








# The First Genome of the Balearic Shearwater (*Puffinus mauretanicus*) Provides a Valuable Resource for Conservation Genomics and Sheds Light on Adaptation to a Pelagic lifestyle

Cristian Cuevas-Caballé <sup>1,†</sup>, Joan Ferrer Obiol <sup>1,2,†</sup>, Joel Vizueta <sup>1,3</sup>, Meritxell Genovart <sup>4</sup>, Jacob Gonzalez-Solis <sup>5</sup>, Marta Riutort <sup>1,\*</sup>, and Julio Rozas <sup>1,\*</sup>

<sup>1</sup>Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Spain

<sup>2</sup>Department of Environmental Science and Policy, Università degli Studi di Milano (UniMi), Milan, Italy

<sup>3</sup>Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark

<sup>4</sup>Mediterranean Institute for Advanced Studies (IMEDEA), CSIC-UIB & Centre for Advanced Studies of Blanes (CEAB), CSIC, Esporles, Spain

<sup>5</sup>Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

\*Corresponding authors: E-mails: mriutort@ub.edu; jrozas@ub.edu.

<sup>†</sup>These authors contributed equally to this work.

Accepted: 03 May 2022

## Abstract

The Balearic shearwater (*Puffinus mauretanicus*) is the most threatened seabird in Europe and a member of the most speciose group of pelagic seabirds, the order Procellariiformes, which exhibit extreme adaptations to a pelagic lifestyle. The fossil record suggests that human colonisation of the Balearic Islands resulted in a sharp decrease of the Balearic shearwater population size. Currently, populations of the species continue to be decimated mainly due to predation by introduced mammals and bycatch in longline fisheries, with some studies predicting its extinction by 2070. Here, using a combination of short and long reads, we generate the first high-quality reference genome for the Balearic shearwater, with a completeness amongst the highest across available avian species. We used this reference genome to study critical aspects relevant to the conservation status of the species and to gain insights into the adaptation to a pelagic lifestyle of the order Procellariiformes. We detected relatively high levels of genome-wide heterozygosity in the Balearic shearwater despite its reduced population size. However, the reconstruction of its historical demography uncovered an abrupt population decline potentially linked to a reduction of the neritic zone during the Penultimate Glacial Period (~194–135 ka). Comparative genomics analyses uncover a set of candidate genes that may have played an important role into the adaptation to a pelagic lifestyle of Procellariiformes, including those for the enhancement of fishing capabilities, night vision, and the development of natriuresis. The reference genome obtained will be the crucial in the future development of genetic tools in conservation efforts for this Critically Endangered species.

**Key words:** Balearic shearwater, conservation genomics, comparative genomics, Procellariiformes marine adaptation.

## Significance

The Balearic shearwater (*Puffinus mauretanicus*) is the most threatened seabird in Europe and some studies predict their extinction by 2070. Here we provide a high-quality genome assembly of this species to study critical aspects relevant to its conservation status. We found that despite its low population size, the species harbors relatively high levels of heterozygosity, whereas its historical demography uncovered an abrupt population decline during the Penultimate Glacial Period. In addition, we identified some candidate genes that may have played an important role into the adaptation to a pelagic lifestyle in Procellariiformes. Our data will be useful for future conservation and management plans.

## Introduction

The genomic sequence of a species accumulates valuable information on the evolutionary history, including demographic and selective events, and on the evolution of genes and traits (Jarvis et al. 2014; Foote et al. 2015; Nadachowska-Brzyska et al. 2015; Feng et al. 2020), information that it is also crucial for the emerging field of conservation genomics (Allendorf 2017). The genetic diversity within a species represents a reservoir of adaptive variation that can help populations to cope with environmental variability (Dussex et al. 2021). Understanding the processes that shape genetic diversity and its distribution pattern within species is paramount to assess the conservation status or the factors responsible for a species decline (Brüniche-Olsen et al. 2021). This knowledge can inform the proposal of effective conservation and management plans, as for instance the definition of management units (Funk et al. 2012). In this context, next-generation sequencing techniques allow the analysis of an increased density of markers across the genome, providing unprecedented accuracy in the estimations of population genetic parameters relevant for scientific-based conservation recommendations (Supple and Shapiro 2018).

The Balearic shearwater is a medium-sized pelagic seabird endemic to the Balearic Islands that is listed as Critically Endangered by the IUCN Red List (IUCN 2021). Despite being the most endangered seabird in Europe, its genomic sequence has not been sequenced to date. At present, its population size is undergoing a fast annual decline of 7.4–14% (Genovart et al. 2016) mostly due to bycatch in longline fisheries and predation by invasive mammals in the colonies (Louzao et al. 2004; Arcos et al. 2008). Currently, it has a reduced number of breeding pairs (estimated as ca., 3,200, Arcos, 2011), with a total population size up to 30,000 individuals due to the vast contingent of floaters (Arcos et al. 2012; Arroyo et al. 2016). Genetic studies based on mtDNA and microsatellites found that this species has low levels of genetic diversity and high inbreeding coefficients (Genovart et al. 2012). Although local inbreeding and natal philopatry represent risk factors which may have negative effects on the viability of the species, the most significant threats come from human activities,

and a population viability study based on demographic modeling predicted that the species would become extinct by 2070 (Genovart et al. 2016). Moreover, studies based on mitochondrial markers (Genovart et al. 2005) and also on morphology and migratory behavior (Austin et al. 2019), suggested a possible ongoing hybridization and introgression process between Balearic and Mediterranean (*P. yelkouan*) shearwaters, which may represent an additional threat for the species.

The Balearic shearwater belongs to the most diverse order of seabirds, the Procellariiformes. This order has a worldwide distribution and comprises more than 140 species (IUCN 2021) in four families: petrels and shearwaters (Procellariidae); northern storm petrels (Hydrobatidae); southern storm petrels (Oceanitidae); and albatrosses (Diomedidae). There is a large variation in body mass and lifestyles within the Procellariiformes, ranging from 20 g to 15 kg; from bodies shaped for diving (e.g., short strong wings used for wing-propelled diving) to those fitted for an extremely vagile lifestyle (thin elongated wings), and from a continuous flapping to dynamic soaring flight modes. Despite this diversity, all Procellariiformes show many morphological, physiological, and life-history traits associated with adaptation to a pelagic lifestyle. Such adaptations include prominent salt glands to facilitate the secretion of salt, adaptations to enhance fishing capabilities such as underwater vision and a particularly acute sense of smell, among other traits (Brooke 2004). Even so, there is no compiled study on the genomic signatures of adaptation to a pelagic lifestyle (but see Silva et al. 2020).

Here, we make available a high-quality reference genome for the Balearic shearwater along its structural and functional annotations. This resource allows us to provide important information for the conservation of the species, such as the estimation of genome-wide heterozygosity and historical demography of the species. Benefiting from this new genome assembly and seven additional Procellariiformes genomes (Feng et al. 2020), we also performed a comparative genomics analysis to uncover the putative adaptations to a pelagic lifestyle. The high-quality genome of the most endangered seabird in Europe presented here will be the basis for further population-based conservation genomics studies.

## Results

### Sequencing Data and Genome Assembly and Annotation

Illumina paired-end ( $2 \times 150$  bp) sequencing of the male-MII yielded a throughput of 147.7 Gbp (table 1), representing a mean coverage of  $118\times$ . The five runs of ONT sequencing of the unsexed-Ei resulted in a  $10\times$  coverage with a read N50 of 9,431 bp. RNA sequencing (RNA-seq) of the chick-MII ( $2 \times 100$  bp) yielded 15 Gbp of data.

We obtained a hybrid assembly with MaSuRCA formed by 4,169 scaffolds, with an N50 of 2.1 Mbp, and an assembly length of 1.21 Gbp (table 2 and fig. 1b). The completeness analysis using BUSCO yields a value of 95.9%, and only 0.3% of the complete genes were duplicated and 1.1% were fragmented (table 2). Our *de novo* repeat annotation analysis shows that 9.95% of the genome consists of repetitive regions (supplementary table S1, Supplementary Material online), which is within the range of previously sequenced avian genomes (Zhang et al. 2014). Among repeat elements, long interspersed nuclear elements were the most abundant (4.45% of the genome). The genome annotation process resulted in a total of 21,959 protein-coding genes, of which 18,769 (85.5%) have at least one gene ontology (GO)-associated term, and 19,218 (87.5%) have hits across the surveyed curated databases (supplementary table S2, Supplementary Material online).

Blood transcriptome assembly from the chick-MII resulted in 224,904 transcripts (supplementary table S3, Supplementary Material online). However, BUSCO completeness was only 62.4%, which was far below genome completeness, probably due to the RNA coming from a single and not very transcriptionally active tissue.

The assembly of the mitogenome of *Puffinus mauretanicus* resulted in a single contig of 19,885 bp long, with a coverage (Illumina reads) of  $371\times$ , which is around three times higher than the coverage of the nuclear genome. This mitogenome has the same gene order as other published Procellariiformes' mitogenomes (supplementary fig. S1, Supplementary Material online). The mitogenome has two copies of the *nad6* gene, as predicted in *P. lherminieri* (Torres et al. 2019); the later feature was also confirmed analyzing the mean coverage (Illumina reads) across these genes (supplementary table S4, Supplementary Material online).

**Table 1**

Sequencing Data, Library Information, and Samples Used in this Study

Library	Total Number of Base Pairs	Number of Reads	Coverage	Individual Code	Location
HiSeq X Ten – TruSeq DNA PCR Free. $2 \times 150$ bp	143,765,593,200	958,437,288	$118\times$	Male-MII	Sa Cella (Mallorca)
ONT – Ligation kit SQK-LSK109 1D	12,142,789,693	2,576,486	$10\times$	Unsexed-Ei	Sa Conillera (Eivissa)
NovaSeq 6000 – TruSeq RNA Sample Prep Kit v2. $2 \times 100$ bp	14,997,592,000	149,975,920	–	Chick-MII	Conills islet (Mallorca)

### Historical Demography of the Balearic Shearwater

Multiple Sequentially Markovian Coalescent (MSMC2) analysis showed support for a steady growth in  $N_e$  from an originally low  $N_e$  followed by a sudden increase  $\sim 200$  ka (fig. 1c). High effective population size did not last long, and the species suffered a sudden decrease to nearly one tenth of the population coinciding with the end of the glacial period before the last interglacial period (119–128 ka) and a prolonged period of low sea level (fig. 1c). Hereafter,  $N_e$  remained stable until  $\sim 10$  ka ago, as more recent MSMC2 time segments are regarded as being unreliable (Schiffels and Wang 2020).

Genome-wide heterozygosity in *P. mauretanicus* was 0.0024, which is within the range of genome-wide heterozygosities estimated for other Procellariiformes (ranging from 0.0014 in *Thalassarche chlororhynchos* to 0.0037 in *Pelecanoides urinatrix*) (fig. 2a). Among Procellariiformes, small-bodied species ( $<200$  g) tended to have higher mean heterozygosities but also higher variance than large-bodied species ( $>450$  g) (fig. 2b).

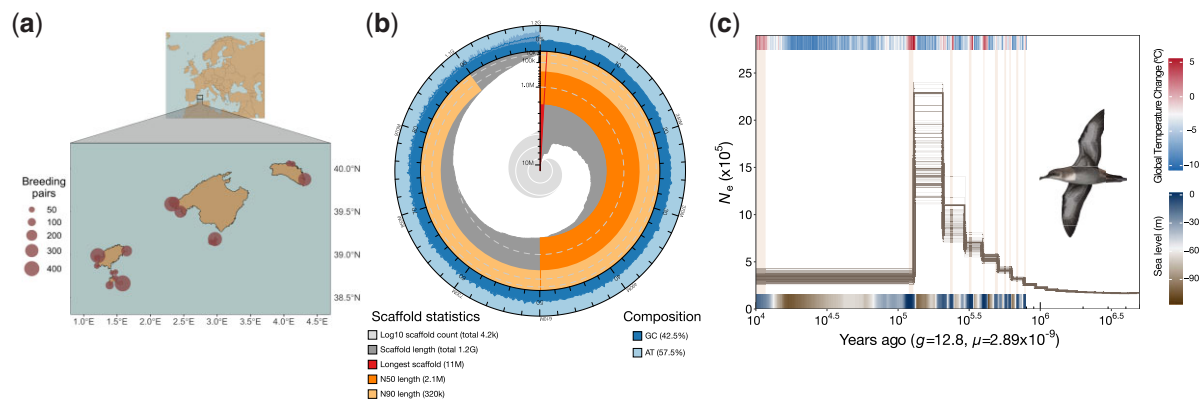
### Phylogenetic Relationships

OrthoFinder analysis estimated 6,172 single-copy (1:1) ortholog genes across the 12 genomes surveyed. With this data we generated three supermatrices: (1) coding sequence (CDS) supermatrix of 10,534,506 bp long to extract

**Table 2**

Balearic Shearwater Genome Assembly Metrics

Assembly length (bp)	1,218,519,395
Number of scaffolds	4,169
Longest scaffold (Mbp)	11.06
N50 (Mbp)	2.13
L50	164
GC content (%)	42.52
Repetitive content (%)	9.95
Mitogenome (bp)	19,855
No. protein-coding genes	21,959
BUSCO %	
Complete	95.9
Single copy	95.6
Duplicated	0.3
Fragmented	1.1
Missing	3.0



**Fig. 1.**—(a) Map depicting the known Balearic shearwater breeding colonies in the Balearic Islands. Circle size is proportional to population size as shown in the legend. Modified from Arcos (2011). (b) Snail plot summarizing genome assembly statistics (Challis et al. 2020). From inside to outside, the light-grey spiral shows the cumulative scaffold count on a log scale with white scale lines depicting changes of order of magnitude. Dark-grey segments show the distribution of scaffold lengths, and the plot radius is scaled to the longest scaffold (shown in red). Orange and light-orange rings represent the N50 and N90 scaffold lengths, respectively. Blue and light-blue rings show GC, AT, and N percentages along the genome assembly. (c) MSMC2 reconstruction of effective population size estimates ( $N_e$ ) over time, estimated using generation time of 12.8 years and mutation rate ( $\mu$ ) of  $2.89 \times 10^{-9}$  substitutions per nucleotide per generation. Light-brown vertical bars represent interglacial periods. Upper panel represents global temperature changes as inferred from the EPICA (European Project for Ice Coring in Antarctica) Dome C ice core (Augustin et al. 2004). Lower panel represents sea level changes inferred from a stack of 57 globally distributed benthic  $\delta^{18}\text{O}$  records (Lisiecki and Raymo 2005). Balearic shearwater illustration by Martí Franch reproduced with permission.

the 4D sites, (2) 4D supermatrix with 1,512,677 4-fold degenerate sites, and (3) the amino acid supermatrix including 3,466,564 sites. Phylogenetic analyses using the 4D and the amino acid supermatrices recovered the same topology as Estandía et al. (2021) with full support at all nodes (ultrafast bootstrap = 100; [supplementary figs. S2 and S3, Supplementary Material](#) online), evidencing that this phylogeny is robust to the use of different phylogenomic markers and to the number of species sampled per taxon. The ultrametric tree (fig. 3) obtained using r8s from the 4D supermatrix ML tree summarizes the recovered topology. In this topology, the Atlantic yellow-nosed albatross (*T. chlororhynchos*, Diomedidae) is the sister group to all the other Procellariiformes. Storm petrels (Hydrobatidae and Oceanitidae) do not constitute a monophyletic group and diving petrels (*P. urinatrix*) are included within Procellariidae.

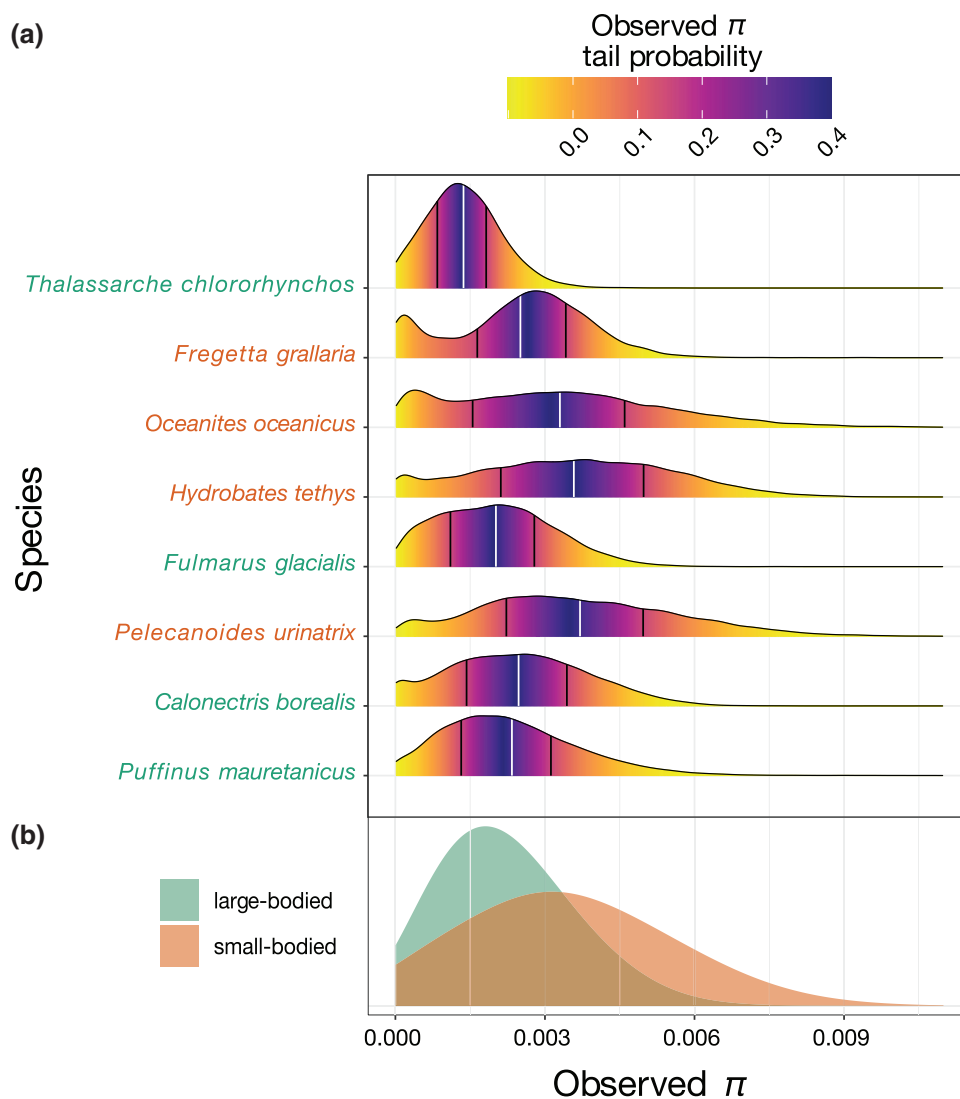
The analysis performed to explicitly account for incomplete lineage sorting (ILS) with ASTRAL using either the individual gene sequences (CDS gene trees) or the individual amino acid sequences (amino acid gene trees), also resulted in the same topology ([supplementary figs. S4 and S5, Supplementary Material](#) online). The normalized quartet score (proportion of input gene tree quartet trees in agreement with the species tree) was 0.78 for CDS gene trees and 0.64 for amino acid gene trees.

### Comparative Genomics and Positive Selection Analyses

To identify genes associated with adaptation to a pelagic lifestyle in the Procellariiformes, we performed a positive

selection analysis across 12 species including eight Procellariiformes species applying the HyPhy aBSREL model. We identified the hallmark of positive selection in 20 out of the 6,172 single-copy orthologs genes, after correcting for multiple testing ([supplementary table S5, Supplementary Material](#) online). We found among these genes enriched GO functions related with striated muscle cell differentiation, nutrient reservoir activity, response to starvation, visual learning, positive regulation of neural retina development, olfactory receptor activity, or natriuresis ([supplementary table S6, Supplementary Material](#) online). We also performed an HyPhy RELAX analysis to assess the intensification/relaxation of natural selection in Procellariiformes, which uncovered a total of 310 genes ([supplementary table S7, Supplementary Material](#) online). The GO terms enriched in these genes include wound healing, response to wounding, inflammatory response, sensory perception of sound, smell and chemical stimulus, neurological system process, defense response, response to stress, camera-type eye development, renal system, and chloride transport among others ([supplementary table S8, Supplementary Material](#) online).

Using OrthoFinder 2.3.8, we identified 182,487 N:N orthogroups across all genes identified in the 12 analyzed genomes. These data, together with the estimated ultrametric tree, were used to estimate gene gains, losses, and number of genes in the ancestral nodes using BadiRate; for the analysis we selected the Free Rates (FR) model, because it was the best-fitted branch model. The analysis was conducted including all orthogroups, and the minimum number of gains and losses per branch is represented



**FIG. 2.**—Comparison of genome-wide heterozygosity among Procellariiformes. (a) Density plots showing the distribution of individual nucleotide diversity ( $\pi$ ) values in nonoverlapping 25 Kb windows for each of the eight Procellariiformes species with an available reference genome. Scientific names of large-bodied (>450 g) and small-bodied (<200 g) species are shown in green and orange, respectively. Color-scale represents  $\pi$  values tail probabilities as shown in the legend. The white line depicts median values and black lines depict 25th and 75th percentiles. (b) Density plots showing the distribution of  $\pi$  values in large-bodied and small-bodied species groups.

in figure 3. Our results showed a tendency to gain genes in Procellariiformes (+442/−34), whereas the branch leading to albatrosses (Diomedidae) showed an opposite effect, with a noticeable loss of genes (+464/−3258); the branch leading to the rest of the Procellariiformes (+379/−15) is in the line of the general behavior of the tubenoses (supplementary table S9, Supplementary Material online). Within the order, families Oceanitidae and Hydrobatidae present the same trend, with the branch leading to *Hydrobates tethys* presenting a stronger gene loss balance (+325/−966) than the branch leading to the ancestor of Oceanitidae (*Oceanites oceanicus* and *Fregatta grallaria* [+182/−414]).

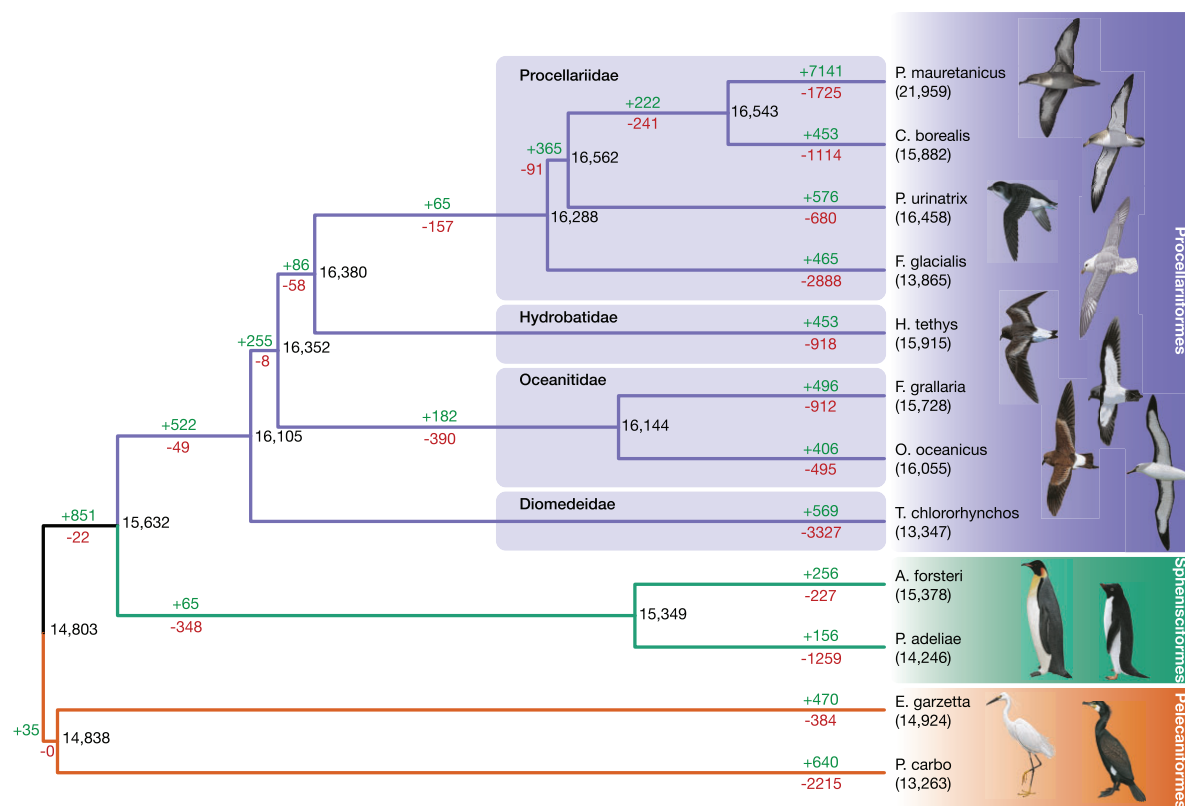
We identified three gene families significantly expanded in the branch leading to the Procellariiformes (supplementary table S9, Supplementary Material online). These families encode zinc finger proteins (OG0000000), olfactory receptors (OG0000084), and avian histones (OG0000224).

## Discussion

### A High-Quality Genome Assembly for the Most Endangered Seabird in Europe

The assembly length and the GC content of the Balearic shearwater hybrid assembly presented here are similar to





**Fig. 3.**—Ultrametric tree based on the 4D CDS ML tree calibrated with r8s. Minimum number of gains (green) and losses (red) per branch are represented according to BadiRate analysis. Numbers in ancestral nodes and in the tips (in parenthesis) indicate the inferred number of genes. Illustrations of seabird species were reproduced with permission from Lynx Edicions and Martí Franch.

those reported in the seven Procellariiformes genomes released by the Bird 10,000 Genomes Project (B10K) (Feng et al. 2020). Repetitive content is remarkably higher (+33.4%) in the Balearic shearwater in comparison to the other genomes of the order, but within the range of avian genomes (Zhang et al. 2014). This up to a third increase in repetitive sequence can be due to the fact that we included a Procellariiform (*Calonectris borealis*) repeat library before running RepeatMasker, achieving a more precise library that encloses clade related repeats that are present in the genome but not found by the *de novo* RepeatModeler library. The genome assembly completeness (BUSCO 95.9%) is slightly higher than the obtained for other recently published bird genomes (Prost et al. 2019; Feng et al. 2020), and even higher than genome assemblies including optical mapping (Peñalba et al. 2020). Despite not being a chromosome-scale assembly, contiguity is also quite high (N50 2.1 Mbp), and higher than recent avian MaSuRCA hybrid assemblies (Gan et al. 2019; Leroy et al. 2019).

The retrieved proteome (21,959 protein-coding genes) is similar to previous genomes (Liu et al. 2021; Recuerda et al. 2021), but higher than the B10K 2020 genomes used in the comparative studies in this work (mean of 16K). This is probably due to the B10K annotation pipeline being fully

based on homology, while we also used *de novo* prediction. The functional annotation quality in terms of genes having at least a GO term (85.9%) is comparable to recent chromosome-scale genomes (Recuerda et al. 2021).

The mitogenome of *P. mauretanicus* spans 19,885 bp, exhibiting the same order and the *nad6* gene duplication observed in *P. Iherminieri* (Torres et al. 2019). We did not find any *cob* duplication as it occurs in the Diomedelidae family (Abbott et al. 2005). Our result supports the hypothesis that *nad6* duplication could be widespread in Procellariiformes (Torres et al. 2019), and, like *cob*, could have undergone various events of deletion or addition during the diversification of the order. Nevertheless, because some of the reported duplications could be artificial (Urantówka et al. 2020; Formenti et al. 2021), to fully identify the true number of gene duplications/deletions will require additional and specific experimental analyses.

### Balearic Shearwater Heterozygosity Levels and Historical Demography

Current levels of intraspecific heterozygosity is a relevant parameter to determine the adaptive capacity of a population (or species) (Ørsted et al. 2019). Because the Balearic

shearwater is categorized as Critically Endangered by the IUCN, we could naively expect low heterozygosity levels in the species when compared with other Procellariiformes. However, the fossil record suggests that the Balearic shearwater had a very large population (>30,000 pairs) until the arrival of human settlers in the Balearic Islands (Alcover et al. 1991), which hunted shearwaters (Ramis 2018) and introduced invasive mammals that also preyed on them (Pinya and Carretero 2011). In line with Genovart et al. (2007) results using mtDNA markers, we observed relatively high genome-wide heterozygosity levels, suggesting that the very recent demographic decline in the species is not yet visible in its genetic diversity.

Regarding the historical demography, our MSMC2 analysis shows an increase in  $N_e$  in the Balearic shearwater from around 1 Ma to later expand to reach high population sizes, until around 150,000 ya, when it suddenly suffered a sharp decline, resulting in lower  $N_e$  values maintained until 10,000 years ago. Because current MSMC2 analysis is based on the analysis of a single genome, we could not reliably infer more recent events (Schiffels and Wang 2020). The Plio-Pleistocene eustatic variations resulted in a loss of neritic zones as sea level regressed (Pimiento et al. 2017), this may represent a loss of coastal habitat availability, which added to other oceanographic alterations (changes in ocean circulation or productivity) may have been the drivers of great population losses in marine megafauna, including seabirds. In the case of the Balearic shearwater, the MSMC2 analysis shows an abrupt decay of  $N_e$  associated with a long period of low sea level during the Penultimate Glacial Period (~194–135 ka) which may have resulted in an important loss of neritic zones. However, the particular timing of the sharp  $N_e$  decline should be interpreted with caution as it depends on both the mutation rate and generation time estimates; therefore, underestimations or overestimations of these parameters would result in biased timings (Nadachowska-Brzyska et al. 2015).

In view of the current critical population declines affecting the Balearic shearwater populations, understanding its impacts on current genetic diversity of the species and among colonies will be crucial to assess the conservation status of the Balearic shearwater. Future ongoing research, using a more powerful population genomics approach, will allow the reconstruction of more recent demographic histories of the species and to test the fossil-based hypotheses of a recent loss of population due to human colonization of the island, as well as why heterozygosity values have not decayed.

#### Association Between Body Size and Heterozygosity in Procellariiformes

We observed a strong association between body size and heterozygosity in the Procellariiformes. Small-bodied

species (*O. oceanicus*, *F. grallaria*, *H. tethys*, and *P. urinatrix*) have higher heterozygosities and higher variance in heterozygosity levels across the genome than large-bodied species (*Fulmarus glacialis*, *C. borealis*, *P. mauretanicus* and *T. chlororhynchos*). This pattern is in line with the negative correlations reported between heterozygosity and body size across animals from many taxa (Romiguier et al. 2014; Brüniche-Olsen et al. 2018, 2021). Such well-supported correlations are likely due to the interplay between  $N_e$  and  $\mu$ , both of which are correlated with body size (Romiguier et al. 2014; Brüniche-Olsen et al. 2018). Smaller animals generally live at higher population densities, which tend to be associated with higher  $N_e$ . Indeed, the small-bodied species included here have a higher number of breeding pairs (200,000–7,000,000) than large-bodied species (3,193–400,000) (Billerman et al. 2020). In addition, the small Procellariiformes lineages included here, also show higher  $\mu$  (Nunn and Stanley 1998; Estandía et al. 2021), showing the likely interplay of  $N_e$  and  $\mu$  at driving the association between body mass and heterozygosity. Island species, such as most Procellariiformes, tend to show lower levels of heterozygosity and increased numbers of deleterious mutations, probably due to a lower ability of natural selection to efficiently remove weakly deleterious mutations in small population (Leroy et al. 2021). Because standing genetic variation is the main source for rapid adaptation (Barrett and Schluter 2008; Jamie and Meier 2020), species with low levels of heterozygosity might not be able to rapidly adapt to environmental changes. Due to their low levels of heterozygosity, large-bodied Procellariiformes may be particularly vulnerable to environmental change. This could represent an additional threat for these species that are already suffering from anthropogenic threats such as predation by invasive alien species and fisheries bycatch (Dias et al. 2019). Future studies should investigate the accumulation of deleterious mutations in Procellariiformes to understand the potential responses of procellariiform species to environmental changes.

#### Adaptation to a Pelagic Lifestyle in Procellariiformes

Our selection inference uncovered 20 genes evolving under positive selection in Procellariiformes, and are therefore potential candidates involved in the adaptation of the order to a pelagic lifestyle. Indeed, the function of these genes reveals biological processes related to striated muscle cell differentiation, response to starvation, and nutrient reservoir activity, that may be related to the high energy expenditure during the vast distances they cover in the open ocean, whereas visual related genes could be related with underwater vision to fish and night vision (Hayes et al. 1991; Martin and De 1991; Mitkus et al. 2016). Positive selection of genes related to natriuresis also makes sense for

Procellariiformes because this biological process plays a key role to maintain the osmotic equilibrium in a sodium-rich environment like the ocean (Goldstein 2001; Gutiérrez 2014), which Procellariiformes perform thanks to the development of salt glands (modified nasal glands engaged in secretion of salts). Olfactory receptors, also found here among enriched GOs of positively selected genes, showed signature of adaptive evolution in shearwaters (Silva et al. 2020), and are crucial to Procellariiformes for navigation (Gagliardo et al. 2013; Pollonara et al. 2015; Padget et al. 2017), partner recognition and mating (Bonadonna and Nevitt 2004; Strandh et al. 2012; Hoover et al. 2018), finding their own burrows (Bonadonna and Bretagnolle 2002), or foraging (Nevitt et al. 1995; Nevitt 2008; Bastos et al. 2020; Sin et al. 2022).

We also identified genes with intensified natural selection in Procellariiformes, for which the GOs annotations (supplementary table S8, Supplementary Material online) are similar and coherent to those in the candidate set of genes with positive selection in all tubenoses, or, in other words, related to the adaptation of the order to a pelagic lifestyle. For example, molecular functions such as sensory perception of sound, smell and chemical stimulus, neurological system process, camera-type eye development are related with oceanic navigation. On the other hand, functions such as homeostatic process, renal system, renal response, chloride transport, and regulation of ion transport point to the need of maintaining osmotic equilibrium. We also found intensified natural selection in genes participating in functions related to immune response (like inflammatory response, defense response, response to wounding, wound healing, positive regulation of phagocytosis, etc.), accompanied by a relaxation of natural selection in regulators of blood constituents, induction of bacterial agglutination, regulation of antigen processing and presentation, viral budding via host ESCRT complex, or macrophage antigen processing and presentation. As exposure to parasites in Procellariiformes is high (Khan et al. 2019) and their life-history traits favor parasite maintenance within populations (McCoy et al. 2016), we hypothesize that the tuning between the intensification and relaxation of natural selection in multiple biological processes and molecular functions related with immune response could have emerged following an arms race-like model. For example, as many parasites of tubenoses are blood-feeding, the intensified natural selection on the thrombin-activated receptor signaling pathway (GO:0070493), could be an evolutionary response to counter the anticoagulant activity that most blood-feeding parasites present (Bensaoud et al. 2018).

Among the gene families expanded in the branch of Procellariiformes, the one encoding olfactory receptors is remarkable as it is coherent not only with the finding of a gene with positive selection in all Procellariiformes with

the same functional annotation (g16276.t1 in the *P. mauretanicus* reference annotation, supplementary table S4, Supplementary Material online) but also with three olfactory receptors genes with intensification selection in the same branch (g14377.t1, g16276.t1, and g17936.t1 in *P. mauretanicus* reference annotation, supplementary table S8, Supplementary Material online). This triple evidence highlights the importance of how the adaptation to a pelagic life resulted in the enhancement of the olfactory function in Procellariiformes, as discussed above. Moreover, similar results of positive selection in olfactory genes were obtained in *C. borealis* (Silva et al. 2020). Physiologically, tubenoses have one of the largest olfactory bulb to brain size ratio of all birds (Cobb 1968).

## Conclusions

Our study highlights the utility of the hybrid assembly strategy using Illumina and ONT at recovering high-quality genome assemblies, especially regarding contiguity and completeness. Comparative genomics analyses identified candidate genes under selection to have played a major role in the adaptation of the Procellariiformes to a pelagic lifestyle such as changes in sensory perception, navigation, natriuresis, and physiological adaptations. The high-quality genome presented in this work will be a great tool for future population genomic analyses, that will reveal with more precision the genetic variability of the species, its recent demographic history and the potential introgression with its sister species, the Mediterranean shearwater (*P. yelkouan*). The data obtained will be of great help in future conservation and management plans for the species.

## Materials and Methods

### Sampling, DNA and RNA Extraction, and Sequencing

We sampled two Balearic shearwater adults and one chick. Adults were sampled on Sa Cella colony, Mallorca (male) and on Sa Conillera, Eivissa (unsexed) in 2004, whereas the chick was sampled on Conills islet (Mallorca) in July 2019. From here on the animals will be referred to as male-MII, unsexed-Ei, and chick-MII, respectively. Special permits to obtain the samples were issued by Conselleria de Medi Ambient, Agricultura i Pesca (Govern de les Illes Balears, Spain).

We extracted DNA from blood samples preserved in absolute ethanol for both adults. The DNA extraction for the male-MII was performed with DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's instructions, and with Blood & Cell Culture DNA Mini Kit (Qiagen) for the unsexed-Ei. RNA was extracted from the chick-MII's blood cells preserved in RNAlater 1:5 using the RNeasy Mini Kit (Qiagen) according to the manufacturer's protocols. We performed the quality control with gel electrophoresis



and NanoDrop One (Thermo Fisher Scientific, Waltham, MA, USA), and the quantification with an Invitrogen Qubit Fluorometer 2.0 (Broad Range kit).

We obtained the reference genome combining short-read and long-read sequencing libraries, and using RNA-seq data to assist with the annotation. First, an Illumina TruSeq DNA PCR Free library (insert size = 350 bp) was prepared by Macrogen (South Korea) using DNA from male-MII, and sequenced using two HiSeq X Ten runs ( $2 \times 150$  bp). Second, long-read libraries were prepared, from the DNA of unsexed-Ei, using the Ligation kit SQK-LSK109 1D from ONT (Oxford Nanopore Technologies) (N50 of 9431 bp) at Centro Nacional de Análisis Genómico, Spain and sequenced through five runs of MinION on FLO-MIN106 flow cells. Third, we prepared RNA-seq libraries from the chick-MII's RNA using the TruSeq RNA Sample Prep Kit v2 with Ribo-Zero, and we sequenced the libraries on a NovaSeq 6000 ( $2 \times 100$  bp) (Macrogen, South Korea).

### Genome Assembly

We performed a *de novo* hybrid genome assembly with MaSuRCA 3.3.1 (Zimin et al. 2017), using short (Illumina) and long (ONT) reads. Before the assembly step, we filtered the ONT reads with a Phred quality score ( $Q \geq 5$ ) using the NanoFilter software included in NanoPack (De Coster et al. 2018). Paired-end Illumina reads were parsed into MaSuRCA without any preprocessing, as adapters and errors are handled by the QuORUM error corrector (Marçais et al. 2015), which is part of the MaSuRCA pipeline. MaSuRCA was run applying the following parameters: fragment mean (422), fragment stdev (312), and estimated genome size (1.2 Gbp). The resulting assembly was screened for contaminants with BlobTools v1.0 (Laetsch and Blaxter 2017) `-x bestsumorder`. Assembly completeness was assessed with BUSCO 4.0.2 (Seppey et al. 2019) using the 8,338 single-copy conserved genes in *aves\_odb10* database (Kriventseva et al. 2019).

### Transcriptome Assembly

We trimmed RNA-seq raw reads for adapters with BBDuk (<https://sourceforge.net/projects/bbmap/>) ( $k = 17$ , `tpe` option), and used STAR 2.7.3a (Dobin et al. 2013) to map the filtered reads to the newly assembled reference genome. We obtained the transcriptome assembly with Trinity 2.8.6 (Grabherr et al. 2011) using the genome-guided `bam` mode (`-genome_guided_max_intron 82945`). Transcripts were clustered with CD-HIT (Li and Godzik 2006; Fu et al. 2012) 4.8.1 (`-c 0.98`) and CDS were predicted with TransDecoder 5.5.0 (<https://github.com/TransDecoder/>).

### Mitogenome Assembly

We trimmed adapters from Illumina raw reads with BBDuk ( $k = 23$ , `tpe` option), before using them as input to NOVOPlasty 2.7.2 (Dierckxsens et al. 2017). The *P. Iherminieri* mitogenome (MH206163.1) was used as seed using the following parameters: Genome Range (16,000–24,000), Insert size (422), Insert range (1.74) and Insert range strict (1.3). The annotation was performed using the MITOS WebServer (Bernt et al. 2013).

### Repeat Annotation

We generated a *de novo* repeat library of the genome with RepeatModeler—1.0.11 (Smit et al. 2021) on scaffolds  $> 100$  kbp. This library was combined with all avian and ancestral consensus repeats from Dfam\_Consensus-20181026 (Storer et al. 2021), RepBase-20181026 (Jurka et al. 2005), and the repeat annotation of the Cory's shearwater (*C. borealis*) (Feng et al. 2020), which represents the most closely related sequenced genome. Redundancies among libraries were removed with the script ReannTE\_MergeFasta.pl (<https://github.com/4ureliek/ReannTE>). We then ran RepeatMasker 4.0.7 (Smit et al. 2021) using the combined library as a reference, with the following parameters: `-xsmall -e ncbi -s -gccalc -no_is -gff`.

### Structural and Functional Annotation

We performed the structural annotation with BRAKER 2.1.2 (<https://github.com/Gaius-Augustus/BRAKER>) (`-etp-mode`) using data from both the Cory's shearwater proteome (Feng et al. 2020), and the RNA-Seq data generated in this work. Because the inclusion of RNA-Seq data appeared detrimental, we excluded this piece of information to perform the final annotation using the soft-masked genome with BRAKER 2.1.2 (`-prg = gth -trainFromGth`).

We made the functional annotation of the predicted genes using a similarity-based approach. We determined the protein domains with InterProScan 5.31–70.0 (Jones et al. 2014), used BLASTP (Altschul et al. 1990; Camacho et al. 2009) (`-evalue 1e-5; -max_target_seqs 10`) against the Swiss-Prot database (Boutet et al. 2016) and the Cory's shearwater and the Zebra finch reference (UP000007754) proteomes. Transcripts were annotated in the same manner. We also annotated the ncRNAs using cmscan from INFERNAL 1.1.2 (Nawrocki et al. 2009) with the covariance models (CMs) from the Rfam 14.1 database, and tRNA genes using tRNAscan-SE 2.0.5 (Chan and Lowe 2019).

### Demographic History

We used the Multiple Sequentially Markovian Coalescent 2 (MSMC2) (Schiffels and Wang 2020) to infer the historical demography of the Balearic shearwater. MSMC2

implements a MSMC model, which allows the estimation of the effective population size ( $N_e$ ) over time. To generate input files for MSMC2, first we mapped cleaned Illumina short reads from male-MII to the reference genome using BWA-MEM 0.7.17 (Li and Durbin 2009). We then subsampled the bam file to only include scaffolds larger than 1 Mbp (343 scaffolds spanning 71.8% of the assembled genome), as recommended in Gower et al. (2018). Second, we called SNPs using samtools mpileup 1.9 (options: -q 20 -Q 20 -C 50) and then bcftools call 1.9 (options -c -V indels) (Danecek et al. 2021) and, for each scaffold, we generated a VCF file and a mask file in bed format, containing the regions on the scaffold that were sufficiently covered using the bamCaller.py script provided in the msmc-tools package (<https://github.com/stschiff/msmc-tools>). Third, we generated a mappability mask using GenMap (Pockrandt et al. 2020) with a k-mer size of 150 bp and allowing for up to two mismatches. We masked sites with a mappability score of  $<0.5$  and we also masked sites within annotations of repetitive regions. In total, we masked 9.37% of the sites in the 343 scaffolds larger than 1 Mbp. Fourth, we generated input files for MSMC2 with the script generate\_multihetsep.py, also provided in the msmc-tools package, by using the VCF and mask files as input files. Fifth, multiple sequentially markovian coalescent (MSMC) for two haplotypes was run with MSMC2 with time patterning specified as -p 1 \* 4 + 30 \* 2 + 1 \* 4 + 1 \* 6 + 1 \* 10. Additional time patterning settings were used to assess the effect of the number of segments on the shape of the MSMC2 curves, but this had little effect on the shape.

We obtained the confidence intervals for  $N_e$  estimates by running 100 bootstraps using multihetsep\_bootstrap.py. We scaled time and  $N_e$  using a generation time for the Balearic shearwater of 12.8 years (Genovart et al. 2016) and the Northern fulmar (*F. glacialis*) mutation rate ( $\mu = 2.89 \times 10^{-9}$  substitutions per nucleotide per generation, Nadachowska-Brzyska et al. 2015).

### Genome-Wide Heterozygosity

We estimated genome-wide heterozygosities using information of a single individual from all eight Procellariiformes species studied. We applied the Robinson et al. (2019) method, with minor modifications to take genome fragmentation into consideration, because we included genome assemblies with varying amounts of contiguity. The DNA sequence data (genome assemblies and whole-genome sequencing data) were downloaded from NCBI (PRJNA261828, PRJNA545868; Jarvis et al. 2014; Feng et al. 2020). For each species, adapter-trimmed reads were aligned to its genome assembly using BWA-MEM (Li 2013), bam files were merged using Picard-Tools (<http://broadinstitute.github.io/picard/>) and

variants were called using the GATK 4.1.9 HaplotypeCaller and GenotypeGVCFs (McKenna et al. 2010). Sites with a coverage  $<1/3 \times$  or  $>2 \times$  of the average coverage depth (of the particular genome) were filtered out using VCFtools 0.1.15 (Danecek et al. 2011). We computed per-site heterozygosity as the proportion of heterozygous sites per total number of called genotypes within a single individual in nonoverlapping 25 Kb windows across each scaffold. Windows with  $<50\%$  of net sites (those excluding missing or filtered sites), were excluded from the analysis.

### Orthology Inference

We performed the phylogenomics and comparative genomics analyses including information from 12 species with an available genome assembly: eight Procellariiformes (*P. mauretanicus*, *T. chlororhynchos*, *H. tethys*, *O. oceanicus*, *F. grallaria*, *P. urinatrix*, *F. glacialis*, and *C. borealis*), and four outgroups, *Aptenodytes forsteri*, *Pygoscelis adeliae* (Sphenisciformes); *Egretta garzetta*, *Phalacrocorax carbo* (Pelecaniformes). We inferred orthologous genes across the proteomes of these 12 species using OrthoFinder 2.3.8 (Emms and Kelly 2019) with default parameters.

### Inference of Phylogenetic Relationships

In order to perform selection and gene family evolution analyses, we performed phylogenetic analyses using the inferred orthologous genes. Despite the Procellariiformes phylogeny has been recently resolved using UCE data (Estandía et al. 2021), we performed phylogenetic analyses to evaluate the robustness of this phylogeny when using different phylogenomic datasets. We built a multiple sequence alignment (MSA) for each 1:1 orthologs with PRANK v.100802 (Löytynoja 2014), using both CDS (-codon -noxml -notree -F) and amino acid sequences (-noxml -notree -F). Individual alignments were concatenated with catfasta2phym1 v1.1.0 (<https://github.com/nylander/catfasta2phym1>) to create a CDS supermatrix and an amino acid supermatrix. Only locus with data for all the 12 species has been considered. Fourfold degenerate sites (4D) for the CDS supermatrix were extracted with MEGA X (Kumar et al. 2018). We performed unpartitioned maximum likelihood (ML) phylogenetic analyses using IQ-TREE 1.6.12 (Nguyen et al. 2015) (-bb 1000) both for 4D and amino acid supermatrices. Optimal models of sequence evolution were obtained with ModelFinder (Kalyaanamoorthy et al. 2017) according to Bayesian information criterion), and the resulting best-fit models were GTR + F+R2 for 4D and HIVb + F+R3 for amino acid supermatrix. Node support was obtained with Ultrafast Bootstrap (Hoang et al. 2018).

To explicitly account for ILS under the Multispecies Coalescent Model, we inferred the species tree using the

summary coalescent approach, as implemented in ASTRAL-III 5.6.3 (Zhang et al. 2018). We first obtained all gene trees (for each 1:1 orthologous genes) using IQ-TREE 1.6.12, and inferred the species tree and its normalized score (from both CDS and amino acid gene trees) using ASTRAL-III.

We generated an ultrametric tree with r8s v.1.81 (Sanderson 2003) using the 4D supermatrix ML tree. We used four calibration points (in myr): root (max\_age = 84 min\_age = 73, Braun et al. 2011), most recent common ancestor (MRCA) of Spheniscidae (min\_age = 12.6, Subramanian et al. 2013), MRCA Procellariiformes (min\_age = 49, Claramunt and Cracraft 2015), MRCA Procellariidae (min\_age = 14, Prum et al. 2015), retrieved from TimeTree (Kumar et al. 2017). We used the penalized likelihood method and the Truncated Newton algorithm, and the smoothing parameter was set to 100.

### Positive Selection Analysis

We evaluated the selective constraints of genes that could be associated with pelagic lifestyle. For this purpose, we performed the analysis with HyPhy 2.5 (Kosakovsky Pond et al. 2005), using Procellariiformes data (1:1 MSAs). Before the analysis, nonreliable positions across all 1:1 orthologs MSAs were filtered with ZORRO (Wu et al. 2012) (default options; MSA with average quality < 5 were filtered). We used the aBSREL method (Smith et al. 2015) to test for positive selection, and the RELAX method (Wertheim et al. 2015) to test for relaxed/intensified selection. We also performed a GO enrichment analysis of the candidate genes using the GOstats (Falcon and Gentleman 2007) R package against the background GOs of 1:1 orthologs.

### Gene Family Evolution

We estimated gene turnover rates, the number of gene gains and losses across the phylogeny lineages, and inferred gene family contractions and expansions using BadiRate 1.7 (Librado et al. 2012). For the analysis we first inferred the orthogroups with OrthoFinder 2.3.8, and we used the calibrated ultrametric tree estimated with r8s. We tested, under the Birth-Death-Innovation model for turnover rates, several biological relevant hypotheses with three different branch models: FR, global rates, and branch-specific rates, and chose the best model based on the lowest Akaike information criterion value. To ensure an appropriate convergence we ran multiple times each model.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

We are grateful to David García and Maite Louzao for kindly providing samples and the Govern Illes Balears for research permits (CEP19/2019). We would like to thank Josephine R. Paris for advice on analyses and for her comments on an early version of this manuscript and Martí Franch for his wonderful shearwater illustrations. This research was supported by Fundación Banco Bilbao Vizcaya (Spain), Project 062-17; by the Ministerio de Economía y Competitividad of Spain, projects CGL2016-78530-R, PGC2018-093924-B-100, and PID2019-103947GB.

## Author Contributions

J.F.O., J.G.-S., M.R., and J.R. conceived the study. C.C.-C., J.F.O., M.G., and J.G.-S. performed samplings. C.C.-C., J.F.O. performed wet lab work. C.C.-C., J.F.O., and J.V. performed the bioinformatic analyses. C.C.-C., J.F.O., M.R., and J.R. interpreted the genomic data. M.G., J.G.-S. discussed on biological data. C.C.-C., J.F.O., M.R., and J.R. drafted the first version of the manuscript. All authors revised and approved the final version of the manuscript.

## Data Availability

The whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the Bioproject PRJNA780920, and BioSamples SAMN24039388, SAMN23492024, and SAMN23212142. The raw reads are also available in the Sequence Read Archive (SRA) under the Bioproject accession. Other relevant datasets, such as those including the structural and functional annotations, are available in [https://github.com/molevol-ub/Puffinus\\_mauretanicus\\_genome](https://github.com/molevol-ub/Puffinus_mauretanicus_genome), and in the [Supplementary Material](#) online.

## Literature Cited

- Abbott CL, Double MC, Trueman JWH, Robinson A, Cockburn A. 2005. An unusual source of apparent mitochondrial heteroplasmy: duplicate mitochondrial control regions in *Thalassarche albatrosses*. *Mol Ecol*. 14:3605–3613.
- Alcover JA, Bover P, Seguí B. 1991. Paleoeecologia de les illes. In: Alcover JA, editor. *Ecologia de les Illes*, Monografies de la Societat d'Història Natural de les Balears. p. 169–204.
- Allendorf FW. 2017. Genetics and the conservation of natural populations: allozymes to genomes. *Mol Ecol*. 26:420–430.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.
- Arcos JM (compiler) 2011. International species action plan for the Balearic shearwater, *Puffinus mauretanicus*. SEO/BirdLife & BirdLife International.
- Arcos JM et al. 2012. New estimates at sea suggest a larger global population of the Balearic Shearwater *Puffinus mauretanicus*. In: *Proceedings of the 13th Medmaravis Pan-Mediterranean Symposium*.

- Arcos JM, Louzao M, Oro D. 2008. Fisheries ecosystem impacts and management in the Mediterranean: seabirds point of view. *Am. Fish. Soc. Symp.* 587–596.
- Arroyo GM, et al. 2016. New population estimates of a critically endangered species, the Balearic Shearwater *Puffinus mauretanicus*, based on coastal migration counts. *Bird Conserv Int.* 26:87–99.
- Augustin L, et al. 2004. Eight glacial cycles from an Antarctic ice core. *Nature* 429:623–628.
- Austin RE, et al. 2019. Patterns of at-sea behaviour at a hybrid zone between two threatened seabirds. *Sci Rep.* 9:14720.
- Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. *Trends Ecol Evol.* 23(1):38–44.
- Bastos R, et al. 2020. Oceans of stimuli: an individual-based model to assess the role of olfactory cues and local enhancement in seabirds' foraging behaviour. *Anim Cogn.* 23(4):629–642.
- Bensaoud C, et al. 2018. De novo assembly and annotation of *Hyalomma dromedarii* tick (Acari: Ixodidae) sialotranscriptome with regard to gender differences in gene expression. *Parasites Vectors* 11(1):1–16.
- Bernt M, et al. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol.* 69:313–319.
- Billerman SM, Keeney BK, Rodewald PG, Schulenberg TS. 2020. *Birds of the world*. Ithaca (NY): Cornell Laboratory of Ornithology. <https://birdsoftheworld.org/bow/home>.
- Bonadonna F, Bretagnolle V. 2002. Smelling home: a good solution for burrow-finding in nocturnal petrels? *J Exp Biol.* 205:2519–2523.
- Bonadonna F, Nevitt GA. 2004. Partner-specific odor recognition in an Antarctic seabird. *Science* 306:835.
- Boutet E, et al. 2016. Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: how to use the entry view. In: *Methods in molecular biology*. Vol. 1374. Humana Press Inc. p. 23–54.
- Braun EL, et al. 2011. Homoplastic microinversions and the avian tree of life. *BMC Evol Biol.* 11:141.
- Brooke M. 2004. *Albatrosses and petrels across the world*. Oxford University Press.
- Brüniche-Olsen A, Kellner KF, Anderson CJ, DeWoody JA. 2018. Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conserv Genet.* 19(6):1295–1307.
- Brüniche-Olsen A, Kellner KF, Belant JL, DeWoody JA. 2021. Life-history traits and habitat availability shape genomic diversity in birds: implications for conservation. *Proc R Soc B.* 288: 20211441.
- Camacho C, et al. 2009. BLAST+: architecture and applications. *BMC Bioinform.* 10:1–9.
- Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit – interactive quality assessment of genome assemblies. *G3 Genes|Genomes|Genetics* 10:1361–1374.
- Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. In: *Methods in molecular biology*. Vol. 1962. Humana Press Inc. p. 1–14.
- Claramunt S, Cracraft J. 2015. Evolutionary ecology: a new time tree reveals Earth history's imprint on the evolution of modern birds. *Sci Adv.* 1:e1501005.
- Cobb S. 1968. The size of the olfactory Nulb in 108 species of Birds. *Auk* 85:55–61.
- Danecek P, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.
- Danecek P, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10(2):giab008.
- De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34:2666–2669.
- Dias MP, et al. 2019. Threats to seabirds: a global assessment. *Biol Conserv.* 237:525–537.
- Dierckxens N, Mardulyn P, Smits G. 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18.
- Dobin A, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.
- Dussex N, et al. 2021. Population genomics of the critically endangered kāāō. *Cell Genomics* 1:100002.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:1–14.
- Estandía A, et al. 2021. Substitution rate variation in a robust Procellariiform seabird phylogeny is not solely explained by body mass, flight efficiency, population size or life history traits. *bioRxiv*. doi:10.1101/2021.07.27.453752.
- Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* 23:257–258.
- Feng S, et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature* 587:252–257.
- Footo AD, et al. 2015. Convergent evolution of the genomes of marine mammals. *Nat Genet.* 47(3):272–275.
- Formenti G, et al. 2021. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biol.* 22:1–22.
- Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150–3152.
- Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. 2012. Harnessing genomics for delineating conservation units. *Trends Ecol Evol.* 27: 489–496.
- Gagliardo A, et al. 2013. Oceanic navigation in Cory's shearwaters: evidence for a crucial role of olfactory cues for homing after displacement. *J Exp Biol.* 216:2798–2805.
- Gan HM, et al. 2019. Genomic evidence of neo-sex chromosomes in the eastern yellow robin. *Gigascience* 8:1–10.
- Genovart M, et al. 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.
- Genovart M, Oro D, Juste J, Bertorelle G. 2007. What genetics tell us about the conservation of the critically endangered Balearic Shearwater? *Biol Conserv.* 137:283–293.
- Goldstein DL. 2001. Water and salt balance in seabirds. In: *Biol Mar Birds*. CRC Press. p. 485–502.
- Gower G, et al. 2018. Population size history from short genomic scaffolds: how short is too short? *bioRxiv*. doi:10.1101/382036.
- Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29: 644–652.
- Gutiérrez JS. 2014. Living in environments with contrasting salinities: a review of physiological and behavioural responses in waterbirds. *Ardeola* 61:233–256.
- Hayes B, Martin GR, Brooke M de L. 1991. Novel area serving binocular vision in the retinae of procellariiform seabirds. *Brain Behav Evol.* 37:79–84.
- Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 35:518–522.



- Hoover B, et al. 2018. Ecology can inform genetics: disassortative mating contributes to MHC polymorphism in Leach's storm-petrels (*Oceanodroma leucorhoa*). *Mol Ecol*. 27:3371–3385.
- IUCN. 2021. The IUCN red list of threatened species. Version 2021-1. Available from: <https://www.iucnredlist.org>. Accessed July 2, 2021.
- Jamie GA, Meier JI. 2020. The persistence of polymorphisms across species radiations. *Trends Ecol Evol*. 35(9):795–808.
- Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* (80-). 346: 1320–1331.
- Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240.
- Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110:462–467.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14:587–589.
- Khan JS, et al. 2019. Parasites of seabirds: a survey of effects and ecological implications. *Adv Mar Biol*. 82:1–50.
- Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
- Kriventseva EV, et al. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*. 47:D807–D811.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 35:1547–1549.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 34: 1812–1819.
- Laetsch DR, Blaxter ML. 2017. BlobTools: interrogation of genome assemblies. *F1000Research* 6:1287.
- Leroy T, et al. 2019. A bird's white-eye view on avian sex chromosome evolution. *bioRxiv*. <https://doi.org/10.1101/505610>.
- Leroy T, et al. 2021. Island songbirds as windows into evolution in small populations. *Curr Biol*. 31:1303–1310.e4.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint*. doi:10.48550/arXiv.1303.3997.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
- Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* 28: 279–281.
- Lisiecki LE, Raymo ME. 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic  $\delta^{18}\text{O}$  records. *Paleoceanography* 20(1). doi: 10.1029/2004PA001071.
- Liu J, et al. 2021. A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes. *Genome Res*. 31:497–511.
- Louzao M, Arcos JM, Hyrenbach DK, de Sola LG, Oro D. 2004. Resultados preliminares sobre el hábitat de alimentación de la Pardela Balear en el Levante Ibérico Peninsular. *Anu Ornitològic les Balear Rev d'observació Estud. i Conserv dels aucells*. 61–67.
- Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. *Methods Mol Biol*. 1079:155–170.
- Marçais G, Yorke JA, Zimin A. 2015. Quorum: an error corrector for Illumina reads. *PLoS One* 10:e0130821.
- Martin GR, De M. 1991. The eye of a procellariiform seabird, the Manx shearwater, *Puffinus puffinus*: visual fields and optical structure. *Brain Behav Evol*. 37:65–78.
- McCoy KD, et al. 2016. The role of seabirds of the Iles Eparses as reservoirs and disseminators of parasites and pathogens. *Acta Oecologica (Montrouge, Fr.)* 72:98–109.
- McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20:1297–1303.
- Mitkus M, Nevitt GA, Danielsen J, Kelber A. 2016. Vision on the high seas: spatial resolution and optical sensitivity in two procellariiform seabirds with different foraging strategies. *J Exp Biol*. 219: 3329–3338.
- Nadachowska-Brzyska K, Li C, Smeds L, Zhang G, Ellegren H. 2015. Temporal dynamics of avian populations during pleistocene revealed by whole-genome sequences. *Curr Biol*. 25:1375–1380.
- Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25:1335–1337.
- Nevitt GA. 2008. Sensory ecology on the high seas: the odor world of the procellariiform seabirds. *J Exp Biol*. 211:1706–1713.
- Nevitt GA, Veit RR, Kareiva P. 1995. Dimethyl sulphide as a foraging cue for Antarctic Procellariiform seabirds. *Nature* 376(6542): 680–682.
- Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32:268–274.
- Nunn GB, Stanley SE. 1998. Body size effects and rates of cytochrome b evolution in tube-nosed seabirds. *Mol Biol Evol*. 15:1360–1371.
- Ørsted M, Hoffmann AA, Sverrisdóttir E, Nielsen KL, Kristensen TN. 2019. Genomic variation predicts adaptive evolutionary responses better than population bottleneck history. *PLoS Genet*. 15: e1008205.
- Padgett O, Dell'Arciccia G, Gagliardo A, González-Solis J, Guilford T. 2017. Anosmia impairs homing orientation but not foraging behaviour in free-ranging shearwaters. *Sci Rep*. 7(1):1–12.
- Peñalba JV, et al. 2020. Genome of an iconic Australian bird: high-quality assembly and linkage map of the superb fairy-wren (*Malurus cyaneus*). *Mol Ecol Resour*. 20:560–578.
- Pimiento C, et al. 2017. The Pliocene marine megafauna extinction and its impact on functional diversity. *Nat Ecol Evol*. 1(8): 1100–1106.
- Pinya S, Carretero MA. 2011. The Balearic herpetofauna: a species update and a review on the evidence. *Acta Herpetol*. 6:59–80.
- Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. 2020. GenMap: ultra-fast computation of genome mappability. *Bioinformatics* 36(12):3687–3692.
- Pollonara E, et al. 2015. Olfaction and topography, but not magnetic cues, control navigation in a pelagic seabird: displacements with shearwaters in the Mediterranean Sea. *Sci Rep*. 5(1):1–10.
- Prost S, et al. 2019. Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise. *Gigascience* 8:1–12.
- Prum RO, et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526(7574): 569–573.
- Ramis D. 2018. Animal exploitation in the early prehistory of the Balearic Islands. *J Isl Coast Archaeol*. 13:265–278.
- Recuerda M, et al. 2021. Chromosome-level genome assembly of the common Chaffinch (*Aves: Fringilla coelebs*): a valuable resource for evolutionary biology. *Genome Biol Evol*. 13:evab034.
- Robinson JA, et al. 2019. Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Sci Adv*. 5:757–786.



- Romiguier J, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515(7526): 261–263.
- Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.
- Schiffels S, Wang K. 2020. MSMC and MSMC2: the multiple sequentially Markovian coalescent. *Methods Mol Biol.* 2090:147–166.
- Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In: *Methods in molecular biology*. Vol. 1962. Humana Press Inc. p. 227–245.
- Silva M, et al. 2020. Signature of adaptive evolution in olfactory receptor genes in Cory's Shearwater supports molecular basis for smell in procellariiform seabirds. *Sci Rep.* 10(1):1–11.
- Sin SYW, Cloutier A, Nevitt G, Edwards SV. 2022. Olfactory receptor subgenome and expression in a highly olfactory procellariiform seabird. *Genetics.* 220(2):iyab210. doi:10.1093/genetics/iyab210.
- Smit AF, Hubley R, Green P. 2021. RepeatMasker. Available from: [www.repeatmasker.org](http://www.repeatmasker.org). Accessed January 9, 2021.
- Smith MD, et al. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol.* 32:1342–1353.
- Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* 12:1–14.
- Strandh M, et al. 2012. Major histocompatibility complex class II compatibility, but not class I, predicts mate choice in a bird with highly developed olfaction. *Proc R Soc B Biol Sci.* 279:4457–4463.
- Subramanian S, Beans-Picón G, Swaminathan SK, Millar CD, Lambert DM. 2013. Evidence for a recent origin of penguins. *Biol Lett.* 9: 20130748.
- Supple MA, Shapiro B. 2018. Conservation of biodiversity in the genomics era. *Genome Biol.* 19:1–12.
- Torres L, et al. 2019. Evidence for a duplicated mitochondrial region in Audubon's shearwater based on MinION sequencing. *Mitochondrial DNA Part A* 30:256–263.
- Urantówka AD, Krocak A, Mackiewicz P. 2020. New view on the organization and evolution of Palaeognathae mitogenomes poses the question on the ancestral gene rearrangement in Aves. *BMC Genomics* 21(1):1–25.
- Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol.* 32:820–832.
- Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS One* 7:e30288.
- Zhang G, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science (80-)* 346: 1311–1320.
- Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* 19:15–30.
- Zimin AV, et al. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* 27:787–792.

**Associate editor:** Cristina Vieira