King penguin demography since the last glaciation inferred from genome-wide data

Emiliano Trucchi¹, Paolo Gratton², Jason D. Whittington¹,³,⁴,⁵, Robin Cristofari¹,³,⁴,⁵, Yvon Le Maho³,⁴, Nils Chr Stenseth¹ and Céline Le Bohec¹,³,⁴,⁵

¹Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, PO Box 1066, Blindern, Oslo 0316, Norway
²Department of Biology, University of Rome ’Tor Vergata’, Via della Ricerca Scientifica, Rome 00133, Italy
³Institut Pluridisciplinaire Hubert Curien, Physiologie et Ethologie, Université de Strasbourg, 23 Rue Becquerel, Strasbourg Cedex 02 67087, France
⁴Centre National de la Recherche Scientifique (UMR 7178 and LIA-647 BioSensib), 23 Rue Becquerel, Strasbourg Cedex 02 67087, France
⁵Centre Scientifique de Monaco (LIA-647 BioSensib), 8 Quai Antoine 1er, Monaco 98000, Principality of Monaco

How natural climate cycles, such as past glacial/interglacial patterns, have shaped species distributions at the high-latitude regions of the Southern Hemisphere is still largely unclear. Here, we show how the post-glacial warming following the Last Glacial Maximum (ca 18 000 years ago), allowed the (re)colonization of the fragmented sub-Antarctic habitat by an upper-level marine predator, the king penguin Aptenodytes patagonicus. Using restriction site-associated DNA sequencing and standard mitochondrial data, we tested the behaviour of subsets of anonymous nuclear loci in inferring past demography through coalescent-based and allele frequency spectrum analyses. Our results show that the king penguin population breeding on Crozet archipelago steeply increased in size, closely following the Holocene warming recorded in the Epica Dome C ice core. The following population growth can be explained by a threshold model in which the ecological requirements of this species (year-round ice-free habitat for breeding and access to a major source of food such as the Antarctic Polar Front) were met on Crozet soon after the Pleistocene/Holocene climatic transition.

1. Introduction

Environmental variation through time is one of the drivers of the evolutionary process and a key mechanism in shaping biodiversity. Both short- and long-term climate shifts dramatically affect the geographical distribution of species/populations according to their own dispersal abilities and ecological requirements [1,2]. Understanding how past fluctuations impacted the demography of key species in different ecosystems is essential for predicting the response of communities to ongoing and future change, including anthropogenic-induced climate forcing [3,4]. Using genomic data, we investigate the impact of the last glaciation on the past demography of a king penguin (Aptenodytes patagonicus) colony from the most important breeding areas for this species (the Crozet archipelago).

While the biological impacts of recent Quaternary glaciation events on the largely terrestrial Northern Hemisphere have been relatively well documented [3], the oceanic Southern Hemisphere, particularly at higher latitudes, remains poorly described [5] owing to its relative inaccessibility. Recent hypotheses suggested that winter sea ice during the Last Glacial Maximum (LGM) could have extended as far north as South Georgia, Crozet, Kerguelen and Macquarie Islands, while Falkland, Gough and New Zealand Islands may have acted as ice-free refugia (figure 1 [6]). As southern ocean conditions fluctuated between glacial and interglacial periods, local species had to contend with changes in...
both habitat and resource availability. In particular, the distribution of most seabirds is primarily constrained by the spatial location of suitable breeding sites on land and foraging areas in the ocean [7]. For a population to thrive, productive foraging areas must exist within reach of a suitable breeding site. Climate oscillations affect both these key ecological constraints. On one side, these fluctuations cause glaciers to expand and retreat, thus covering or exposing potential breeding sites, on the other side, they affect the location and intensity of oceanic currents and fronts that in turn determine marine primary production. According to Thatje et al. [8] the large-scale Southern Ocean productivity may have been nearly shut down during the LGM, at least at the latitude where it is currently present. Highly dispersive species negatively affected by glacial conditions could have moved northwards to more suitable ice-free breeding habitats during the LGM, while low-dispersive ones may have been strongly reduced in population size or may even have gone extinct [9]. Therefore, post-glacial warming and the subsequent retreat of glaciers and sea ice, probably favoured the re-colonization of Antarctic and sub-Antarctic territories by dispersive species dependent on ice-free breeding sites [5,6,10].

Although some marine mammals or seabirds from Antarctic and sub-Antarctic areas have shown dispersal responses to long-term climatic variations (e.g. southern elephant seal *Mirounga leonina* [11] or Adélie penguin *Pygoscelis Adeliae* [12]), the future rate of environmental changes induced by a warming climate may outpace the ability of most upper-trophic-level predators to adjust. In this context, projections simulated according to future global warming scenarios given by the latest Intergovernmental Panel on Climate Change Fourth Assessment Report [13] reveal that the genus *Aptenodytes* may be under serious threat of extinction before the end of the twenty-first century [14,15]. Indeed, an abrupt increase in sea surface temperature (SST) has been demonstrated to be detrimental to king penguin survival and reproductive rates in populations breeding in the Crozet archipelago in the southern Indian Ocean [14,16]. This is caused by the longer distance the individuals have to swim to get to their main foraging ground (the Antarctic Polar Front during the summer season) whose latitudinal location is influenced by SST (the higher the SST, the higher the latitude [17]). However, when SST was lower, and food resources probably closer, the Crozet Islands were probably not suitable for the king penguin to complete their breeding cycle as this species needs year-round ice-free grounds for reproduction.

Although ca half of the global king penguin population is resident in the Crozet basin, and this top predator represents one of the largest components of the sub-Antarctic marine ecosystem as measured by biomass and energy flux [18], the long-term response of this species to the warming period following the LGM remains entirely unknown. Here, we use a genomic-based demographic reconstruction to test whether the king penguin population from the Crozet
archipelago (‘La Baie du Marin’ colony, Possession Island) was strongly affected by the colder conditions characterizing the LGM. We tested the applicability of genome-wide data produced by restriction site-associated DNA (RAD) sequencing [19,20] within a multi-locus coalescent-based framework (Extended Bayesian Skyline Plot, EBSP [21]) that aims to average the stochasticity of mutation and drift across tens of genomic loci and to overcome the idiosyncrasy of the results obtained from one or a few molecular markers.

We first demonstrated that genomic regions with different levels of polymorphism consistently show a similar pattern of diversity, and that a consistent demographic signature is recognizable across the entire genome, thus providing evidence that a reduced subset of anonymous genomic loci can be safely employed to estimate past population trends. Historic demography was then inferred by employing a random selection of RAD loci sequenced in multiple individuals and time calibrated according to the mitochondrial substitution rate as estimated in the Adélie penguin Control Region (CR) (in substitutions site$^{-1}$Myr$^{-1}$; median = 0.55, 95% interval = 0.29–0.88 [22]). Our data show that the king penguin population from Crozet was strongly reduced on this potentially ice-covered sub-Antarctic archipelago before the end of the LGM, or king penguins may have been totally absent, but they (re)colonized the island as soon as the ecological conditions required by this species were met during the following Holocene warming.

2. Results

ILLUMINA sequencing of a paired-end RAD library from eight king penguins yielded ca 65 million, 100 bp reads. Front reads only were used in our analyses and are available on GenBank at the Sequence Read Archive (Run Num.: SRR942341). After quality filtering, trimming of the last 5 bp, and barcode sorting, 101 115 anonymous loci (each as 95 bp nucleotide sequence) with 50 5 bp, and barcode sorting, 101 115 anonymous loci (each as 95 bp nucleotide sequence) were aligned in an unreferenced catalogue. GC content of this dataset was estimated as 50.45%. According to our quality criteria, the catalogue was further filtered to 66 172 loci (of which 31 452 were polymorphic) matching all eight individuals without missing data. This dataset was then used in downstream statistical analyses.

First, we tested whether genomic regions with different degrees of polymorphism carried consistent information about the demographic history of the king penguin population from Crozet. Polymorphic RAD loci were sorted into six different classes according to the number of single nucleotide polymorphisms (SNPs) observed (from 1 to 6 SNPs), and separate analyses were performed for each category: 20 458 loci with 1 SNP; 7750 loci with 2 SNPs; 2417 loci with 3 SNPs; 652 loci with 4 SNPs; 154 loci with 5 SNPs; and 21 loci with 6 SNPs. None of the loci with more than 6 SNPs passed our quality criteria. The signature of the past demography in the 1-SNP class was investigated using the allele frequency spectrum (AFS) (figure 2a) in 0a0i [23]. A likelihood ratio test strongly supported a 2-epoch sudden demographic expansion model over a constant-size model ($\chi^2 = 8063; p < 0.0001$). When scaled by our estimated mutation rate for the 1-SNP class (see below), estimates of the parameters from the best model indicate a demographic expansion from ca 1400 to ca 7000 breeders around 18 500 years ago. The mismatch distribution density was then checked for the loci in the 2- to 6-SNP classes (figure 2b). A unimodal distribution of mismatches, characteristic of a recent population expansion, was typical across loci in all SNP classes, though more evident in those classes with higher number of SNPs. Notably, the number of outlier loci (in this case, loci showing high frequency of mismatch at the highest number of differences) was negligible in all classes, highlighting the consistent mismatch distributions across loci, which is expected in a population that has recently undergone a demographic expansion. The same 2- to 6-SNP classes of loci were then used to reconstruct the demographic function through time by a multi-locus coalescent-based Bayesian approach (EBSP [21]).

Highly consistent results were obtained using separate datasets with 2, 3 or 4–6-SNP loci (electronic supplementary material, figure S1a–c). However, loci from 4- to 6-SNP classes...
were chosen for further analyses because of their higher information content.

Ten independent datasets including 50 loci selected at random from 4 to 6-SNP classes were tested in order to check the consistency of the results over multiple loci. No significant linkage among the selected loci was detected in any dataset: the minimum p-value in the set of pairwise comparisons analysed in each dataset (0.004–0.006) was about two orders of magnitude higher than the Bonferroni-corrected threshold (0.00004) for a set of 50 loci. The substitution rate for each locus was set as a wide uninformative uniform prior (0.005–0.5 substitutions site\(^{-1}\)).

The substitution rates estimated in BEAST for loci with different SNP class across the 66 172 loci.

Figure 3. Past demographic trend of the king penguin colony of 'La Baie du Marin' on Possession Island, Crozet archipelago: median value (red) and 95% confidence interval (pale red). Trend of temperature anomalies recorded in the EPICA Dome C ice core (black [25]). The LGM period is given in green.
3. Discussion

(a) Using restriction site-associated DNA sequencing data in a coalescent-based framework

Our analyses empirically supported that RAD sequencing data can be effectively used for inferring past demography in a coalescent-based framework and that a small subset of these data (i.e. 50 loci only) is sufficient to describe the post-glacial history of the king penguin colony breeding on Crozet. This result is particularly important because, though thousands of markers can now be easily and cost-effectively sequenced in any biological system [20], many existing analytical tools for demographic inference are not yet optimized for genome-level data. Recently published methods, like the pairwise sequentially Markovian coalescence [25] or DiCal [26] require phased diploid genomic data over a long fragment of the chromosome that are not readily available for most non-model species. The development of new algorithms to exploit SNP data in coalescent-based frameworks (i.e. ‘SNPs and amplified fragment length polymorphisms phylogenies’ [27]) are paving the way for a full use of genome-scan data in phylogeography and population genetics, although linkage needs to be taken into account for extensive SNP datasets.

We found that EBSP analysis is not really optimized for handling high numbers of loci (very high parametrization) and this may result in slow convergence and poor mixing of the MCMC chains. We estimated that 3–5 billion iterations would be needed to get effective sample size above 200 for all model parameters when analysing our dataset of 50 loci and eight individuals. Nevertheless, our EBSP demographic reconstructions were highly consistent with each other, with those produced by the diffusion approximation of the AFS (figure 2a) and with standard single-locus mtDNA analyses (electronic supplementary material, figure S3). The AFS analysis, based on approximation to the neutral Wright–Fisher diffusion of an allele [23], has been strongly criticized as theoretically unfit to distinguish between competing population histories even if strongly simplified assumptions (panmixia, infinite site mutation model and neutrality) are met in the studied system [28]. In particular, it seems that compensating events in population dynamics are impossible to detect, thus making the analysis of complex histories unreliable. In our test case, the different analytical approaches converged on a quite simple history of sudden population growth making it a likely suitable case for AFS analysis. Both EBSP and AFS analyses, based on independent datasets (1-SNP class versus 4- to 6-SNP classes loci), produced largely similar inferences of the time and trend of the past demography of our king penguin colony. A major difference between results from the two methods was in the estimate of the current effective population size ($N_e = 7000$ from AFS analysis versus $N_e = 170,000$ from EBSP). Both estimates are inconsistent with the direct count of breeding birds on ‘La Baie du Marin’ colony, Possession Island (i.e. 32,000 breeding birds [29]). However, it is likely that a certain degree of gene flow exists within the Possession Island, and that ‘La Baie du Marin’ colony is part of a meta-population at the island level. According to the most recent estimate by direct count, the breeding population of Possession Island is ca 150,000 individuals [29]. Therefore, the population size we estimated from genomic data may reflect this larger meta-population, rather than the single colony. A study investigating the level of connectivity and gene flow among the colonies of Possession Island is currently ongoing.

In our test case, the demographic trend inferred from a single mitochondrial locus was consistent with a similar analysis carried out on 50 nuclear loci (figure 3 and electronic supplementary material, figure S3). Nevertheless, analysing a fairly large number of (presumably) unlinked loci should be considered as best practice as it allows us to average the stochasticity of mutation and drift over many independent genealogies, thus avoiding the risk of a biased demographic inference owing to the idiosyncratic history of a single marker. In addition, when a large number of loci are genotyped, many fewer samples are needed to estimate population-level statistics and obtain demographic inferences [30,31].

Thanks to the inclusion of the mitochondrial marker in our genomic analyses, we were also able to propose an average substitution rate for the loci included in our RAD dataset. Our estimated rate is about one order of magnitude faster than previously reckoned in other bird species [32]. However, recent evidence showed that genomic substitution rates can be faster than formerly estimated from pedigree studies [33,34]. Moreover, the RAD sequencing protocol may introduce biases in the genome representation: when using a digestion enzyme like SbfI, whose restriction site is rich in GC, a selection for GC-rich regions occurs. Indeed, the GC content in our dataset was 50.45% that corresponds to the highest estimates for chicken microchromosomes [35]. Hence, our dataset is probably centred on GC-rich regions (e.g. microchromosomes), which are characterized by increased levels of gene density, recombination rate, number of CpG sites, methylation and mutation rate [36]. Excess of hypermutable CpG nucleotides in GC-rich sequences has been identified as one possible explanation of increased mutation rates [37] so that GC-rich microchromosomes can show a 1.2–1.3× faster substitution rate than the average of the genome [38]. This effect probably contributes, at least in part, at explaining the high mutation rate estimated for our dataset. Even if further investigation is clearly necessary, our interpretation sounds like a cautionary tale about the bias in the selection of genomic regions when using sequence-based restriction enzymes.

(b) King penguin population history during the Last Glacial Maximum

Our genomic data from the Crozet king penguin colony contain a consistent signal of demographic expansion at all levels of variability, with very few outliers in the mismatch analysis. Our demographic reconstruction strongly supports a sudden population expansion following the LGM and starting ca 15,000 years ago. Therefore, our results show that even a cold-adapted species, such as the king penguin, was limited by colder glacial conditions, and benefited from post-glacial warming, which offered suitable breeding sites and foraging conditions in the Crozet region.

Like several other top-level predators in the Antarctic and Sub-Antarctic, king penguins depend on marine ecosystems for food and on terrestrial habitats for reproduction. This species needs a combination of two ecological factors that are both influenced by temperature: (i) year-round ice-free breeding grounds, and (ii) access to major feeding grounds (e.g. the Antarctic Polar Front) within swimming distance and compatible with birds’ breeding duties (exchanging the egg with
of 140 individuals were chosen for mtDNA CR analysis, and eight of these were randomly selected for RAD sequencing analysis. Total DNA was extracted from the filter papers using standard methods and controlled for quantity and quality.

(b) Restriction site-associated DNA sequencing
Genomic DNA from eight king penguins was individually bar-coded, pooled and genotyped by RAD sequencing [19] in one library sequenced on an ILLUMINA HiSeq2000 at the Norwegian Sequencing Centre, University of Oslo. Raw reads were trimmed, demultiplexed and aligned in an unreferenced catalogue using the STACKS software pipeline [47] running on the server facility on the ABEL cluster, University of Oslo. Further quality filtering using custom bash and python scripts was applied to produce the final dataset used in downstream statistical analyses (detailed protocol in the electronic supplementary material).

(c) Genome-wide demographic analysis and calibration
Genomic RAD fragments (95 bp) were sorted into six different classes according to the number of SNPs observed (from 1 to 6 SNPs) and separate analyses were performed for each category. Minor allele frequency spectra were calculated by functions available in the R package ‘adegenet’ [48] using loci included in the 1-SNP class. The signature of the past demography in this class of polymorphism was investigated using the APS analysis in Haplo [23]. Using a diffusion approximation of the APS, this analysis allows demographic inferences to be made from genetic data for testing alternative demographic scenarios in a maximum-likelihood framework. A sudden growth in population size was tested against the null hypothesis of constant population size using the ‘2_epoch’ and the ‘snc’ functions, respectively. The mismatch distribution density (average mismatch distribution of pairwise differences) was then analysed to check for the same demographic pattern in the 2- to 6-SNP classes. Functions included in the R package ‘ape’ [49] and the mismatch distribution density (average mismatch distribution of pairwise differences) was then analysed to check for the same demographic pattern in the 2- to 6-SNP classes. Functions included in the R package ‘ape’ [49] and the R standard boxplot function were used to estimate and plot the mismatch distribution density in each SNP class. Random selections of 50–100 loci in the 2- to 6-SNP classes were used to infer the past demography of the king penguin population using the coalescent-based multi-locus analysis implemented in BEAST v. 1.7.4 [50], setting the EBSP [21] as the tree prior model (see the electronic supplementary material for details). The robustness of the approach was tested with respect to: (i) the number of SNPs per locus, (ii) the different random selection of loci, and (iii) the number of loci included in the random selection. All analyses were run on the Bioportal facility (now LifePortal) of the ABEL cluster, University of Oslo. An estimate proposed for the substitution rate of the mitochondrial CR in the Adélie penguins (0.55 substitutions site$^1$ Myr$^{-1}$ [22]) was used to calibrate our demographic reconstruction. We then plotted the population trend of the last 35 000 years together with the trend of temperature anomalies (in $\Delta$C) as inferred by the analysis of the EPICA Dome C ice core [24]. Concerning the calibration of the mean genome-wide substitution rate: first, the mean of the MCMC posterior median values for each SNP class included in the selected EBSP analysis (4- to 6-SNP classes) was calculated; then, a linear regression was used to infer the substitution rate of those SNP classes excluded from the final EBSP analysis (0- to 3-SNP classes); finally, we calculated the mean genomic substitution rate weighting each SNP class according to the frequency (number of loci) of each class.

(d) Mitochondrial DNA control region analysis
Partial sequences of the CR (354 bp) were amplified and sequenced in 140 samples according to the protocol published in Heupink et al. [51]. Summary molecular statistics (Haplotype diversity: $Hd$; nucleotide diversity: $n$), demographic parameters
as Tajima’s D and Fu’s Fs, and the mismatch distribution of pairwise differences were calculated. This dataset was used to infer the king penguin past demography employing the Bayesian Skyride plot [52] reconstruction, which produces estimates of population size through time and its associated credibility intervals, combining both phylogenetic and coalescent uncertainties, as implemented in BEAST v. 1.7.4. The estimate of the CR substitution rate proposed for Adélie penguins (0.55 substitutions site\(^{-1}\)Myr\(^{-1}\) [22]) was used to calibrate our analyses.

All animals in this study were handled only once in order to mark them with a plastic tag (FloTag®), to blood-sample them, and to conduct morphological measurements. All procedures employed during this fieldwork were approved by the Ethical Committee of the French Polar Institute (Institut Paul Emile Victor – IPEV) and conducted in accordance with its guidelines, also complying with French laws including those relating to conservation and welfare. Authorizations to enter the breeding site (permits n°2009-57 issued on the 26th of August 2009) and handle birds (permits n°2009-59 issued on the 29th of August 2009) were delivered first by the French “Ministère de l’Aménagement du Territoire et de l’Environnement” and then by the Terras Australes et Antarctiques Françaises (TAAF).

Handled animals were removed from the colony in order to minimize the disturbance to neighbouring birds and taken to a few meters away for manipulation. They were hooded to reduce their stress and manipulations lasted between 5 and 10 minutes. All blood-sampling (quantities adjusted according to the age of the chick) and tagging material was sterilized (either sealed, or through chemical sterilization). Moreover, Vertedin soap and alcoholic antiseptic solutions were used to disinfect the skin before bleeding and tagging. Flesh wounds did not seem infected thereafter (personal observations on a subset of recaptured birds).

**Acknowledgements.** We are very grateful to Hélène Gachot-Neveu and Magali Beaugé for their help in the laboratory. We also thank M. Le Vaillant, O. Prudhomme and M. Ripoche for the sample collection. We are very grateful to Sissel Jentoft and core members of the Centre for Ecological and Evolutionary Synthesis, University of Oslo, for the support in implementing the RAD sequencing platform and to Sanne Boessenkool for helpful comments on early versions of the manuscript. We also thank the editors and two anonymous referees for their very helpful comments.

**Data accessibility.** RADseq data: Sequence Read Archive (Run Num.: SRR942341). MIDNA data: GenBank accession nos. KF530582–KFS530702.

**Funding statement.** This study was supported by the Institut Polaire Français Paul-Emile Victor (Programme 137—ECOPHY), the Centre National de la Recherche Scientifique (Programme Zone Atelier de Recherches sur l’Environnement Antarctique et Subantarctique), the Fondation de France/Fondation Ars Cuttoli (to J.D.W.) and Marie Curie Intra European Fellowships (FP7-PEOPLE-IEF-2008, European Commission; project no. 235962 to C.L.B. and FP7-PEOPLE-IEF-2010, European Commission; project no. 252252 to E.T.).

**References**


