn M.W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23:3133–3157.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution. 24:332–340.

Del Hoyo J., Collar N.J., Christie D.A., Elliott A., Fishpool L.D.C. 2014. HBW and BirdLife International Illustrated Checklist of the Birds of the World: non-passerines. Lynx Edicions Barcelona, España.

Deppe L., Rowley O., Rowe L.K., Shi N., Gooday O. 2017. Investigation of fallout events in Hutton's shearwaters (Puffinus huttoni) associated with artificial lighting. .

De Queiroz K. 2005. Ernst Mayr and the modern concept of species. Proceedings of the National Academy of Sciences. 102:6600–6607.

De Queiroz K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886.

Dias M.P., Martin R., Pearmain E.J., Burfield I.J., Small C., Phillips R.A., Yates O., Lascelles B., Borboroglu P.G., Croxall J.P. 2019. Threats to seabirds: A global assessment. Biol. Conserv. 237:525–537.

Díaz-Arce N., Arrizabalaga H., Murua H., Irigoien X., Rodríguez-Ezpeleta N. 2016. RAD-seq derived genome-wide nuclear markers resolve the phylogeny of tunas. Mol. Phylogenet. Evol. 102:202–207.

Doolittle R.F., Feng D.F., Tsang S., Cho G., Little E. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science. 271:470–477.

Doolittle W.F., Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc. Natl. Acad. Sci. U. S. A. 104:2043–2049.

DPIPWE. 2018. Short-Tailed Shearwater (Muttonbird). Hobart TAS: Department of Primary Industries, Parks, Water and Environment. Government of Tasmania.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

Eaton D.A.R., Hipp A.L., González-Rodríguez A., Cavender-Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution. 69:2587–2601.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Edwards S.V., Arctander P., Wilson A.C. 1991. Mitochondrial resolution of a deep branch in the genealogical tree for perching birds. Proc. Biol. Sci. 243:99–107.

Edwards S.V., Kingan S.B., Calkins J.D., Balakrishnan C.N., Jennings W.B., Swanson W.J., Sorenson M.D. 2005. Speciation in birds: genes, geography, and sexual selection. Proc. Natl. Acad. Sci. U. S. A. 102 Suppl 1:6550–6557.

Edwards S.V., Potter S., Schmitt C.J., Bragg J.G., Moritz C. 2016. Reticulation, divergence, and the phylogeography–phylogenetics continuum. Proceedings of the National Academy of Sciences. 113:8025–8032.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 6:1–10.

Emerson K.J., Merz C.R., Catchen J.M., Hohenlohe P. a., Cresko W. a., Bradshaw W.E., Holzapfel C.M. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. U. S. A. 107:16196–16200.

Ericson P.G.P., Anderson C.L., Britton T., Elzanowski A., Johansson U.S., Källersjö M., Ohlson J.I., Parsons T.J., Zuccon D., Mayr G. 2006. Diversification of Neoaves: integration of molecular sequence data and fossils. Biol. Lett. 2:543–547.

Eriksson A., Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. U. S. A. 109:13956–13960.

Estandia A. 2019. Genome-wide phylogenetic reconstruction for Procellariiform seabirds is robust to molecular rate variation. MSc Dissertation. Durham University.

Ewart K.M., Lo N., Ogden R., Joseph L., Ho S.Y.W., Frankham G.J., Eldridge M.D.B., Schodde R., Johnson R.N. 2020. Correction: Phylogeography of the iconic Australian red-tailed black-cockatoo (Calyptorhynchus banksii) and implications for its conservation. Heredity . 125:167.

Excoffier L., Foll M. 2011. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics. 27:1332–1334.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.

Farris J.S. 1970. Methods for Computing Wagner Trees. Syst. Biol. 19:83–92.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Systematic Biology. 27:401–410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Feng Y.-J., Blackburn D.C., Liang D., Hillis D.M., Wake D.B., Cannatella D.C., Zhang P. 2017. Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. Proc. Natl. Acad. Sci. U. S. A. 114:E5864–E5870.

Ferchaud A.-L., Hansen M.M. 2016. The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: Three-spine sticklebacks in divergent environments. Mol. Ecol. 25:238–259.

Fitch W.M., Margoliash E. 1967. Construction of phylogenetic trees. Science. 155:279–284.

Flood R.L., Fisher E.A. 2020. North Atlantic seabirds: shearwaters, Jouanin's & White-chinned Petrels. .

Flouri T., Jiao X., Rannala B., Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35:2585–2593.

Flouri T., Jiao X., Rannala B., Yang Z. 2020. A Bayesian Implementation of the Multispecies Coalescent Model with Introgression for Phylogenomic Analysis. Mol. Biol. Evol. 37:1211–1223.

Fraser D.J., Bernatchez L. 2001. Adaptive evolutionary conservation: towards a unified concept for defining conservation units. Mol. Ecol. 10:2741–2752.

Friesen V.L. 2015. Speciation in seabirds: why are there so many species... and why aren't there more? J. Ornithol. 156:27–39.

Friesen V.L., Burg T.M., McCoy K.D. 2007a. Mechanisms of population differentiation in seabirds: Invited review. Mol. Ecol. 16:1765–1785.

Friesen V.L., Smith A.L., Gomez-Diaz E., Bolton M., Furness R.W., González-Solís J., Monteiro L.R. 2007b. Sympatric speciation by allochrony in a seabird. Proceedings of the National Academy of Sciences. 104:18589–18594.

Gangloff B., Zino F., Shirihai H., González-Solís J., Couloux A., Pasquet E., Bretagnolle V. 2013. The evolution of north-east Atlantic gadfly petrels using statistical phylogeography. Mol. Ecol. 22:495–507.

Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80:231–266.

Genovart M., Arcos J.M., Álvarez D., McMinn M., Meier R., Wynn R., Guilford T., Oro D. 2016. Demography of the critically endangered Balearic shearwater: the impact of fisheries and time to extinction. J. Appl. Ecol.

Genovart M., Bécares J., Igual J.-M., Martínez-Abraín A., Escandell R., Sánchez A., Rodríguez B., Arcos J.M., Oro D. 2018. Differential adult survival at close seabird colonies: The importance of spatial foraging segregation and bycatch risk during the breeding season. Glob. Chang. Biol. 24:1279–1290.

Genovart M., Juste J., Contreras-Díaz H., Oro D. 2012. Genetic and phenotypic differentiation between the critically endangered balearic shearwater and neighboring colonies of its sibling species. J. Hered. 103:330–341.

Genovart M., Juste J., Oro D. 2005. Two sibling species sympatrically breeding: A new conservation concern for the critically endangered Balearic shearwater. Conserv. Genet. 6:601–606.

Gill F., Donsker D., Rasmussen P. 2020. IOC World Bird List (v10. 1). doi : 10.14344/IOC.ML.10.1.

Gillespie R.G., Baldwin B.G., Waters J.M., Fraser C.I., Nikula R., Roderick G.K. 2012. Long-distance dispersal: a framework for hypothesis testing. Trends Ecol. Evol. 27:47–56.

Gil-Velasco M., Rodriguez G., Menzie S., Arcos J.M. 2015. Plumage variability and field identification of Manx, Yelkouan and Balearic Shearwaters. 108.

Gómez-Díaz E., González-Solís J., Peinado M.A. 2009. Population structure in a highly pelagic seabird, the Cory's shearwater Calonectris diomedea: An examination of genetics, morphology and ecology. Mar. Ecol. Prog. Ser. 382:197–209.

Gómez-Díaz E., González-Solís J., Peinado M.A., Page R.D.M. 2006. Phylogeography of the Calonectris shearwaters using molecular and morphometric data. Mol. Phylogenet. Evol. 41:322–332.

González-Solís J., Croxall J.P., Oro D., Ruiz X. 2007. Trans-equatorial migration and mixing in the wintering areas of a pelagic seabird. Front. Ecol. Environ. 5:297–301.

Gotch A.F. 1979. Albatrosses, Fulmars, Shearwaters, and Petrels. Latin Names Explained A Guide to the Scientific Classifications of Reptiles, Birds & Mammals.:191–192.

Granadeiro J.P., Monteiro L.R., Furness R.W. 1998. Diet and feeding ecology of Cory$\backslash$'s shearwater Calonectris diomedea in the Azores, north-east Atlantic. Mar. Ecol. Prog. Ser. 166:267–276.

Gray R.D., Drummond A.J., Greenhill S.J. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science. 323:479–483.

Gray S.M., McKinnon J.S. 2007. Linking color polymorphism maintenance and speciation. Trends Ecol. Evol. 22:71–79.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A draft sequence of the Neandertal genome. Science. 328:710–722.

Grémillet D., Ponchon A., Paleczny M., Palomares M.-L.D., Karpouzi V., Pauly D. 2018. Persisting Worldwide Seabird-Fishery Competition Despite Seabird Community Decline. Curr. Biol. 28:4009–4013.e2.

Guilford T., Meade J., Willis J., Phillips R., Boyle D., Roberts S., Collett M., Freeman R., Perrins C.M. 2009. Migration and stopover in a small pelagic seabird, the Manx shearwater Puffinus puffinus: insights from machine learning. Proceedings of the Royal Society B: Biological Sciences. 276:1215–1223.

Hackett S.J., Kimball R.T., Reddy S., Bowie R.C.K., Braun E.L., Braun M.J., Chojnowski J.L., Cox W.A., Han K.-L., Harshman J., Huddleston C.J., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Steadman D.W., Witt C.C., Yuri T. 2008. A phylogenomic study of birds reveals their evolutionary history. Science. 320:1763–1768.

Halanych K.M., Bacheller J.D., Aguinaldo A.M., Liva S.M., Hillis D.M., Lake J.A. 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science. 267:1641–1643.

Haller B.C., Messer P.W. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Mol. Biol. Evol. 36:632–637.

Han F., Lamichhaney S., Grant B.R., Grant P.R., Andersson L., Webster M.T. 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Res. 27:1004–1015.

Heath T.A., Moore B.R. 2014. Bayesian inference of species divergence times. Bayesian phylogenetics: methods algorithms, and applications.:277–318.

Heidrich P., Amengual J., Wink M. 1998. Phylogenetic relationships in mediterranean and North Atlantic shearwaters (Aves: Procellariidae) based on nucleotide sequences of mtDNA. Biochem. Syst. Ecol. 26:145–170.

Hejase H.A., Salman-Minkov A., Campagna L., Hubisz M.J., Lovette I.J., Gronau I., Siepel A. 2020. Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. Proc. Natl. Acad. Sci. U. S. A.

Hendry A.P., Day T. 2005. Population structure attributable to reproductive time: isolation by time and adaptation by time. Mol. Ecol. 14:901–916.

Hey J. 2006. On the failure of modern species concepts. Trends Ecol. Evol. 21:447–450.

Hohenlohe P.A., Bassham S., Etter P.D., Stiffler N., Johnson E.A., Cresko W.A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6.

Holmes N.D., Spatz D.R., Oppel S., Tershy B., Croll D.A., Keitt B., Genovesi P., Burfield I.J., Will D.J., Bond A.L., Wegmann A., Aguirre-Muñoz A., Raine A.F., Knapp C.R., Hung C.-H., Wingate D., Hagen E., Méndez-Sánchez F., Rocamora G., Yuan H.-W., Fric J., Millett J., Russell J., Liske-Clark J., Vidal E., Jourdan H., Campbell K., Springer K., Swinnerton K., Gibbons-Decherong L., Langrand O., Brooke M. de L., McMinn M., Bunbury N., Oliveira N., Sposimo P., Geraldes P., McClelland P., Hodum P., Ryan P.G., Borroto-Páez R., Pierce R., Griffiths R., Fisher R.N., Wanless R., Pasachnik S.A., Cranwell S., Micol T., Butchart S.H.M. 2019. Globally important islands where eradicating invasive mammals will benefit highly threatened vertebrates. PLoS One. 14:e0212128.

Hosegood J., Humble E., Ogden R., de Bruyn M., Creer S., Stevens G.M.W., Abudaya M., Bassos-Hull K., Bonfil R., Fernando D., Foote A.D., Hipperson H., Jabado R.W., Kaden J., Moazzam M., Peel L.R., Pollett S., Ponzo A., Poortvliet M., Salah J., Senn H., Stewart J.D., Wintner S., Carvalho G. 2020. Phylogenomics and species delimitation for effective conservation of manta and devil rays. Mol. Ecol.

Ho S.Y.W. 2014. The changing face of the molecular evolutionary clock. Trends Ecol. Evol. 29:496–503.

Ho S.Y.W., Phillips M.J. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. 58:367–380.

Howard H. 1971. Pliocene avian remains from Baja California. Los Angeles County Museum of Natural History.

Howell S.N.G. 2012. Petrels, Albatrosses, and Storm-Petrels of North America: A Photographic Guide. Princeton University Press.

Hughes L.C., Ortí G., Huang Y., Sun Y., Baldwin C.C., Thompson A.W., Arcila D., Betancur-R R., Li C., Becker L., Bellora N., Zhao X., Li X., Wang M., Fang C., Xie B., Zhou Z., Huang H., Chen S., Venkatesh B., Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. U. S. A. 115:6249–6254.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Irwin D.E., Milá B., Toews D.P.L., Brelsford A., Kenyon H.L., Porter A.N., Grossen C., Delmore K.E., Alcaide M., Irwin J.H. 2018. A comparison of genomic islands of differentiation across three young avian species pairs. Mol. Ecol.

Jamie G.A., Meier J.I. 2020. The Persistence of Polymorphisms across Species Radiations. Trends Ecol. Evol. 0.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., Fonseca R.R. da, Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F., Brumfield R.T., Mello C., Lovell P.V., Wirthlin M., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D., Zhou Q., Perelman P., Driskell A.C., Ruby G., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker K., Jønsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole Genome Analyses Resolve the Early Branches in the Tree of Life of Modern Birds. Science. 346:1126–1138.

Jay P., Whibley A., Frézal L., Rodríguez de Cara M.Á., Nowell R.W., Mallet J., Dasmahapatra K.K., Joron M. 2018. Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion. Curr. Biol. 28:1839–1845.e3.

Jenkins M. 2003. Prospects for Biodiversity. Science. 302:1175–1177.

Jenkins T.L., Ellis C.D., Triantafyllidis A., Stevens J.R. 2019. Single nucleotide polymorphisms reveal a genetic cline across the north-east Atlantic and enable powerful population assignment in the European lobster. Evol. Appl. 12:1881–1899.

Jones H.P., Tershy B.R., Zavaleta E.S. 2008. Severity of the effects of invasive rats on seabirds: a global review. Conservation.

Jones J.C., Fan S., Franchini P., Schartl M., Meyer A. 2013. The evolutionary history of Xiphophorus fish and their sexually selected sword: A genome-wide approach using restriction site-associated DNA sequencing. Mol. Ecol. 22:2986–3001.

Jouanin C., Mougin J.L. 1979. Order Procellariiformes. Check-list of birds of the world. 1:48–121.

Jukes T.H., Cantor C.R. 1969. Evolution of Protein Molecules. Mammalian Protein Metabolism.:21–132.

Jun Inoue P.C.J.D.&. Z.Y. 2010. The Impact of the Representation of Fossil Calibrations on Bayesian Estimation of Species Divergence Times. Syst. Biol. 59:74–89.

Kamm J., Terhorst J., Durbin R., Song Y.S. 2019. Efficiently inferring the demographic history of many populations with allele count data. J. Am. Stat. Assoc. 0:1–42.

Kearns A.M., Restani M., Szabo I., Schrøder-Nielsen A., Kim J.A., Richardson H.M., Marzluff J.M., Fleischer R.C., Johnsen A., Omland K.E. 2018. Genomic evidence of speciation reversal in ravens. Nat. Commun. 9:906.

Kennedy M., Page R.D.M. 2002. Seabird Supertrees: Combining Partial Estimates of Procellariiform Phylogeny. Auk. 119:88–108.

Kidd M.G., Friesen V.L. 1998. ANALYSIS OF MECHANISMS OF MICROEVOLUTIONARY CHANGE IN CEPPHUS GUILLEMOTS USING PATTERNS OF CONTROL REGION VARIATION. Evolution. 52:1158–1168.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Kingman J.F.C. 1982. The coalescent. Stochastic Process. Appl. 13:235–248.

Knowles L.L., Carstens B.C. 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56:887–895.

Knowles L.L., Huang H., Sukumaran J., Smith S.A. 2018. A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. Am. J. Bot. 105:376–384.

Koonin E.V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–338.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics.:1–3.

Kumar S. 2005. Molecular clocks: four decades of evolution. Nat. Rev. Genet. 6:654–662.

Kuroda N. 1954. On the Classification and Phylogeny of the Order Tubinares, Particularly the Shearwaters (Puffinus), with Special Considarations [ie Considerations] on Their Osteology and Habit Differentiation. Kuroda.

Kutschera V.E., Bidon T., Hailer F., Rodi J.L., Fain S.R., Janke A. 2014. Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol. Biol. Evol. 31:2004–2017.

Landis M.J., Matzke N.J., Moore B.R., Huelsenbeck J.P. 2013. Bayesian analysis of biogeography when the number of areas is large. Syst. Biol. 62:789–804.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.

Langley C.H., Fitch W.M. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.

Lartillot N. 2013. Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. Mol. Biol. Evol. 30:489–502.

Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: Conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.

Leaché A.D., Fujita M.K., Minin V.N., Bouckaert R.R. 2014. Species delimitation using genome-wide SNP Data. Syst. Biol. 63:534–542.

Lee M.S.Y., Palci A. 2015. Morphological Phylogenetics in the Genomic Age. Curr. Biol. 25:R922–9.

Lescak E.A., Bassham S.L., Catchen J., Gelmond O., Sherbick M.L., von Hippel F.A., Cresko W.A. 2015. Evolution of stickleback in 50 years on earthquake-uplifted islands. Proc. Natl. Acad. Sci. U. S. A.:201512020.

Lesecque Y., Mouchiroud D., Duret L. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Mol. Biol. Evol. 30:1409–1419.

Liebers D., Helbig A.J., De Knijff P. 2001. Genetic differentiation and phylogeography of gulls in the Larus cachinnans--fuscus group (Aves: Charadriiformes). Mol. Ecol. 10:2447–2462.

Li H., Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature. 475:493–496.

Lindblad-Toh K., Garber M., Zuk O., Lin M.F., Parker B.J., Washietl S., Kheradpour P., Ernst J., Jordan G., Mauceli E., Ward L.D., Lowe C.B., Holloway A.K., Clamp M., Gnerre S., Alföldi J., Beal K., Chang J., Clawson H., Cuff J., Di Palma F., Fitzgerald S., Flicek P., Guttman M., Hubisz M.J., Jaffe D.B., Jungreis I., Kent W.J., Kostka D., Lara M., Martins A.L., Massingham T., Moltke I., Raney B.J., Rasmussen M.D., Robinson J., Stark A., Vilella A.J., Wen J., Xie X., Zody M.C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin J., Bloom T., Chin C.W., Heiman D., Nicol R., Nusbaum C., Young S., Wilkinson J., Worley K.C., Kovar C.L., Muzny D.M., Gibbs R.A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree A., Dihn H.H., Fowler G., Jhangiani S., Joshi V., Lee S., Lewis L.R., Nazareth L.V., Okwuonu G., Santibanez J., Warren W.C., Mardis E.R., Weinstock G.M., Wilson R.K., Genome Institute at Washington University, Delehaunty K., Dooling D., Fronik C., Fulton L., Fulton B., Graves T., Minx P., Sodergren E., Birney E., Margulies E.H., Herrero J., Green E.D., Haussler D., Siepel A., Goldman N., Pollard K.S., Pedersen J.S., Lander E.S., Kellis M. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 478:476–482.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360:36–53.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.

Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

Longo S.J., Faircloth B.C., Meyer A., Westneat M.W., Alfaro M.E., Wainwright P.C. 2017. Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. Mol. Phylogenet. Evol. 113:33–48.

Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature.

Lowry D.B., Hoban S., Kelley J.L., Lotterhos K.E., Reed L.K., Antolin M.F., Storfer A. 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol. Ecol. Resour. 17:142–152.

Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Maguire K.C., Stigall A.L. 2008. Paleobiogeography of Miocene Equinae of North America: A phylogenetic biogeographic analysis of the relative roles of climate, vicariance, and dispersal. Palaeogeogr. Palaeoclimatol. Palaeoecol. 267:175–184.

Mancera E., Bourgon R., Brozzi A., Huber W., Steinmetz L.M. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature. 454:479–485.

Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32:244–257.

Martin S.H., Jiggins C.D. 2017. Interpreting the genomic landscape of introgression. Curr. Opin. Genet. Dev. 47:69–74.

Matzke N.J. 2013. BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. R package, version 0. 2. 1:2013.

Matzke N.J. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Syst. Biol. 63:951–970.

Mayden R.L. 1997. A Hierarchy of Species Concepts: The Denouement in the Saga of the Species Problem. In: Claridge M.F., Dawah H.A., Wilson M.R., editors. Species: The units of diversity,. Chapman & Hall. p. 381–423.

Mayr E. 1963. Animal Species and. Evolution.:521.

Mayr E. 1999. Systematics and the Origin of Species, from the Viewpoint of a Zoologist. Harvard University Press.

McCormack J.E., Heled J., Delaney K.S., Peterson A.T., Knowles L.L. 2011. Calibrating divergence times on species trees versus gene trees: implications for speciation history of Aphelocoma jays. Evolution. 65:184–202.

McGowen M.R., Tsagkogeorga G., Williamson J., Morin P.A., Rossiter A.S.J. 2020. Positive Selection and Inactivation in the Vision and Hearing Genes of Cetaceans. Mol. Biol. Evol. 37:2069–2083.

Meek M.H., Baerwald M.R., Stephens M.R., Goodbla A., Miller M.R., Tomalty K.M.H., May B. 2016. Sequencing improves our ability to study threatened migratory species: Genetic population assignment in California's Central Valley Chinook salmon. Ecol. Evol. 6:7706–7716.

Mendes F.K., Hahn M.W. 2016. Gene Tree Discordance Causes Apparent Substitution Rate Variation. Systematic Biology. 65:711–721.

Miller L. 1961. Birds from the Miocene of Sharktooth Hill, California. Condor. 63:399–402.

Miller M., Dunham J., Amores A., Cresko W., Johnson E. 2007. genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17:240–248.

Milot E., Weimerskirch H., Bernatchez L. 2008. The seabird paradox: Dispersal, genetic structure and population dynamics in a highly mobile, but philopatric albatross species. Mol. Ecol. 17:1658–1673.

Mirarab S., Warnow T. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 31:i44–i52.

Moritz C. 1994. Defining "Evolutionarily Significant Units" for conservation. Trends Ecol. Evol. 9:373–375.

Morris-Pocock J.A., Anderson D.J., Friesen V.L. 2011. Mechanisms of global diversification in the brown booby (Sula leucogaster) revealed by uniting statistical phylogeographic and multilocus phylogenetic methods. Mol. Ecol. 20:2835–2850.

Mullis K.B. 1990. The unusual origin of the polymerase chain reaction. Sci. Am. 262:56–61, 64–5.

Munilla I., Genovart M., Paiva V.H., Velando A. 2016. Colony Foundation in an Oceanic Seabird. PLoS One. 11:1–24.

Nachman M.W., Payseur B.A. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367:409–421.

NewmaN J., ScOTT D., Bragg C., McKechnie S., Moller H., Fletcher D. 2009. Estimating regional population size and annual harvest intensity of the sooty shearwater in New Zealand. N. Z. J. Zool. 36:307–323.

Newton L.G., Starrett J., Hendrixson B.E., Derkarabetian S., Bond J.E. 2020. Integrative species delimitation reveals cryptic diversity in the southern Appalachian Antrodiaetus unicolor (Araneae: Antrodiaetidae) species complex. Mol. Ecol. 29:2269–2287.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Nielsen E.E., Cariani A., Mac Aoidh E., Maes G.E., Milano I., Ogden R., Taylor M., Hemmer-Hansen J., Babbucci M., Bargelloni L., Bekkevold D., Diopere E., Grenfell L., Helyar S., Limborg M.T., Martinsohn J.T., McEwing R., Panitz F., Patarnello T., Tinti F., Van Houdt J.K.J., Volckaert F.A.M., Waples R.S., FishPopTrace consortium, Albin J.E.J., Vieites Baptista J.M., Barmintsev V., Bautista J.M., Bendixen C., Bergé J.-P., Blohm D., Cardazzo B., Diez A., Espiñeira M., Geffen A.J., Gonzalez E., González-Lavín N., Guarniero I., Jeráme M., Kochzius M., Krey G., Mouchel O., Negrisolo E., Piccinetti C., Puyet A., Rastorguev S., Smith J.P., Trentini M., Verrez-Bagnis V., Volkov A., Zanzi A., Carvalho G.R. 2012. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nat. Commun. 3:851.

Nunn G.B., Stanley S.E. 1998. Body size effects and rates of cytochrome b evolution in tube-nosed seabirds. Mol. Biol. Evol. 15:1360–1371.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. Mol. Biol. Evol. 34:2101–2114.

Oliveros C.H., Field D.J., Ksepka D.T., Barker F.K., Aleixo A., Andersen M.J., Alström P., Benz B.W., Braun E.L., Braun M.J., Bravo G.A., Brumfield R.T., Chesser R.T., Claramunt S., Cracraft J., Cuervo A.M., Derryberry E.P., Glenn T.C., Harvey M.G., Hosner P.A., Joseph L., Kimball R.T., Mack A.L., Miskelly C.M., Peterson A.T., Robbins M.B., Sheldon F.H., Silveira L.F., Smith B.T., White N.D., Moyle R.G., Faircloth B.C. 2019. Earth history and the passerine superradiation. Proc. Natl. Acad. Sci. U. S. A. 116:7916–7925.

Olson S.L. 1985. The fossil record of birds. .

Olson S.L. 2010. Stasis and turnover in small shearwaters on Bermuda over the last 400 000 years (Aves: Procellariidae: Puffinus lherminieri group): SMALL BERMUDA SHEARWATERS. Biol. J. Linn. Soc. Lond. 99:699–707.

Olson S.L., Rasmussen P.C., Others. 2001. Miocene and Pliocene birds from the Lee Creek Mine, North Carolina. Smithson. Contrib. Paleobiol. 90:233–365.

Oro D., Aguilar J.S., Igual J.M., Louzao M. 2004. Modelling demography and extinction risk in the endangered Balearic shearwater. Biol. Conserv. 116:93–102.

Oro D., Ruxton G.D. 2001. The formation and growth of seabird colonies: Audouin's gull as a case study. J. Anim. Ecol. 70:527–535.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Parham J.F., Donoghue P.C.J., Bell C.J., Calway T.D., Head J.J., Holroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patané J.S.L., Smith N.D., Tarver J.E., Van Tuinen M., Yang Z., Angielczyk K.D., Greenwood J.M., Hipsley C.A., Jacobs L., Makovicky P.J., Müller J., Smith K.T., Theodor J.M., Warnock R.C.M., Benton M.J. 2012. Best practices for justifying fossil calibrations. Syst. Biol. 61:346–359.

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. 2012. Ancient admixture in human history. Genetics. 192:1065–1093.

Pease J.B., Hahn M.W. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. Evolution. 67:2376–2384.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 7.

Pimiento C., Griffin J.N., Clements C.F., Silvestro D., Varela S., Uhen M.D., Jaramillo C. 2017. The Pliocene marine megafauna extinction and its impact on functional diversity. Nat Ecol Evol. 1:1100–1106.

Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Price T., Others. 2008. Speciation in birds. Roberts and Co.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Pyle P., Welch A.J., Fleischer R.C. 2011. A New Species of Shearwater (Puffinus) Recorded from Midway Atoll, Northwestern Hawaiian Islands. Condor. 113:518–527.

Ramos R., Paiva V.H., Zajková Z., Precheur C., Fagundes A.I., Jodice P.G.R., Mackin W., Zino F., Bretagnolle V., González-Solís J. 2020. Spatial ecology of closely related taxa: the case of the little shearwater complex in the North Atlantic Ocean. Zool. J. Linn. Soc.

Rando J.C., Alcover J.A. 2008. Evidence for a second western Palaearctic seabird extinction during the last Millennium: the Lava Shearwater Puffinus olsoni. Ibis . 150:188–192.

Rannala B. 2015. The art and science of species delimitation. Curr. Zool. 61:846–853.

Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.

Ravinet M., Faria R., Butlin R.K., Galindo J., Bierne N., Rafajlović M., Noor M.A.F., Mehlig B., Westram A.M. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. 30:1450–1477.

Rayner M.J., Hauber M.E., Steeves T.E., Lawrence H. a., Thompson D.R., Sagar P.M., Bury S.J., Landers T.J., Phillips R. a., Ranjard L., Shaffer S. a. 2011. Contemporary and historical separation of transequatorial migration between genetically distinct seabird populations. Nat. Commun. 2:332.

Ree R.H., Sanmartín I. 2018. Conceptual and statistical problems with the DEC+ J model of founder-event speciation and its comparison with DEC via model selection. J. Biogeogr. 45:741–749.

Ree R.H., Smith S.A. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Syst. Biol. 57:4–14.

Reis M., Donoghue P.C.J., Yang Z. 2015. Bayesian molecular clock dating of species divergences in the genomics era. Nat. Rev. Genet. 17:71–80.

Rexer-Huber K., Veale A.J., Catry P., Cherel Y. 2019. Genomics detects population structure within and between ocean basins in a circumpolar seabird: The white-chinned petrel. Molecular.

Rexer-Huber K., Veale A.J., Catry P., Cherel Y., Dutoit L., Foster Y., McEwan J.C., Parker G.C., Phillips R.A., Ryan P.G., Stanworth A.J., van Stijn T., Thompson D.R., Waters J., Robertson B.C. 2019. Genomics detects population structure within and between ocean basins in a circumpolar seabird: The white-chinned petrel. Mol. Ecol. 28:4552–4572.

Ristow D., Wink M. 1981. Sexual dimorphism of Cory's Shearwater. .

Rockman M.V. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution. 66:1–17.

Rodríguez A., Arcos J.M., Bretagnolle V., Dias M.P., Holmes N.D., Louzao M., Provencher J., Raine A.F., Ramírez F., Rodríguez B., Ronconi R.A., Taylor R.S., Bonnaud E., Borrelle S.B., Cortés V., Descamps S., Friesen V.L., Genovart M., Hedd A., Hodum P., Humphries G.R.W., Le Corre M., Lebarbenchon C., Martin R., Melvin E.F., Montevecchi W.A., Pinet P., Pollet I.L., Ramos R., Russell J.C., Ryan P.G., Sanz-Aguilar A., Spatz D.R., Travers M., Votier S.C., Wanless R.M.,

Woehler E., Chiaradia A. 2019. Future Directions in Conservation Research on Petrels and Shearwaters. Frontiers in Marine Science. 6:94.

Rodríguez A., Holmes N.D., Ryan P.G., Wilson K.-J., Faulquier L., Murillo Y., Raine A.F., Penniman J.F., Neves V., Rodríguez B., Others. 2017. Seabird mortality induced by land-based artificial lights. Conserv. Biol. 31:986–1001.

Rodríguez A., Rodríguez B., Curbelo Á.J. 2012. Factors affecting mortality of shearwaters stranded by light pollution. Animal.

Rodríguez A., Rodríguez B., Montelongo T., Garcia-Porta J., Pipa T., Carty M., Danielsen J., Nunes J., Silva C., Geraldes P., Medina F.M., Illera J.C. 2020. Cryptic differentiation in the Manx Shearwater hinders the identification of a new endemic subspecies. J. Avian Biol. n/a.

Rodríguez F., Oliver J.L., Marín A., Medina J.R. 1990. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology. 142:485–501.

Rojas D., Warsi O.M., Dávalos L.M. 2016. Bats (Chiroptera: Noctilionoidea) Challenge a Recent Origin of Extant Neotropical Diversity. Syst. Biol. 65:432–448.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Romiguier J., Ranwez V., Douzery E.J.P., Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20:1001–1009.

Ronquist F. 1997. Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography. Syst. Biol. 46:195–203.

Ronquist F., Sanmartín I. 2011. Phylogenetic Methods in Biogeography. .

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hohna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol. 61:539–542.

Rosenberg N.A., Nordborg M. 2002. Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms. Nat. Rev. Genet. 3:380–390.

Rubin B.E.R., Ree R.H., Moreau C.S. 2012. Inferring phylogenies from RAD sequence data. PLoS One. 7.

Sackton T.B., Grayson P., Cloutier A., Hu Z., Liu J.S., Wheeler N.E., Gardner P.P., Clarke J.A., Baker A.J., Clamp M., Edwards S.V. 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. Science. 364:74–78.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.

Sangster G., Collinson J.M., Helbig A.J., Knox A.G., Parkin D.T. 2005. Taxonomic recommendations for British birds: third report†. Ibis . 147:821–826.

Schiffels S., Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46:919–925.

Schumer M., Cui R., Powell D.L., Rosenthal G.G., Andolfatto P. 2016. Ancient hybridization and genomic stabilization in a swordtail fish. Mol. Ecol. 25:2661–2679.

Scotland R.W., Olmstead R.G., Bennett J.R. 2003. Phylogeny reconstruction: the role of morphology. Syst. Biol. 52:539–548.

Seehausen O., Butlin R.K., Keller I., Wagner C.E., Boughman J.W., Hohenlohe P.A., Peichel C.L., Saetre G.-P., Bank C., Brännström Å., Brelsford A., Clarkson C.S., Eroukhmanoff F., Feder J.L., Fischer M.C., Foote A.D., Franchini P., Jiggins C.D., Jones F.C., Lindholm A.K., Lucek K., Maan M.E., Marques D.A., Martin S.H., Matthews B., Meier J.I., Möst M., Nachman M.W., Nonaka E.,

Rennison D.J., Schwarzer J., Watson E.T., Westram A.M., Widmer A. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–192.

Shaffer S.A., Tremblay Y., Weimerskirch H., Scott D., Thompson D.R., Sagar P.M., Moller H., Taylor G.A., Foley D.G., Block B.A., Costa D.P. 2006. Migratory shearwaters integrate oceanic resources across the Pacific Ocean in an endless summer. Proc. Natl. Acad. Sci. U. S. A. 103:12799–12802.

Shoji A., Dean B., Kirk H., Freeman R., Perrins C.M., Guilford T. 2016. The diving behaviour of the Manx Shearwater Puffinus puffinus. Ibis . 158:598–606.

Silva M.C., Matias R., Wanless R.M., Ryan P.G., Stephenson B.M., Bolton M., Ferrand N., Coelho M.M. 2015. Understanding the mechanisms of antitropical divergence in the seabird W hite-faced S torm-petrel (P rocellariiformes: P elagodroma marina) using a multilocus approach. Mol. Ecol. 24:3122–3137.

Silva M.F., Smith A.L., Friesen V.L., Bried J., Hasegawa O., Coelho M.M., Silva M.C. 2016. Mechanisms of global diversification in the marine species Madeiran Storm-petrel Oceanodroma castro and Monteiro's Storm-petrel O. monteiroi: insights from a multi-locus approach. Mol. Phylogenet. Evol. 98:314–323.

Slatkin M. 1996. In Defense of Founder-Flush Theories of Speciation. The American Naturalist. 147:493–505.

Smith B.T., Harvey M.G., Faircloth B.C., Glenn T.C., Brumfield R.T. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63:83–95.

Solís-Lemus C., Ané C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genet. 12:1–21.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: A package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Spatz D.R., Holmes N.D., Reguero B.G., Butchart S.H.M., Tershy B.R., Croll D.A. 2017. Managing Invasive Mammals to Conserve Globally Threatened Seabirds in a Changing Climate. Conservation Letters. 10:736–747.

Speidel L., Forest M., Shi S., Myers S.R. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. 51:1321–1329.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

Stange M., Sánchez-Villagra M.R., Salzburger W., Matschiner M. 2018. Bayesian Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports Miocene Closure of the Panamanian Isthmus. Syst. Biol. 67:681–699.

Sternkopf V., Liebers-Helbig D., Ritz M.S., Zhang J., Helbig A.J., de Knijff P. 2010. Introgressive hybridization and the evolutionary history of the herring gull complex revealed by mitochondrial and nuclear DNA. BMC Evol. Biol. 10:348.

Storey AA., Lien J. 1985. Development of the First North American Colony of Manx Shearwaters. Auk. 102:395–401.

Suh A., Smeds L., Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13:1–18.

Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. U. S. A. 114:1607–1612.

Taylor R.S., Bolton M., Beard A., Birt T., Deane-Coe P., Raine A.F., González-Solís J., Lougheed S.C., Friesen V.L. 2019. Cryptic species and independent origins of allochronic populations within a seabird species complex (Hydrobates spp.). Mol. Phylogenet. Evol. 139:106552.

Telford M.J., Copley R.R. 2011. Improving animal phylogenies with genomic data. Trends Genet. 27:186–195.

Templeton A.R. 2008. The reality and importance of founder speciation in evolution. Bioessays. 30:470–479.

Thanou E., Sponza S., Nelson E.J., Perry A., Wanless S., Daunt F., Cavers S. 2017. Genetic structure in the European endemic seabird, Phalacrocorax aristotelis, shaped by a complex interaction of historical and contemporary, physical and nonphysical drivers. Mol. Ecol. 26:2796–2811.

Tigano A., Damus M., Birt T.P., Morris-Pocock J.A., Artukhin Y.B., Friesen V.L. 2015. The Arctic: Glacial Refugium or Area of Secondary Contact? Inference from the Population Genetic Structure of the Thick-Billed Murre (Uria lomvia), with Implications for Management. Journal of Heredity. 106:238–246.

Tobias J.A., Seddon N., Spottiswoode C.N., Pilgrim J.D., Fishpool L.D.C., Collar N.J. 2010. Quantitative criteria for species delimitation. Ibis . 152:724–746.

Tonini J., Moore A., Stern D., Shcheglovitova M., Ortí G. 2015. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. PLoS Curr. 7.

Tonzo V., Papadopoulou A., Ortego J. 2019. Genomic data reveal deep genetic structure but no support for current taxonomic designation in a grasshopper species complex. Mol. Ecol. 28:3869–3886.

Turner T.L., Hahn M.W., Nuzhdin S.V. 2005. Genomic islands of speciation in Anopheles gambiae. PLoS Biol. 3:e285.

Vargas P., Kayman M.A. 2014. The Tree of Life. Sinauer.

Vianna J.A., Fernandes F.A.N., Frugone M.J., Figueiró H.V., Pertierra L.R., Noll D., Bi K., Wang-Claypool C.Y., Lowther A., Parker P., Le Bohec C., Bonadonna F., Wienecke B., Pistorius P., Steinfurth A., Burridge C.P., Dantas G.P.M., Poulin E., Simison W.B., Henderson J., Eizirik E., Nery M.F., Bowie R.C.K. 2020. Genome-wide analyses reveal drivers of penguin diversification. Proc. Natl. Acad. Sci. U. S. A.

Vianna J.A., Noll D., Dantas G.P.M., Petry M.V., Barbosa A., González-Acuña D., Le Bohec C., Bonadonna F., Poulin E. 2017. Marked phylogeographic structure of Gentoo penguin reveals an ongoing diversification process along the Southern Ocean. Mol. Phylogenet. Evol. 107:486–498.

Vijay N., Bossu C.M., Poelstra J.W., Weissensteiner M.H., Suh A., Kryukov A.P., Wolf J.B.W. 2016. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat. Commun. 7:13195.

Wagner C.E., Keller I., Wittwer S., Selz O.M., Mwaiko S., Greuter L., Sivasundar A., Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–798.

Wang X., Liang D., Jin W., Tang M., Liu S., Zhang P. 2020. Out of Tibet: Genomic Perspectives on the Evolutionary History of Extant Pikas. Mol. Biol. Evol. 37:1577–1592.

Warham J. 1990. The Petrels: Their Ecology and Breeding Systems. A&C Black.

Weber C.C., Boussau B., Romiguier J., Jarvis E.D., Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol. 15:549.

Weimerskirch H., Jouventin P., Mougin J.L., Stahl J.C., Van B.M. 1985. Banding recoveries and the dispersal of seabirds breeding in French Austral and Antarctic Territories. Emu. 85:22–33.

Weimerskirch H., Sagar P.M. 1996. Diving depths of Sooty Shearwaters Puffinus griseus. Ibis . 138:786–788.

Weissensteiner M.H., Bunikis I., Catalán A., Francoijs K.-J., Knief U., Heim W., Peona V., Pophaly S.D., Sedlazeck F.J., Suh A., Warmuth V.M., Wolf J.B.W. 2020. Discovery and population genomics of structural variation in a songbird genus. Nat. Commun. 11:3403.

Welch A.J., Olson S.L., Fleischer R.C. 2014. Phylogenetic relationships of the extinct St Helena petrel, Pterodroma rupinarum Olson, 1975 (Procellariiformes: Procellariidae), based on ancient DNA : St Helena petrel, Pterodroma rupinarum. Zool. J. Linn. Soc. 170:494–505.

Wen D., Nakhleh L. 2018. Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data. Syst. Biol. 67:439–457.

Wen D., Yu Y., Hahn M.W., Nakhleh L. 2016. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol. Ecol. 25:2361–2372.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet. Syst. Biol. 67:735–740.

Woese C.R., Fox G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. 74:5088–5090.

Wolf J.B.W., Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. Nat. Rev. Genet. 18:87–100.

Wolf J.B.W., Lindell J., Backström N. 2010. Speciation genetics: current status and evolving approaches. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365:1717–1733.

Wragg G. 1985. The comparative biology of Fluttering Shearwater and Hutton's Shearwater and their relationship to other shearwater species. .

Wright S. 1943. Isolation by Distance. Genetics. 28:114–138.

Wu C.-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Yamamoto T., Takahashi A., Katsumata N., Sato K., Trathan P.N. 2010. At-sea distribution and behavior of streaked shearwaters (Calonectris leucomelas) during the nonbreeding period. Auk. 127:871–881.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z., Rannala B. 2012. Molecular phylogenetics: Principles and practice. Nat. Rev. Genet. 13:303–314.

Younger J.L., Clucas G.V., Kao D., Rogers A.D., Gharbi K., Hart T., Miller K.J. 2017. The challenges of detecting subtle population structure and its importance for the conservation of emperor penguins. Mol. Ecol. 26:3883–3897.

Zardoya R., Meyer A. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Molecular Biology and Evolution. 13:933–942.

Zarza E., Faircloth B.C., Tsai W.L.E., Bryson R.W., Klicka J., McCormack J.E. 2016. Hidden histories of gene flow in highland birds revealed with genomic markers. Mol. Ecol.:5144–5157.

Zeller D., Cashion T., Palomares M., Pauly D. 2018. Global marine fisheries discards: A synthesis of reconstructed data. Fish Fish . 19:30–39.

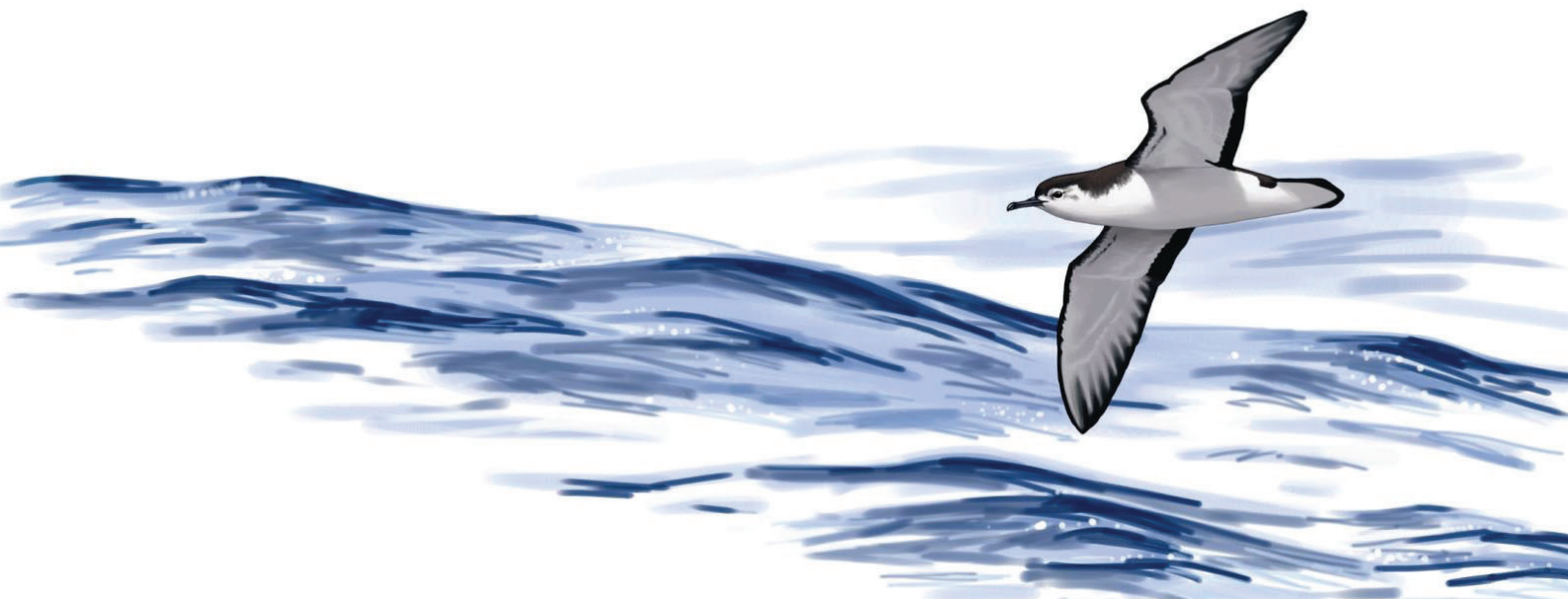
Zheng Y., Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics. 19:10.

Zhong B., Liu L., Penny D. 2014. The multispecies coalescent model and land plant origins: a reply to Springer and Gatesy. Trends Plant Sci. 19:270–272.

Zuckerkandl E., Pauling L. 1965. Evolutionary Divergence and Convergence in Proteins. In: Bryson V., Vogel H.J., editors. Evolving Genes and Proteins. Academic Press. p. 97–166.

Žydelis R., Small C., French G. 2013. The incidental catch of seabirds in gillnet fisheries: A global review. Biol. Conserv.

Appendices



Appendix I

Supplementary Information for:

Integrating Sequence Capture and Restriction-Site Associated DNA Sequencing to Resolve Recent Radiations of Pelagic Seabirds

Joan Ferrer-Obiol, Helen F. James, R. Terry Chesser, Vincent Bretagnolle, Jacob González-Solís, Julio Rozas, Marta Riutort, Andreanna J. Welch

This appendix includes:

Supplementary Information Text

Data Assembly

Three samples did not pass quality control due to insufficient DNA, limited number of reads or contamination issues (P. auricularis_1, P. myrtae_1 and P. bryanni_1) (Table S1).

UCE Assembly Summary Statistics

We used the assembly_get_trinity_coverage.py script from the PHYLUCE pipeline and custom scripts in order to report summary statistics for the UCE assembly per individual (total number of assembled loci, mean sequencing coverage, mean locus length and on target sequence percentage). We calculated and visualised the correlations between different UCE summary statistics using R to assess if they were potentially impacted by the quality of the samples. We observed significant positive correlations between all summary statistics for the UCE assembly (total number of assembled loci, mean sequencing coverage, mean locus length and on target sequence percentage), except for one comparison (total number of assembled loci vs mean locus length) (Fig. S1). Interestingly, the two samples with the lowest number of reads consistently showed very low values for all summary statistics. Therefore, low sequencing yield not only results in low sequencing coverage but also in a lower number of assembled loci, that are shorter on average, and in a lower percentage of sequencing on target.

PE-ddRAD-Seq Quality Control and Filtering

The quality of the sequence PE-ddRAD reads was checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). We removed reads with adapter contamination using TRIMMOMATIC. As the samples were collected at different times, from different sources and by different fieldworkers, we suspected that some reads might come from environmental contaminants such as bacteria and fungi. To remove such sequences we used BOWTIE2 v2.2.3 (Langmead and Salzberg 2012) to align both PE raw reads to a selection of the non-redundant NCBI nucleotide database containing bacteria, fungi and virus sequences. We kept the unpaired reads that failed to align and we filtered out reads only present in one of the two PE files using the

sync_paired_end_reads.py script (https://github.com/sdwfrost/viral-ngs-source). Next, we quality-filtered and demultiplexed reads using PROCESS_RADTAGS in STACKS2 v2.41 (Rochette et al. 2019). We removed reads with uncalled bases (-c) and low quality scores (-q), allowed for barcode rescue (-r), and set a stringent quality threshold of 20 for the average quality score within sliding windows (-s 20). Samples with an abnormally low number of reads were discarded.

PE-ddRAD-Seq Data Set Optimisation for Phylogenomic Analyses

To obtain PE-ddRAD datasets optimised for phylogenomic analyses, we performed a two-step approach to assembling RAD loci using two different pipelines, STACKS2 and PYRAD v3.0.66 (Eaton 2014), and a total of seven different parameterisations (Fig. 1 and Table S2). STACKS has been developed to be suitable for population genomics analyses (Catchen et al. 2013) while PYRAD was originally developed for phylogenetic analyses (Eaton 2014). The two pipelines significantly differ in the method employed for detecting homologous loci with one of the main differences being that originally STACKS did not allow for gapped alignments while PYRAD was designed to assemble data for phylogenetic studies containing divergent species using global alignment clustering, which may include indel variation. Nonetheless, STACKS has been capable of gapped assemblies since version 1.38 (2016) and has been widely used for phylogenetic analyses (Wagner et al. 2013; Díaz-Arce et al. 2016; Wang et al. 2017; Brandrud et al. 2019).

We performed a two-step approach for optimising assembly of RAD loci for phylogenomic analyses. First, we used a method for optimising *de novo* assembly of loci using STACKS (Paris et al. 2017; Rochette and Catchen 2017) which consists of varying each of the two key parameters (M: within-individual distance parameter and n: between-individual distance parameter; Catchen et al. 2011) separately using DENOVO_MAP.PL. We also assessed the frequency of putative sequencing errors in the data by examining the number and proportion of singletons in the dataset whilst varying the m parameter (stack-depth parameter).

Correctly setting M requires a balance: set it too low and alleles from the same locus will not collapse; set it too high and paralogous or repetitive loci will incorrectly merge together. Increasing values for M contributed new broadly shared, and therefore likely

real, polymorphic loci (i.e. loci found in 80% of the samples) (Fig. S2a). We selected the value of M where the tendency of new polymorphic loci when increasing M becomes linear (M = 5). At this phase, the addition of new polymorphic loci due to collapsing alleles from the same locus levels out and most of the new polymorphic loci added are likely to be paralogous or repetitive loci. Choosing the value for n involves a trade-off between setting it too low and failing to find homologous loci in different samples, and setting it too high and collapsing true loci that have similar sequences, and is particularly important to optimise in phylogenomic studies. We followed the same procedure that we used for M to optimise n, fixing the value for M to the optimal value. The tendency of the curve became linear at n = 8 (Fig. S2b). Finally, the number of raw reads required to form an allele is governed by m. Low values of m resulted in larger datasets that contained a higher number of singleton alleles (Fig. S3a). The decrease of the number of singletons when increasing m from 3 to 10 was nearly linear, indicating that we were mostly removing true singletons when increasing m. However, the proportion of singletons in the dataset stabilised at m values between 5 and 7 (Fig. S3b).

From these analyses, we selected five STACKS and two PYRAD parameterisations to evaluate the effects of common issues in RAD-Seq for phylogenetic reconstruction: STACKS optimal (m3 M5 n8), STACKS default (m3, M2, n1), STACKS higher n (m3, M5, n15), STACKS higher m (m7, M5, n8), STACKS REFMAP, PYRAD CLUST 89 and PYRAD CLUST 94. STACKS optimal was the result of optimising m, M and n using denovo_map.pl. STACKS default used the default STACKS parameterization optimised for population genomics' analyses. STACKS higher n was chosen to evaluate the effects of adding extremely variable loci which may contain important phylogenetic information at the expense of adding more paralogs. STACKS higher m was used as a more stringent parameterisation to avoid including sequencing errors as real alleles, but at the expense of removing real low coverage alleles. STACKS REFMAP used reference based identification of PE-ddRAD loci. For this analysis, we aligned the catalog loci from the STACKS higher m parameterisation to the Cory's shearwater (*Calonectris borealis*) genome (Feng et al. 2020) using BWA-MEM v0.7.17 (Li 2013) and we integrated alignment positions using stacks-integrate-alignments. For PYRAD analyses, we used the merged reads protocol. Prior to clustering, we merged the read pairs with PEAR (Zhang et al. 2014). PYRAD was run with the following

parameters: Mindepth=7, Datatype=ddRAD, MaxSH=5 with two separate runs at each Wclust=0.89 and 0.94 (parameterisations PYRAD CLUST 89 and PYRAD CLUST 94, respectively). The clustering thresholds were chosen to be comparable with STACKS higher n and optimal (and higher M) parameterizations, respectively.

We extracted PE-ddRAD loci *in silico* from the outgroups using the python script Digital_RADs.py from DaCosta and Sorenson (2014). All analyses were run downsampling high coverage samples (>60x) to prevent an excess of error arising from PCR duplicates. We also trimmed reads to the same length of 145 bp for forward and 149 bp for reverse reads. In STACKS, when building the set of putative loci using USTACKS, we set the parameter m=1 for the outgroups to account for the coverage of 1x from *in silico* extracted PE-ddRAD loci. We used the Bayesian genotype caller (BGC; Maruki and Lynch 2015, 2017) for SNP calling using the STACKS2 program GSTACKS.

Secondly, we assessed the effect of the pipelines and different parameterisations when assembling the PE-ddRAD tags on two different metrics: the number of parsimony informative sites (PIS) per locus relative to total locus length; and the sum of bootstrap branch support (BS). We chose to relativise the number of PIS because variation in locus length among PE-ddRAD loci is relatively low; however, a relative metric would be inappropriate for phylogenomic data sets that display more substantial variation in locus length (McClean et al. 2019). The second metric was used to measure the phylogenetic resolution provided by the datasets. We prepared three alignments that differed in the level of missing data for each chosen parameterisation (n=3x7; Table S2): no more than 5%; 25%; and 35% missing data. We computed locus length and the number of PIS per locus using the PHYLUCE script phyluce_align_get_informative_sites.py and we calculated the proportion of PIS for every locus (i.e. PIS/locus length) in every data set. To test the phylogenetic resolution of the data resulting from each alignment, we inferred a maximum likelihood (ML) phylogeny using an unpartitioned concatenated approach in RAXML v. 8.0.19 (Stamatakis 2014). We used the GTRGAMMA model along with 1000 rapid bootstrap replicates and 30 thorough ML searches for each run. We analysed the resulting trees and calculated the sum of BS for all branches.

The size of the datasets selected for the second optimisation step differed by nearly one order of magnitude (Table S2). The level of missing data had the strongest impact on dataset size: allowing a maximum of 5% of missing data yielded between 2415 to 7269 PE-ddRAD loci, allowing a maximum of 25% yielded between 6276 to 14464 loci, and allowing a maximum of 35% yielded between 8599 to 17604 loci. With similar parameterisations, STACKS yielded approximately twice the number of loci than PYRAD for each level of missing data, but PYRAD loci had a higher number of PIS per locus (Fig. S4a and Table S2). Different parameterisations had a minor effect on the number of PIS per locus with the exception of the default parameters in STACKS that yielded loci with a much lower number of PIS. With STACKS we reconstructed a small number of very variable loci (proportion of PIS per locus > 0.15) when allowing 25% or 35% of missing data. As it would be expected, the most variable loci were obtained with the higher n parameterisation. PYRAD datasets did not contain any loci with a proportion of PIS per locus higher than 0.15. Phylogenetic analyses using the different datasets and levels of missing data resulted in highly resolved and congruent phylogenies. Nonetheless, the general trend was an increase in resolution when reducing missing data from a maximum of 35% to a maximum of 25% and thereafter, a decrease when reducing it to a maximum of 5% (Fig. S4b). This general trend could be explained because increasing missing data allows the inclusion of more variable loci with high phylogenetic informativeness that may improve the phylogenies, but when increasing it too much we may also introduce more paralogs and thus result in more noise. Phylogenetic analyses using datasets with a higher amount of missing data (35% and 25%) tended to yield higher bootstrap supports on recent splits whilst analyses using datasets with a low level of missing data (5%) yielded higher bootstrap supports on more ancient splits (Fig. S5). PYRAD datasets with a maximum of 25% missing data yielded the highest overall resolution but Stacks higher m datasets were the most consistent across different levels of missing data. We selected STACKS higher m parameterisation datasets (PE ddRAD 75% and 95%) for downstream analyses due to their consistency in phylogenetic resolution across different levels of missing data, due to their more stringent filtering of putative sequencing errors and because STACKS allowed an easy inclusion of *in silico* extracted outgroups.

Marker Distribution and Genomic Context

Because the genomic distribution of phylogenomic markers can assist with understanding their informativeness across phylogenetic scales, we compared the distributions and genomic context of PE-ddRAD and UCE loci. We used BLASTN (Altschul et al. 1997) to map both sets of loci to the Balearic Shearwater (BaSh; *Puffinus mauretanicus*) draft genome assembly (Cuevas-Caballé et al. 2019) and to the most closely related chromosome-level genome assembly, the Anna's Hummingbird (AnHu; *Calypte anna*; Korlach et al. 2017), which diverged between 62.7 to 71.1 Ma (Jarvis et al. 2015). Due to the old divergence time between shearwaters and AnHu, we also mapped the PE-ddRAD markers to the AnHu genome using a liftover approach. Briefly, we used SATSUMA2 (https://github.com/bioinfologics/satsuma2) to align the BaSh draft genome assembly to the AnHu genome. We then used KRAKEN (Zamani et al. 2014) to translate PE-ddRAD loci BaSh genomic coordinates to AnHu coordinates. Finally, we plotted the genomic distributions of both sets of markers when mapped to the 20 longest scaffolds of the BaSh assembly and to the AnHu chromosomes using an R script modified slightly from Harvey et al. (2016).

Using BLASTN, 97.7% of the UCE and 95.4% of the PE-ddRAD loci successfully mapped to the *P. mauretanicus* draft genome assembly. When mapped to the more distant *C. anna* chromosome-level genome assembly, an even higher percentage of UCE loci successfully mapped (99.4%) which contrasted with the low percentage of successfully mapped PE-ddRAD loci (30.6%). Using the liftover approach, we managed to improve the percentage of successfully mapped PE-ddRAD loci to the *C. anna* genome assembly to 77.5%. We thus show that to better understand the genomic context of RAD markers, the liftover approach works best when no chromosome-level genomes of closely related species are available, especially in groups with highly syntenic genomes such as birds.

For each locus, we calculated the distance to the next closest locus on the AnHu assembly using the closest-features program from the BEDOPS tool set v. 2.4.36 (Neph et al. 2012). We applied a permutation procedure (Bioconductor package REGIONER; Gel et al. 2016) to determine if the level of clustering was greater than expected if the loci were

randomly distributed across the genome. REGIONER randomises query and reference region-sets over the genome for each chromosome and the MEANDISTANCE function calculates the mean distance from every region in the query to the closest region in the reference.

The distribution of distances to the next closest locus showed that UCE loci had a higher level of clustering than PE-ddRAD loci (median UCE = 16.2 kbp, median PE-ddRAD = 51.3 kbp; Fig. 2, Fig. S6 and S7). However, both UCE and PE-ddRAD loci were more clustered along the AnHu genome than expected by chance (Z = -31.96, P = 0.0002 for UCE; Z = -10.01, P = 0.0002 for PE-ddRAD; Fig. S8). Researchers should be aware of the high levels of clustering among UCE loci when building datasets for coalescent analyses that assume free recombination between loci.

We performed genome-wide association analyses between PE-ddRAD, UCE and protein-coding genes using the MEANDISTANCE and the NUMOVERLAPS functions in REGIONER. For each statistical analysis, 5000 permutations were performed.

PE-ddRAD loci were closer to protein-coding genes (51.4 ± 94.0 kbp) than UCEs (73.8 ± 104.8 kbp). Permutation tests showed that PE-ddRAD loci were significantly closer to protein-coding genes (Z = -8.53, P = 0.0002) whilst the opposite was true for UCEs (Z = 5.17, P = 0.0002) (Fig. S9). However, the number of overlaps with protein-coding genes was significantly higher than expected from the random sets for both PE-ddRAD (Z = 4.55, P = 0.0002) and UCEs (Z = 5.25, P = 0.0002) (Fig. S10). UCE and PE-ddRAD loci were also significantly closer to each other (Z = -13.17, P = 0.0002) and overlapped slightly more than expected by chance (Z = 2.10, P = 0.0296).

RAD-Seq loci are not necessarily dispersed randomly throughout the genome, in part due to a different preponderance of restriction enzyme cut sites depending on a region's base composition (DaCosta and Sorenson 2014). In our case, the association between PE-ddRAD loci and protein-coding genes may be driven by the fact that one of the restriction enzymes we used had 100% GC content and thus, GC-rich areas are likely to harbour higher concentrations of restriction enzyme cut sites.

Phylogenetic Analyses

UCE Dataset Partitioning

Concatenated unpartitioned analyses can converge to a tree other than the species tree (Roch and Steel 2015) and produce highly supported but incorrect nodes in the tree (Kainer and Lanfear 2015). To explore if this applied to our dataset, we estimated best-fit partitioning schemes for our 75% complete UCE IUPAC consensus matrix using the Sliding-Window Site Characteristics (SWSC-EN) method described in Tagliacollo and Lanfear (2018). SWSC-EN uses a sliding-window approach to divide each UCE into data blocks with similar site-entropy to generate partitions that account for heterogeneity in rates and patterns of molecular evolution within each UCE. We then used the output from SWSC-EN to construct an input file for PARTITIONFINDER2 (Lanfear et al. 2017). Finally, we estimated the optimal partition scheme using: the rcluster algorithm; equal weighting for overall rates, base frequencies, model parameters and the alpha parameter; and model selection by Bayesian information criterion (BIC). PARTITIONFINDER2 yielded 131 partitions for the 75% complete matrix. Finally, we estimated bayesian (EXABAYES) and ML (RAXML-NG) phylogenies using the partitioned 75% complete matrix.

Divergence Time Estimation Priors and Substitution Model

To inform the regene_gamma prior in MCMCTREE during the first step of the dos Reis and Yang (2011) procedure, we estimated mean substitution rates for the UCE and the PE-ddRAD alignments using the BASEML programme in the PAML package with a strict molecular clock. We used the function BIRTHDEATH from the APE package (Paradis and Schliep 2019) in R to calculate birth and death rates. The birth–death process with λ = 0.172 (birth rate) μ = 0.083 (death rate) and ρ = 0.75 (fraction of species sampled) was used to construct the prior on node ages. We applied the HKY85 + GAMMA model with five rate categories.

Introgression Analyses

Split Networks

To better visualise patterns of genealogical discordance and potential areas of reticulate evolution, we computed phylogenetic networks for each genus using the Neighbour-Net approach (Bryant and Moulton 2004), implemented in SPLITSTREE version 5 (Huson and Bryant 2006). We computed pairwise absolute genetic distance between all pairs of samples using the dist.dna function from the APE R package. NeighbourNet networks were computed from the distance matrix using default parameters.

Evaluation of Potential Causes of Significant Patterson's D-statistic

For those cases with a significant Patterson's D statistic, we extracted SNPs with strong signatures of introgression (i.e. SNPs with ABBA configuration and fixed within species, ABBA SNPs) to evaluate alternative potential causes of these signatures.

Firstly, we checked if derived alleles (B alleles) were also present in other outgroup species and could thus be a signature of shared ancestral variation. Secondly, we counted the number of SNPs and the number of haplotypes per locus in loci with ABBA SNPs, and we compared them with the averages for all loci to evaluate if ABBA patterns arose preferentially in high mutational areas where convergent mutations were more likely. Thirdly, we calculated the proportion of ABBA patterns that most likely arose by W-to-S mutations to assess the potential role of gBGC in generating these patterns; and finally, we calculated individual heterozygosities to evaluate the levels of genetic variation of potentially introgressed compared to non-introgressed species.

Phylogenetic Network Analyses

We reconstructed phylogenetic networks using the maximum pseudolikelihood method implemented in SNAQ (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). This method accommodates ILS and gene flow under the multispecies network coalescent model (MSNC). We ran phylogenetic networks for each of the datasets (PE-ddRAD, UCE and TENT) and independently for *Puffinus* and *Ardenna* genera, to reduce computation time and likely improve the accuracy of the inferred networks. In both

cases, we used the time-calibrated topology, pruned to only contain taxa in each genus, as a starting topology.

The inclusion of multiple individuals per species allows to calculate population sizes and might improve performance in SNAQ although further simulation studies are needed to test this assumption (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). Here, we decided to run the analyses with one individual per taxon because we were not interested in calculating population sizes.

The reduced number of variable positions per locus in PE-ddRAD and UCE datasets is likely to preclude the reconstruction of well-resolved gene trees. In order to explore the impact of contracting low support branches on phylogenetic networks reconstruction, we used three different BS thresholds (0, 10, 50) to collapse low support branches in the gene trees estimated for ASTRAL-III analyses for both PE-ddRAD 75% and UCE 75% data. We then used the function READTREES2CF in the PHYLONETWORKS julia package to calculate concordance factors (CF) tables from the gene trees. We also used an alternative method that allows calculation of CF tables directly from SNP data using the R function SNPS2CF (Olave and Meyer 2020).

Fossil Calibrations

The shearwater fossil record is limited although it is the richest among Procellariidae (Olson 1985). Nonetheless, most of the fossils are either too young to be of use for calibration, lacking precise and accurate age estimates or too fragmented to be placed in the phylogeny with confidence. The latter has been particularly challenging for selecting minimum bounds for the age of the root as some of the oldest shearwater fossils have not been evaluated in detail and have even been reassigned from non-procellariiform genera (Brodkorb 1962). To evaluate the effect of truncation of calibration densities on time estimates (Barba-Montoya et al. 2017; Dos Reis et al. 2018), we used two different calibration strategies. Strategy A (SA) used the 4 calibrations whereas strategy B (SB) left the *Ardenna-Calonectris* node with no calibration as the branch separating that clade from the genus *Puffinus* was very short and resulted in slightly truncated prior densities. Due to the difficulty of determining a maximum age for shearwaters based on the fossil record, we performed analyses using a much older

maximum age to examine the sensitivity of age estimates to the use of much older priors. Specifying maximum bounds is a difficult task, mainly because absence of fossil evidence is not evidence that a clade did not exist in a point in time (Ho and Phillips 2009). To assess the effect of not setting maximum bounds on time estimates, we also conducted analyses only setting minimum ages for each calibration, excepting the root. Here, we follow the best practices (Parham et al. 2012) to provide justifications for dates assigned to the 4 fossil calibrations used in this study and their phylogenetic placement.

Calibration 1: Root (SA, SB)

Maximum age: 23.03 Ma (38 Ma)

Minimum age: 15.2 Ma

Fossil taxon and specimen: *Puffinus mitchelli*; No. 58184, Mus. Paleo., Univ. Calif., Berkeley; the distal half of a right humerus. *Puffinus priscus*; No. 58185, Mus. Paleo., Univ. Calif., Berkeley; the distal third of a left humerus.

Locality: Locality V-2401, Sharktooth Hill, Temblor Formation, Bakersfield, California, USA.

Phylogenetic justification: The oldest shearwater fossil species was believed to be *Puffinus micraulax* from the Hawthorne Formation in Florida, lower Miocene (Brodkorb 1963). However, the Hawthorne Group consists of several formations (early Miocene-early Pliocene) with the lower Miocene ones being rarely exposed (Scott 1988) which makes *P. micraulax* unlikely to be from deposits of the early Miocene. Other older *Puffinus* species have been described, such as *Puffinus arvernensis* (Milne-Edwards in Shufeldt 1896), which may resemble *Pterodroma* more than *Puffinus* (Olson 1985), and *Puffinus raemdonckii* (Van Beneden 1871), which was originally described as a gull and redefined as a *Puffinus* by (Brodkorb 1962) although it needs to be re-examined before it can be definitely attributed to *Puffinus* (Olson 1985). For calibration, we used *P. mitchelli* and *P. priscus* (Miller 1961) because after revaluation, the age of Sharktooth Hill was older (Pyenson et al. 2009) and the dating was more precise than the age of the Hawthorne Formation layer where *P. micraulax* was found. The flattened humeri of *P. mitchelli* and *P. priscus* place them in the genus *Puffinus* although it is not clear whether they represent the stem or the crown of *Puffinus*.

Age justification: We based the minimum age based on the proximal limit of the Sharktooth Hill Bone Bed (15.2-15.9 Ma) where *P. mitchelli* and *P. priscus* were found. The bonebed was dated based on magnetostratigraphy and biostratigraphy which limited the bonebed to Chron 5Br and the first appearance datum of the diatom *Denticulopsis lauta* (Pyenson et al. 2009).

We based the maximum age on the limit of the early Miocene due to the absence of shearwater fossils before the Miocene apart from *P. raemdonckii,* which, if it is indeed a shearwater, is likely to represent a stem group. We also used a much older maximum age (38 Ma) based on the maximal age of the fossil *Argyrodyptes microtarsus* (Ameghino 1905), which is one of the oldest representatives of the Procellariidae and closely related to the shearwaters based on morphology (Agnolin 2007).

Calibration 2: MRCA of *Calonectris* and *Ardenna* (SA)

Maximum age: 23.03 Ma (38 Ma)

Minimum age: 14 Ma

Fossil taxon and specimen: *Calonectris kurodai*; USNM 237220; right humerus lacking part of the shaft. *Ardenna conradi*; No. 13360 Academy of Natural Sciences in Philadelphia; distal half of a left humerus.

Locality: Calvert Formation Bed 14, Westmore-land County, Virginia, USA (*Calonectris kurodai*). Maryland, USA (*Ardenna conradi*).

Phylogenetic justification: *Calonectris kurodai* is the oldest fossil attributable to the genus *Calonectris* and it is much smaller than any living congeneric species (Olson 2009), suggesting its placement at the stem of *Calonectris*. *Ardenna conradi* (Marsh 1870) is similar in size and form to *Ardenna gravis* from which Wetmore (1926) could distinguish it only by careful comparison. Thus, it is safe to place this fossil within *Ardenna*. Species in the genus *Ardenna* typically present osteological characters of the humerus that are intermediate between the aerially adapted *Calonectris* and *Puffinus* (adapted for underwater propulsion). Thus, it seems likely that the high similarity of *A. conradi* with *A. gravis* may be due to morphological stasis in the *Ardenna* lineage.

Age justification: Calvert Formation Bed 14 belongs to the Middle Miocene (Langhian). We used the proximal age of the Langhian age as a minimum age.

Calibration 3: MRCA of *Calonectris leucomelas* and *Calonectris diomedea* (SA, SB)

Maximum age: 14 Ma

Minimum age: 3.7 Ma

Fossil taxon and specimen: *Calonectris aff. diomedea*; USNM 215433, 366013; Distal ends of right humeri. *Calonectris aff. borealis*; USNM 501506; Distal two-thirds of right humerus lacking most of ectepicondylar spur.

Locality: Lee Creek Mine, Yorktown Formation, Aurora, North Carolina, USA.

Phylogenetic justification: The high similarity of these two fossil taxa to present day *Calonectris diomedea* and *Calonectris borealis* suggested that these two lineages could have been separated for some 5 Ma (Olson et al. 2001). Nonetheless, more recent molecular work disagreed and estimated the coalescence of these lineages plus *Calonectris edwardsii* between 900,000 and 700,000 years ago (Gómez-Díaz et al. 2006). To accommodate both perspectives, we designated these fossils as the oldest members of crown *Calonectris*.

Age justification: Based on ostracod and foraminifera biostratigraphy, (Hazel 1983) revised the age of the Yorktown Formation as early Pliocene and not younger than 3.7 Ma which is the age that we use as a minimum for the calibration. The maximum age is based on a safe 14 Ma which is the oldest age known for stem *Calonectris*. From 14 Ma to the early Pliocene the only fossil records of *Calonectris* shearwaters are of species either much smaller (*C. kurodai*) or much larger (*C. krantzi*; Olson et al. 2001) than crown *Calonectris* species and thus we consider 14 Ma as a safe maximum bound.

Calibration 4: MRCA of *Ardenna bulleri* and *Ardenna pacifica* (SA, SB)

Maximum age: 14 Ma

Minimum age: 3.11 Ma

Fossil taxon and specimen: *Ardenna sp. (aff. pacifica)*; LB12868 Land Vertebrates collection of the Auckland War Memorial Museum; skull with minor compression and damage to external processes.

Locality: road cut on Ridge Road North, Mataroa, North Island, New Zealand.

Phylogenetic justification: (Henderson and Gill 2010) diagnosed this fossil as *Calonectris/Ardenna/Puffinus* based on apomorphic characters and found extensive morphometric similarity to *A. pacifica* (the fossil overlaps with this species in 11 out of the 12 fossil dimension or ratio measurements reported). We consider this description as adequate for assigning the fossil to the crown of *A. pacifica-A. bulleri*.

Age justification: The sediments where the fossil was found accumulated during the Gauss Chron and, in particular, within the normal subchron (3.220-3.110 Ma) lying between the Mammoth and Kaena Subchrons (Turner et al. 2005).

References

Agnolin F.L. 2007. Argyrodyptes microtarsus Ameghino, 1905: un petrel (Procellariiformes) del Eoceno-Oligoceno de Argentina. Studia Geológica Salmanticensia 43:207–213.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Ameghino F. 1905. Enumeración de los impennes fósiles de Patagonia y de la Isla Seymour. Anales del Museo Nacional de Buenos Aires 6:97-167.

Barba-Montoya J., Dos Reis M., Yang Z. 2017. Comparison of different strategies for using fossil calibrations to generate the time prior in Bayesian molecular clock dating. Mol. Phylogenet. Evol. 114:386–400.

Brandrud M.K., Paun O., Lorenz R., Baar J., Hedrén M. 2019. Restriction-site associated DNA sequencing supports a sister group relationship of Nigritella and Gymnadenia (Orchidaceae). Mol. Phylogenet. Evol. 136:21–28.

Brodkorb P. 1962. The systematic position of two Oligocene birds from Belgium. Auk 79:706–707.

Brodkorb P. 1963. MIOCENE BIRDS FROM THE HAWTHORNE FORMATION. Quarterly Journal of the Florida Academy of Sciences 26:159–167.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., Roychoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Bryant D., Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H., De Koning D.-J. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. Genes|Genomes|Genetics 1:171–182.

Catchen J.M, Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. 22:3124–3140.

Cuevas-Caballé C., Ferrer-Obiol, J., Genovart, M., Rozas, J., González-Solís, J., Riutort, M. 2019. Conservation genomics applied to the Balearic shearwater. G10K-VGP/EBP 2019. doi: 10.13140/RG.2.2.15751.21923.

DaCosta J.M., Sorenson M.D. 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PLoS One 9:e106713.

Díaz-Arce N., Arrizabalaga H., Murua H., Irigoien X., Rodríguez-Ezpeleta N. 2016. RAD-seq derived genome-wide nuclear markers resolve the phylogeny of tunas. Mol. Phylogenet. Evol. 102:202–207.

dos Reis M., Gunnell G.F., Barba-Montoya J., Wilkins A., Yang Z., Yoder A.D. 2018. Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. Syst. Biol. 67:594–615.

dos Reis M., Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. 28:2161–2172.

Eaton D.A.R. 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 30:1844–1849.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Syst. Biol.. 27:401–410.

Feng S. et al. In press. Dense sampling of bird diversity increases power of comparative genomics. Nature

Gel B., Díez-Villanueva A., Serra E., Buschbeck M., Peinado M.A., Malinverni R. 2016. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics 32:289–291.

Gómez-Díaz E., González-Solís J., Peinado M.A., Page R.D.M. 2006. Phylogeography of the Calonectris shearwaters using molecular and morphometric data. Mol. Phylogenet. Evol. 41:322–332.

Harvey M.G., Smith B.T., Glenn T.C., Faircloth B.C., Brumfield R.T. 2016. Points of View Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Syst. Biol. 65:910–924.

Hazel J.E. 1983. Age and correlation of the Yorktown (Pliocene) and Croatan (Pliocene and Pleistocene) formations at the Lee Creek Mine. Smithson. Contrib. Paleobiol. 53:81–199.

Henderson N., Gill B.J. 2010. A mid-Pliocene shearwater skull (Aves: Procellariidae: Puffinus) from the Taihape Mudstone, central North Island, New Zealand. N.Z. J. Geol. Geophys. 53:327–332.

Ho S.Y.W., Phillips M.J. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. 58:367–380.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Alfaro-Núñez A., Narula N., Liu L., Burt D., Ellegren H., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G., Avian Phylogenomics Consortium. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4:4.

Korlach J., Gedman G., Kingan S.B., Chin C.-S., Howard J.T., Audet J.-N., Cantin L., Jarvis E.D. 2017. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience 6:1–16.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2017. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Mol. Biol. Evol. 34:772–773.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN].

Marsh C. 1870. ART. XXV.--Notice of some Fossil Birds, from the Cretaceous and Tertiary Formations of the United States. American Journal of Science and Arts (1820-1879) 49:205.

Maruki T., Lynch M. 2015. Genotype-Frequency Estimation from High-Throughput Sequencing Data. Genetics 201:473–486.

Maruki T., Lynch M. 2017. Genotype Calling from Population-Genomic Sequencing Data. G3: Genes|Genomes|Genetics 7:1393–1404.

McClean B., Bell K., Allen J., Helgen K., Cook J. 2019. Impacts of Inference Method And Dataset Filtering On Phylogenomic Resolution In A Rapid Radiation of Ground Squirrels (Xerinae: Marmotini). Syst. Biol. 68:298–316.

Miller L. 1961. Birds from the Miocene of Sharktooth Hill, California. Condor 63:399–402.

Neph S., Kuehn M.S., Reynolds A.P., Haugen E., Thurman R.E., Johnson A.K., Rynes E., Maurano M.T., Vierstra J., Thomas S., Sandstrom R., Humbert R., Stamatoyannopoulos J.A. 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28:1919–1920.

Olave M., Meyer A. 2020. Implementing Large Genomic Single Nucleotide Polymorphism Data Sets in Phylogenetic Network Reconstructions: A Case Study of Particularly Rapid Radiations of Cichlid Fish. Syst. Biol. syaa005.

Olson S.L. 1985. The fossil record of birds. In: Farner, D. S. et al. (eds), Avian Biology. Academic Press, pp. 79–238.

Olson S.L. 2009. A new diminutive species of shearwater of the genus Calonectris (Aves: Procellariidae) from the Middle Miocene Calvert Formation of Chesapeake Bay. Proceedings of the Biological Society of Washington 122:466–470.

Olson S.L., Rasmussen P.C., Others. 2001. Miocene and Pliocene birds from the Lee Creek Mine, North Carolina. Smithson. Contrib. Paleobiol. 90:233–365.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528.

Parham J.F., Donoghue P.C.J., Bell C.J., Calway T.D., Head J.J., Holroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patané J.S.L., Smith N.D., Tarver J.E., Van Tuinen M., Yang Z., Angielczyk K.D., Greenwood J.M., Hipsley C.A., Jacobs L., Makovicky P.J., Müller J., Smith K.T., Theodor J.M., Warnock R.C.M., Benton M.J. 2012. Best practices for justifying fossil calibrations. Syst. Biol. 61:346–359.

Paris J.R., Stevens J.R., Catchen J.M. 2017. Lost in parameter space: a road map for stacks. Methods Ecol. Evol. 8:1360–1373.

Pyenson N.D., Irmis R.B., Lipps J.H., Barnes L.G., Mitchell E.D., McLeod S.A. 2009. Origin of a widespread marine bonebed deposited during the middle Miocene Climatic Optimum. Geology 37:519–522.

Rochette N.C., Catchen J.M. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. Nat. Protoc. 12:2640–2659.

Rochette N.C., Rivera-Colón A.G., Catchen J.M. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28:4737–4754.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100:56–62.

Scott T.M. 1988. The lithostratigraphy of the Hawthorn Group (Miocene) of Florida. Florida Geological Survey.

Shufeldt R.W. 1896. Fossil Bones of Birds and Mammals from Grotto Pietro Tamponi and Grive-St. Alban. Proceedings of the Academy of Natural Sciences of Philadelphia 48:507–516.

Smith B.T., Harvey M.G., Faircloth B.C., Glenn T.C., Brumfield R.T. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63:83–95.

Solís-Lemus C., Ané C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genet. 12:e1005896.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: A package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Tagliacollo V.A., Lanfear R. 2018. Estimating Improved Partitioning Schemes for Ultraconserved Elements. Mol. Biol. Evol. 35:1798–1811.

Turner G.M., Kamp P.J.J., McIntyre A.P., Hayton S., McGuire D.M., Wilson G.S. 2005. A coherent middle Pliocene magnetostratigraphy, Wanganui Basin, New Zealand. J. R. Soc. N. Z. 35:197–227.

Van Beneden P.J. 1871. Les oiseaux de l'argile rupélienne et du crag d'Anvers. Bull. Acad. R. Belg. 2:11.

Wagner C.E., Keller I., Wittwer S., Selz O.M., Mwaiko S., Greuter L., Sivasundar A., Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–798.

Wang X., Ye X., Zhao L., Li D., Guo Z., Zhuang H. 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (Poaceae: Bambusoideae). Sci. Rep. 7:1–11.

Wetmore A. 1926. Observations on Fossil Birds Described from the Miocene of Maryland. Auk 43:462–468.

Zamani N., Sundström G., Meadows J.R.S., Höppner M.P., Dainat J., Lantz H., Haas B.J., Grabherr M.G. 2014. A universal genomic coordinate translator for comparative genomics. BMC Bioinformatics 15:227.

Zhang J., Kobert K., Flouri T., Stamatakis A. 2014. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620.

Supplementary Figures

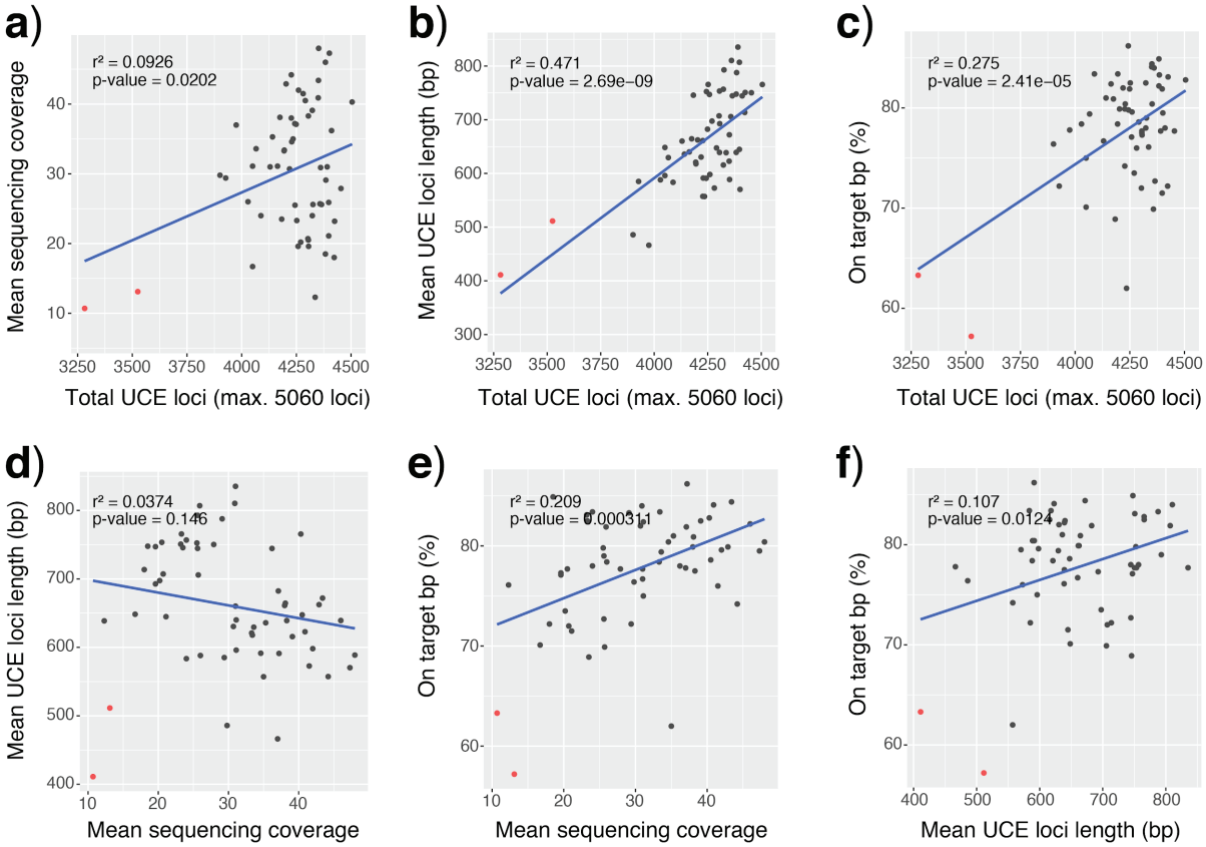


Figure S1 Correlations between UCE assembly summary statistics. a) Mean sequencing coverage vs. Total number of assembled loci, b) Mean UCE loci length vs. Total number of assembled loci, c) On target sequence percentage vs. Total number of assembled loci, d) Mean UCE loci length vs. Mean sequencing coverage, e) On target sequence percentage vs. Mean sequencing coverage, and f) On target sequence percentage vs. Mean UCE loci length. The two samples with the lowest number of reads are shown in red.

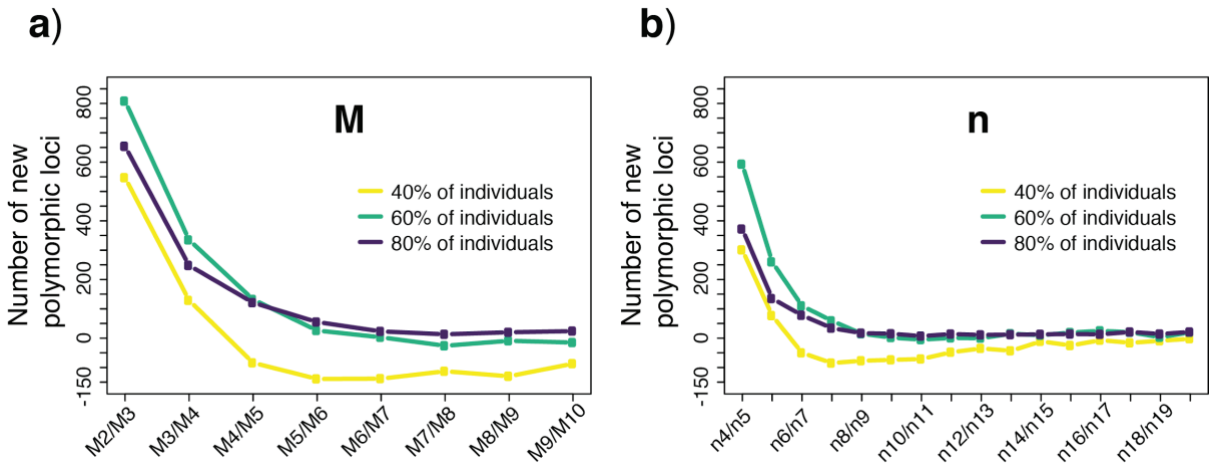


Figure S2 Plots of the number of new polymorphic loci added for each iteration of STACKS parameters a) M and b) n for loci present in at least 40% (yellow), 60% (green) and 80% (purple) of the samples.

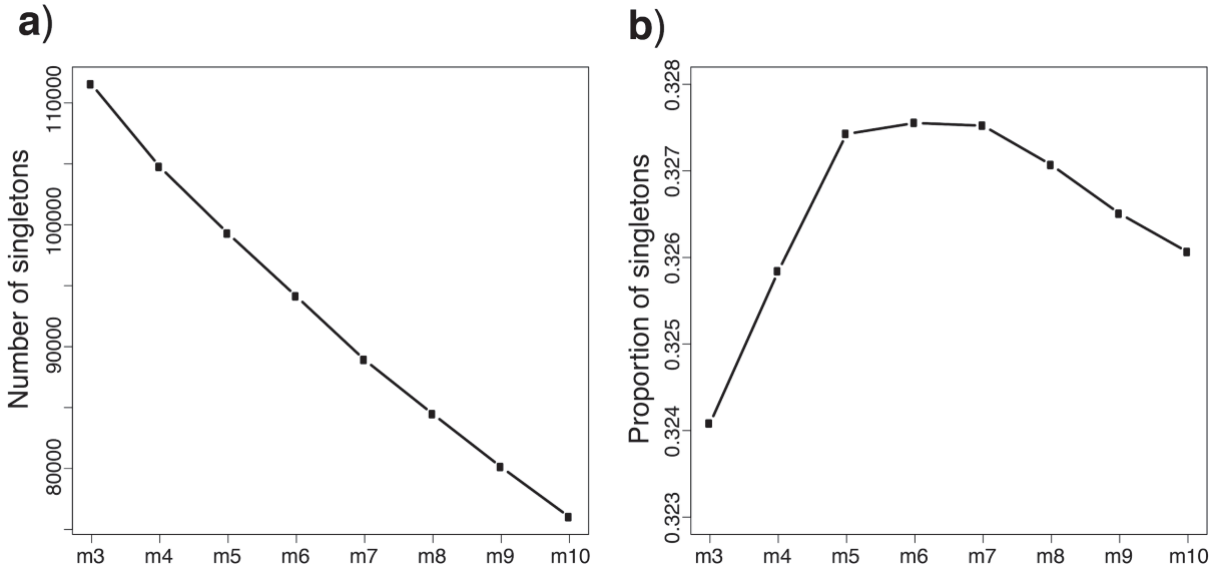


Figure S3 Plots of a) the number of singletons and b) the proportion of singletons in the STACKS catalog while iterating values for the minimum number of raw reads required to form a stack (m).

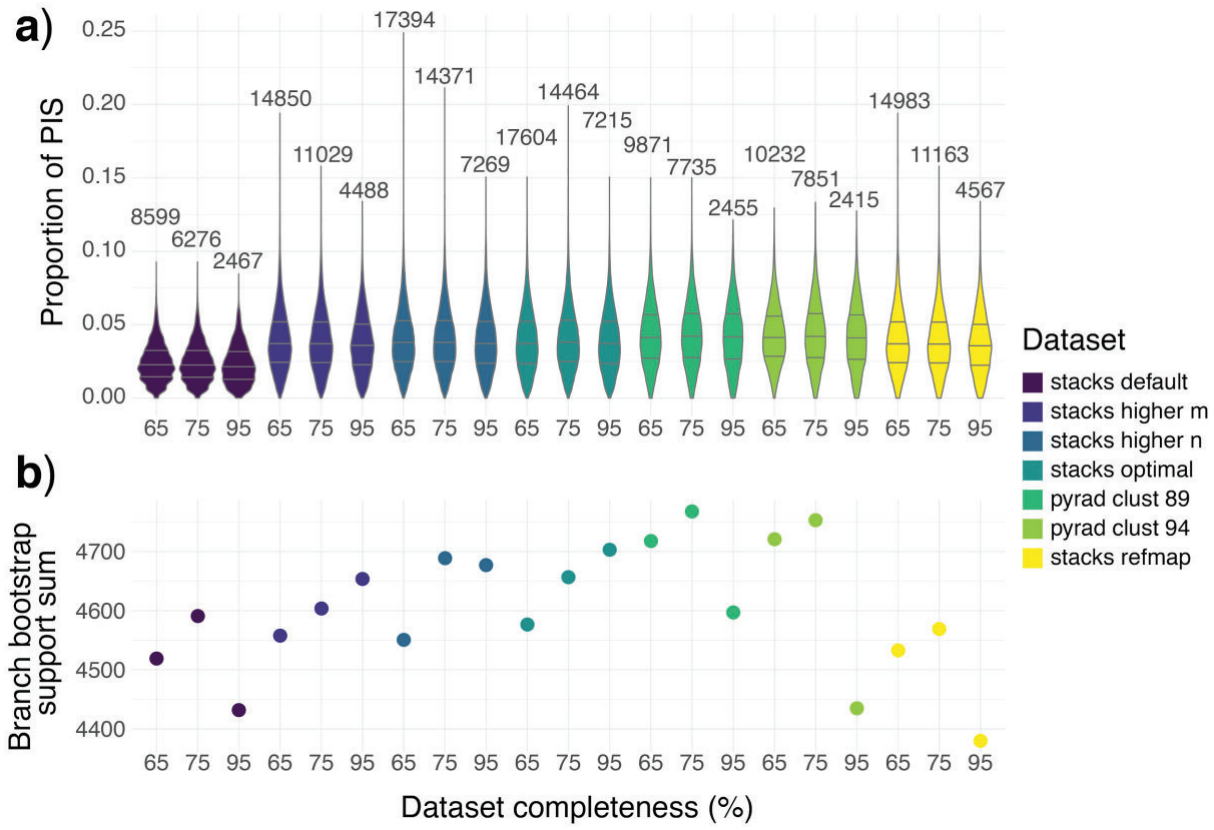


Figure S4 PE-ddRAD dataset optimisation for phylogenomic analyses. a) Violin plots of the number of parsimony informative sites (PIS) relative to locus length for each dataset and b) sum of branch bootstrap support for different parameterisations and levels of missing data. The numbers on top of the violin plots are the number of PE-ddRAD loci of the dataset.

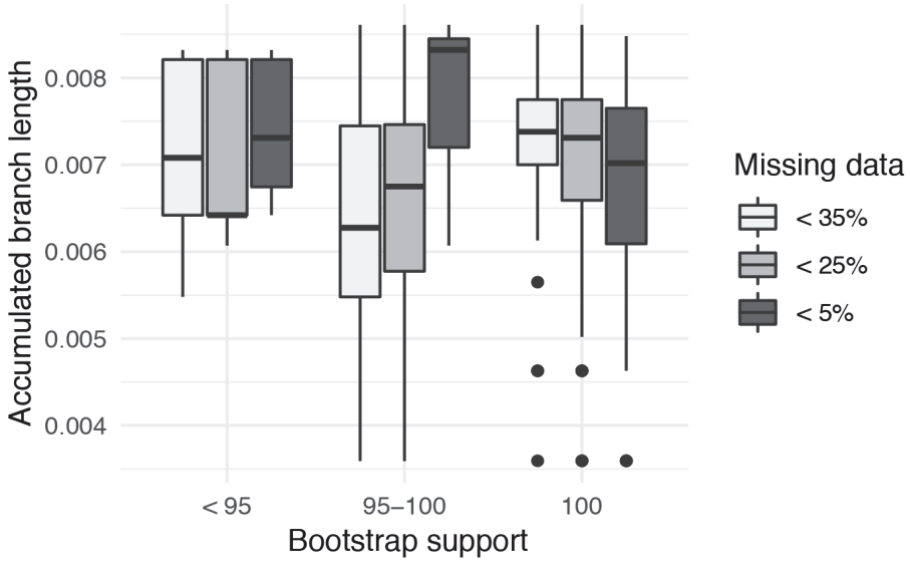


Figure S5 Distribution of root to node branch lengths for low supported (BS < 95), not fully supported (BS = 95-100) and fully supported (BS = 100) branches in ML analyses with PE-ddRAD datasets with varying levels of missing data (< 5%, < 25% and <35%).

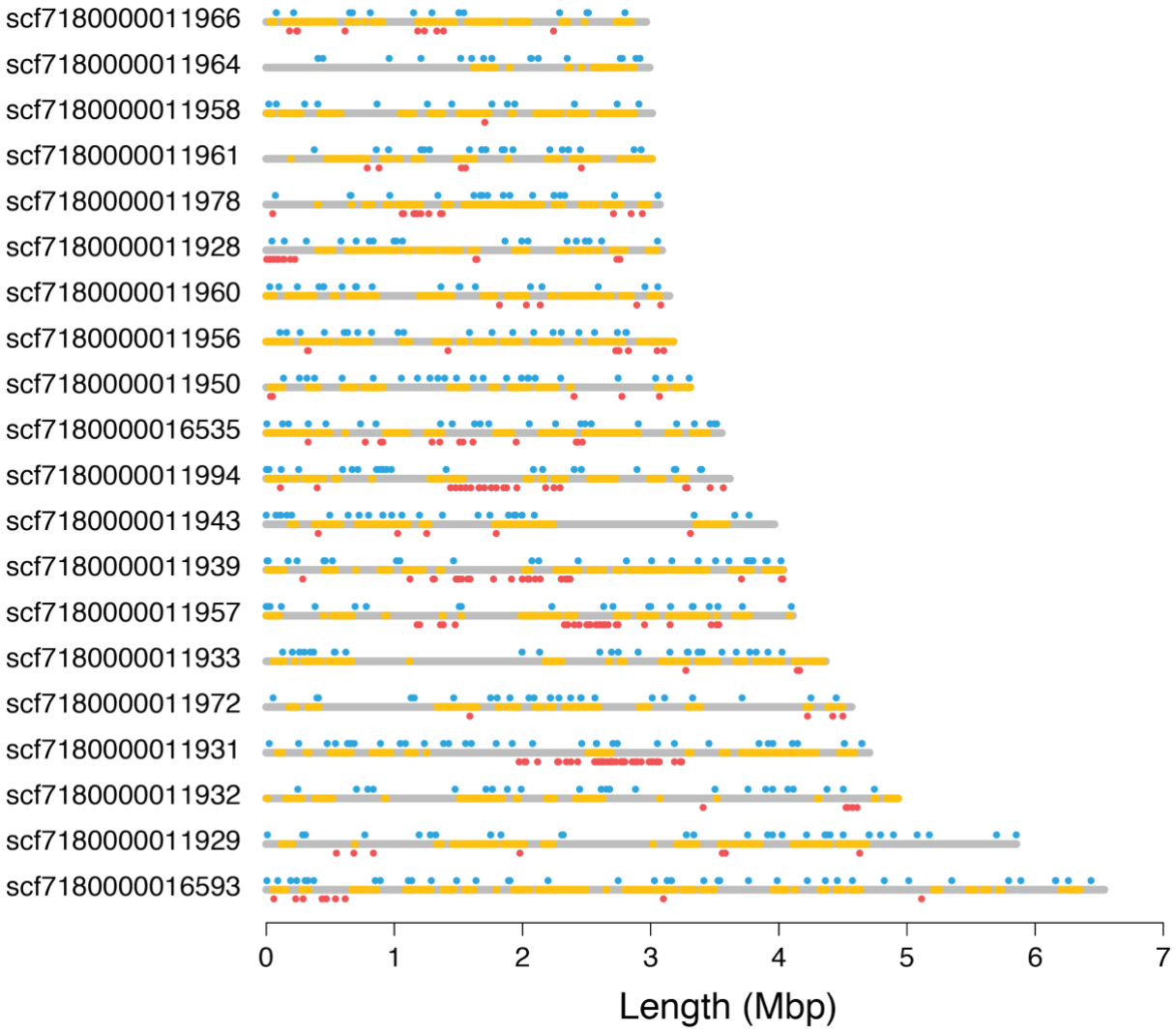


Figure S6 Genomic distribution of PE-ddRAD (blue) and ultraconserved elements (red) when mapped to the 20 longest scaffolds of the draft genome assembly for *Puffinus mauretanicus*. Scaffolds are represented as grey lines and the yellow spots on the scaffolds represent the location of protein-coding genes as in Harvey et al. (2016).

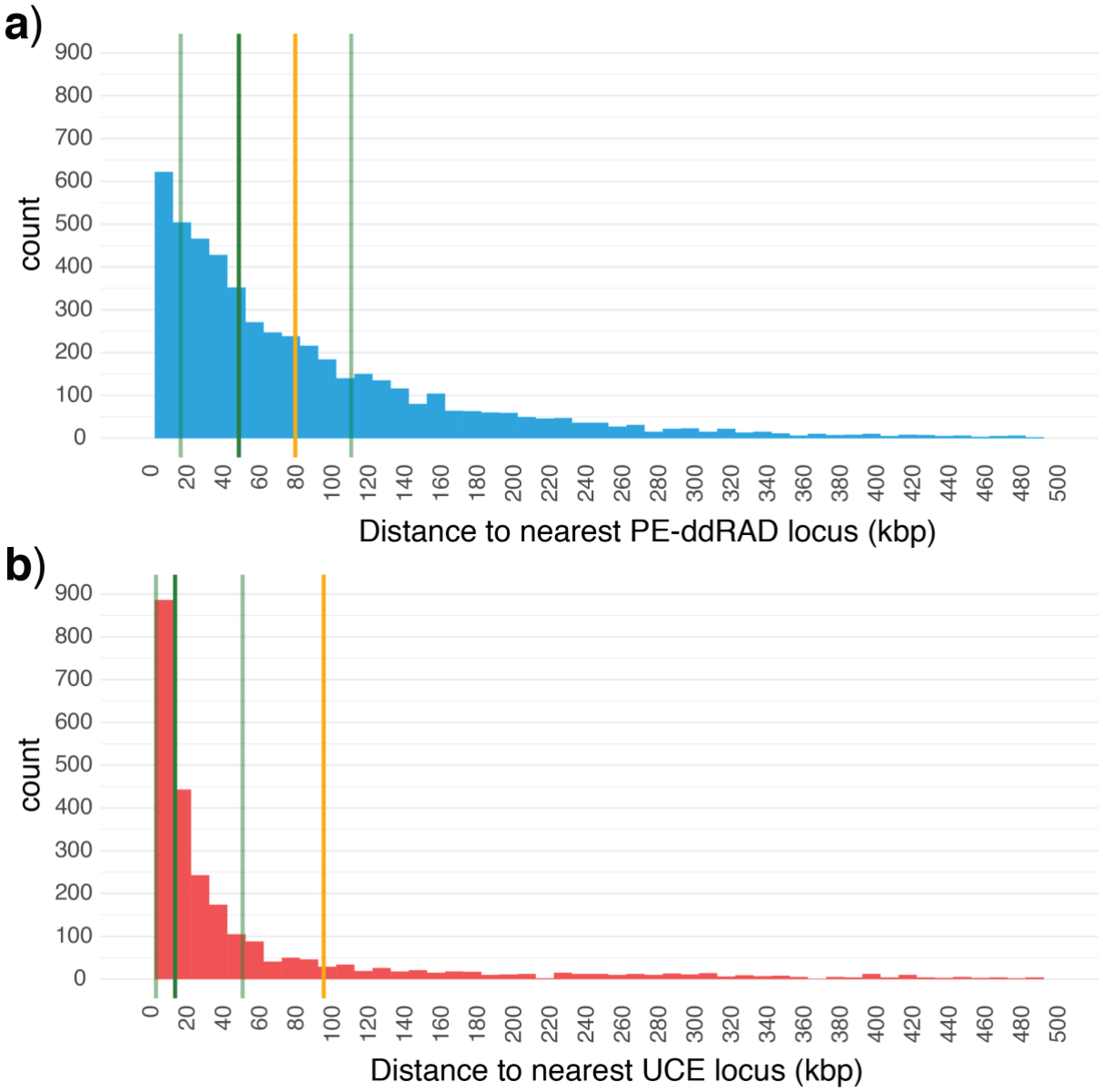


Figure S7 Histograms of the distance to the next closest locus for a) PE-ddRAD loci and b) UCEs based on the positions where they mapped to the *Calypte anna* genome. Vertical lines denote the mean (yellow), the median (dark green) and the upper and lower quantiles (dull green).

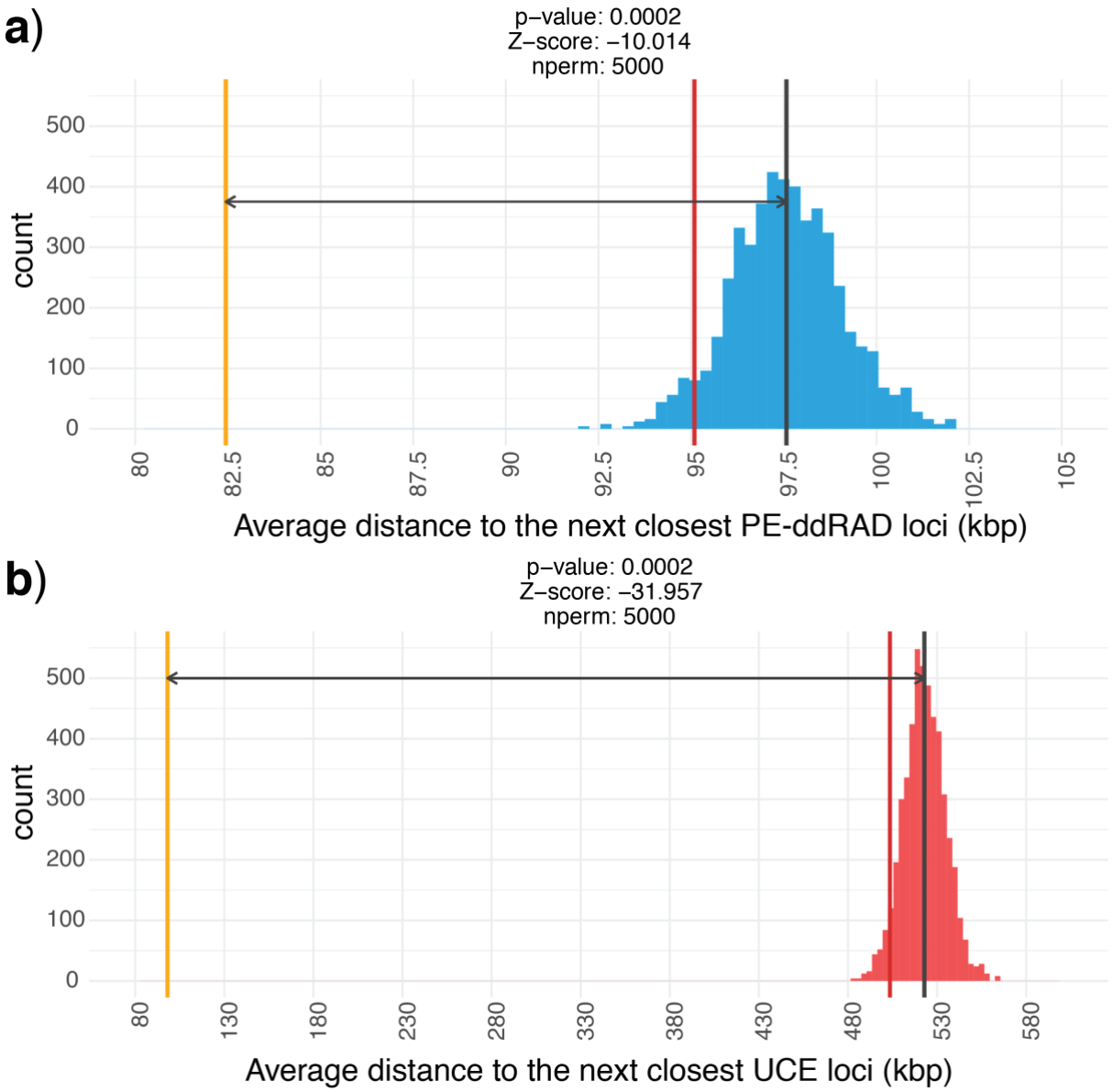


Figure S8 Histograms of the average distance to the next closest locus from 5000 permutations of a) PE-ddRAD and b) UCE loci randomly distributed along the *Calypte anna* genome. Vertical lines denote the mean (black), the lower quantile (red) and the observed mean (yellow).

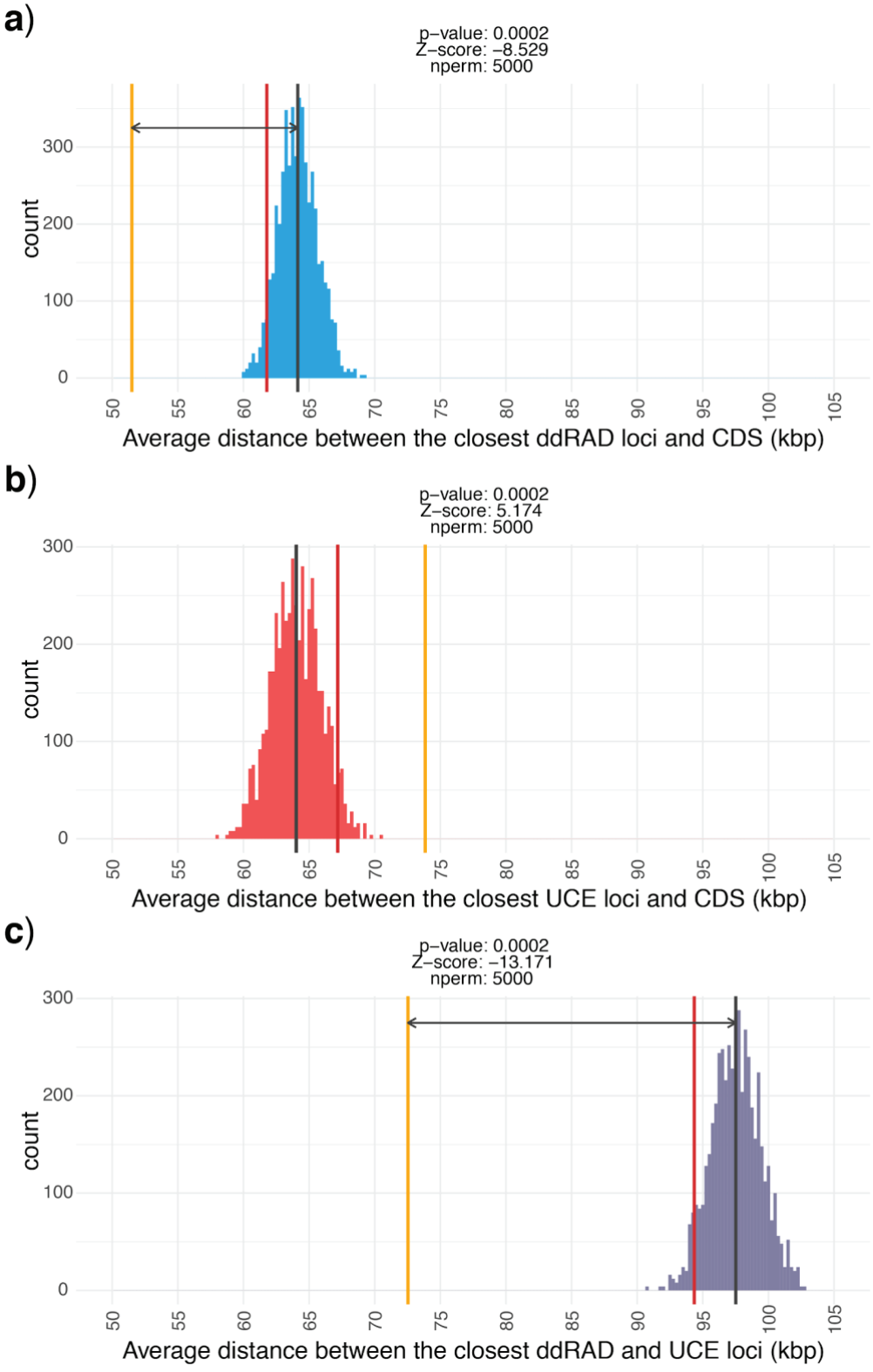


Figure S9 Histograms of the average distance between the closest a) PE-ddRAD locus and coding sequence, b) UCE and coding sequence and c) PE-ddRAD locus and UCE from randomly distributing identical sets of loci 5000 times along the *Calypte anna* genome. Vertical lines denote the mean (black), the lower or upper quantile (red) and the observed mean (yellow).

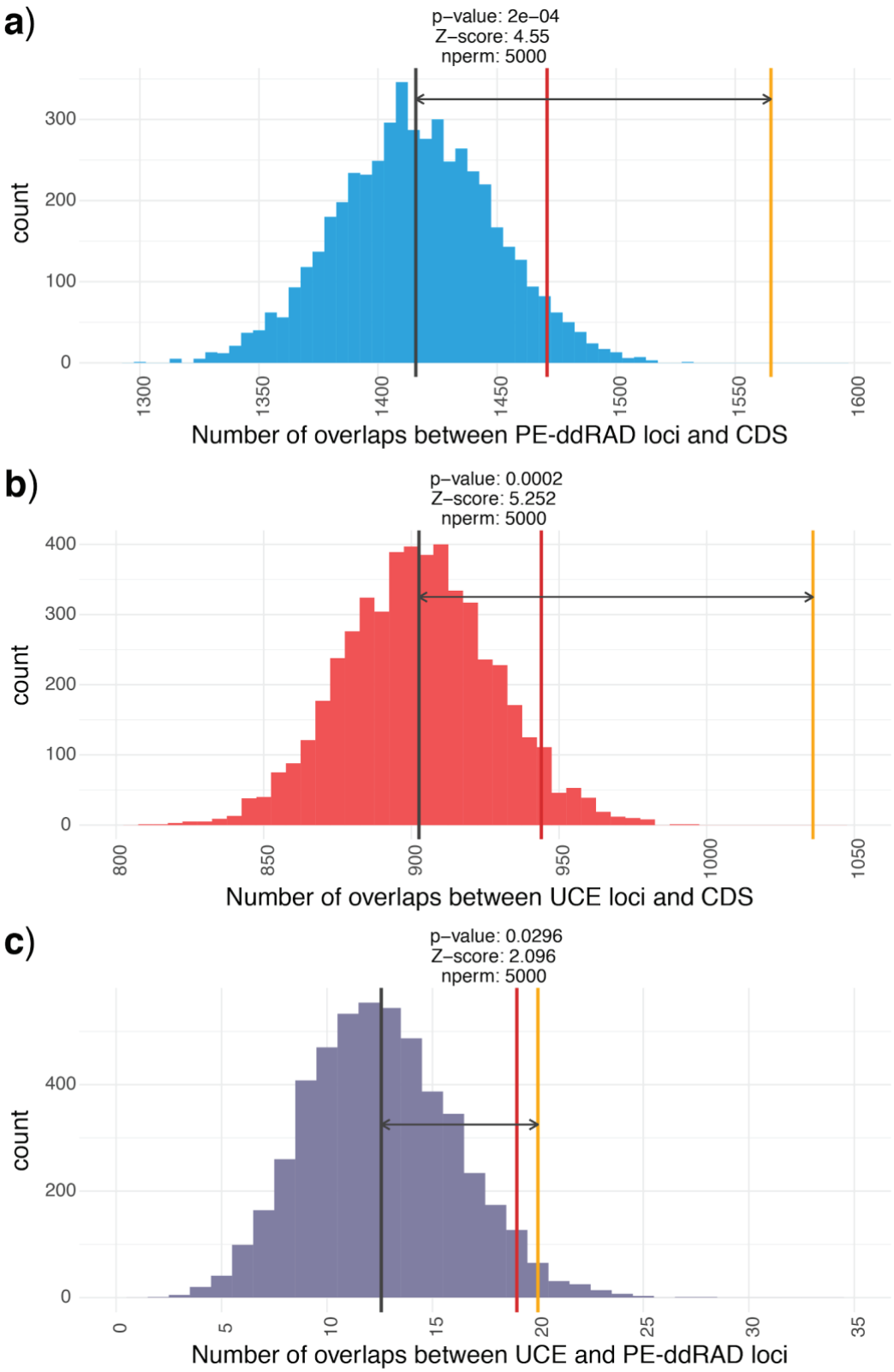


Figure S10 Histograms of the average number of overlaps between a) PE-ddRAD locus and coding sequences, b) UCE and coding sequences and c) PE-ddRAD locus and UCE from randomly distributing identical sets of loci 5000 times along the *Calypte anna* genome. Vertical lines denote the mean (black), the lower or upper quantile (red) and the observed mean (yellow).

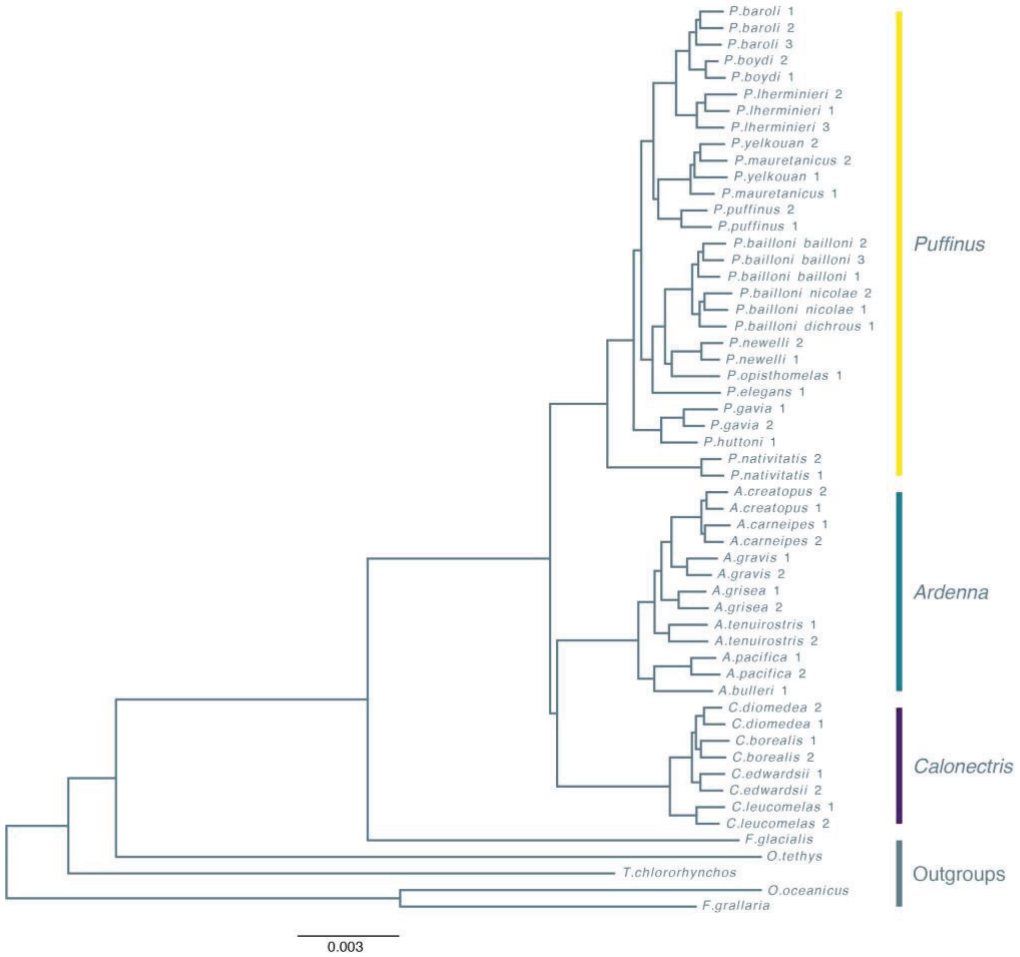


Figure SI1 Maximum-likelihood RAXML-NG tree derived from the 75% complete concatenated UCE matrix with all the outgroups. For the main phylogenomic analyses, we decided to only use *Fulmarus glacialis* as an outgroup to avoid long-branch attraction (Felsenstein 1978) and systematic error due to highly divergent outgroup taxa.

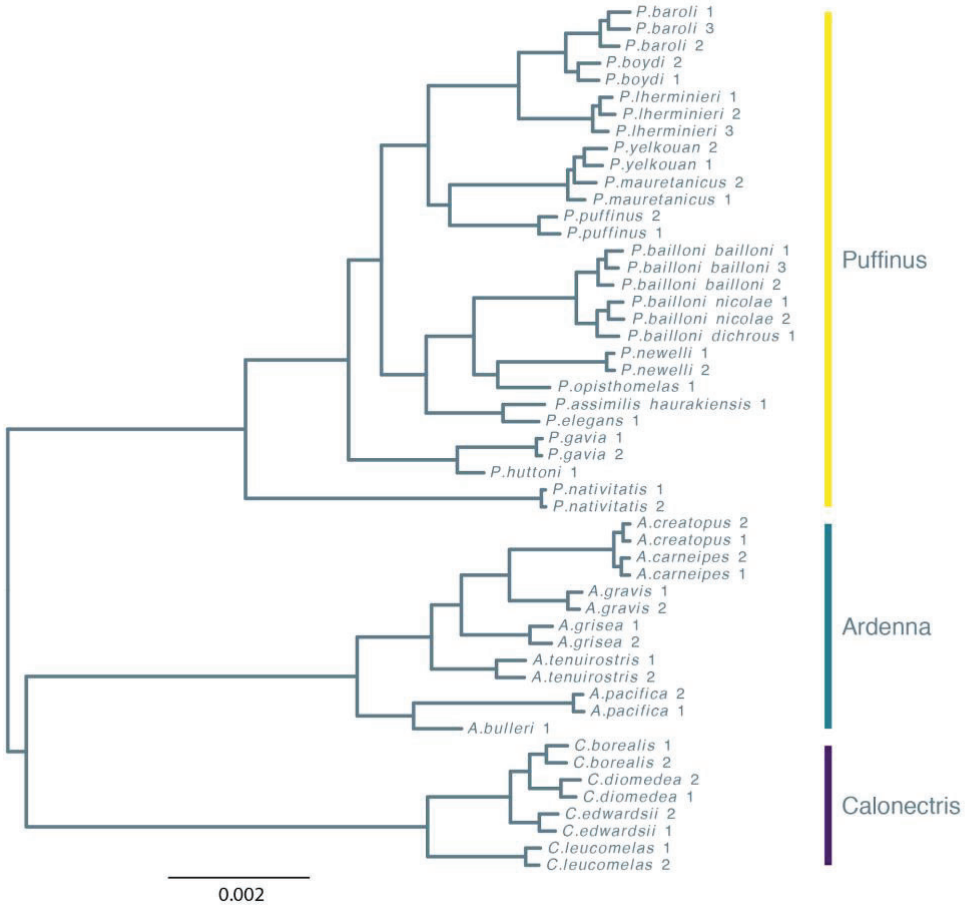


Figure S12 Maximum-likelihood RAXML-NG tree derived from the 75% complete concatenated PE-ddRAD matrix including the ingroup taxon *P. assimilis haurakiensis*.

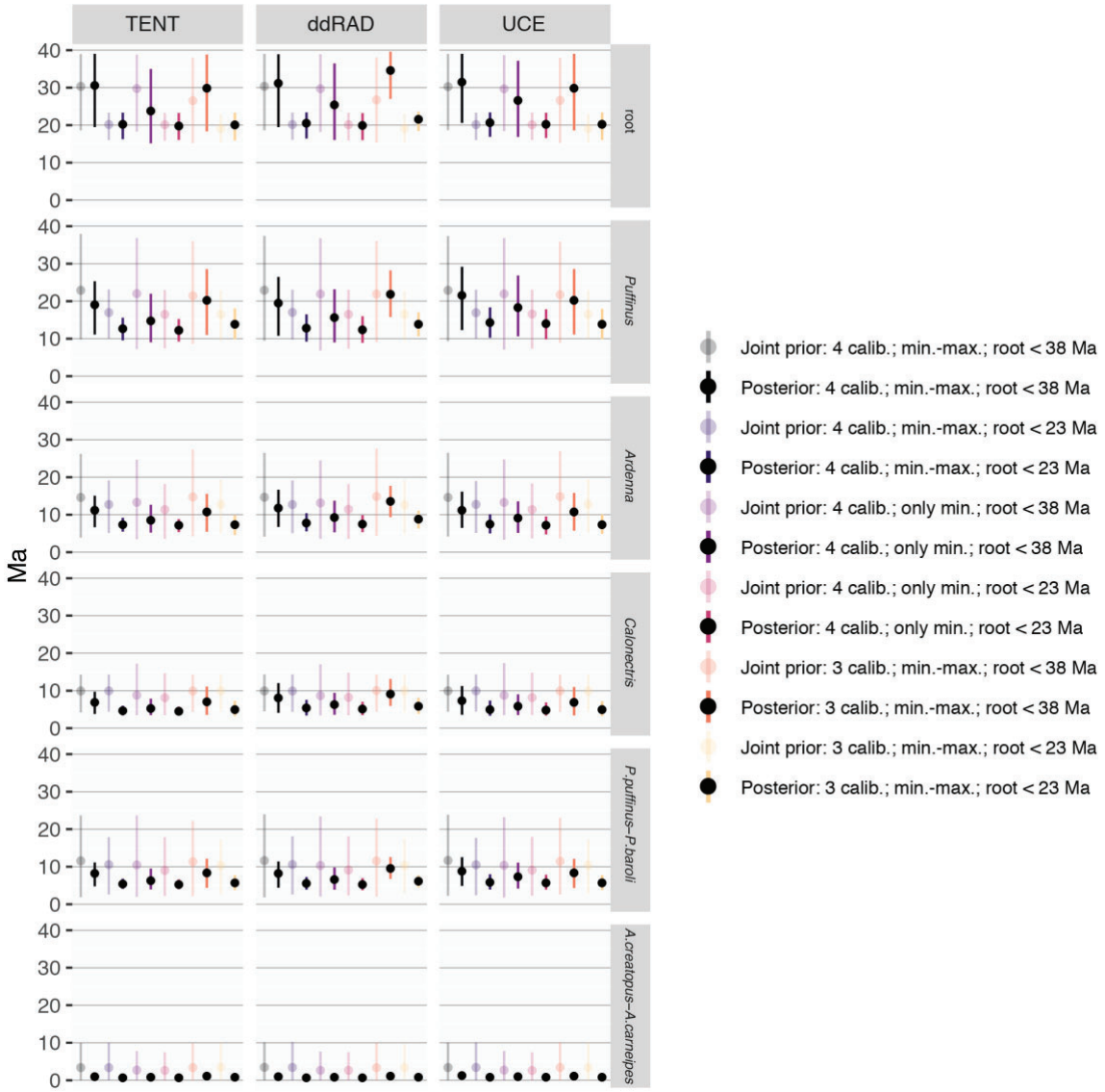


Figure SI3 Comparison of divergence time estimates of key nodes in the shearwater phylogeny across different data types (PE-ddRAD, UCE and TENT), two calibration schemes (3 and 4 calibrations), calibrations with only minimum bounds versus with minimum and maximum bounds and setting the root maximum bound at 38 Ma versus 23 Ma. Bars represent 95% prior or posterior credibility intervals (CI).

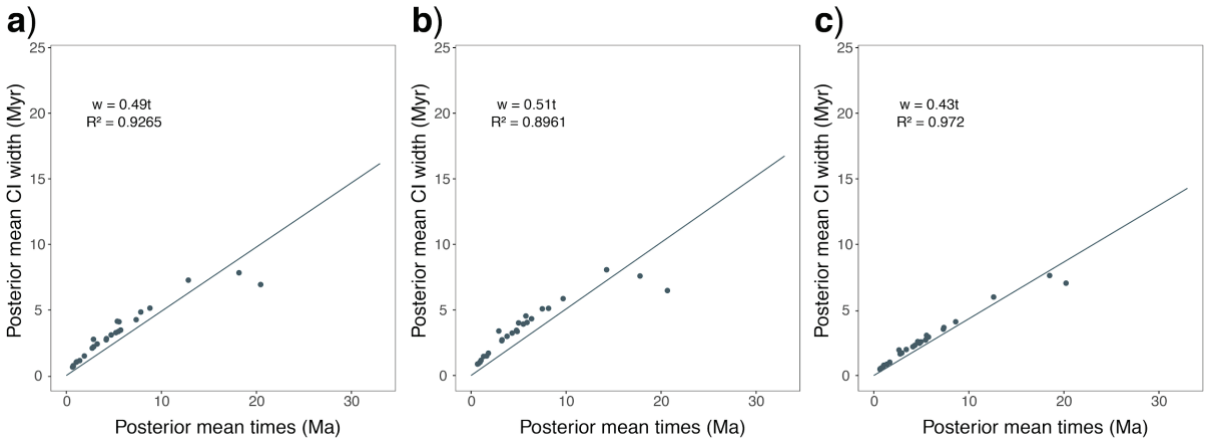


Figure S14 Infinite-sites plots for a) PE-ddRAD, b) UCE and c) 2-partition combined PE-ddRAD and UCE data. The x-axis shows posterior mean divergence times (Ma) and the y-axis the 95% posterior credibility intervals (Myr). The solid line shows the regression through the origin fitted to all the dots. The equations of the regression line and the coefficients of determination (R^2) are shown for each plot.

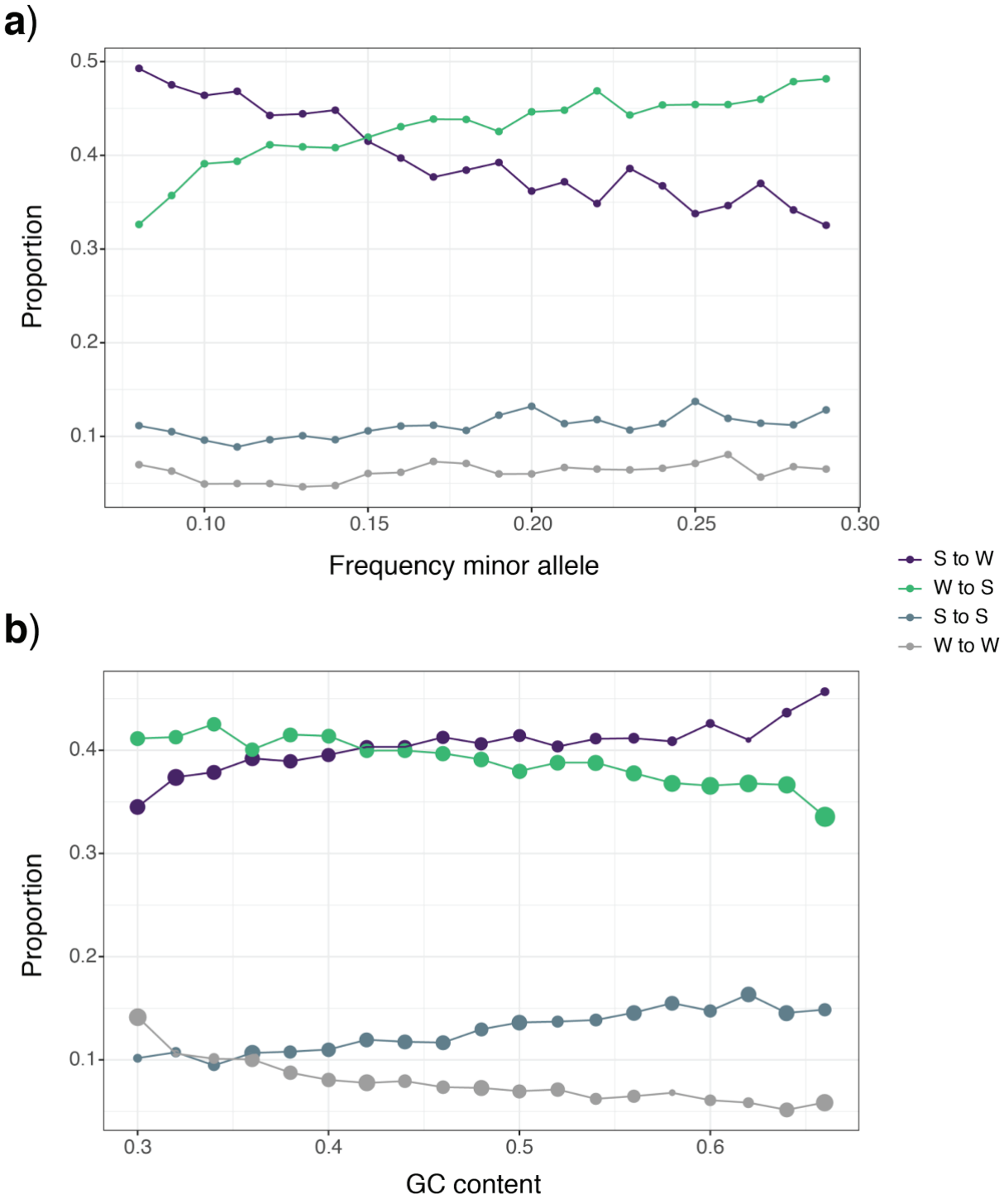


Figure S15 Relative site-frequency spectra for different mutation categories at PE-ddRAD loci. a) Proportion of each mutation category for a given minor allele frequency and b) for a given GC content in the locus. Sizes of dots in b) are proportional to the frequency of the minor allele.

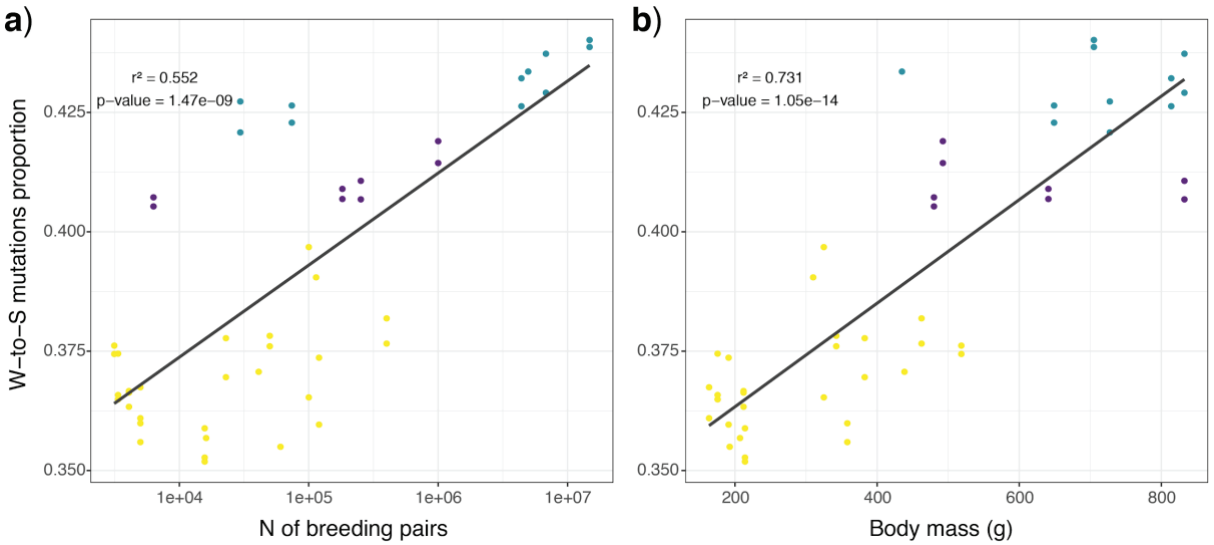


Figure S16 Correlations between the overall proportion of putative W-to-S mutations and a) the number of breeding pairs and b) the average body mass per taxon. Dots are colored depending on the genera: yellow (*Puffinus*), purple (*Calonectris*) and teal (*Ardenna*).

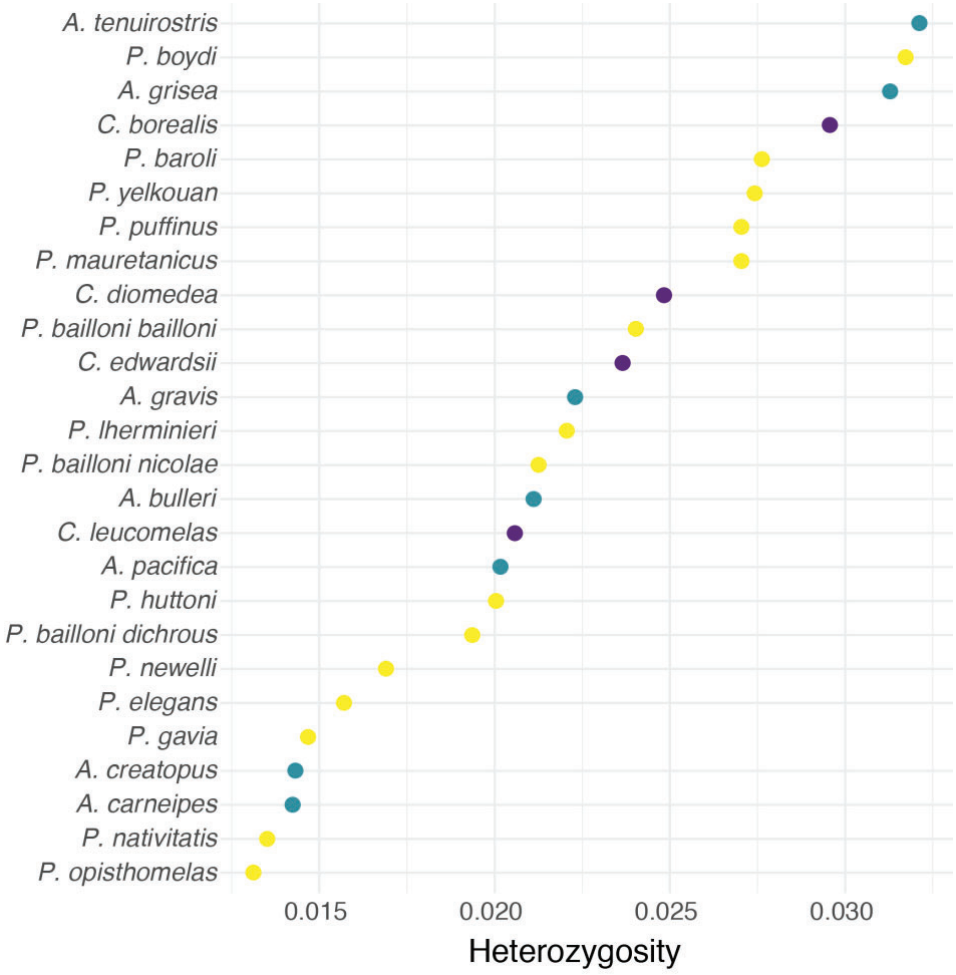


Figure S17 Average individual heterozygosity per taxon. Dots are coloured depending on the genera: yellow (*Puffinus*), purple (*Calonectris*) and teal (*Ardenna*).

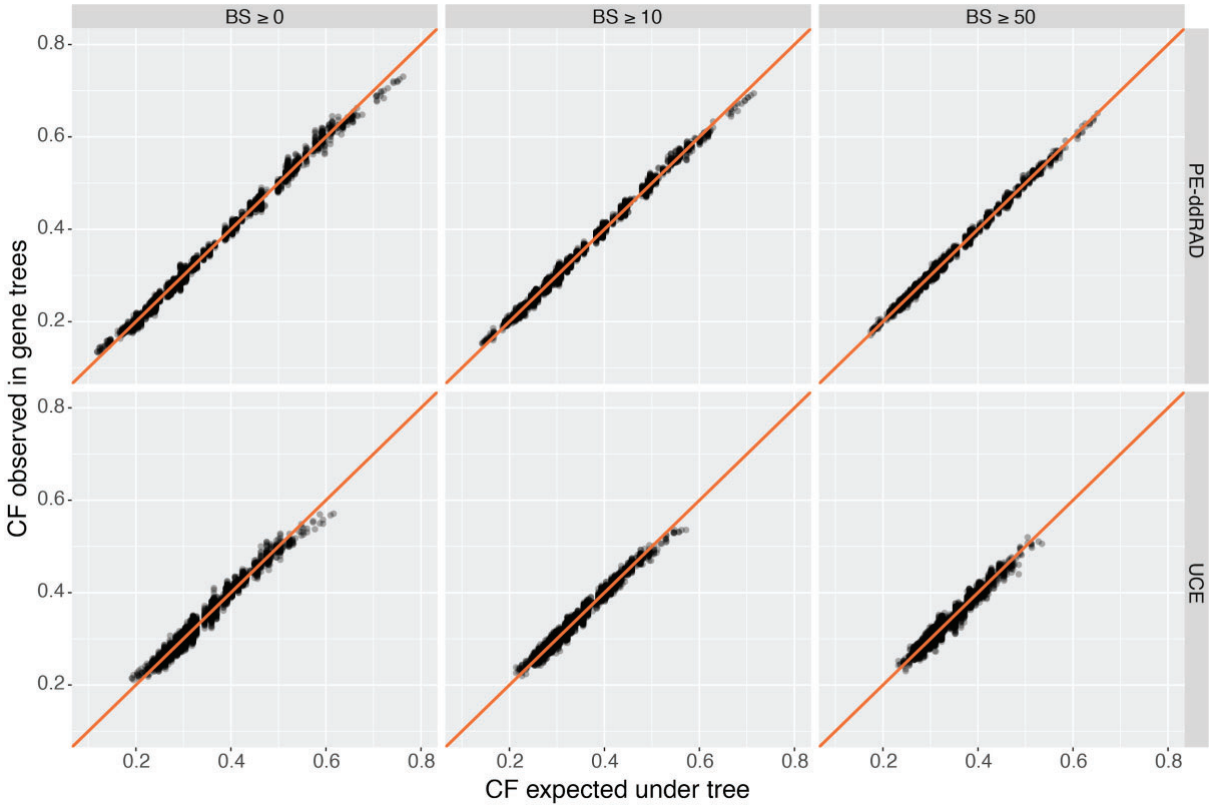


Figure S18 Observed versus expected concordance factors for *Puffinus* data calculated from the species tree under the MSC using PE-ddRAD (panels above) and UCE (panels below) gene trees under different BS thresholds. Observed values at the extremes tend to the expected values when increasing the BS threshold for the gene trees.

Supplementary Tables

Table S1 Samples used in this study, their sample ID, localities and summary statistics of UCE and PE-ddRAD sequencing per sample.

Taxon	Sample ID	Locality	PE-ddRAD reads	UCE reads	PE-ddRAD contigs	PE-ddRAD coverage (x̄)	UCE contigs	UCE coverage (x̄)
Calonectris leucomelas	CLeu1	Mikura Islands, Japan	2,540,681	1,990,286	27,773	66.1	4,361	30.9
Calonectris leucomelas	CLeu2	Mikura Islands, Japan	3,773,702	-	28,109	91.9	-	
Calonectris leucomelas	CLeu2	Mikura Islands, Japan	-	2,181,051	-		4,249	37.1
Calonectris edwardsii	CEdw1	Curral Velho, Cape Verde	450,711	2,190,603	22,492	16.6	4,227	38.0
Calonectris edwardsii	CEdw2	Curral Velho, Cape Verde	914,467	1,262,146	26,154	33.4	4,383	18.5
Calonectris borealis	CBor1	Montaña Clara, Canary Islands	1,528,248	2,139,079	24,867	48.1	4,350	40.9
Calonectris borealis	CBor2	Montaña Clara, Canary Islands	1,060,583	1,470,492	24,258	34.8	4,425	23.2
Calonectris diomedea	CDio1	Menorca, Balearic Islands	1,773,647	1,110,846	25,089	55.4	4,088	24.0
Calonectris diomedea	CDio2	Menorca, Balearic Islands	918,281	-	24,476	30.4	-	
Calonectris diomedea	CDio2	Off Cape Hatteras, USA	-	1,867,956	-		4,365	25.6
Ardenna bulleri	ABul1	North Pacific Ocean	320,868	2,133,156	23,390	12.0	4,260	42.0
Ardenna pacifica	APac1	Sand Island, Johnston Atoll	821,754	1,375,420	23,989	27.2	4,270	20.2
Ardenna pacifica	APac2	Sand Island, Johnston Atoll	1,582,710	-	32,482	52.0	-	
Ardenna pacifica	APac2	Kauai, Hawaii, USA	-	1,690,182	-		4,383	46.0
Ardenna tenuirostris	ATen1	Woolamai, Victoria, Australia	1,756,109	-	27,580	47.8	-	
Ardenna tenuirostris	ATen1	Magadan, Russia	-	1,690,182	-		4,183	35.3
Ardenna tenuirostris	ATen2	Woolamai, Victoria, Australia	1,609,633	1,902,018	28,097	44.0	4,141	23.5
Ardenna grisea	AGri1	Kidney island, Falklands, UK	1,849,813	-	28,328	51.0	-	
Ardenna grisea	AGri1	Sand Island, Johnston Atoll	-	1,264,537	-		3,901	29.8
Ardenna grisea	AGri2	Kidney island, Falklands, UK	1,280,890	1,984,435	26,168	38.6	4,385	29.1
Ardenna gravis	AGra1	Gough Island, Tristan da Cunha	1,345,709	1,621,299	26,414	39.5	4,050	31.1
Ardenna gravis	AGra2	Gough Island, Tristan da Cunha	518,516	-	25,478	16.7	-	
Ardenna gravis	AGra2	Gough Island, Tristan da Cunha	-	1,825,470	-		4,336	12.3
Ardenna carneipes	ACar1	North Pacific Ocean	272,873	759,422	22,863	13.0	3,525	13.1
Ardenna carneipes	ACar2	Ocean off Albany, Australia	309,011	2,274,464	23,334	15.5	4,305	38.3
Ardenna creatopus	ACre1	Juan Fernández, Chile	310,127	2,295,537	22,138	14.5	4,281	41.5
Ardenna creatopus	ACre2	North Pacific Ocean	375,000	2,613,080	22,871	16.4	4,291	40.5
Puffinus nativitatis	PNat1	Sand Island, Johnston Atoll	1,162,644	1,278,699	24,188	38.7	4,029	26.0
Puffinus nativitatis	PNat2	Sand Island, Johnston Atoll	889,479	-	23,968	30.5	-	
Puffinus nativitatis	PNat2	North Pacific Ocean	-	1,235,475	-		4,306	19.6
Puffinus huttoni	PHut1	New Zealand	885,934	1,839,659	27,000	28.0	4,243	37.2
Puffinus gavia	PGav1	Near Raglon, New Zealand	345,693	595,063	22,423	15.7	3,282	10.7

Table SI continued

Taxon	Sample ID	Locality	PE-ddRAD	UCE reads	PE-ddRAD contigs	PE-ddRAD coverage (x̄)	UCE contigs	UCE coverage (x̄)
Puffinus gavia	PGav2	Off Coromandel, New Zealand	1,000,461	1,999,847	27,281	31.6	4,454	27.9
Puffinus assimilis	PAHa1	Hauraki Gulf, New Zealand	1,067,567	-	26,363	36.0	-	
Puffinus elegans	PEle1	Gough Island, Tristan da Cunha	343,997	1,735,216	21,274	44.2	4,252	23.3
Puffinus bryanni	PBry1	Bonin Islands, Japan	105,225	-	1,824	6.6	-	
Puffinus auricularis	PAur1	Socorro Island, Mexico	-	-	-		-	
Puffinus opisthomelas	POpi1	Point Fermin, USA	538,481	2,659,494	24,726	22.4	4,401	47.3
Puffinus newelli	PNNe1	Kauai Island, Hawaii, USA	1,211,221	-	24,450	39.1	-	
Puffinus newelli	PNNe2	Kauai Island, Hawaii, USA	1,159,743	-	24,006	38.6	-	
Puffinus newelli	PNNe1	Kauai Island, Hawaii, USA	-	2,429,498	-		4,351	43.4
Puffinus newelli	PNNe2	Kauai Island, Hawaii, USA	-	1,870,561	-		4,358	25.7
Puffinus myrtae	PMyr1	Rapa Island, French Polynesia	3,548	-	3	6.1	-	
Puffinus bailloni bailloni	PBBa1	Réunion, France	751,761	-	24,792	29.6	-	
Puffinus bailloni bailloni	PBBa1	Réunion, France	-	1,426,196	-		4,305	48.0
Puffinus bailloni bailloni	PBBa2	Réunion, France	900,258	2,664,728	25,135	34.5	4,352	20.5
Puffinus bailloni bailloni	PBBa3	Amsterdam island, TAAF	449,478	1,994,719	25,764	20.1	4,398	21.1
Puffinus bailloni dichrous	PBDi1	Ua Pou, French Polynesia	492,522	-	25,190	31.7	-	
Puffinus bailloni dichrous	PBDi1	Ua Pou, French Polynesia	-	1,964,493	-		4,194	33.3
Puffinus bailloni nicolae	PBNi1	Seychelles	646,609	1,682,650	26,382	27.2	4,195	33.4
Puffinus bailloni nicolae	PBNi2	Seychelles	346,376	-	24,364	16.5	-	
Puffinus bailloni nicolae	PBNi2	Seychelles	-	2,683,371	-		4,505	40.3
Puffinus puffinus	PPuf1	Copeland Islands, UK	1,239,378	-	27,615	66.0	-	
Puffinus puffinus	PPuf1	Bay Center, Washington, USA	-	2,549,079	-		4,411	42.9
Puffinus puffinus	PPuf2	Heimaey Island, Iceland	2,005,299	2,566,705	26,120	57.3	4,203	36.2
Puffinus mauretanicus	PMau1	Sa Conillera, Balearic Islands	2,970,283	1,748,061	27,905	74.5	4,326	25.6
Puffinus mauretanicus	PMau2	Sa Conillera, Balearic Islands	2,575,842	2,255,284	27,816	66.8	4,175	38.1
Puffinus yelkouan	PYel1	Port-Cros, France	1,419,012	1,890,416	27,627	40.8	4,130	31.0
Puffinus yelkouan	PYel2	Port-Cros, France	1,588,402	1,043,807	25,475	47.5	4,050	16.7
Puffinus lherminieri	PLLh1	near Oregon Inlet, USA	1,235,309	1,609,852	25,632	37.4	4,323	24.0
Puffinus lherminieri	PLLh2	Panama	643,128	2,848,408	25,445	30.7	4,235	25.9
Puffinus lherminieri	PLLh3	Martinique, France	611,121	-	25,233	35.9	-	
Puffinus boydi	PLBo1	Ilheu de Cima, Cape Verde	3,069,189	1,618,663	27,229	76.7	3,927	29.4
Puffinus boydi	PLBo2	Ilheu de Cima, Cape Verde	2,131,712	1,492,582	26,817	56.1	4,303	20.7

Table SI continued

Taxon	Sample ID	Locality	PE-ddRAD reads	UCE reads	PE-ddRAD contigs	PE-ddRAD coverage ($\bar{x}$)	UCE contigs	UCE coverage ($\bar{x}$)
Puffinus baroli	PLBa1	Vila, Azores, Portugal	2,705,037	1,267,078	31,839	61.8	4,258	19.6
Puffinus baroli	PLBa2	Vila, Azores, Portugal	3,570,311	1,787,339	28,300	80.8	4,229	34.6
Puffinus baroli	PLBa3	Tenerife, Canary Islands	1,669,996	2,336,564	24,807	50.6	4,391	31.0
Fulmarus glacialis	FGla	Denmark	-	-	31,318	-	-	
Fulmarus glacialis	FGla	USA	-	1,711,428	-		4,244	39.1
Thalassarche chlororhynchos	TChl	Rocha, Uruguay	-	-	28,382	-	-	
Thalassarche chlororhynchos	TChl	Gough Island, Tristan da Cunha	-	2,189,344	-		4,423	31.1
Oceanites oceanicus	OOce	Gulf of Mexico, USA	-	-	31,803	-	-	
Oceanites oceanicus	OOce	San Antonia, Valparaíso, Chile	-	1,506,577	-		4,219	30.7

Table S2 PE-ddRAD datasets used for the second step of the optimisation process.

Software	Parameterisation	Taxon coverage (%)	m	M	n	Number of loci	Sum of branch bootstrap support	Mean Proportion of PIS per locus
Stacks 2	default	65	3	2	1	8,599	4519	0.0245
Stacks 2	default	75	3	2	1	6,276	4591	0.0243
Stacks 2	default	95	3	2	1	2,467	4432	0.0231
Stacks 2	optimal	65	3	5	8	17,604	4577	0.0406
Stacks 2	optimal	75	3	5	8	14,464	4657	0.0404
Stacks 2	optimal	95	3	5	8	7,215	4703	0.0391
Stacks 2	higher m	65	7	5	8	14,850	4558	0.0396
Stacks 2	higher m	75	7	5	8	11,029	4604	0.0393
Stacks 2	higher m	95	7	5	8	4,488	4654	0.0377
Stacks 2	higher n	65	3	5	15	17,394	4551	0.0404
Stacks 2	higher n	75	3	5	15	14,371	4689	0.0403
Stacks 2	higher n	95	3	5	15	7,269	4677	0.0392
Stacks 2	refmap	65	7	5	8	14,983	4533	0.0392
Stacks 2	refmap	75	7	5	8	11,163	4569	0.0392
Stacks 2	refmap	95	7	5	8	4,567	4380	0.0371
PyRAD	clust 89	65	Mindepth=7	—	Wclust=0.89	9,871	4718	0.0428
PyRAD	clust 89	75	Mindepth=7	—	Wclust=0.89	7,735	4768	0.0434
PyRAD	clust 89	95	Mindepth=7	—	Wclust=0.89	2,455	4597	0.0424
PyRAD	clust 94	65	Mindepth=7	—	Wclust=0.94	10,232	4721	0.0429
PyRAD	clust 94	75	Mindepth=7	—	Wclust=0.94	7,851	4753	0.0433
PyRAD	clust 94	95	Mindepth=7	—	Wclust=0.94	2,415	4435	0.0419

Notes: m, M and n correspond to STACKS parameters. For PYRAD, analogous parameter values are shown.

Table S3 Phylogenomic analyses and their discordances with the TENT EXABAYES 75% tree. Each row represents a different analysis, and each column represents an internode. Empty cells denote full support for that internode in that analysis, numbers show bootstrap support or posterior probabilities depending on the analysis and when alternative branches were supported, these are shown before writing their support.

Marker type	Min. Taxon coverage	Phylogenetic analysis	*A.grisea(A.gravis(A.carneipes, A.creatopus))*	*A.gravis(A.carneipes, A.creatopus)*	*(A.bulleri, A.pacifica)*	*(C.borealis,C.diomedea)*
	100	SNAPP	(A.grisea,A.tenuirostris)0.32	0.31	0.62	(C.borealis,C.edwardsii)0.35
	75	raxml-ng_contig				
	75	raxml-ng_iupac				
	75	raxml-ng_iupac_partitioned				99
	75	exabayes_contig				
	75	exabayes_iupac				
	75	exabayes_iupac_partitione				
	75	astral	(A.grisea,A.tenuirostris)1	0.66		(C.borealis,C.edwardsii)1
UCE	75	astral_BS10	(A.grisea,A.tenuirostris)0.39			(C.borealis,C.edwardsii)1
	75	SVDQuartets	99	99		97
	95	raxml-ng_contig				
	95	raxml-ng_iupac				99
	95	exabayes_contig				
	95	exabayes_iupac				
	95	astral	(A.grisea,A.tenuirostris)0.98	0.52		(C.borealis,C.edwardsii)1
	95	astral_BS10	(A.grisea,A.tenuirostris)0.78			(C.borealis,C.edwardsii)0.89
	95	SVDQuartets	95			92
	100	SNAPP	0.95			0.82
	75	raxml-ng_iupac				99
	75	exabayes_iupac				
	75	astral	(A.grisea,A.tenuirostris)1			
	75	astral_BS10				
ddRAD	75	SVDQuartets				
	95	raxml-ng_iupac	97			98
	95	exabayes_iupac				
	95	astral	(A.grisea,A.tenuirostris)1			
	95	astral_BS10	0.97			
	95	SVDQuartets				93
	75	raxml-ng_iupac				
	75	exabayes_iupac				
TENT	75	astral	(A.grisea,A.tenuirostris)1			
	75	astral_BS10				
	75	SVDQuartets				

Table S3 continued

Marker type	Min. Taxon coverage	Phylogenetic analysis	*((P.gavia,P.huttoni)(P.all-nativitatis)*	*(P.Atlant,P.pacif-indian)*
	100	SNAPP	P.elegans(P.gavia,P.huttoni)0.39	(P.Indian-pacific,((P.gavia,P.huttoni)P.elegans))0.31
	75	raxml-ng_contig		
	75	raxml-ng_iupac		
	75	raxml-ng_iupac_partitioned		
	75	exabayes_contig		
	75	exabayes_iupac		
	75	exabayes_iupac_partitioned		
	75	astral		0.92
UCE	75	astral_BS10		
	75	SVDQuartets		
	95	raxml-ng_contig		
	95	raxml-ng_iupac		
	95	exabayes_contig		
	95	exabayes_iupac		
	95	astral		0.83
	95	astral_BS10		
	95	SVDQuartets		
	100	SNAPP		0.98
	75	raxml-ng_iupac		
	75	exabayes_iupac		
	75	astral		
	75	astral_BS10		
ddRAD	75	SVDQuartets		
	95	raxml-ng_iupac		
	95	exabayes_iupac		
	95	astral		0.99
	95	astral_BS10		
	95	SVDQuartets		
	75	raxml-ng_iupac		
	75	exabayes_iupac		
TENT	75	astral		
	75	astral_BS10		
	75	SVDQuartets		

Table S3 continued

Marker type	Min. Taxon coverage	Phylogenetic analysis	*(P.opisthomelas,P.newelli)*	*P.bailloni(P.opisthomelas ,P.newelli)*	*(P.b.nicolae, P.b.dichrous)*	P.atlantic
	100	SNAPP	0.56		0.96	
	75	raxml-ng_contig				
	75	raxml-ng_iupac				
	75	raxml-ng_iupac_partitioned				
	75	exabayes_contig				
	75	exabayes_iupac				
	75	exabayes_iupac_partitioned				
	75	astral	0.66			
UCE	75	astral_BS10				
	75	SVDQuartets	97			
	95	raxml-ng_contig				
	95	raxml-ng_iupac				
	95	exabayes_contig				
	95	exabayes_iupac				
	95	astral	P.opisthomelas(P.newelli,P.bailloni)0.83	(P.bailloni,P.newelli)0.67		0.99
	95	astral_BS10				
	95	SVDQuartets	96		95	
	100	SNAPP	P.opisthomelas(P.newelli,P.bailloni)0.70			
	75	raxml-ng_iupac				
	75	exabayes_iupac				
	75	astral				
	75	astral_BS10				
ddRAD	75	SVDQuartets	95			
	95	raxml-ng_iupac	98			
	95	exabayes_iupac				
	95	astral	0.94			
	95	astral_BS10				
	95	SVDQuartets	74			
	75	raxml-ng_iupac				
	75	exabayes_iupac				
TENT	75	astral				
	75	astral_BS10				
	75	SVDQuartets				

Table S3 continued

Marker type	Min. Taxon coverage	Phylogenetic analysis	*P.puffinus(P.mauretanicus,P.yelkouan)*	*(P.baroli, P.boydi)*	*P.baroli*	*P.boydi*	*P.puffinus*	*P.mauretanicus*
	100	SNAPP	0.6	0.88		NA		NA
	75	raxml-ng_contig	62					0
	75	raxml-ng_iupac	84					0
	75	raxml-ng_iupac_partitioned	82					0
	75	exabayes_contig						0
	75	exabayes_iupac						0
	75	exabayes_iupac_partitioned						0
	75	astral	0.94					0.99
UCE	75	astral_BS10	0.88					0.99
	75	SVDQuartets	58				94	0
	95	raxml-ng_contig	86					54
	95	raxml-ng_iupac	68					0
	95	exabayes_contig						0
	95	exabayes_iupac						
	95	astral	0.78					
	95	astral_BS10	0.98					
	95	SVDQuartets	71		91		98	
	100	SNAPP	0.88			NA		NA
	75	raxml-ng_iupac	94			94		0
	75	exabayes_iupac						0.81
	75	astral	0.99					
	75	astral_BS10						0.99
ddRAD	75	SVDQuartets	67				98	46
	95	raxml-ng_iupac	63			0		25
	95	exabayes_iupac	P.puffinus(P.lherminieri(P.boydi,P.baroli))1					0
	95	astral	0.74					
	95	astral_BS10	0.54					0.85
	95	SVDQuartets	55			98		0
	75	raxml-ng_iupac	92					79
	75	exabayes_iupac						
TENT	75	astral						
	75	astral_BS10						
	75	SVDQuartets	48					68

Table S3 continued

Marker type	Min. Taxon coverage	Phylogenetic analysis	*P.bailloni bailloni*	*P.bailloni nicolae*	*A.grisea*	*A.creatopus*	*A.carneipes*	*C.diomedea*	*C.borealis*	*C.edwardsii*
	100	SNAPP	NA	NA		NA	NA		NA	NA
	75	raxml-ng_contig								
	75	raxml-ng_iupac		47		96	82		90	
	75	raxml-ng_iupac_partitioned		51		95	84		87	
	75	exabayes_contig								
	75	exabayes_iupac								
	75	exabayes_iupac_partitioned								
	75	astral	0.96			0	NA			
UCE	75	astral_BS10				0	NA			
	75	SVDQuartets	98	81		54	41		70	
	95	raxml-ng_contig		0		57	0		97	
	95	raxml-ng_iupac				96	96			
	95	exabayes_contig								
	95	exabayes_iupac		0			0			
	95	astral	0.65	0.98		0	NA		0.99	
	95	astral_BS10	0.98	0.99		0	NA		0.98	
	95	SVDQuartets	89	81	99	68	43	91	0	98
	100	SNAPP	NA	NA		NA	NA		NA	NA
	75	raxml-ng_iupac					98		99	
	75	exabayes_iupac								
	75	astral		0.98					0.98	
	75	astral_BS10				0.99	0.75			
ddRAD	75	SVDQuartets		98		85	0		95	
	95	raxml-ng_iupac		99		97	0		94	96
	95	exabayes_iupac								
	95	astral		0.95			0.84		0.85	0.99
	95	astral_BS10				0.86	0.75			
	95	SVDQuartets		87		61	0		67	
	75	raxml-ng_iupac					99			
	75	exabayes_iupac								
TENT	75	astral				0	0			
	75	astral_BS10					0			
	75	SVDQuartets				97	85		99	

Table S4 Bayesian selection of relaxed-clock models using the stepping-stones method.

Data set	Model	log mL (SE)	Pr	BF
UCE subset 1	SC	-32104.60 (0.150)	0.0014	0.0014
	IR	**-32098.04 (0.659)**	**0.9986**	—
	AR	-32112.95 (0.850)	3.3×10^{-7}	3.3×10^{-7}
UCE subset 2	SC	-31892.77 (0.106)	0.2585	0.3526
	IR	**-31891.73 (0.615)**	**0.7331**	—
	AR	-31896.19 (0.704)	0.0084	0.0115
ddRAD subset 1	**SC**	**-39907.34 (0.184)**	**0.9262**	—
	IR	-39910.57 (0.290)	0.0368	0.0398
	AR	-39910.56 (0.407)	0.0370	0.0400
ddRAD subset 2	SC	-39793.94 (0.199)	1.9×10^{-7}	1.9×10^{-7}
	IR	**-39778.48 (0.484)**	**0.9997**	—
	AR	-39786.77 (0.459)	0.0003	0.0003

Table S5 Summary of ABBA SNPs in the two cases of introgression inferred by Patterson's *D* Statistic analyses. Number of fixed ABBA SNPs, presence of B allele in outgroups species, average number of SNPs and haplotypes per ABBA locus and the proportion of ABBA patterns driven by W-to-S mutations.

Introgression event	N ABBA SNPs	B allele in outgroup species	N SNPs / ABBA locus	N haplotypes / ABBA locus	Proportion ABBA patterns generated by W-to-S mutations
A. tenuirostris - A. grisea	20	4	13.5 (14.5, 0.5452)	15.9 (18.0, 0.2722)	**0.65 (0.43, 0.0257)**
P. boydi - P. lherminieri	16	5	18.1 (15.2, 0.1247)	22.6 (18.9, 0.1278)	0.44 (0.36, 0.2546)

Notes: total average values and P-values of the observed values in ABBA loci are shown in parentheses.

Table S6 Log pseudo-likelihood values for the best network under different numbers of hybridization events (h) using *Puffinus* and *Ardenna* gene trees (with different bootstrap support (BS) thresholds) and SNPs. The last three columns show log pseudo-likelihoods and inheritance probabilities for candidate networks based on the *D*-Statistic results. Results are shown for PE-ddRAD, UCE and TENT datasets.

Genus	Marker type	Input data	Optimized network h = 0	Optimized network h = 1	Optimized network h = 2	Candidate network	Candidate network (Pres / Abs)	Inheritance Probability
Puffinus	PE-ddRAD	gene trees (BS ≥ 0)	40.22	38.07	31.83	*P. lherminieri* —> *P. boydi*	34.12 (Pres)	0.265
						P. boydi —> *P. lherminieri*	34.20 (Abs)	0.102
		gene trees (BS ≥ 10)	29.85	—	—	*P. lherminieri* —> *P. boydi*	25.50 (—)	0.272
						P. boydi —> *P. lherminieri*	25.20 (—)	0.082
		gene trees (BS ≥ 50)	21.07	19.32	18.55	*P. lherminieri* —> *P. boydi*	20.41 (Abs)	0.171
						P. boydi —> *P. lherminieri*	19.58 (Abs)	0.109
		SNPs	725.64	626.08	566.95	*P. lherminieri* —> *P. boydi*	710.42 (Abs)	0.083
						P. boydi —> *P. lherminieri*	712.05 (Abs)	0.065
	UCE	gene trees (BS ≥ 0)	61.49	57.62	52.52	*P. lherminieri* —> *P. boydi*	66.62 (Abs)	0.162
						P. boydi —> *P. lherminieri*	62.58 (Abs)	0.016
		gene trees (BS ≥ 10)	45.02	—	—	*P. lherminieri* —> *P. boydi*	41.49 (—)	0
						P. boydi —> *P. lherminieri*	42.14 (—)	0.022
		gene trees (BS ≥ 50)	43.95	41.21	38.09	*P. lherminieri* —> *P. boydi*	63.44 (Abs)	0.350
						P. boydi —> *P. lherminieri*	45.68 (Abs)	0
		SNPs	3532.71	3080.21	2852.19	*P. lherminieri* —> *P. boydi*	3502.49 (Abs)	0.145
						P. boydi —> *P. lherminieri*	3515.72 (Abs)	0.079
	TENT	gene trees (BS ≥ 0)	35.44	32.42	30.68	*P. lherminieri* —> *P. boydi*	32.42 (Pres)	0.130
						P. boydi —> *P. lherminieri*	33.39 (Abs)	0.078
		gene trees (BS ≥ 10)	22.07	—	—	*P. lherminieri* —> *P. boydi*	24.57 (—)	0.158
						P. boydi —> *P. lherminieri*	21.15 (—)	0.161
		gene trees (BS ≥ 50)	17.10	16.04	15.40	*P. lherminieri* —> *P. boydi*	24.71 (Abs)	0.161
						P. boydi —> *P. lherminieri*	16.91 (Abs)	0.086
		SNPs	593.26	566.17	545.27	*P. lherminieri* —> *P. boydi*	573.74 (Pres)	0.115
						P. boydi —> *P. lherminieri*	576.62 (Abs)	0.113
Ardenna	PE-ddRAD	gene trees (BS ≥ 0)	4.03	0.86	—	*A. tenuirostris* —> *A. grisea*	2.58 (Abs)	0.072
						A. grisea —> *A. tenuirostris*	2.58 (Abs)	0.074
		gene trees (BS ≥ 50)	1.26	0.82	—	*A. tenuirostris* —> *A. grisea*	1.02 (Abs)	0.050
						A. grisea —> *A. tenuirostris*	1.02 (Abs)	0.083
	UCE	gene trees (BS ≥ 0)	1.50	0.96	—	*A. tenuirostris* —> *A. grisea*	1.09 (Abs)	0.080
						A. grisea —> *A. tenuirostris*	1.09 (Abs)	0.104
		gene trees (BS ≥ 50)	1.04	0.77	—	*A. tenuirostris* —> *A. grisea*	0.99 (Abs)	0.095
						A. grisea —> *A. tenuirostris*	1.00 (Abs)	0.100
	TENT	gene trees (BS ≥ 0)	2.01	0.60	—	*A. tenuirostris* —> *A. grisea*	1.02 (Abs)	0.052
						A. grisea —> *A. tenuirostris*	1.02 (Abs)	0.058
		gene trees (BS ≥ 50)	0.53	0.33	—	*A. tenuirostris* —> *A. grisea*	0.38 (Abs)	0.029
						A. grisea —> *A. tenuirostris*	0.37 (Abs)	0.029

Appendix II

Supplementary Information for:

Paleoceanographic Changes in the Late Pliocene Promoted Rapid Diversification in Pelagic Seabirds

Joan Ferrer-Obiol, Helen F. James, R. Terry Chesser, Vincent Bretagnolle, Jacob González-Solís, Julio Rozas, Andreanna J. Welch, Marta Riutort

This appendix includes:

Supplementary Figures S1 to S6

Supplementary Tables S1 to S4

Supplementary Figures

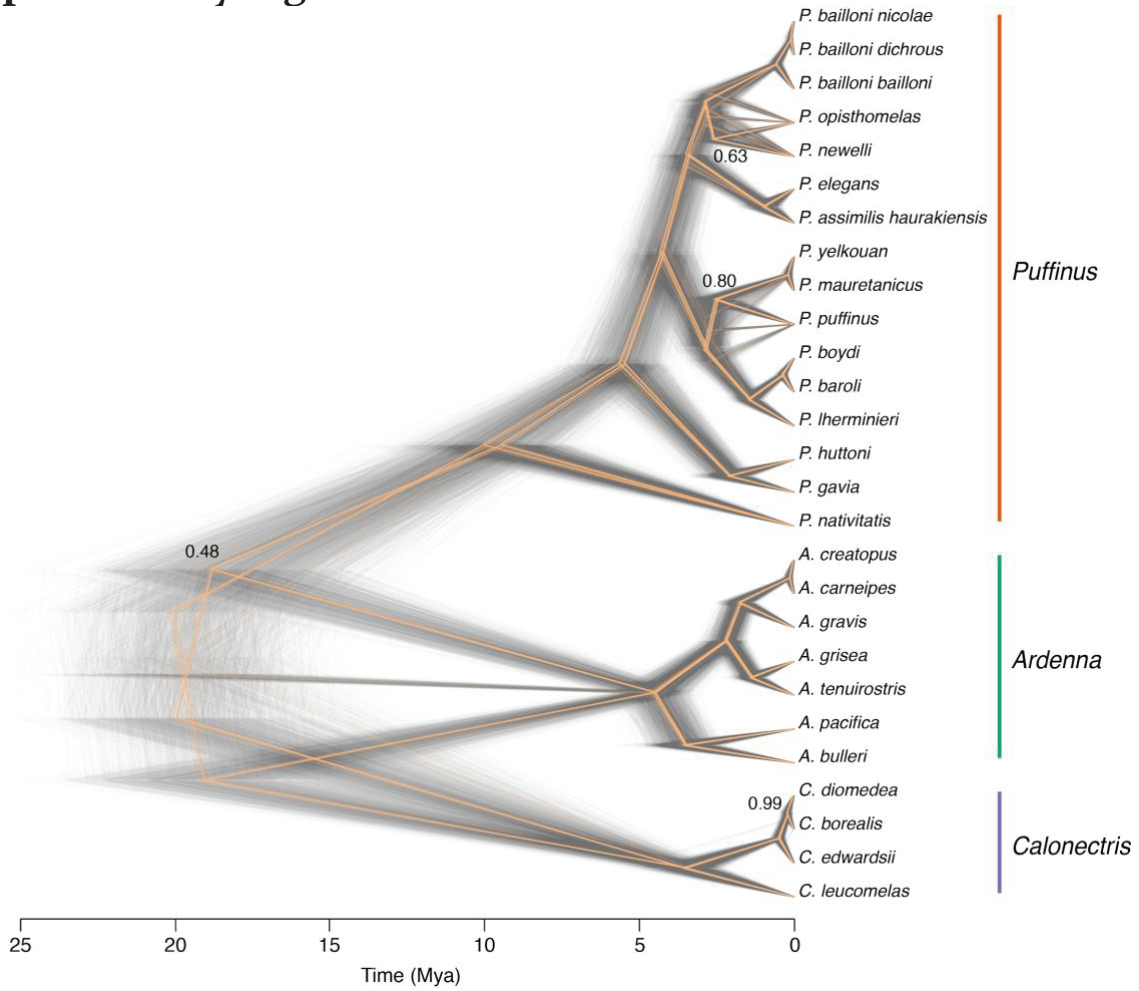


Figure S1 Time-calibrated species tree under the MSC model using a constraint on the root age and no constraints on the topology. Node supports are shown for nodes with bayesian posterior probability (BPP) < 1.

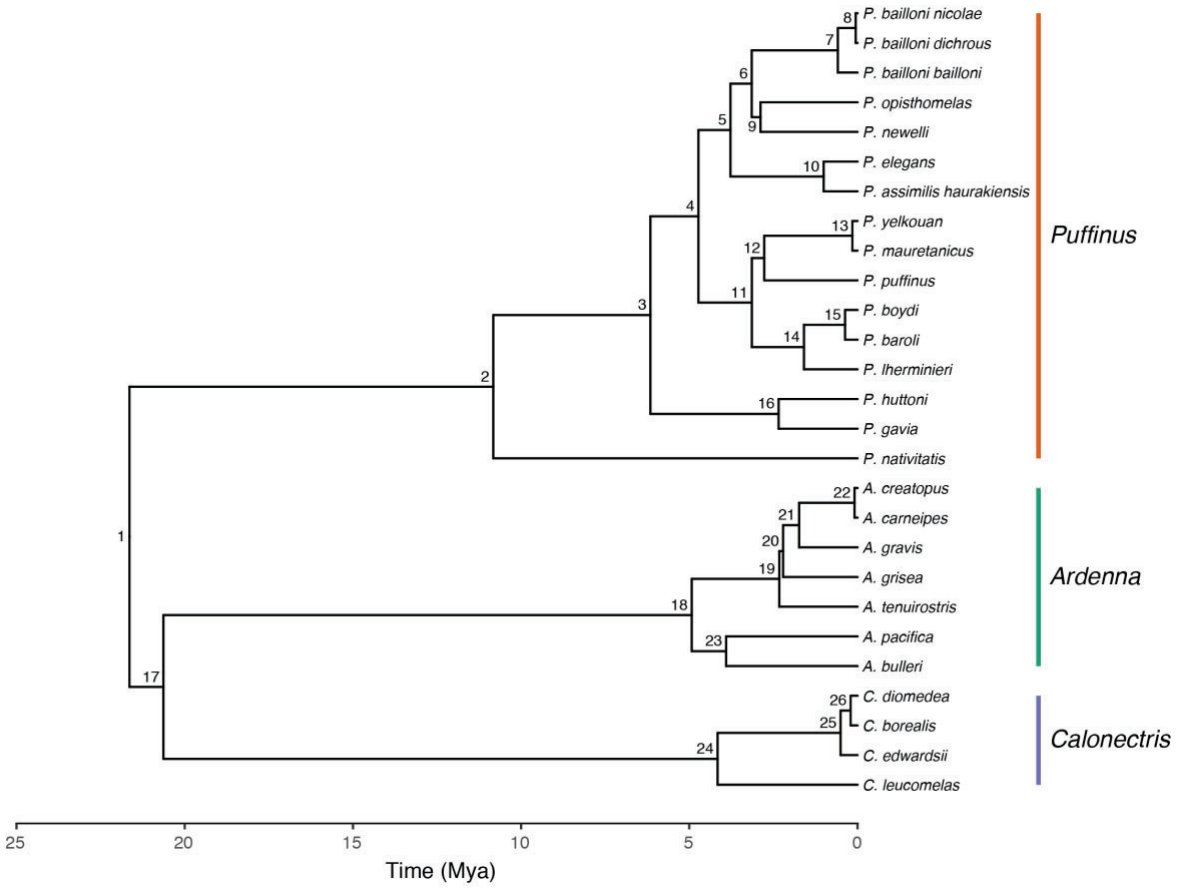


Figure S2 Time-calibrated maximum-clade-credibility summary tree using three fossil calibrations and a fixed topology. Posterior estimates of divergence times are summarized in Table S2. Nodes are numbered to allow comparison to node support and age estimates summarised in Table S2.

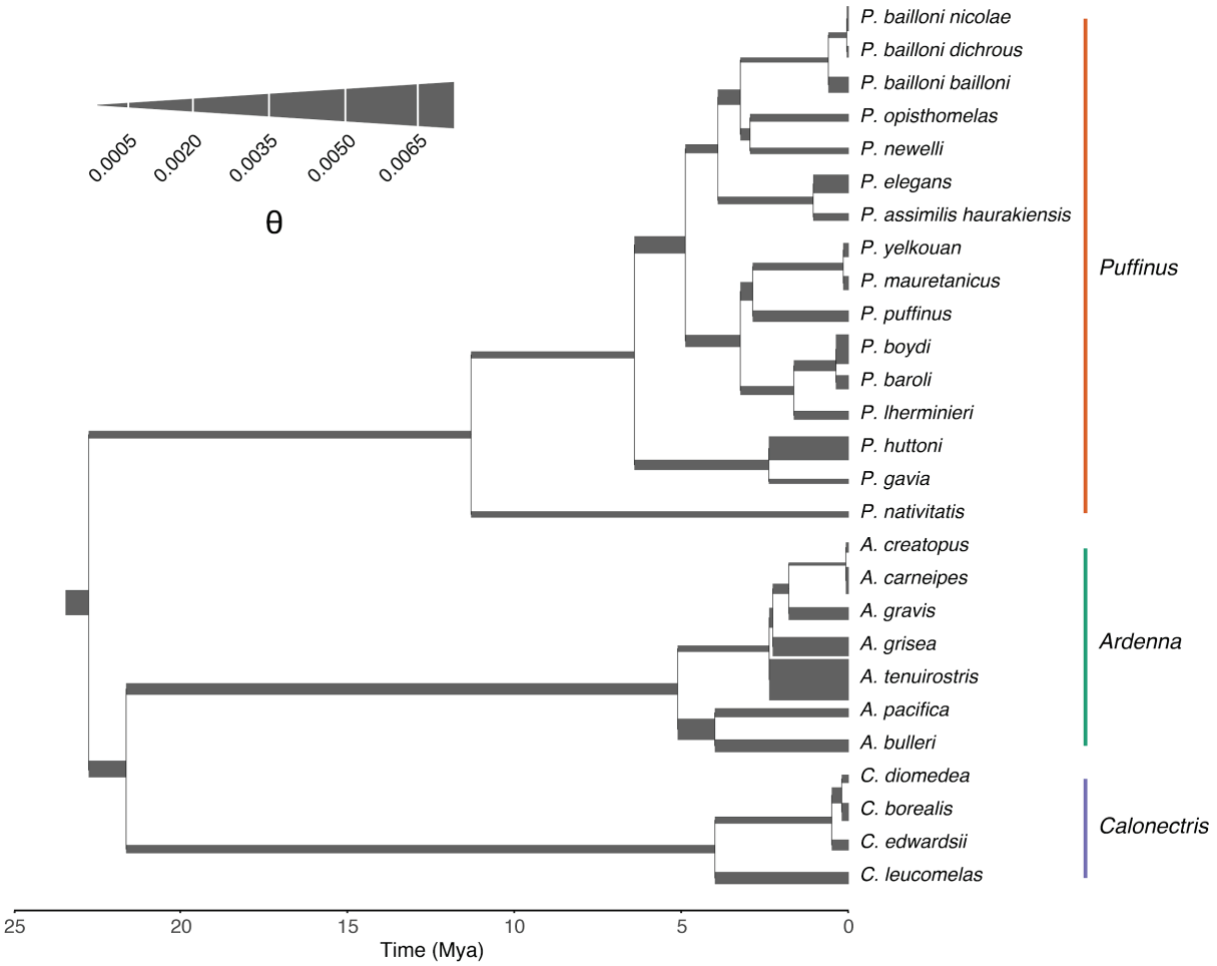


Figure S3 Time-calibrated maximum-clade-credibility summary tree using a constraint on the root age and a fixed topology. Branch widths are proportional to the estimated value of θ from the SNAPP analysis without any age constraint.

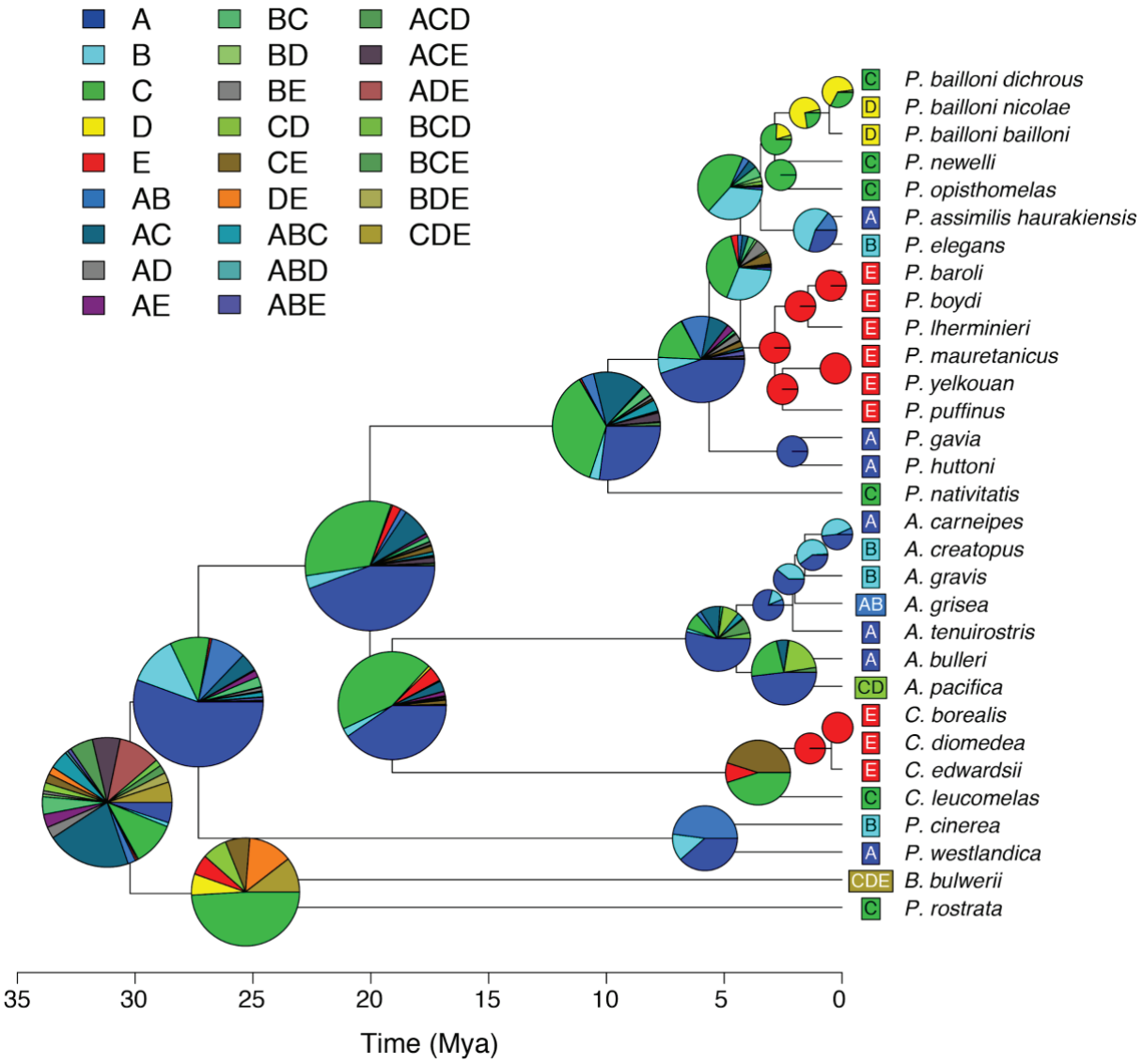


Figure S4 Ancestral ranges estimated under the DIVALIKE + j model using a dispersal matrix restricting dispersal between areas connected by main historical and present ocean currents in BIOGEOBEARS. Estimates are shown as pie charts at nodes and as boxes at tips coded according to the map in Figure 1.

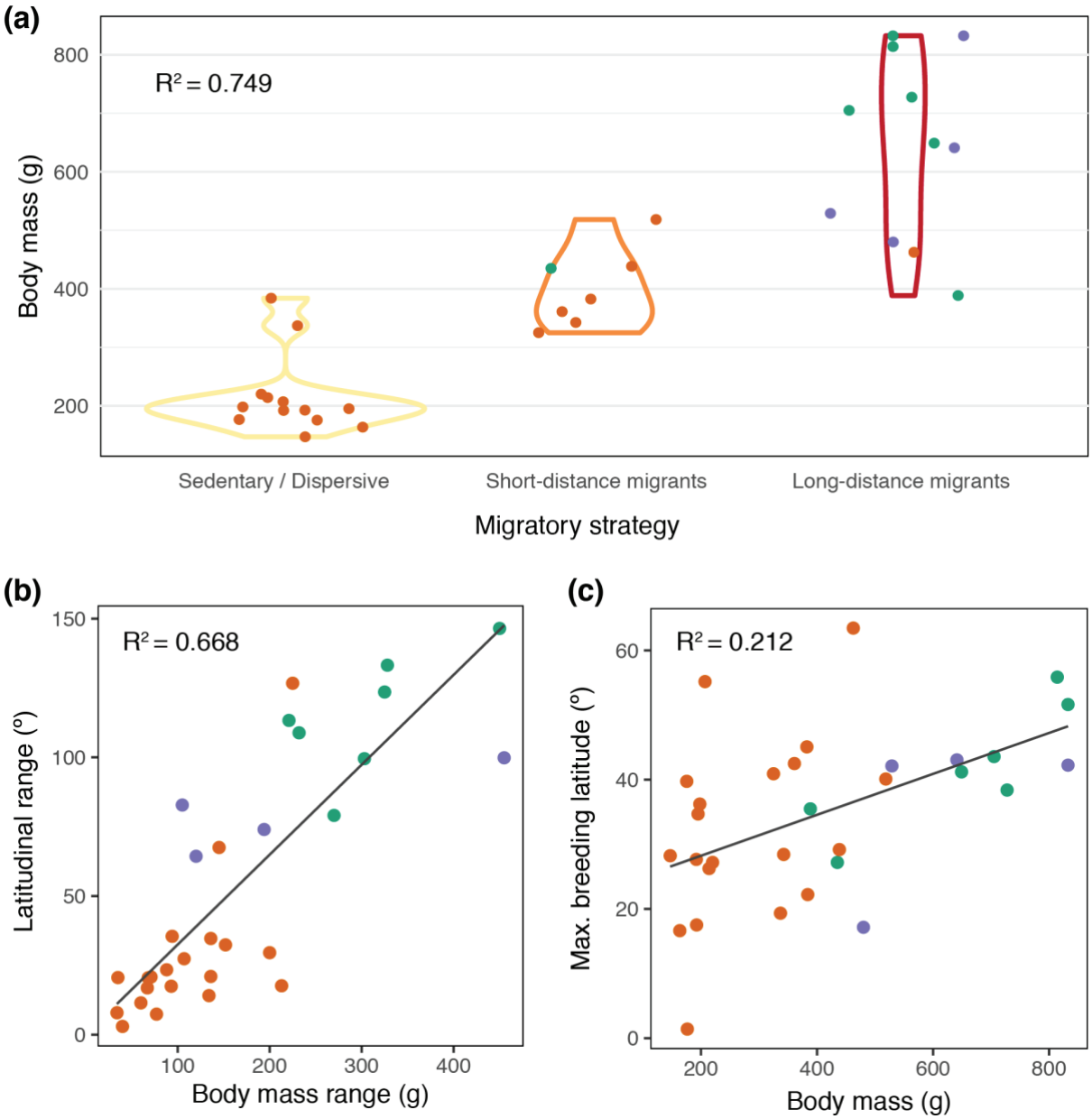


Figure S5 PGLS results. (a) Violin plots showing the relationship between mean body mass and the migratory strategy. (b) Correlation between latitudinal range and body mass range and (c) correlation between maximum breeding latitude and mean body mass. Adjusted R^2 values from PGLS analyses are shown. Orange dots = Puffinus, green dots = Ardenna, blue dots = Calonectris.

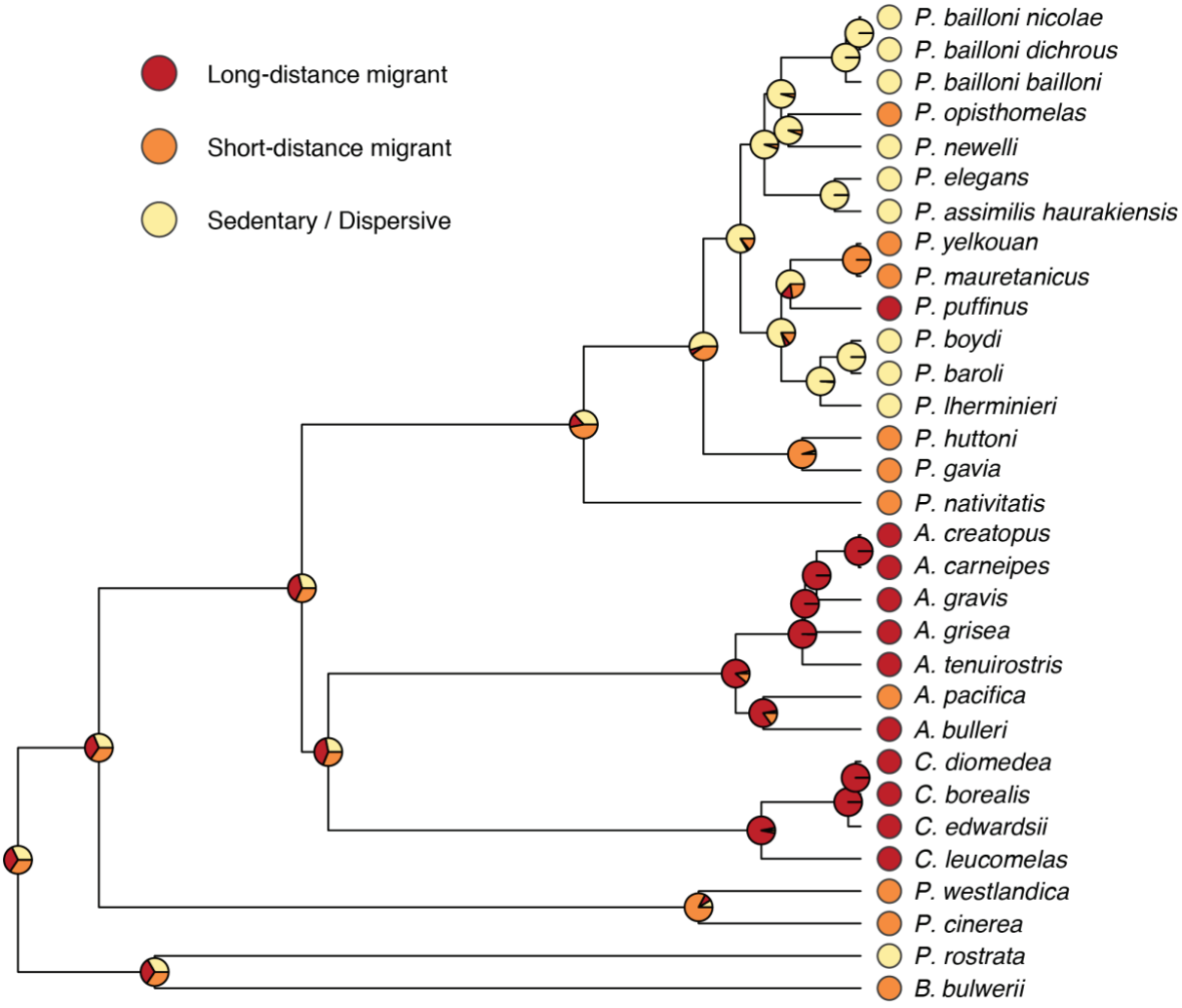


Figure S6 Ancestral state reconstruction of migratory strategy. We considered long-distance migrants those species with at least a percentage of the individuals undertaking trans equatorial migrations, short-distance migrants those species that are not present around the breeding areas during the non-breeding period but do not undertake trans equatorial migrations, and sedentary / dispersive those species that remain around their breeding grounds year-round.

Supplementary Tables

Supplementary Tables

Table S1 Samples used in this study, their sample ID (voucher numbers in case of museum samples), localities and PE-ddRAD sequencing summary statistics for each sample. Means and standard deviations for samples that passed quality control are shown at the bottom.

Taxon	Sample ID	Locality	Number of reads	Total PE-ddRAD contigs	Mean PE-ddRAD coverage
Calonectris leucomelas	González-Solís-CLeu13	Mikura Islands, Japan	2,540,681	27,773	66.1
Calonectris leucomelas	González-Solís-CLeu15	Mikura Islands, Japan	3,773,702	28,109	91.9
Calonectris edwardsii	González-Solís-7500105	Curral Velho, Cape Verde	450,711	22,492	16.6
Calonectris edwardsii	González-Solís-7500660	Curral Velho, Cape Verde	914,467	26,154	33.4
Calonectris edwardsii	González-Solís-7500115	Curral Velho, Cape Verde	815,067	24,932	29.1
Calonectris edwardsii	González-Solís-7500117	Curral Velho, Cape Verde	198,428	8,702	14.1
Calonectris borealis	González-Solís-6168533	Montaña Clara, Canary Islands, Spain	1,528,248	24,867	48.1
Calonectris borealis	González-Solís-6168542	Montaña Clara, Canary Islands, Spain	1,060,583	24,258	34.8
Calonectris borealis	González-Solís-6108848	Terreros Island, Almeria, Spain	1,243,371	23,969	45.7
Calonectris borealis	González-Solís-6108711	Terreros Island, Almeria, Spain	1,727,008	25,447	59.0
Calonectris diomedea	González-Solís-6073669	Menorca, Balearic Islands, Spain	1,773,647	25,089	55.4
Calonectris diomedea	González-Solís-6059707	Menorca, Balearic Islands, Spain	918,281	24,476	30.4
Calonectris diomedea	González-Solís-6114652	Chafarinas Islands, Spain	2,174,270	26,412	74.9
Calonectris diomedea	González-Solís-6120531	Chafarinas Islands, Spain	2,312,728	28,014	76.6
Ardenna bulleri	USNM-613921	North Pacific Ocean	320,868	23,390	12.0
Ardenna pacifica	Austin-Ppac76	Sand Island, Johnston Atoll, USA	821,754	23,989	27.2
Ardenna pacifica	Austin-Ppac77	Sand Island, Johnston Atoll, USA	356,025	17,505	15.9
Ardenna pacifica	Martínez-Gómez-LMG	Revillagigedo Islands, Colima, Mexico	1,582,710	32,482	52.0
Ardenna tenuirostris	González-Solís-Pten56	Woolamai, Victoria, Australia	1,756,109	27,580	47.8
Ardenna tenuirostris	González-Solís-Pten59	Woolamai, Victoria, Australia	1,609,633	28,097	44.0
Ardenna grisea	González-Solís-Pgri1	Kidney island, Falklands, UK	1,849,813	28,328	51.0
Ardenna grisea	González-Solís-Pgri2	Kidney island, Falklands, UK	1,280,890	26,168	38.6
Ardenna gravis	González-Solís-7A01702	Gough Island, Tristan da Cunha, UK	1,345,709	26,414	39.5
Ardenna gravis	González-Solís-89521089	Gough Island, Tristan da Cunha, UK	518,516	25,478	16.7
Ardenna gravis	González-Solís-89521088	Gough Island, Tristan da Cunha, UK	438,792	20,032	17.9
Ardenna carneipes	UWBM-85403	North Pacific Ocean	272,873	22,863	13.0
Ardenna carneipes	AMNH-DOT17805	Ocean off Albany, Western Australia	309,011	23,334	15.5
Ardenna creatopus	AMNH-DOT3131	Juan Fernández, Chile	310,127	22,138	14.5
Ardenna creatopus	UWBM-61948	North Pacific Ocean	375,000	22,871	16.4
Ardenna creatopus	UWBM-55743	North Pacific Ocean	766,873	20,762	31.7
Puffinus nativitatis	Austin-Pnat85	Sand Island, Johnston Atoll, USA	1,162,644	24,188	38.7
Puffinus nativitatis	Austin-Pnat82	Sand Island, Johnston Atoll, USA	889,479	23,968	30.5

Table S1 continued.

Taxon	Sample ID	Locality	Number of reads	Total PE-ddRAD	Mean PE-ddRAD coverage
Puffinus huttoni	LSU-B23388	New Zealand	885,934	27,000	28.0
Puffinus gavia	KU-14876	Near Raglon, New Zealand	345,693	22,423	15.7
Puffinus gavia	UWBM-82796	Off Coromandel, New Zealand	1,000,461	27,281	31.6
Puffinus assimilis haurakiensis	Gaskin-MK012	Hauraki Gulf, New Zealand	1,067,567	26,363	36.0
Puffinus assimilis haurakiensis	Gaskin-MK027	Hauraki Gulf, New Zealand	1,101,638	26,108	37.7
Puffinus elegans	Nunn-LS-2	Gough Island, Tristan da Cunha, UK	343,997	12,627	19.6
Puffinus elegans	Nunn-LS-5	Gough Island, Tristan da Cunha, UK	439,734	13,477	24.5
Puffinus elegans	González-Solís-5H44297	Gough Island, Tristan da Cunha, UK	1,303,103	23,707	47.4
Puffinus opisthomelas	LSU-B19402	Point Fermin, California, USA	538,481	24,726	22.4
Puffinus newelli	Austin-Panw107	Kauai Island, Hawaii, USA	1,211,221	24,450	39.1
Puffinus newelli	Austin-Panw110	Kauai Island, Hawaii, USA	1,159,743	24,006	38.6
Puffinus bailloni bailloni	Bretagnolle-Pufflhe14_2254	Réunion, France	751,761	24,792	29.6
Puffinus bailloni bailloni	Bretagnolle-Pufflhe14_2182	Réunion, France	900,258	25,135	34.5
Puffinus bailloni bailloni	MNHN-1990-796	Amsterdam island, TAAF, France	449,478	19,124	18.5
Puffinus bailloni dichrous	Bretagnolle-Plher_Uapou_VII_98_1	Ua Pou, French Polynesia, France	492,522	20,185	19.6
Puffinus bailloni dichrous	Bretagnolle_Plher_Uapou	Ua Pou, French Polynesia, France	920,145	21,505	35.2
Puffinus bailloni nicolae	Bretagnolle-GE50901	Seychelles	646,609	26,382	27.2
Puffinus bailloni nicolae	Bretagnolle-GE5045	Seychelles	346,376	24,364	16.5
Puffinus puffinus	González-Solís-ET01744	Copeland Islands, UK	1,239,378	27,615	66.0
Puffinus puffinus	González-Solís-477631	Heimaey Island, Iceland	2,005,299	26,120	57.3
Puffinus mauretanicus	Louzao-PMau15	Sa Conillera, Balearic Islands, Spain	2,970,283	27,905	74.5
Puffinus mauretanicus	Louzao-PMau18	Sa Conillera, Balearic Islands, Spain	2,575,842	27,816	66.8
Puffinus yelkouan	González-Solís-FT67724	Port-Cros, France	1,419,012	27,627	40.8
Puffinus yelkouan	González-Solís-FT67739	Port-Cros, France	1,588,402	25,475	47.5
Puffinus lherminieri	Austin-PIB-20918	near Oregon Inlet, USA	1,235,309	25,632	37.4
Puffinus lherminieri	USNM-620721	Panama	643,128	25,445	30.7
Puffinus lherminieri	Bretagnolle-FX14746	Martinique, France	611,121	25,233	35.9
Puffinus lherminieri	Bretagnolle-FX21591	Martinique, France	1,191,084	26,000	40.3
Puffinus boydi	González-Solís-5500446	Ilheu de Cima, Cape Verde	3,069,189	27,229	76.7
Puffinus boydi	González-Solís-5500458	Ilheu de Cima, Cape Verde	2,131,712	26,817	56.1
Puffinus boydi	González-Solís-5500108	Ilheu Raso, Cape Verde	920,680	25,553	31.3
Puffinus boydi	González-Solís-5500109	Ilheu Raso, Cape Verde	1,145,889	26,214	37.9

Table S1 continued.

Taxon	Sample ID	Locality	Number of reads	Total PE-ddRAD contigs	Mean PE-ddRAD coverage
Puffinus baroli	González-Solís-I008097	Vila, Azores, Portugal	2,705,037	31,839	61.8
Puffinus baroli	González-Solís-I008071	Vila, Azores, Portugal	3,570,311	28,300	80.8
Puffinus baroli	Austin-Pabr91	Tenerife, Canary Islands, Spain	1,669,996	24,807	50.6
Puffinus baroli	Austin-Pabr94	Tenerife, Canary Islands, Spain	1,443,790	24,634	51.4
		mean	1,227,032	24,621	39
		sd	815,798	3,837	19

Table S2 Node support and age estimates for time-calibrated trees using SNAPP. Results for concatenation analyses in Chapter I are shown for comparison. Divergence times 95% HPD intervals are shown in parentheses. When a node was not recovered, node supports and age estimates for the alternative recovered topology are shown in parentheses. Nodes are numbered as in Figure S2.

Node	Node support (BPP)				Age estimates (Mya)			
	concatenation	MSC fixed topology + calibrations	MSC fixed topology	MSC SNAPP topology	concatenation	MSC fixed topology + calibrations	MSC fixed topology	MSC SNAPP topology
1	—	—	—	—	21.56 (18.46-23.59)	21.64 (18.17-24.83)	20.09 (16.29-23.86)	19.99 (16.06-23.81)
2	1.00	1.00	1.00	1.00	13.90 (10.63-17.03)	10.82 (8.81-12.84)	9.98 (7.65-12.26)	9.69 (7.55-12.15)
3	1.00	1.00	1.00	1.00	9.77 (7.36-12.07)	6.16 (4.94-7.42)	5.67 (4.47-7.14)	5.53 (4.27-6.26)
4	1.00	1.00	1.00	1.00	8.19 (6.27-10.26)	4.73 (3.89-5.71)	4.32 (3.41-5.42)	4.24 (3.29-5.26)
5	1.00	1.00	1.00	1.00	6.36 (4.70-7.99)	3.78 (3.09-4.64)	3.46 (2.64-4.26)	3.41 (2.57-4.24)
6	1.00	1.00	1.00	1.00	4.68 (3.38-6-01)	3.14 (2.50-3.79)	2.87 (2.23-3.57)	2.84 (2.16-3.53)
7	1.00	1.00	1.00	1.00	1.58 (1.02-2.20)	0.58 (0.39-0.78)	0.54 (0.35-0.75)	0.53 (0.35-0.74)
8	1.00	1.00	1.00	1.00	0.83 (0.47-1.22)	0.05 (0.00-0.14)	0.05 (0.00-0.14)	0.05 (0-0.14)
9	1.00	fixed	fixed	0.63	3.62 (2.49-4.90)	2.88 (3.09-4.64)	2.61 (1.97-3.29)	2.60 (1.94-3.31)
10	1.00	1.00	1.00	1.00	—	1.01 (0.52-1.49)	0.94 (0.46-1.39)	0.93 (0.46-1.43)
11	1.00	1.00	1.00	1.00	6.11 (4.48-7.68)	3.14 (2.59-3.82)	2.87 (2.26-3.54)	2.81 (2.15-3.49)
12	— (1.00)	fixed	fixed	0.80	5.26 (3.78-6.74)	2.77 (2.19-3.46)	2.54 (1.89-3.17)	2.50 (1.81-3.18)
13	1.00	1.00	1.00	1.00	0.82 (0.44-1.22)	0.15 (0.03-0.27)	0.15 (0.04-0.28)	0.15 (0.03-0.27)
14	1.00	1.00	1.00	1.00	3.05 (1.98-4.11)	1.59 (1.25-1.94)	1.45 (1.08-1.84)	1.41 (1.04-1.81)
15	1.00	1.00	1.00	1.00	1.17 (0.67-1.72)	0.37 (0.23-0.54)	0.34 (0.20-0.48)	0.33 (0.19-0.48)
16	1.00	1.00	1.00	1.00	3.20 (1.79-4.69)	2.34 (1.65-3.00)	2.11 (1.46-2.84)	2.06 (1.38-2.76)
17	1.00	fixed	fixed	— (0.48)	19.72 (16.28-23.02)	20.63 (17.13-23.80)	19.10 (15.15-22.86)	— (18.84 (15.17-22.69))
18	1.00	1.00	1.00	1.00	8.77 (6.32-11.07)	4.93 (4.02-5.93)	4.52 (3.53-5.64)	4.47 (3.39-5.53)
19	1.00	1.00	1.00	1.00	5.85 (4.22-7.46)	2.32 (1.88-2.82)	2.10 (1.64-2.62)	2.15 (1.61-2.70)
20	1.00	fixed	fixed	— (1)	4.75 (3.37-6.16)	2.21 (1.76-2.68)	2.01 (1.55-2.51)	— (1.35 (0.94-1.79))
21	1.00	1.00	1.00	1.00	3.24 (2.19-4.38)	1.74 (1.33-2.13)	1.59 (1.17-1.97)	1.72 (1.26-2.19)
22	1.00	1.00	1.00	1.00	0.73 (0.41-1.08)	0.09 (0.00-0.21)	0.08 (0-0.18)	0.09 (0.00-0.22)
23	1.00	1.00	1.00	1.00	6.23 (4.18-8.40)	3.91 (3.14-4.76)	3.54 (2.54-4.58)	3.50 (2.51-4.49)
24	1.00	1.00	1.00	1.00	5.87 (3.80-8.10)	4.16 (3.70-4.88)	3.54 (2.55-4.57)	3.50 (2.46-4.58)
25	1.00	1.00	1.00	1.00	2.11 (1.36-2.92)	0.50 (0.35-0.68)	0.45 (0.27-0.66)	0.44 (0.28-0.63)
26	1.00	1.00	0.99	0.99	1.23 (0.72-1.79)	0.20 (0.08-0.35)	0.19 (0.06-0.32)	0.19 (0.06-0.32)

Table S3 Phylogenetic generalized least squares linear models of the relationship between body size measures and ecological predictors.

			Response variables				
			Mean body mass	Range body mass	Wing length	Body length	Mean body mass / (Wingspan x Body length)
Predictors	Minimum breeding latitude	λ	0.79	0.41	0.93	0.89	0
		$\beta \pm$ SE	5.96 (2.56)	1.80 (1.68)	0.84 (0.51)	0.10 (0.07)	0.0015 (0.0004)
		F value	5.44	1.15	2.72	2.15	11.62
		FDR	0.0536	0.2930	0.1692	0.2043	0.0060
		R^2	0.129	0.005	0.054	0.037	0.262
	Mean breeding latitude	λ	0.77	0.43	0.95	0.91	0
		$\beta \pm$ SE	5.22 (1.81)	1.78 (1.18)	0.46 (0.39)	0.07 (0.05)	0.0012 (0.0003)
		F value	8.34	2.26	1.41	1.68	15.81
		FDR	0.0209	0.2043	0.2579	0.2284	0.0021
		R^2	0.197	0.040	0.013	0.022	0.330
	Maximum breeding latitude	λ	0.79	0.43	0.94	0.91	0
		$\beta \pm$ SE	5.27 (1.75)	2.7 (1.16)	0.48 (0.36)	0.07 (0.05)	0.0013 (0.0003)
		F value	9.05	5.52	1.79	1.80	19.97
		FDR	0.0180	0.0536	0.2253	0.2253	0.0007
		R^2	0.212	0.131	0.026	0.026	0.387
	Latitudinal range	λ	0.80	0.00	0.96	0.93	0.84
		$\beta \pm$ SE	2.61 (0.84)	2.10 (0.27)	0.33 (0.18)	0.05 (0.02)	0.0006 (0.0002)
		F value	9.61	61.39	3.18	4.25	12.08
		FDR	0.0172	1×10^{-7}	0.1420	0.0880	0.0058
		R^2	0.223	0.668	0.068	0.098	0.270
	Migratory strategy	λ	0	0	0.84	0.79	0.645
		$\beta \pm$ SE	—	—	—	—	—
		F value	45.70	15.46	7.37	10.86	8.59
		FDR	1×10^{-7}	0.0002	0.0385	0.0155	0.0051
		R^2	0.749	0.491	0.298	0.397	0.336

Table S4 Covariance analysis between parameters of the substitution process and body mass and the number of breeding pairs in COEVOL.

Covariance	dS	GC*	N pairs	Body mass
dS	291	-5.53	-60.3	-6.25
GC*	—	1.76	9.08	-0.203
N pairs	—	—	100	0.581
Body mass	—	—	—	1.51
Correlation				
dS	—	-0.232	-0.328	-0.277
GC*	—	—	0.684	-0.162
N pairs	—	—	—	0.0256
Body mass	—	—	—	—
Posterior Probability				
dS	—	0.27	0.16	0.21
GC*	—	—	0.94	0.34
N pairs	—	—	—	0.52
Body mass	—	—	—	—

Appendix III

Supplementary Information for:

Neutral Processes Shape Landscapes of Divergence in a Speciation Continuum of Pelagic Seabirds

Joan Ferrer Obiol, Jose M Herranz, Josephine R. Paris, James R. Whiting, Jacob González-Solís, Julio Rozas, Marta Riutort

This appendix includes:

Supplementary Methods

Phylogenetic Analyses

To infer the phylogenetic relationships of the studied taxa and to evaluate the monophyly of the clusters identified by FINERADSTRUCTURE, we estimated phylogenies using concatenation and coalescent methods. *C. borealis* and *P. nativitatis* were used as outgroups. For concatenation analyses, we used the MPI version of EXABAYES v.1.5 (Aberer et al. 2014) and RAXML-NG v.0.6.0 (Kozlov et al. 2019) to estimate unpartitioned Bayesian and maximum-likelihood (ML) phylogenies, respectively. We ran two independent EXABAYES runs with four coupled chains for 1,000,000 generations and we assessed runs for stationarity in TRACER v.1.7 (Rambaut et al. 2018) by checking for effective sample sizes > 300 for all model parameters. A consensus tree from the two independent runs was generated using the CONSENSE programme from the EXABAYES package (burnin: 25%). We conducted 50 ML tree searches in RAXML-NG with the GTR+G substitution model using 25 random and 25 parsimony-based starting trees. Following the best tree search, we generated 500 non-parametric bootstrap replicates. Convergence was checked using the RAXML-NG --bsconverge command with a cutoff value of 0.03 and branch support values were mapped onto the best-scoring ML tree using the RAXML-NG --support command.

For SNAPP, forward (u) and backward (v) mutation rate parameters were set to 1, and not sampled. We used the default value 10 for the coalescent rate parameter and we sampled it during the MCMC run. We used uninformative priors as we do not assume strong *a priori* knowledge about the parameters. A gamma distribution was used for the tree height prior λ. Two replicates were run for 100,000 burn-in iterations, followed by 1,000,000 MCMC. Tree and parameter estimates were sampled every 1000 MCMC generations. Convergence and stationarity were confirmed (effective sample sizes > 300) using TRACER. Posterior distributions of run replicates were merged to generate maximum-clade-credibility (MCC) trees with node heights set to mean age estimates with TREEANNOTATOR (Heled and Bouckaert 2013). MCC trees were visualised in DENSITREE v.2.2.7 (Bouckaert 2010). For ASTRAL-III, we used RAXML v.8 (Stamatakis 2014) to estimate gene trees for each PE-ddRAD locus in the 65% complete dataset running

100 rapid bootstrap replicates followed by a thorough ML search. We then used ASTRAL-III to estimate a species trees from the best-scoring ML gene trees and bootstrap replicates. Following Zhang et al. (2018), we contracted very low support branches (BS < 10) on gene trees prior to species tree estimation. Branch supports were inferred using local posterior probabilities (Sayyari and Mirarab 2016).

To estimate divergence times, we applied the MSC approach of Stange et al. (2018) implemented in SNAPP. To avoid the inclusion of potentially introgressed individuals, we performed the analysis including only two individuals per taxon from the most geographically distant populations. Because SNAPP assumes a single nucleotide substitution rate, we performed the analysis only with transitions to reduce rate heterogeneity. To remove potentially linked SNPs we only included transitions separated by > 5 Kbp, which resulted in datasets with no missing data of 5403 transitions. We followed Stange et al. (2018) in specifying an age constraint on the root as a normally-distributed calibration density with a mean of 2.87 Mya and a standard deviation (SD) of 0.39 (Chapter II). We conducted three replicate runs, each of 1,500,000 Markov-chain Monte Carlo (MCMC) iterations after 100,000 burn-in iterations. Post-treatment of this analysis was performed as outlined above for standard SNAPP analysis.

Species Delimitation

We followed the BFD* protocol (Leaché et al. 2014), using a matrix of 500 SNPs with no missing data, to rank ten competing species delimitation hypotheses (SDH) based on the five most popular world bird lists (IOC v.10.2: Gill et al. 2020; Clements v2019: Clements et al. 2019; HBW & Birdlife International: del Hoyo et al. 2014; Howard & Moore v.4.1: Christidis 2014; Peters: Peters 1931), and also using the genetic clustering and phylogenetic analyses performed here (Table 2). For each SDH, we conducted species tree estimation and calculated marginal likelihoods estimates (MLE) using SNAPP. We set mutation rates u and v to 1.0. The prior for the expected genetic divergence (θ) was set with a mean (alpha/beta) of 0.002 using a gamma distribution θ ~ G(2, 1000) and the gamma hyperprior for the speciation rate parameter (lambda) was set to G(2, 200). For MLE calculation, we performed path sampling analyses with 40

steps for 100,000 iterations after a pre-burnin of 12,000 iterations and setting alpha to 0.3. Every analysis was run twice using different seeds to assess consistency in marginal likelihood estimation.

Because the amount of SNPs included in the analysis has the potential to impact model ranks when using BFD* (Leaché et al. 2014), we also performed additional analyses using 2000 SNPs with no missing data. To ensure computational tractability using this larger amount of SNPs, we performed analyses separately for each of the three main groups detected by our phylogenetic and populations structure analyses. Similar to Ciezarek et al. (2019), for each analysis, we included all the individuals in the group and five individuals, designated as a separate species, of its sister group as outgroups. Each analysis was run for 48 steps at a chain length of 200,000, following 50,000 pre-burn-in iterations. For both types of analyses, models were ranked by their MLE and MLEs were compared using Bayes Factors (Kass and Raftery 1995).

Liftover Approach

We aligned the *P. mauretanicus* scaffolds to *C. anna* chromosomes using Satsuma2 (https://github.com/bioinfologics/satsuma2) and used Kraken (Zamani et al. 2014) to translate PE-ddRAD loci *P. mauretanicus* genomic coordinates to *C. anna* coordinates. We then incorporated the new coordinates into the STACKS2 catalog using custom Python scripts and integrated alignment positions to the catalog using STACKS-INTEGRATE-ALIGNMENTS. Such an approach inevitably results in some erroneous assignments due to chromosomal rearrangements between *P. mauretanicus* and *C. anna*. However, genome shuffling is low in birds (Zhang et al. 2014), which are characterised by high levels of synteny, and thus erroneous assignments should only marginally affect the results.

References

Aberer A.J., Kobert K., Stamatakis A. 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–2556.

Bouckaert R.R. 2010. DensiTree: Making sense of sets of phylogenetic trees. Bioinformatics. 26:1372–1373.

Christidis L. 2014. The Howard and Moore Complete Checklist of the Birds of the World, version 4.1.

Ciezarek A.G., Osborne, O.G., Shipley, O.N., Brooks, E.J., Tracey, S.R., McAllister, J.D., Gardner, L.D., Sternberg, M.J.E., Block, B., Savolainen, V. 2019. Phylotranscriptomic Insights into the Diversification of Endothermic Thunnus Tunas. Mol. Biol. Evol. 36: 84–96.

Clements J.F., Schulenberg T.S., Iliff M.J., Billerman S.M., Fredericks T.A., Sullivan B.L., Wood C.L. 2019. The eBird/Clements Checklist of Birds of the World: v2019.

del Hoyo J., Collar N. J., Christie D. A., Elliott A. , Fishpool L. D. C., Boesman P., Kirwan G. M. 2014. HBW and BirdLife International Illustrated Checklist of the Birds of the World, Volume 1, Lynx Edicions in association with BirdLife International, Barcelona, Spain and Cambridge, UK.

Gill F., Donsker D., Rasmussen P. 2020. IOC World Bird List (v10. 1). doi : 10.14344/IOC.ML.10.1.

Heled J., Bouckaert R.R. 2013. Looking for trees in the forest: summary tree from posterior samples. BMC Evol. Biol. 13:221.

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

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35:4453–4455.

Leaché A.D., Fujita M.K., Minin V.N., Bouckaert R.R. 2014. Species delimitation using genome-wide SNP Data. Syst. Biol. 63:534–542.

Peters J.L. 1986. Check-list of Birds of the World, 1931-1986. Harvard University Press/Museum of Comparative Zoology.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. 67:901–904.

Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stange M., Sánchez-Villagra M.R., Salzburger W., Matschiner M. 2018. Bayesian Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports Miocene Closure of the Panamanian Isthmus. Syst. Biol. 67:681–699.

Zamani N., Sundström, G., Meadows, J.R.S., Höppner, M.P., Dainat, J., Lantz, H., Haas, B.J., Grabherr, M. G. 2014. A universal genomic coordinate translator for comparative genomics. BMC Bioinformatics. 15:227.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:15–30.

Zhang G., Li C., Li Q., Li B., Larkin D.M., Lee C. ... Ganapathy G. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 346: 1311–1321.

Supplementary Figures

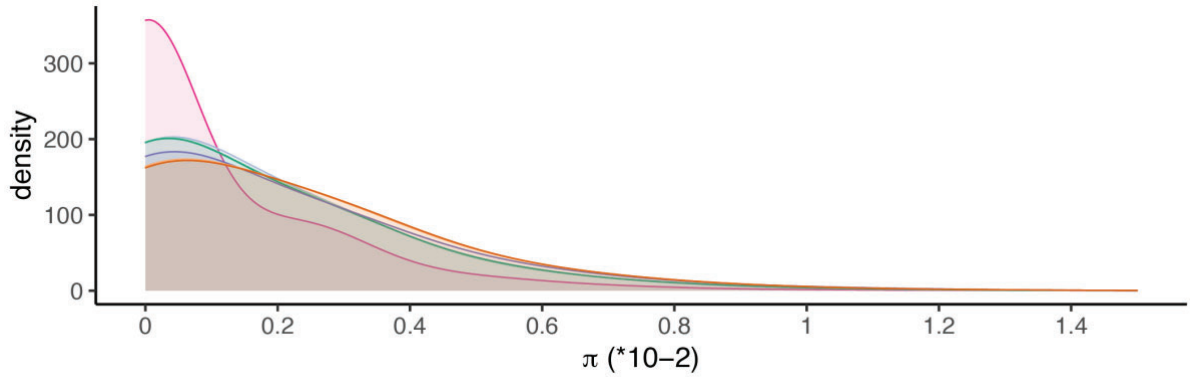


Figure S1 Smoothed distributions of per-locus π estimates for seven shearwater taxa using the function geom_smooth from the R package ggplot2. Each colour represents a different taxon: *P. mauretanicus* (dark green), *P. yelkouan* (light green), *P. p. canariensis* (dark orange), *P. p. puffinus* (light orange), *P. baroli* (light purple) and *P. boydi* (dark purple). Note the higher proportion of low π values in *P. lherminieri*.

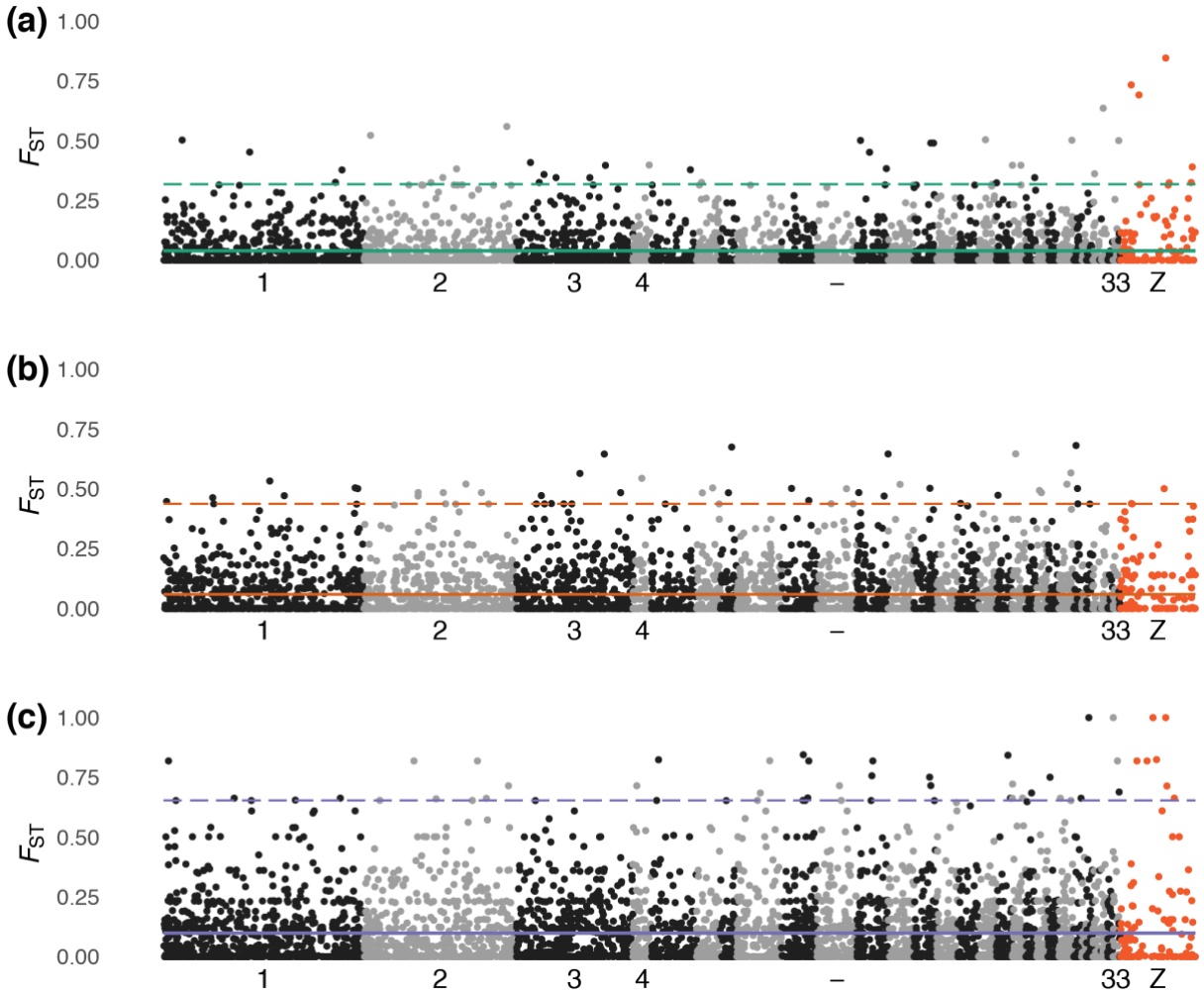


Figure S2 Pairwise genetic differentiation (F_{ST}) per locus across the Anna's hummingbird genome (*Calypte anna*) between (a) *P. mauretanicus* and *P. yelkouan*, (b) *P. p. puffinus* and *P. p. canariensis*, and (c) *P. boydi* and *P. baroli*. Alternating shading denotes the different chromosomes and the Z chromosome is coloured in red for easier visualisation. Horizontal continuous lines mark the mean genome-wide F_{ST} and horizontal dashed lines the 95th percentile.

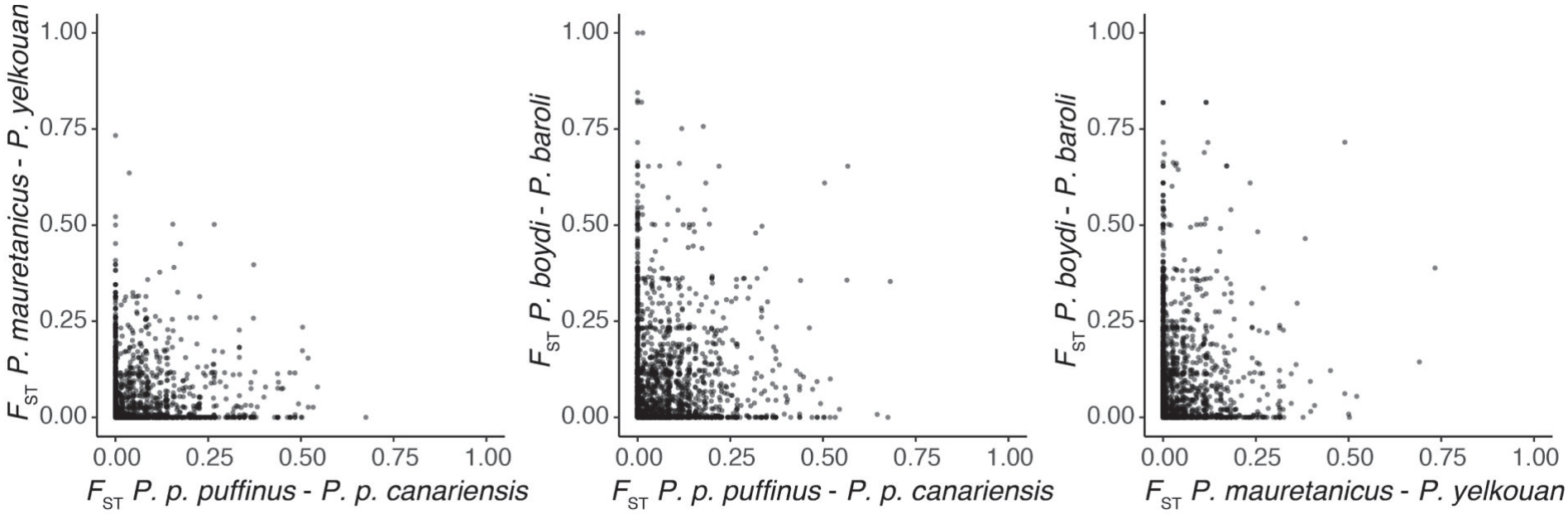


Figure S3 Per-locus F_{ST} values show no association between taxon pairs. Each dot represents a locus.

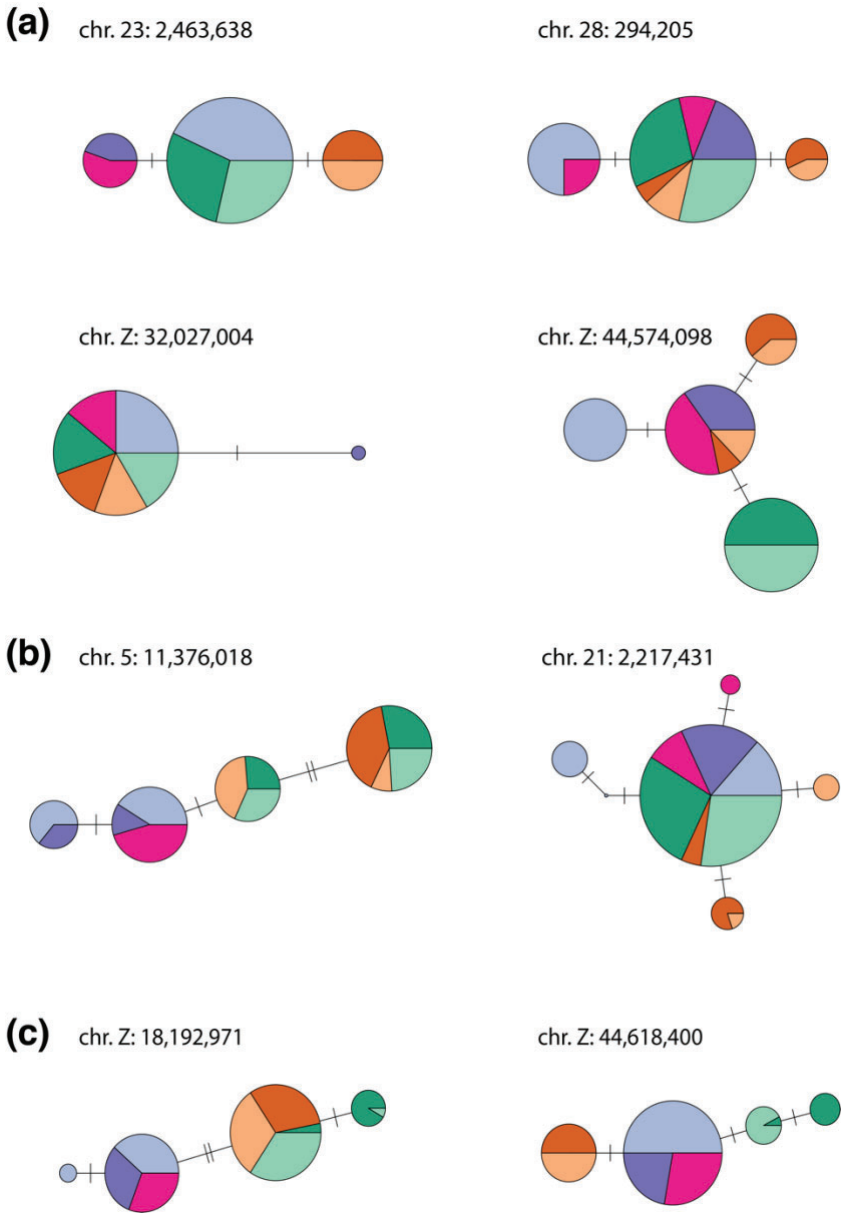


Figure S4 Haplotype networks showing allelic variation in (a) the four highest differentiation outlier loci in the *P. boydi* – *P. baroli* comparison, (b) the two highest differentiation outlier loci in the *P. p. puffinus* – *P. p. canariensis* comparison, and (c) the two highest differentiation outlier loci in the *P. mauretanicus* – *P. yelkouan* comparison. Every colour represents a different taxon: *P. mauretanicus* (dark green), *P. yelkouan* (light green), *P. p. canariensis* (dark orange), *P. p. puffinus* (light orange), *P. baroli* (light purple), *P. boydi* (dark purple) and *P. lherminieri* (pink).

Supplementary Tables

Table S1 Samples used in this study, their sample ID, localities and PE-ddRAD sequencing summary statistics per sample.

Species	Sample ID	Locality	Number of reads	Total PE-ddRAD contigs assembled	Mean coverage
Puffinus puffinus puffinus	PPuf1	Heimaey island, Iceland	2,005,299	26,120	67.7
Puffinus puffinus puffinus	PPuf2	Heimaey island, Iceland	778,361	24,058	26.6
Puffinus puffinus puffinus	PPuf3	Isle of Rum, Scotland	2,381,735	25,117	77.6
Puffinus puffinus puffinus	PPuf4	Isle of Rum, Scotland	1,072,341	25,297	34.0
Puffinus puffinus puffinus	PPuf5	Copeland Island, Northern Ireland, UK	1,239,378	25,738	39.3
Puffinus puffinus puffinus	PPuf6	Copeland Island, Northern Ireland, UK	805,981	26,094	25.7
Puffinus puffinus puffinus	PPuf7	Azores	1,386,533	27,013	41.7
Puffinus puffinus canariensis	PPuf8	Madeira	1,219,234	26,267	37.8
Puffinus puffinus canariensis	PPuf9	Madeira	1,077,130	26,473	33.1
Puffinus puffinus canariensis	PPuf10	La Palma, Canary Islands, Spain	1,347,508	26,732	42.1
Puffinus puffinus canariensis	PPuf11	La Palma, Canary Islands, Spain	1,064,049	27,061	32.2
Puffinus puffinus canariensis	PPuf12	Tenerife, Canary Islands, Spain	767,542	25,404	24.4
Puffinus mauretanicus	PMau1	Conillera, Ibiza, Balearic Islands, Spain	2,970,283	27,477	88.5
Puffinus mauretanicus	PMau2	Conillera, Ibiza, Balearic Islands, Spain	2,575,842	27,060	78.6
Puffinus mauretanicus	PMau3	Sa Cella, Mallorca, Balearic Islands, Spain	1,819,147	27,073	54.6
Puffinus mauretanicus	PMau4	Sa Cella, Mallorca, Balearic Islands, Spain	950,584	26,617	29.9
Puffinus mauretanicus	PMau5	Menorca, Balearic Islands, Spain	1,408,770	26,181	43.8
Puffinus mauretanicus	PMau6	Menorca, Balearic Islands, Spain	1,680,132	27,084	50.4
Puffinus yelkouan	PYel1	Port Cross, France	1,419,012	25,486	45.5
Puffinus yelkouan	PYel2	Port Cross, France	1,588,402	23,993	53.7
Puffinus yelkouan	PYel3	Tunis	714,450	25,611	23.7
Puffinus yelkouan	PYel4	Tunis	463,132	24,453	16.0
Puffinus yelkouan	PYel5	Croatia	956,002	25,585	31.0
Puffinus yelkouan	PYel6	Croatia	893,718	25,049	29.9
Puffinus lherminieri loyemilleri	PLhe1	Near Oregon Inlet, USA	1,235,309	24,060	41.5
Puffinus lherminieri loyemilleri	PLhe2	Near Oregon Inlet, USA	1,183,604	27,245	33.7
Puffinus lherminieri loyemilleri	PLhe3	Panama	643,128	25,179	20.0
Puffinus lherminieri lherminieri	PLhe4	Martinique	611,121	22,529	22.3
Puffinus lherminieri lherminieri	PLhe5	Martinique	1,191,084	26,226	38.1
Puffinus baroli	PBar1	Vila, Azores, Portugal	2,705,037	30,605	73.0
Puffinus baroli	PBar2	Vila, Azores, Portugal	3,570,311	28,991	97.3
Puffinus baroli	PBar3	Madeira, Portugal	1,159,455	26,024	36.5
Puffinus baroli	PBar4	Selvagems, Portugal	655,819	23,808	21.8
Puffinus baroli	PBar5	Selvagems, Portugal	1,115,408	26,081	34.5
Puffinus baroli	PBar6	Tenerife, Canary Islands, Spain	1,669,996	23,887	58.6
Puffinus baroli	PBar7	Tenerife, Canary Islands, Spain	1,443,790	24,129	49.6
Puffinus baroli	PBar8	Lanzarote, Canary Islands, Spain	983,575	26,096	31.5
Puffinus baroli	PBar9	Lanzarote, Canary Islands, Spain	984,368	26,182	32.0
Puffinus boydi	PBoy1	Ilheu de Cima, Cape Verde	3,069,189	27,664	91.0
Puffinus boydi	PBoy2	Ilheu de Cima, Cape Verde	2,131,712	26,493	64.7
Puffinus boydi	PBoy3	Ilheu Raso, Cape Verde	920,680	26,479	29.5
Puffinus boydi	PBoy4	Ilheu Raso, Cape Verde	1,145,889	26,874	36.2
Puffinus nativitatis	PNat1	Sand Island, Johnston Atoll, USA	1,162,644	24,188	38.7
Puffinus nativitatis	PNat2	Sand Island, Johnston Atoll, USA	889,479	23,968	30.5
Calonectris borealis	CBor1	Montaña Clara, Lanzarote, Canary	1,528,248	24,867	48.1
Calonectris borealis	CBor2	Montaña Clara, Lanzarote, Canary	1,060,583	24,258	34.8

Table S2 Datasets used for each analysis performed in this study. The number of loci and number of SNPs is reported.

Analysis	Minimum % of individuals required	Other filtering options	Number of loci	Number of SNPs
PCA, DAPC, ADMIXTURE	95	> 5 Kbp between SNPs	7,695	7,695
fineRADstructure (all), TreeMix	75		8,049	63,492
fineRADstructure (*puffinus*)	75		16,339	47,186
fineRADstructure (*mauretanicus-yelkouan*)	75		14,940	37,299
fineRADstructure (*lherminieri-boydi-baroli*)	75		15,903	57,117
Phylogenetic analyses, SplitsTree, DSuite	65		15,525	141,767
Population genomics analyses	75	≥ 75% of taxa	7,106	16,545
BFD* (all)	100		500	500
BFD* (subsets)	100		2000	2,000
BPP (subset 1)	100	≥ 4 SNPs per locus	500	4,972
BPP (subset 2)	100	≥ 1 SNPs per locus	500	2,948

Table S3 Results of species delimitation analyses using BFD* with 2000 SNPs for each of the three main groups of taxa included in this study. For each species delimitation hypothesis, marginal likelihood estimates (MLE), Bayes factors (compared to the current taxonomy) and rank are shown.

Taxa	Model	Species	ML	Rank	2lnBF
P. puffinus	Current taxonomy	1	-20498.1	2	-
	Split *puffinus* / *canariensis* (with Madeira)	2	-20446.5	1	-103.2
P. lherminieri / *P. baroli* / *P. boydi*	Current taxonomy	3	-35261.1	1	-
	Lump *baroli* and *boydi*	2	-35611.9	2	701.5
	Lump *lherminieri*, *baroli* and *boydi*	1	-37535.1	3	4548.2
P. mauretanicus / *P. yelkouan*	Current taxonomy	2	-27506.8	1	-
	Lump	1	-27513.5	3	11.1
	Reassign Menorca	2	-27508.0	2	2.4

Table S4 Number of polymorphic loci, significant BLAST hits against the *P. mauretanicus* annotated proteins and the number of SNPs with synonymous and non-synonymous substitutions per taxon.

Taxon	Polymorphic loci	Significant BLAST hits on *P. mauretanicus* proteins	SNPs with synonymous substitutions	SNPs with non-synonymous substitutions
P. baroli	6138	126	95	31
P. boydi	5665	128	102	26
P. lherminieri	3611	85	59	26
P. mauretanicus	5476	108	87	21
P. yelkouan	5441	109	90	19
P. puffinus puffinus	6397	127	108	26
P. puffinus canariensis	6068	123	97	19

Table S5 Number of F_{ST} outliers for each of the taxon pairs: *P. mauretanicus* versus *P. yelkouan* (MaYe), *P. p. puffinus* versus *P. p. canariensis* (PuCa) and *P. boydi* versus *P. baroli* (BoBa). We report the number of observed overlaps in F_{ST} outliers between taxon pairs and the number of expected overlaps, with 95% confidence intervals in parentheses, based on permutation tests.

F_{st} outliers (95% percentile)			Overlaps in F_{st} outliers				
MaYe	PuCa	BoBa		MaYe - PuCa	MaYe - BoBa	PuCa-BoBa	MaYe - PuCa - BoBa
191	219	224	Observed	6	5	5	1
			Expected	5.9 (2-10)	6.1 (3-10)	6.8 (3-11)	0.2 (0 - 1)

Appendix IV

Other Publications

Feng S., Stiller J., Deng Y., Armstrong J., Fang Q., Reeve A.H., Xie D., Chen G., Guo C., Faircloth B.C., Petersen B., Wang Z., Zhou Q., Diekhans M., Chen W., Andreu-Sánchez S., Margaryan A., Howard J.T., Parent C., Pacheco G., Sinding M.-H.S., Puetz L., Cavill E., Ribeiro Â.M., Eckhart L., Fjeldså J., Hosner P.A., Brumfield R.T., Christidis L., Bertelsen M.F., Sicheritz-Ponten T., Tietze D.T., Robertson B.C., Song G., Borgia G., Claramunt S., Lovette I.J., Cowen S.J., Njoroge P., Dumbacher J.P., Ryder O.A., Fuchs J., Bunce M., Burt D.W., Cracraft J., Meng G., Hackett S.J., Ryan P.G., Jønsson K.A., Jamieson I.G., da Fonseca R.R., Braun E.L., Houde P., Mirarab S., Suh A., Hansson B., Ponnikas S., Sigeman H., Stervander M., Frandsen P.B., van der Zwan H., van der Sluis R., Visser C., Balakrishnan C.N., Clark A.G., Fitzpatrick J.W., Bowman R., Chen N., Cloutier A., Sackton T.B., Edwards S.V., Foote D.J., Shakya S.B., Sheldon F.H., Vignal A., Soares A.E.R., Shapiro B., González-Solís J., **Ferrer-Obiol J.**, Rozas J., Riutort M., Tigano A., Friesen V., Dalén L., Urrutia A.O., Székely T., Liu Y., Campana M.G., Corvelo A., Fleischer R.C., Rutherford K.M., Gemmell N.J., Dussex N., Mouritsen H., Thiele N., Delmore K., Liedvogel M., Franke A., Hoeppner M.P., Krone O., Fudickar A.M., Milá B., Ketterson E.D., Fidler A.E., Friis G., Parody-Merino Á.M., Battley P.F., Cox M.P., Lima N.C.B., Prosdocimi F., Parchman T.L., Schlinger B.A., Loiselle B.A., Blake J.G., Lim H.C., Day L.B., Fuxjager M.J., Baldwin M.W., Braun M.J., Wirthlin M., Dikow R.B., Ryder T.B., Camenisch G., Keller L.F., DaCosta J.M., Hauber M.E., Louder M.I.M., Witt C.C., McGuire J.A., Mudge J., Megna L.C., Carling M.D., Wang B., Taylor S.A., Del-Rio G., Aleixo A., Vasconcelos A.T.R., Mello C.V., Weir J.T., Haussler D., Li Q., Yang H., Wang J., Lei F., Rahbek C., Gilbert M.T.P., Graves G.R., Jarvis E.D., Paten B., Zhang G. 2020. Dense sampling of bird diversity increases power of comparative genomics. Nature. 587:252–257. https://doi.org/10.1038/s41586-020-2873-9

Soliño L., **Ferrer-Obiol J.**, Navarro-Herrero L., González-Solís J., Costa P.R. 2019. Are pelagic seabirds exposed to amnesic shellfish poisoning toxins? Harmful Algae. 84:172–180. https://doi.org/10.1016/j.hal.2019.03.014

Gil-Velasco M., Rouco M., **Ferrer J.**, García-Tarrasón M. 2018. Observaciones de Aves Raras en España, 2016. Ardeola. 65:97–139. https://doi.org/10.13157/arla.65.1.2018.rb

Gil-Velasco M., Rouco M., Ferrer J., García-Tarrasón M. 2017a. Observaciones de Aves Raras en España, 2015. Ardeola. 64:397–442. https://doi.org/10.13157/arla.64.2.2017.rb

Gil-Velasco M., Rouco M., **Ferrer J.**, García-Tarrasón M. 2017b. Observaciones de aves raras en España, 2014. Ardeola. 64:161–235. https://doi.org/10.13157/arla.64.1.2017.rb