Complex population structure of the Atlantic puffin revealed by whole genome analyses

Oliver Kersten 1✉, Bastiaan Star 1, Deborah M. Leigh 2, Tycho Anker-Nilssen 3, Hallvard Strøm 4, Jóhannis Danielsen 5, Sébastien Descamps 4, Kjell E. Erikstad 6,7, Michelle G. Fitzsimmons 8, Jérôme Fort 9, Erpur S. Hansen 10, Mike P. Harris 11, Martin Irestedt 12, Oddmund Kleven 3, Mark L. Mallory 13, Kjetill S. Jakobsen 1 & Sanne Boessenkool 1

The factors underlying gene flow and genomic population structure in vagile seabirds are notoriously difficult to understand due to their complex ecology with diverse dispersal barriers and extensive periods at sea. Yet, such understanding is vital for conservation management of seabirds that are globally declining at alarming rates. Here, we elucidate the population structure of the Atlantic puffin (Fratercula arctica) by assembling its reference genome and analyzing genome-wide resequencing data of 72 individuals from 12 colonies. We identify four large, genetically distinct clusters, observe isolation-by-distance between colonies within these clusters, and obtain evidence for a secondary contact zone. These observations disagree with the current taxonomy, and show that a complex set of contemporary biotic factors impede gene flow over different spatial scales. Our results highlight the power of whole genome data to reveal unexpected population structure in vagile marine seabirds and its value for seabird taxonomy, evolution and conservation.
Seabirds are important ecosystem indicators and drivers\textsuperscript{1–3}, and have long had an integral place in human culture and economy\textsuperscript{4–6}. Nevertheless, global seabird numbers have deteriorated by an alarming 70\% since the mid-20th century\textsuperscript{7,8}. These declines pose a serious threat to marine ecosystems, human society, and culture\textsuperscript{9,10,11,12}, highlighting the importance of seabird conservation management. Within such management, the identification of distinct population units, i.e., demographically independent populations with restricted gene flow among them\textsuperscript{11,12}, is a fundamental first step towards optimized conservation\textsuperscript{11,13,14}. Defining such units is, however, difficult for many seabirds because of their complex ecology\textsuperscript{15}. Detailed genomic data including thousands of loci provide new possibilities to assess levels of connectivity and gene flow between distinct breeding populations and, thus, help identify relevant conservation units for seabirds\textsuperscript{15,16}. Indeed, a few recent publications using reduced genomic representation approaches (e.g., RAD-seq) have reported fine-scale structure over various spatial scales\textsuperscript{17–21}. These studies highlight the great potential of genomic data to disentangle barriers to gene flow that would otherwise remain undetected, but remain nonetheless limited due to incomplete sampling of the genome\textsuperscript{22}.

The Atlantic puffin (Fratercula arctica, Linnaeus, 1789, hereafter “puffin”) is an iconic seabird species, prevalent in popular culture\textsuperscript{23} important for tourism\textsuperscript{24,25} and inherently valuable for conservation\textsuperscript{23,16,17}. Puffins are especially valuable for tourism, as they provide a unique opportunity to observe seabirds in their natural habitat. Indeed, many seabirds because of their complex ecology\textsuperscript{15}. Detailed genomic data including thousands of loci provide new possibilities to assess levels of connectivity and gene flow between distinct breeding populations and, thus, help identify relevant conservation units for seabirds\textsuperscript{15,16}. Indeed, a few recent publications using reduced genomic representation approaches (e.g., RAD-seq) have reported fine-scale structure over various spatial scales\textsuperscript{17–21}. These studies highlight the great potential of genomic data to disentangle barriers to gene flow that would otherwise remain undetected, but remain nonetheless limited due to incomplete sampling of the genome\textsuperscript{22}.

### Results

#### Genome assembly and population sequencing

Based on synteny with the razorbill (Alca torda), a total of 13,328 puffin scaffolds were placed into 26 pseudo-chromosomes, leaving 17.06 Mb (1.4\%) unplaced and yielding an assembly of 1.294 Gbp (Supplementary Data 1, Table S1). This assembly contains 4,522 of the 4,915 genes (92.0\%) of complete protein-coding sequences from the avian set of the OrthoDB 9 database (Supplementary Data 1). We also assembled the puffin mitogenome (length of 17,084 bp) with a similar arrangement of genomic elements as other members within the Alcidae\textsuperscript{39,40} (Fig. S1, Table S2). For the 72 resequenced specimens, we analyzed a total of 5.77 billion paired reads, obtaining an average fold-coverage of 7X (range 3.0–10) for the nuclear genome and 591X (5.3–1800) for the mitochondrial genome per specimen (Fig. 1a, Supplementary Data 2). One individual (IOM001) was removed from both datasets (nuclear and mitochondrial) due to a substantially lower number of mapped reads (endogeny) relative to all other samples (Supplementary Data 2) resulting in a large proportion of missing sites (Fig. S2).

Additional filtering produced a final genotype likelihood dataset of 1,093,765 polymorphic nuclear sites and 192 mitochondrial single-nucleotide polymorphisms (SNPs, Supplementary Data 3) in 71 birds (36 males and 35 females).

**Genomic population structure.** Genomic variation across 71 puffin mitogenomes defines 66 polymorphic haplotypes that indicate a recent global population expansion and show no significant population structure (Fig. 1b, Figs. S3, S4, Tables S3, S4). In contrast, we inferred four main population clusters using principal component analysis (PCA) of the nuclear whole-genome dataset (Fig. 1c). Puffins from Spitsbergen are most distinct, while puffins from Bjørnøya are located between Spitsbergen and a larger, central cluster consisting of populations from Norway, Iceland, and the Faroe Islands\textsuperscript{24}. Its breeding range stretches from the Arctic coast and islands of Europe, Russia, Norway, Greenland, and Canada, southward to France and the USA\textsuperscript{26} (Fig. 1a). Puffins have been designated as “vulnerable” to extinction globally and listed as “endangered” in Europe\textsuperscript{29}. Notably, the once world’s largest puffin colony (Røst, Norway) has experienced complete fledging failure during nine of the last 13 seasons and has lost nearly 80\% of its breeding pairs during the last 40 years\textsuperscript{29–31}. Similarly, Icelandic and Faroese puffins have experienced low productivity and negative population growth since 2003\textsuperscript{32}.

Puffins have been broadly classified into three taxonomic groups along a latitudinal gradient based on size, with the smallest puffins found around France, Britain, Ireland and southern Norway (F. a. grabae), intermediate sized puffins around Norway, Iceland, and Canada (F. a. arctica) and the largest puffins found in the High Arctic, e.g. Spitsbergen\textsuperscript{33}, Greenland\textsuperscript{34} and north-eastern Canada\textsuperscript{35} (F. a. naumannii)\textsuperscript{36} (Fig. 1a). Nevertheless, this broad classification into three subspecies has been controversial\textsuperscript{37,38} and the population structure of puffins remains unresolved at all spatial scales\textsuperscript{37}. This knowledge gap obstructs efforts towards an assessment of dispersal barriers, limits our understanding of cause-and-effect dynamics between population trends, ecology and the marine ecosystem, and hinders the development of adapted large-scale conservation actions.

Here, we present the, to the best of our knowledge, first whole-genome analysis of structure, gene flow, and taxonomy of a pelagic, North Atlantic seabird. We generated a de novo draft assembly for the puffin and resequenced 72 individuals across 12 colonies representing the majority of the species’ breeding range (Fig. 1a). Our work suggests that a complex interplay of ecological factors contributes to the range-wide genomic population structure of this vagile seabird.
Fig. 1 Sampling distribution and genomic structure of 71 Atlantic puffin individuals across 12 colonies throughout the breeding range. 

**a** Map presenting the location of the 12 sampling sites. Color shading indicates the breeding range of the species as a whole, as well as the recognized subspecies. **b** Mitochondrial haplotype network based on a maximum likelihood tree generated with IQTree and visualized using Fitchi. It contains 66 unique haplotypes identified by 192 mitogenome-wide SNPs. Sizes of circles are proportional to haplotype abundance. Color legend is provided in (c). Black dots represent inferred haplotypes that were not found in the present sampling. 

**c** Principal component analysis (PCA) using genotype likelihoods at 1,093,765 polymorphic nuclear sites calculated in ANGSD to project the 71 individuals onto PC axes 1 and 2. Each circle represents a sample and colors indicate the different colonies. The percentage indicates the proportion of genomic variation explained by each axis. The color coding of the colonies is consistently used throughout the manuscript. 

**d** CLUMPAK-averaged admixture plots of the best K’s using the same genotype likelihood panel as in (c). Each column represents a sample and colonies are separated by solid white lines. Optimal K’s were determined by the method of Evanno et al.41 (see Fig. S6a) and colors indicate the ancestry fraction to the different clusters. 

The dataset(s) needed to create this figure can be found at https://doi.org/10.6084/m9.figshare.14743242.v1.
**Fig. 2** Phylogenetic reconstruction of individual and colony relationships from 71 Atlantic puffin individuals sampled across 12 colonies throughout the species’ breeding range. **a** An individual-based neighbor-joining tree constructed using pairwise p-distances calculated from genotype likelihoods at 1,093,765 polymorphic nuclear sites. Branch lengths and the outgroup were removed for the zoomed-in section to improve visualization. **b** A population-based maximum likelihood Treemix analysis using allele frequencies at the same 1,093,765 polymorphic nuclear sites as in **(a)**. Both trees are rooted using the razorbill as an outgroup. The tree in **(b)** is visualized with and without the outgroup. Branch lengths are equivalent to a genetic drift parameter. The heatmap indicates the residual fit of the tree displaying the standard error of the covariance between populations. In **(a)** and **(b)**, the color coding of the colonies is consistent with those in Fig. 1 and node labels show bootstrap support >80. The dataset(s) needed to create this Figure can be found at https://doi.org/10.6084/m9.figshare.14743299.v1.
Spitsbergen to Børnøya (likelihood = 792.106; Figs. S9, S10a). Adding additional migration edges to the population-based ML tree does not improve the model fit and such edges are therefore not further interpreted (Figs. S9–S11).

**Genetic diversity, heterozygosity, and inbreeding.** Tajima’s D does not significantly deviate from neutral expectation per colony (Table S3). Nucleotide diversity (\(\pi\)) of penguins is significantly different between colonies, with the Spitsbergen population having significantly lower nucleotide diversity than the global median (Wilcoxon Rank Sum test, \(U = 4824, n_{SPI} = 25, n_{Global} = 300, P = 0.017,\) Table S3). Colonies also differ significantly in levels of heterozygosity (Kruskal–Wallis test, \(n = 12, P = 1 \times 10^{-6}\); Fig. 3a) and inbreeding (Kruskal–Wallis test, \(n = 12, P = 1 \times 10^{-7}\), Fig. 3b), whereby individual inbreeding \((F_{RoH})\) was approximated based on runs of homozygosity (RoH)\(^{42}\). Again, the Spitsbergen colony has significantly lower levels of heterozygosity \((0.00220–0.00223)\) and significantly higher levels of \(F_{RoH}\) values \((0.161–0.172)\), compared to the Faroese and Icelandic colonies \((Dunn test with Holm correction, P < 0.05, n_3 = 6, n_2 = 6)\). The Faroese and Icelandic colonies contain the highest levels of heterozygosity and lowest \(F_{RoH}\) values \((Figs. 3a, b, Fig. S12)\) overall. The remaining colonies display intermediate levels (Fig. 3a, b), although heterozygosity is significantly lower \((Fig. 3a, Fig. S12)\) and inbreeding is significantly higher \((Fig. 3b, Fig. S12)\) on Gull Island and Børnøya compared to the Icelandic and Faroese colonies \((Dunn test with Holm correction, P < 0.05, n_1 = 6, n_2 = 6)\). Moreover, Spitsbergen harbors the most \((an average of 718 per individual)\) and longest RoHs with eight \(≥2.3\) Mbp long \((4.21 \pm 3.02\% of respective chromosome), whereas none of the RoHs in the remaining colonies are \(>2.15\) Mbp long \((Fig. 3c)\). The only exception is a 9.65 Mbp long RoH on pseudo-chromosome 7 \((18\% of chromosome length)\) in an Isle of May individual \((Fig. 3c)\).

**Patterns of gene flow and isolation-by-distance (IBD).** We investigated patterns of gene flow and IBD between the colonies using two-dimensional estimated effective migration surface \((EEMS)\) analyses\(^{43}\). Levels of gene flow between the Icelandic and Faroese colonies and within the Canadian group are high \((3–10× higher than the global average), while intermediate between the Norwegian mainland colonies \((around the global average)\). In contrast, the Spitsbergen colony is split from the remaining colonies by migration rates up to 100× lower than the global average \((Fig. 4a, Fig. S13)\), while additional regions of low gene flow \((2–3× lower than the global average)\) separate the Isle of May, Canadian, and Børnøya colonies from the rest \((Fig. 4a, Fig. S13)\). Geographic distance between all puffin colonies is a poor predictor of pairwise genetic distance, driven by high Slatkin’s linearized \(F_{ST}\) values between Spitsbergen and the other colonies \((Tables S5, S6, Fig. S14)\). Nevertheless, the geographic distance among a subset of puffin colonies is significantly associated with genetic distance as shown by Mantel tests, linear regression model analyses, and distance-based Redundancy Analysis \((dbRDA)\) models \((Fig. 4b, Fig. S14, Tables S5, S6)\). Specifically, by progressively removing the more distant colonies \((Spitsbergen, Isle of May, Børnøya, Canada)\), which are characterized by high Slatkin’s linearized \(F_{ST}\) values at relatively small geographic distances \((Fig. S14)\), the fit of a linear IBD model is significantly improved and the proportion of variance of genetic dissimilarity explained by geographic distance is more than doubled \((Spitsbergen removed: 37.58%; Spitsbergen/Isle of May/Børnøya/Gannet Isl. removed: 84.98%)\) \((Fig. 4b, Fig. S14, Table S5)\). Similarly, the proportion of explained genetic variance by spatial features estimated in global dbRDA models is more than tripled \((All colonies: 18.76%; Spitsbergen/Isle of May/Børnøya removed: 59.87%)\) \((Table S5)\). In all optimized dbRDA models, geographic variables \((IBD)\) contribute significantly to the genetic divergence, while the contribution of the mean sea surface temperature \((isolation-by-environment, IBE)\) is minimal. IBE is only once significantly contributing to the observed genetic variance \((when Spitsbergen was removed)\), yet accounts for less than half of the observed genetic variance \((11.37\%)\) compared to the geographic distance \((28.66\%)\) \((Table S6)\).

**Admixuture on Børnøya.** We specifically tested for patterns of admixture in Børnøya. Significantly negative \(\beta\) statistics \((Z score < –3)\) are found for all unique combinations of the phylogeny \((Spitsbergen, X; Børnøya)\) \((Table S7)\), indicating an admixed colony on Børnøya caused by gene flow between Spitsbergen and the remaining colonies. Similarly, significantly positive D-statistics \((Z score > 3)\) caused by an excess of ABBA sites reveal excessive allele sharing between Spitsbergen and Børnøya \((Fig. S15a)\). The close association and gene flow from Spitsbergen to Børnøya is further confirmed by D-statistics not being significantly different from 0 for the \(((Spitsbergen, Børnøya, Spitsbergen), H3)\), Razorbill topology \((Fig. S15b)\).

**Genetic differentiation.** We assessed genome-wide patterns of genetic differentiation by calculating pairwise \(F_{ST}\) between the four genomic clusters in 50 kb sliding windows. These analyses show that the differentiation between the clusters is driven by increased \(F_{ST}\) in windows across the entire genome, including the presence of several smaller regions with elevated \(F_{ST}\) \((Fig. S16)\). Several of these elevated \(F_{ST}\) regions are present in all pairwise comparisons \((Fig. S16)\), whereas others are specific for certain comparisons, and may be indicative of local adaptation \((Fig. S16)\).

**Discussion**

Barriers to gene flow leading to population structure are notoriously difficult to identify and remain largely unknown for most seabirds\(^{15,44}\). Using whole-genome analyses, we here provide insights into the genetic structure of the Atlantic puffin. First, we identify four main puffin population clusters consisting of \((1)\) Spitsbergen \((High Arctic)\), \((2)\) Canada, \((3)\) Isle of May, and \((4)\) multiple colonies in Iceland, the Faroe Islands, and Norway. Second, we find that within such clusters, genetic differentiation is driven by IBD. Finally, we find evidence for secondary contact between two clusters. These observations show that a complex set of drivers impacts gene flow over different spatial scales \((100–1000s of km)\) between these clusters and the colonies within. In particular, the interplay between overwintering grounds, philopatry, natal dispersal, geographic distance, and potentially ocean regimes appears to explain the genomic differentiation between puffin colonies.

Mature penguins rarely, if ever, change their colonies, resulting in very high colony fidelity once they start breeding\(^{28}\). Immatures, however, have been observed to visit other nearby colonies during the summer and may breed in non-natal colonies\(^{28,46}\). Nevertheless, data on natal philopatry remain scarce, but existing evidence shows rates vary greatly \((38–92\%)\) between colonies\(^{28,46}\). If either breeding or natal philopatry alone drive the puffin population structure, each colony should constitute its own distinct genomic entity and substantial genomic differentiation across the puffin’s entire breeding range would be observed. Yet, philopatry alone cannot explain the presence of the four large-scale population clusters we observe here. Additional factors must therefore promote the distinctiveness of the four clusters. For instance, the Isle of May birds have a largely separate overwintering distribution mainly in the North Sea \((Fig. S17)\)\(^{28,38,47}\). Such potential...
**Fig. 3** Genome-wide heterozygosity, inbreeding, and Runs-of-Homozygosity (RoH) compared between 12 Atlantic puffin colonies across the species’ breeding range. 

- **a** Estimates of individual genome-wide heterozygosity based on the per-sample one-dimensional Site Frequency Spectrum calculated in ANGSD.
- **b** Individual inbreeding coefficients, $F_{\text{RoH}}$, defined as the fraction of the individual genomes falling into RoHs of a minimum length of 150 kb. RoHs were declared as all regions with at least two subsequent 100 kb windows harboring a heterozygosity below $1.435663 \times 10^{-3}$. 
- **c** RoH length distribution across the 12 colonies only including RoHs longer than 500 kb. A single 9.65 Mbp long RoH on pseudo-chromosome 7 in an Isle of May individual required to introduce a break in the y-axis.

In **(a)** and **(b)**, black dots indicate individual sample estimates and black lines the median per colony, while in **(c)**, black dots represent single RoHs. Statistical significance of differences in heterozygosity and $F_{\text{RoH}}$ between populations was assessed with a global Kruskal-Wallis test ($n = 12$). The results of post hoc Dunn tests with Holm corrections are presented in Fig. S12. Error bars show range of values within 1.5 times the interquartile range. Different colonies in all three plots are indicated using the same color code as in Fig. 1. The dataset(s) needed to create this figure can be found at https://doi.org/10.6084/m9..figshare.14743317.v1.
Fig. 4 Estimates of continuous long-distance gene flow and isolation by distance (IBD) across the breeding range of the Atlantic puffin estimated from 71 individuals across 12 colonies. 

**a** Effective migration surfaces inferred by the program EEMS using the average distance between pairs of individuals calculated in ANGSD by sampling the consensus base for each individual at 1,093,765 polymorphic nuclear sites. Darker reds indicate reduced migration across those areas, while darker blues highlight higher migration rates than the global mean. Different colonies are indicated using colors consistent with those in Fig. 1.

**b** Correlation between genetic (Slatkin’s linearized FST) and geographic (Least Cost Path—only over water) distance presented after removing the Spitsbergen, Bjørnøya, Isle of May, and Canadian individuals. The diagonal line visualizes the result of the multiple regression on distance matrices (MRM) analysis (slope and y-intercept). The Mantel test between genetic and geographic distance ($R = 0.775$, $P = 0.012$, $n_{\text{Colonies}} = 7$) was significant and 60.08% of the variance in Slatkin’s linearized $F_{ST}$ was explained by geographic distance (regression coefficient of linear IBD model = $0.76 \times 10^{-6}$, $P = 0.006$, $n_{\text{Colonies}} = 7$). A two-dimensional kernel density estimation (kde2d) highlights dense groups of data points, thus substructure in the genomic landscape pattern. Analyses were conducted and results visualized in R using the ecodist, marmap and MASS packages. The dataset(s) needed to create this figure can be found at https://doi.org/10.6084/m9.figshare.14743323.
geographical separation during the winter season might limit the likelihood of immatures intermixing between the Isle of May and other colonies. Similarly, distinct overwintering distributions have been found to lead to increased genetic diversification in other philopatric seabird species\(^1\), such as the thick-billed murre (\textit{Uria lomvia})\(^2\) and black-browed albatross (\textit{Thalassarche melanophris})\(^3\). The presence of a Canadian cluster can also be largely explained by their winter distribution around Newfoundland\(^4\),\(^5\). There is, however, some fragmentary overlap in the overwintering distribution of the Canadian and Icelandic colonies off southwestern Greenland\(^6\),\(^7\), suggesting that barriers to dispersal of immatures and gene flow in the western Atlantic may be further enforced by the large geographic distance. In contrast, the winter distribution from the colonies in Iceland, Norway, and the Faroe Islands overlaps off the coast of southern Greenland (Fig. S17). This shared overwintering area, combined with the tendency to return to the natal colony and immatures visits to nearby (up to 100 km) colonies during the summer, appears to drive a pattern of IBD among colonies (Fig. 3b). Indeed, IBD has previously been recognized as an important driver of genomic structure in seabirds, for instance in the little auk (\textit{Alle alle})\(^8\) and band-rumped storm-petrel (\textit{Oceanodroma castro})\(^9\). While these illustrated mechanisms provide reasonable explanations for the observed dispersal barriers and population structure based on our current knowledge, validation requires additional evidence, specifically on the winter distribution of immature puffins and natal dispersal rates across colonies covering the entirety of the puffin’s breeding range.

High Arctic puffins from Spitsbergen are genetically the most divergent group within our dataset harboring the highest genomewide differentiation. They are also characterized by significantly lower levels of genetic diversity, greater inbreeding coefficients, and longer and more abundant RoHs compared to other colonies. These observations may either result from a historical bottleneck followed by isolation (e.g., founder effect), local adaptation to their extreme environment, or generally lower effective population sizes. Population abundance estimates of <10,000 breeding pairs on Spitsbergen compared to 500,000 in the West Atlantic, two million on Iceland and more than two million in the boreal East Atlantic potentially indicate a lower effective population size\(^10\). The High Arctic puffins exclusively inhabit harsh, cold-current environments year-round, as they likely stay in an area bounded by the East Greenland ice edge, a latitudinal border at 70° N, and the front between the Barents and Greenland Sea during winter (Fig. S17). They are also substantially larger than birds from lower latitudes\(^11\),\(^12\),\(^13\),\(^14\), following Bergmann’s\(^15\) or James’\(^16\) rule, as has been observed in other seabirds\(^17\),\(^18\). This matches the clinal size variation of puffins that closely track sea temperatures in their breeding areas\(^19\). Despite these distinctions, we find that the relatively small population of puffins on Bjørnøya (<1000 pairs\(^20\)), midway between Spitsbergen and mainland Norway, represents an area of secondary contact between the puffins from the High Arctic and other puffin colonies. Based on D- and the β-statistics, the most likely southern sources are Iceland, the Faroe Islands, Norway, or a combination thereof. Thus, the barriers to gene flow that keep the Spitsbergen colonies distinct do not prevent the formation of a hybrid colony where individuals from the High Arctic and the cluster composed of mainland Norwegian, Icelandic and Faroese colonies meet.

The distinct population structure in the nuclear data is not observed in the mitochondrial genomes, which reveal an abundance of rare alleles and lack of significant population differentiation. The mitogenomic variation suggests that puffins experienced a recent population expansion, possibly out of a refugium after the Last Glacial Maximum. Indeed, it has been shown that mitogenomic variation in seabirds is dominated by historical factors rather than representing contemporary gene flow\(^21\), and a lack of mitogenomic population structure has been observed in many marine birds with high philopatry\(^22\),\(^23\),\(^24\). In contrast to the mitogenomes, the structure in the nuclear data therefore likely originated after the last glacial period and reflects the influence of relatively recent barriers to gene flow in a context of historical demography\(^15\),\(^16\). Such results are relevant for understanding the “seabird paradox”, which contrasts the life-history traits of high philopatry and restricted dispersal in otherwise highly mobile species\(^25\).

Our results have major implications for the conservation management of the Atlantic puffin. The genetic structure we identify in puffins disagrees with the suggestion of three subspecies (\textit{F. a. naumanni}, \textit{F. a. arctica}, \textit{F. a. grabae})\(^26\). Although the genetically distinct Spitsbergen cluster coincides with the classification of morphologically large puffins in the High Arctic (\textit{F. a. naumanni})\(^27\),\(^28\), we observe gene flow from Spitsbergen into Bjørnøya, which has been considered \textit{F. a. arctica}\(^29\). Furthermore, the geographic divide between \textit{F. a. grabae} and \textit{F. a. arctica} lies farther south than previously thought, with the Faroese puffins being genetically closer to \textit{F. a. arctica} than to \textit{F. a. grabae}. Nonetheless, \textit{F. a. grabae} is currently represented by a single colony (Isle of May) in our study and the geographical extent of this genomic cluster needs to be refined by additional sampling, particularly in the western UK, Ireland, and France. Finally, puffins from the Western Atlantic region (e.g., colonies in Canada) form their own distinct genetic cluster that is not recognized within the current classification. Our results do not only warrant a revision of Salomonsen’s taxonomic classification of three subspecies\(^26\), but also highlight the need to acknowledge the four identified clusters as distinct units within the conservation management of puffins\(^11\),\(^13\),\(^14\). Although puffin colonies within clusters are not genetically distinct entities, ecological independence illustrated by contrasting population dynamics across relatively small spatial scales (e.g., western Norway\(^31\)) suggests that higher resolution local management units based on ecological differences should be considered. Nonetheless, the genetically distinct clusters at the outer edges of the puffins’ distribution with putative local adaptations that will not be easily replenished indicate that conservation of these distinct clusters must be a first priority. Finally, our sampling does not cover several outskirts of the puffin’s distribution, such as the U.S., northern Canada, Greenland, Ireland, western UK, France or Russia, and we may therefore still underestimate the true biological and genetic complexity of this species.

In conclusion, our study shows that a complex interplay of barriers to gene flow drives a previously unrecognized population diversification in the iconic Atlantic puffin. So far, much of seabird population genetics research has been based on mitochondrial and microsatellite data\(^15\),\(^16\), which have limited power to characterize contemporary factors that determine population structure and gene flow\(^20\),\(^26\). High-resolution nuclear data are therefore essential to help define evolutionary significant population units, disentangle convoluted ecological relationships, and are particularly important for seabird conservation, which aims to preserve genetic diversity considering profound global population declines\(^28\), and the threat of global warming, which negatively impacts ecosystems worldwide\(^29\).

**Methods**

**Ethical statement.** Feather and blood samples of puffins included in this study were collected and handled under the following permits.

1. Gásseyne, Rast, Hornøya, Bjørnøya (Norway)—FOTS ID #15602 and #15603 from the Norwegian Food Safety Authority for SEATRACK and SEAPOP; Permit 2018/607 from Miljødirektoratet (Norwegian Environment Agency), dated 4 May 2018.
DNA extraction and sequencing. Samples from a total of 72 puffins collected across 12 breeding colonies (Fig. 1a) were made available for the present study by SEAAPop (http://www.seaapop.en/en/seaarck) and ARCTOX (http://www.arctox.cnrs.fr/en/home—Canadian colonies). These samples had been collected between 2012 and 2018 and consisted of blood preserved in EDTA or lysis buffer, or feathers (Supplementary Data 2). We extracted DNA using the DNeasy Blood & Tissue kit (Qiagen) following the manufacturer’s protocol for animal blood or the nail/hair/feathers protocol applying several modifications for improved lysis and DNA yield. Individuals that had no sexing data associated with them were sexed using PCR amplification of specific allozyme loci and visualization via gel electrophoresis. Genomic libraries were built by the Norwegian Sequencing Centre and sequenced on an Illumina HiSeq4000. We processed sequencing reads in PALNGeny v.1.143 by cleaning and filtering in ANGSD v.0.9.0c. The final length of such blocks was evaluated by generating a quantitative approach by evaluating the distribution of explained variance, the log likelihoods, the covariance with an increase in migration edges, and by applying the method of Evanov70 and several different linear threshold models. The topology for \( m_r \) and \( m_{nGSN} \) was evaluated by generating 100 bootstrap replicates. Additional details on the cluster and phylogenetic analyses are given in the Supplementary File.

Genetic diversity, heterozygosity, and inbreeding. We calculated a set of neutrality tests and population statistics in ANGSD using colony-based one-dimensional (1D) fold site frequency spectra (SFS). For each population, genomic cluster, and globally, Tajima’s \( D \) and nucleotide diversity \( (\pi) \) were computed utilizing the per-site \( \theta \) estimates. Individual genome-wide heterozygosity was calculated in ANGSD using individual, folded, 1D SFS. We calculated heterozygosity by dividing the number of polymorphic sites by the number of possible sites. The proportion of RoH within each puffin genome was calculated by calculating local estimates of heterozygosity in 100 kb sliding windows (50 kb slide) following the approach in Sánchez-Barreiro et al.42. We defined the 10% quantile of the average local heterozygosity across all samples as the cutoff for a “low heterozygosity region” (Fig. S5) generating two overlapping RoH files. We calculated an individual inbreeding coefficient based on the RoH, \( F_{\text{RoH}} \), as in Sánchez-Barreiro et al.92 by computing the fraction of the entire genome falling into RoHs, with the entire genome being the total length of windows. Additional details on these analyses can be found in the Supplementary File.

Patterns of gene flow and admixture. Assessing potential patterns of IBD within the breeding range of the puffin, the program EMES80 was used to model the association between genetic and geographic data by visualizing the existing population structure and highlighting regions of higher-than-average and lower-than-average historic gene flow. We calculated a pairwise genetic distance matrix in ADMIXTURE sampling two random loci (two telomeric sites included in the genotype likelihood set (see Nuclear cluster and phylogenetic analyses) for each sample. The matrix was fed into 10 independent runs of EEMS, each consisting of one MCCM chain of six million iterations with a two million iteration burn-in, 9999 thinning iterations, and 1000 underlying demes. Using the EM algorithm for K, the program returned two distinct clusters, and the posterior probability of the data under the model was maximized. We converted pairwise \( F_{ST} \) values to Satlkin’s linearized \( F_{ST} \). Least Cost Path distances (paths over water only) between colony coordinates (latitude/longitude)
were calculated using the R package marmap⁸⁶ and used as geographic distances. We performed the Mantel test (999 permutations) and MRMA analysis with the R package capm⁹⁴. All analyses for IBD were re-run on subsets of colonies by progressively removing the colony from the geographic and genetic distance matrices, whose removal led to the highest increase in the proportion of variance in genetic distance explained by geographic distance in the resulting regression model (Spitsbergen, Isle of May, Bjørnøya and Gannet Isl.).

A distance-based Redundancy Analysis (dbRDA)⁸⁸ was conducted to corroborate the results of the MRMA analyses and Mantel tests and to estimate the relative contribution of IBD and IBE to the observed Atlantic puffin population structure. The dbRDA was run between the genetic distance matrix versus geographic and environmental variables, and for statistically significant global dbRDA models, the most significant variables (geographic or environmental) were selected via a stepwise regression⁹⁹. Those served as input for a reduced dbRDA to explain variation in population genetic structure. This includes versions of any software used, if relevant, and any specific variables or parameters used to generate, test, and process the dataset of this study.

Received: 17 November 2020; Accepted: 1 July 2021; Published online: 29 July 2021

References
1. Otero, X. L., De La Peña-Lastra, S., Pérez-Alberti, A., Ferreira, T. O. & Huerta-Diaz, M. A. Seabird colonies as important global drivers in the nitrogen and phosphorus cycles. Nat. Commun., 9, 246 (2018).
2. Velarde, E., Anderson, D. W. & Ezcurra, E. Seabird clues to ecosystem health. Science 365, 116–117 (2019).
3. Pratt, J. F., Sydeman, W. J. & Wiese, F. Introduction: a modern role for seabirds as indicators. Mar. Ecol. Prog. Ser. 352, 199–204 (2007).
4. Boersma, P. D., Clark, J. & A. & Hillgarth, N. Seabird conservation. In Biology of Marine Birds (eds. Schreiber, E. & Burger, J.), 559–579 (CRC Press Boca Raton, 2002).
5. Derlinger, L. & Wohl, K. Seabird harvest regimes in the circumpolar nations. Conservation of Arctic Flora and Fauna (CAFF), (2001).
6. Merkel, F. & Barry, T. Seabird Harvest in the Arctic. Conservation of Arctic Flora and Fauna (CAFF), (2008).
7. Croxall, J. P. et al. Seabird conservation status, threats and priority actions: a global assessment. Bird Conserv. Int. 22, 1–34 (2012).
8. Palczyn, M., Hammill, F. & Karpouzi, V. & Paudy, D. Population trend of the world’s monitored seabirds, 1950-2010. PLoS ONE 10, e0129342 (2015).
9. Frederiksen, M. Seabirds in the North East Atlantic. Summary of status, trends and anthropogenic impact. TemaNord 587, 21–24 (2010).
10. Chardine, J. & Mendenhall, V. Human Disturbance at Arctic Seabird Colonies. Conservation of Arctic Flora and Fauna (CAFF), (1998).
11. Funk, W. C., McKay, J. K., Hohenlohe, P. A. & Allendorf, F. W. Harnessing genomics for delineating conservation units. Trends Ecol. Evol. 27, 489–496 (2012).
12. Moritz, C. Defining Evolutionarily Significant Units’ for conservation. Trends Ecol. Evol. 9, 373–375 (1994).
13. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genetics and the future of conservation genetics. Nat. Rev. Genet. 11, 697 (2010).
14. Fraser, D. J. & Bernatchez, L. Adaptive evolutionary conservation: towards a unified concept for defining conservation units. Mol. Ecol. 10, 2741–2752 (2001).
15. Friesen, V. L. Speciation in seabirds: why are there so many species… and why aren’t there more? J. Ornithol. 156, 27–39 (2015).
16. Taylor, R. S. et al. Sympatric population divergence within a highly pelagic seabird species complex (Hydrobates spp.). J. Avian Biol. 49, 1–18 (2018).
17. Røn-Huber, K. & et al Genomics detects population structure within and between ocean basins in a circumpolar seabird: the white-chinned petrel. Mol. Ecol. 28, 4552–4579 (2019).
18. Cucco, G. V. & et al Comparative population genetics reveals key barriers to dispersal in Southern Ocean penguins. Mol. Ecol. 27, 4680–4697 (2018).
19. Frugone, M. J. et al. More than the eye can see: Genomic insights into the drivers of genetic differentiation in Royal/Macaroni penguins across the Southern Ocean. Mol. Phylogenet. Evol. 139, 106563 (2019).
20. Cristofari, R. et al. Unexpected population fragmentation in an endangered seabird: the case of the Peruvian diving-petrel. Sci. Rep. 9, 2021 (2019).
21. Tignano, A., Shultz, A. J., Edwards, S. V., Robertson, G. J. & Friesen, V. L. Outlier analyses to test for local adaptation to breeding grounds in a migratory arctic seabird. Ecol. Evol. 7, 2370–2381 (2017).
22. Lowry, D. B. et al. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol. Ecol. Resour. 17, 142–152 (2017).
23. Somvichian-Clausen, A. Behind the stunning photo of a puffin gorging on fish. Nat. Geograph. (2017).
24. Huijbers, E. H. & Einanssor, N. Feasting on Friends: Whales, Puffins, and Tourism in Iceland. In Tourism Experiences and Animal Consumption (ed. Kline, C.), 10–27 (Routledge, 2018).
25. Lund, K. A., Kjartansdóttir, K. & Loftsdóttir, K. ‘Puffin love’: performing and creating Arctic landscapes in Iceland through souvenirs. Tour. Stud. 18, 142–158 (2018).
26. Hodgetts, L. M. Animal bones and human society in the late younger stone age of arctic Norway. (Durham University, 1999).
27. Dove, C. J. & Wickers, S. Identification of bird species used to make a Viking age feather pillow. Arctic 69, 29–36 (2016).
28. Harris, M. P. & Wanless, S. The puffin (T & AD Poyser, Bloomsbury Publishing, 2011).
29. BirdLife International. Fratercula arctica. The IUCN Red List of Threatened Species 2017.
30. Anker-Nilssen, T. et al. Key-site monitoring in Norway 2019, including Svalbard and Jan Mayen. SEAPOP Short Report 1 (ed. Kline, C.) 10
31. Anker-Nilssen, T. et al. Key-site monitoring in Norway 2019, including Svalbard and Jan Mayen. SEAPOP Short Report 1 (ed. Kline, C.) 2019.
32. Nilsen, A. & et al. Key-site monitoring in Norway 2019, including Svalbard and Jan Mayen. SEAPOP Short Report 1 (ed. Kline, C.) 10
33. Walker, S. J. & Meijer, H. J. M. Size variation in mid-Holocene North Atlantic seals. J. Mar. Syst. 34, 269–279 (2001).
34. Burnham, K. K., Burnham, J. L. & Johnson, J. A. Morphological measurements of Atlantic puffin (Fratercula arctica naumanni) in High-Arctic Greenland. Polar Res. 39, https://doi.org/10.3326/polar.v39i2.5142 (2020).
projects for collecting and sharing of samples. In particular, we thank Dave Fitfield and Greg Robertson for the provision of puffin samples from Gull Island, and Árni Ásgeirsson and Robert Arnar Stefánsson for supplying samples from Breiðafjörður. The authors acknowledge support from the National Genomics Infrastructure in Stockholm funded by the Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and the SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing (reference genome) and access to the UPPMAX computational infrastructure. Special thanks goes to the Norwegian Sequencing Centre, UiO (https://www.sequencing.uio.no) for the genomic libraries and resequencing of samples analyzed in this study. The razorbill genome data was made available for this study by Tom Gilbert and the Vertebrate Genome Project. Computation was performed using the resources and assistance from SIGMA2. Albína Pálsdóttir shared scripts for ANGSD, and Emiliano Trucchii advised on the manuscript. Pictures of puffins used in the figures were taken by Annemarie Look.

Author contributions
S.B. and B.S. conceptualized the project. M.I. performed the DNA extraction for the reference genome. O.K. did all other laboratory work. K.S.J. advised on the sequencing strategy and co-supervised O.K. O.K. refined the reference genome assembly. O.K. carried out the population genomic analyses with input from S.B., B.S., and D.M.L. O.K. designed the figures with input from S.B., B.S., and D.M.L. T.A.N., H.S., and E.S.H. advised on colony selection and provided ecological context. T.A.N., H.S., Od.K., S.D., E.S.H., J.F., M.P.H., J.D., K.E., M.L.M., and M.G.F. provided samples. O.K. wrote the paper with S.B., B.S., and D.M.L. All authors read and revised the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02415-4.

Correspondence and requests for materials should be addressed to O.K. or S.B.

Peer review information Communications Biology thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Caitlin Karniski.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.