Supporting information on data analysis

Sampling, DNA extraction, quality assessment

Sampling was conducted during the 2010/2011 summer season in the king penguin colony of ‘La Baie du Marin’ on Possession Island (46°25'S, 51°45'E) in the Crozet Archipelago. Blood (~ 100 µL) was collected from the brachial vein of chicks hatched in the long-term monitored area 'ANTAVIA', transferred to a filter paper (Whatman 113 ®), dried, and later frozen at -20°C. Individuals were randomly selected along a 120m-axis at the periphery of the colony, in order to maximise distance separation. A total of 140 individuals were chosen for mitochondrial DNA Control Region analysis, and 8 of these were randomly selected for restriction site-associated DNA (RAD) sequencing analysis. Total DNA were extracted from the filter papers using a Phenol-Chloroform protocol or the Qiagen DNAase blood & tissue kit according to manufacturer's instructions. After extraction, DNA quantity and quality were tested in each sample by fluorimetric-based measurement (Qubit, Invitrogen) and gel electrophoresis.

MtDNA marker analysis

Partial sequences of the Control Region (354 bp) were amplified and sequenced in 140 samples according to the protocol published in Heupkin *et al.* [1]. PCR products were Sanger-sequenced in the ABI-LAB at the University of Oslo. Sequences were then manually edited and aligned in Bioedit [2]. All new haplotype sequences have been uploaded to GenBank (Accession number: KF530582-KF530720). Summary molecular statistics, demographic parameters and the mismatch distribution of pairwise differences were calculated in DNAsp v5 [3]. This dataset was used to infer the king penguin past demography employing the Bayesian Skyride plot [4], where inferred population history is bounded by credibility intervals that combine phylogenetic and coalescent uncertainties, as implemented in the BEAST 1.7.4 package [5]. Analyses were performed on the Bioportal facility (now LifePortal) running on the ABEL cluster, University of Oslo. A GTR+G+I substitution model was set for the mitochondrial sequence. A relaxed uncorrelated log-normal clock prior was set for the substitution rate to take into account fluctuations of the molecular clock along different branches of the phylogeny; a log-normal priors with mean in the real space of 0.55 substitution/site/Myr respectively were set [6]. The Bayesian Skyline plot was set as coalescent tree prior model. Convergence among three runs, with a MCMC length of 30 million generations for each parameter setting was checked. Effective sample size was checked in Tracer 1.5 [7] and plots of population size change through time were drawn.

RAD sequencing and genome-wide demographic inference

Eight king penguin individuals were pooled and genotyped by RAD sequencing [8] in one library sequenced on an ILLUMINA HiSeq2000, yielding ca. 65 million 100-bp reads. All raw sequence reads are available on GenBank at the Sequence Read Archive (Run Num.: SRR942341).

After quality assessment, samples showing high molecular weight and highly concentrated DNA were employed in next-generation sequencing (NGS) of RADtags [8]. The following RADseq protocol was adopted: (i) approximately 100ng of genomic DNA per sample were digested with the restriction enzyme *Sbf*I (NEB); (ii) each sample was then ligated to a unique barcoded P1 adapter prior to pooling in a single library. The library was then sheared by sonication, and gel electrophoresis of small library aliquots were run after the first 5 cycles (30'' ON – 30'' OFF) and then every 2-3 cycles of sonication; (iii) the target size range fraction (300-500 bp) was achieved after 8 cycles of sonication and was then selected by gel electrophoresis and manual excision; (iv) before size selection on the gel, sonicated libraries were concentrated to 25 µl by DNA capture on magnetic beads (beads solution:DNA = 0.8:1), thus further reducing the carry-over of non-ligated P1 adapters; (v) capture on magnetic beads using the same beads:DNA ratio (0.8:1) was then employed in all following purification steps (after blunt-end repairing, poly-A tailing, P2 adapter ligation and library enrichment by PCR); (vi) PCR amplification was performed in 8 x 12.5 µl aliquots pooled after the amplification in order to reduce amplification bias on few loci due to random drift; (vii) the library was then quantified by a fluorimetric-based method (Qubit, Invitrogen) and molarity was checked on an Agilent Bioanalyzer chip (Invitrogen). A final volume of 20 µl with a DNA concentration of 45 ng/µl was submitted for sequencing on an ILLUMINA HiSeq2000 sequencer at the Norwegian Sequencing Centre, University of Oslo.

Raw reads were then processed using the scripts included in the Stacks package [9] running on our server facility on the ABEL cluster, University of Oslo. Raw reads were quality filtered and grouped according to individual barcodes. Then individual loci were retrieved and SNPs were called by a maximum-likelihood function that excluded likely sequencing errors. Several runs with different settings of read trimming parameter, quality thresholds, mismatches allowed when building the individual and the population catalogs, were performed to check for consistency of the results. The parameters setting used to build the final catalog included: -t 95 and the default values for the quality checking when using "process_radtags.pl"; -m 10, -n 7, -M 3 when running "denovo_map.pl". 101,115 loci with 50X average coverage were aligned in an unreferenced catalog. A table including all loci matching the eight sequenced individuals was built using "export_sql.pl" Stacks script. This table was further filtered by python scripts (available upon request) excluding loci with missing data, with more than 2 alleles per individual, and deleveraged by Stacks algorithm.

Loci were then grouped according to the number of SNPs allowing from 0 to a maximum of 6 substitutions per locus (0 to 6-SNP classes). Loci with 4-6 SNPs were then directly checked through the catalog web-based interface provided by Stacks. Loci with more than 2 SNPs in the last 5 base pairs or with observed heterozygosity higher than 0.6 were blacklisted and removed from the table as likely sequencing errors or paralogous loci. Only those loci hosting 1 single bi-allelic SNP were employed in AFs analysis in order to minimize linkage among the data. Not having a reference genome, we could not exclude loci produced by adjacent genomic regions or by the two sides of each restriction site. Custom python scripts (available upon request) were employed to edit this 1-SNP dataset as a suitable input file for downstream statistical analysis encoding SNPs as 0-2 when homozygote for the two alleles respectively or 1 when heterozygote. On the other hand, loci in 2 to 6-SNP classes were treated as short sequences and locus-by-locus edited using python script as NEXUS format files each containing 16 sequences 95 bp long (two sequences per individual).

Minor allele frequency spectrum was calculated by functions available in the R package "adegenet" [10] using loci included in the 1-SNP class. This information was then passed to the python-based software ∂a∂i [11] that, using a diffusion approximation to the allele frequency spectrum, allows demographic inference from genetic data 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 “two_epoch” and the “snc” functions, respectively. Several runs of likelihood optimization were performed changing the extent of the search by the “fold” parameter in the “dadi.Misc.perturb_params” function. Optimized log-likelihood and Theta values were recorded. In order to calculate effective population size from Theta values produced by ∂a∂I, a total sequence length of 1,943,510 bp (95 bp X 20,458 loci used in this analysis) was used.

Functions included in the R package “ape” [12] and the R standard boxplot function [13] were used to calculate the joint mismatch distribution in the pairwise differences (from here onwards referred to as mismatch distribution density). Calculations were performed and plotted in each 2 to 6-SNP classes separately.

Different random combinations of 50-100 loci in 2 to 6-SNP classes were compared when inferring the past demography of the king penguin population using the coalescent-theory based multi-locus analysis implemented in BEAST 1.7.4. Linkage disequilibrium was tested in all subsets using Genepop [14] with the default setting in the web tool and the Bonferroni correction for multiple tests. 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: 50 loci in 2-SNP class (5 runs), 50 loci in 3-SNP class (5 runs), 50 loci in 4-6-SNP class (10 runs), 10 loci in 4-6-SNP class (1 run), 25 loci in 4-6-SNP class (1 run) and 100 loci in 4-6-SNP class (1 run). Three runs showing hints of multiple optima for the demographic function were discarded. Different settings of the parameters and priors have been explored in preliminary analyses, but the following was the definitive setting: (i) markers were unlinked concerning site substitution model, clock model and tree prior model; (ii) site substitution model was set as a HKY with empirical base frequency; (iii) a strict molecular clock was estimated for each marker with a uniform prior distribution bounded within 0.5 and 0.005 sub/s/Myr; (iv) the Extended Bayesian Skyline Plot (EBSP; [15]) was selected as tree prior model and ploidy of the markers was set accordingly. Fine tuning of operators did not improve our results as running the analyses with longer MCMC simulations; 200 million iterations were set as run length.

Mitochondrial Control Region data were included in the analysis in order to test the consistency of the information provided by the two genetic dataset (genomic and mitochondrial) and to calibrate the genomic substitution rate using the mtDNA Control Region substitution rate as estimated in the Adélie penguin. Site substitution and clock models were set as in the analyses of the mitochondrial marker alone (see below). All analyses were run on the Bioportal facility (now LifePortal) of the Abel cluster, University of Oslo. Results were checked on Tracer 1.5 and plot of the EBSP data were drawn in R [13]. An extensive study on Adélie penguin ancient DNA suggested a fast estimate (0.88 sub/s/Myr; [16]) for the substitution rate of the mitochondrial Control Region. Further analyses confirmed this high figure but it was downscaled to 0.55 sub/s/Myr [6]. We used the more conservative 0.55 sub/s/Myr for our calibrated demography. A generation time of 11.49 years (Le Bohec *in prep*) was used to convert the population size estimates on the EBSP (given by default in effective population size * generation time). We then plotted the population trend for the last 35,000 years together with the trend of temperature anomalies as inferred by the analysis of the EPICA Dome C ice core [17]. Concerning the calibration of the mean genome-wide substitution rate: first, the mean of the median values in each SNP-ratio class included in the EBSP analysis was calculated (4 to 6-SNP classes); then, a linear regression was used to infer the substitution rate of those SNP classes excluded from the EBSP analysis (0 to 3-SNP classes); finally, we calculated the mean genomic substitution rate weighting each SNP class accordingly with the frequency (number of loci) of each class.

References

1. Heupink T. H., van den Hoff J., Lambert D. M. 2012 King penguin population on Macquarie Island recovers ancient DNA diversity after heavy exploitation in historic times. *Biol. Lett.* **8**, 586–589.

2. Hall T. A. 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.* **41**, 95-98.

3. Librado P., Rozas J. 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452.

4. Minin V. N., Bloomquist E. W. & Suchard M. A. 2008 Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Mol Biol Evol* **25**, 1459–1471.

5. Drummond A. J. & Rambaut A 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol*

Biol **7**, 214.

6. Millar, C. D., Dodd, A., Anderson, J., Gibb, G. C., Ritchie, P. A., Baroni, C., Woodhams M. D., Hendy M. D., Lambert D. M. 2008 Mutation and evolutionary rates in Adélie penguins from the Antarctic. *PLoS Genetics*, **4**, e1000209.

7. Rambaut A. & Drummond A. 2005 "Tracer version 1.3: a program for analyzing results from Bayesian MCMC programs such as BEAST and MrBayes." Distributed by the authors at: http://evolove. zoo. ox. ac. uk/software. html

8. Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. & Johnson, E. A. 2008 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* **3**, e3376. (doi:10.1371/journal.pone.0003376)

9. Catchen J. M., Amores A., Hohenlohe P., Cresko W. & Postlethwait J. H. 2011 Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3* **1**, 171–182.

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

11. Gutenkunst R. N., Hernandez R. D., Williamson S. H. & Bustamante C. D. 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet* **5**, e1000695.

12. Paradis E., Claude J. & Strimmer K. 2004 APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290.

13. R Development Core Team 2011 R Foundation for Statistical Computing, Vienna, Austria URL http://www.R-project.org.

14. Raymond M. & Rousset F. 1995 GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Heredity, 86, 248-249

15. Heled J. & Drummond A.J. 2008 Bayesian inference of population size history from multiple loci. *BMC Evol Biol* **8**, 289.

16. Lambert, D. M., Ritchie, P. A., Millar, C. D., Holland, B., Drummond, A. J., & Baroni, C. 2002 Rates of evolution in ancient DNA from Adélie penguins. *Science*, **295**, 2270-2273.

17. Jouzel, J., Masson-Delmotte, V., Cattani, O., Dreyfus, G., Falourd, S., Hoffmann, G., Minster B., Nouet J., Barnola J. M., Chappellaz J. *et al* 2007 Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science*, **317**, 793-796.

Figure S1. Demographic reconstructions of the Crozet king penguin colony employing the Extended Bayesian Skyline Plot analysis. Consistency in the pattern inferred is compared among different data selections including 50 loci chosen at random from different classes of variation: a) 2 SNPs, 4 independent datasets; b) 3 SNPs, 4 independent datasets; c) 4-6 SNPs, 9 independent datasets. In order to facilitate comparison of uncalibrated EBSP reconstruction and solely for visualization purpose, all runs were scaled to have the same value of the demographic function at t0. To do so, we divided all demographic estimates by the ratio Ki = Ni(t0)/max(Ni(t0)), where Ni(t0) is the median value from the posterior distribution of the demographic function at t0 for each run. Correspondingly, the time intervals of each run were multiplied by the same ratio Ki, reflecting the assumption that the actual population size is the same across all runs. Median (black), 95% HPD lower (red) and upper (green) values are reported.

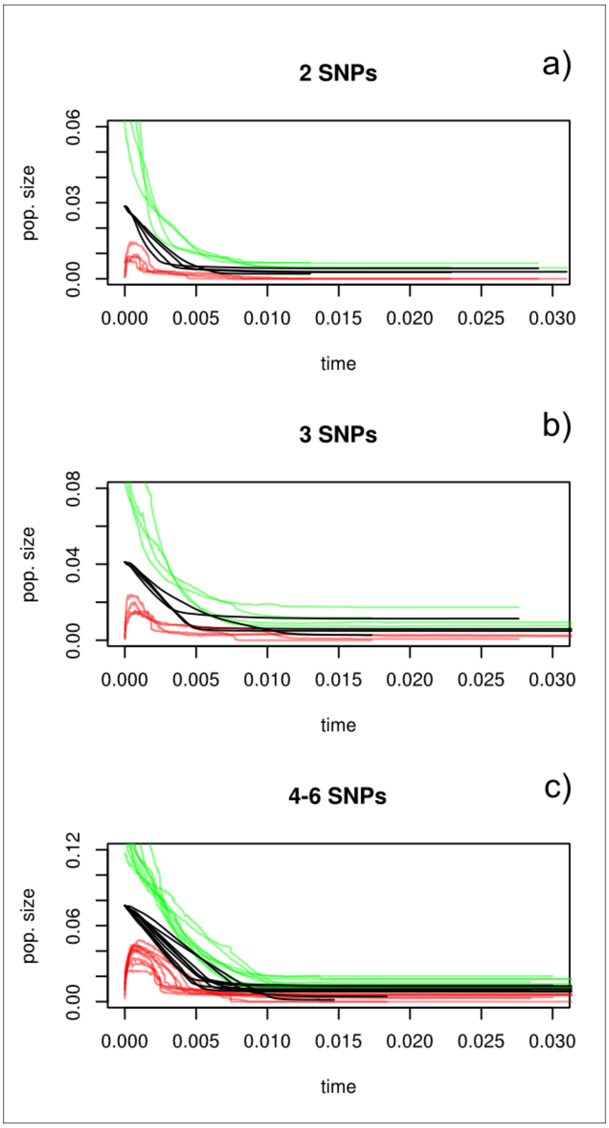


Figure S2. Demographic reconstructions of the Crozet king penguin colony employing the Extended Bayesian Skyline Plot analysis. Consistency of the inference is compared across three nested datasets (10, 25 and 50 loci) randomly selected from the 4-6-SNP class. Population size and time are unscaled. Median (solid line) and 95% upper and lower (filled areas) values are reported.

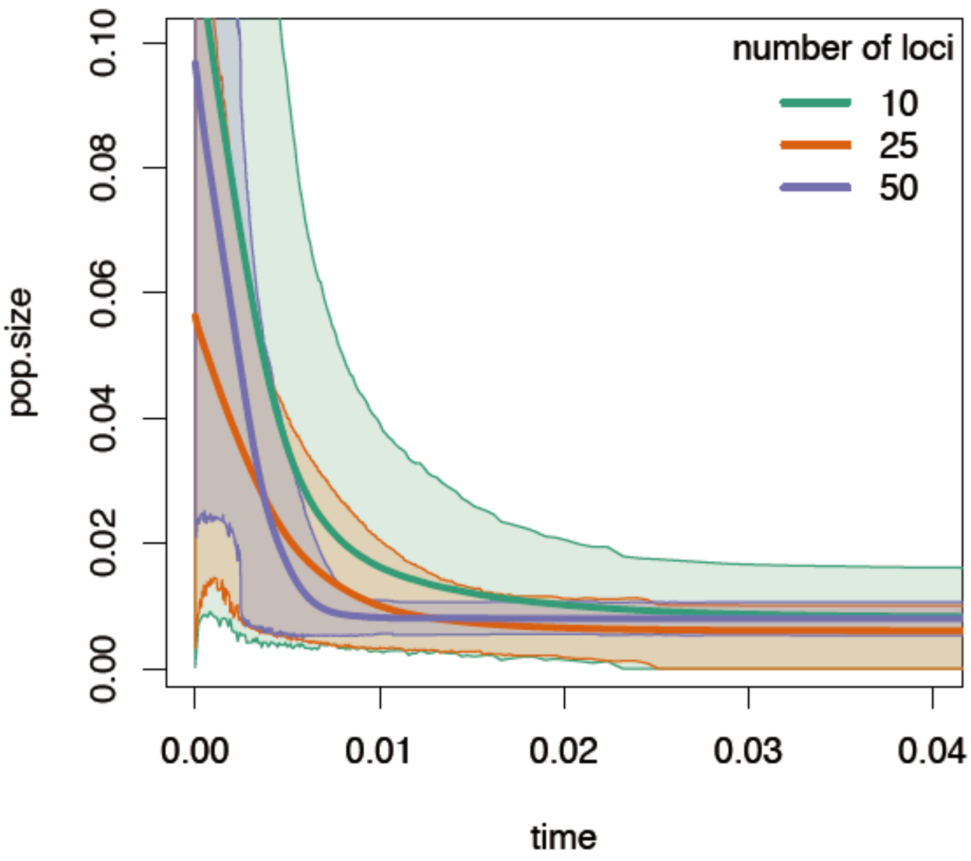
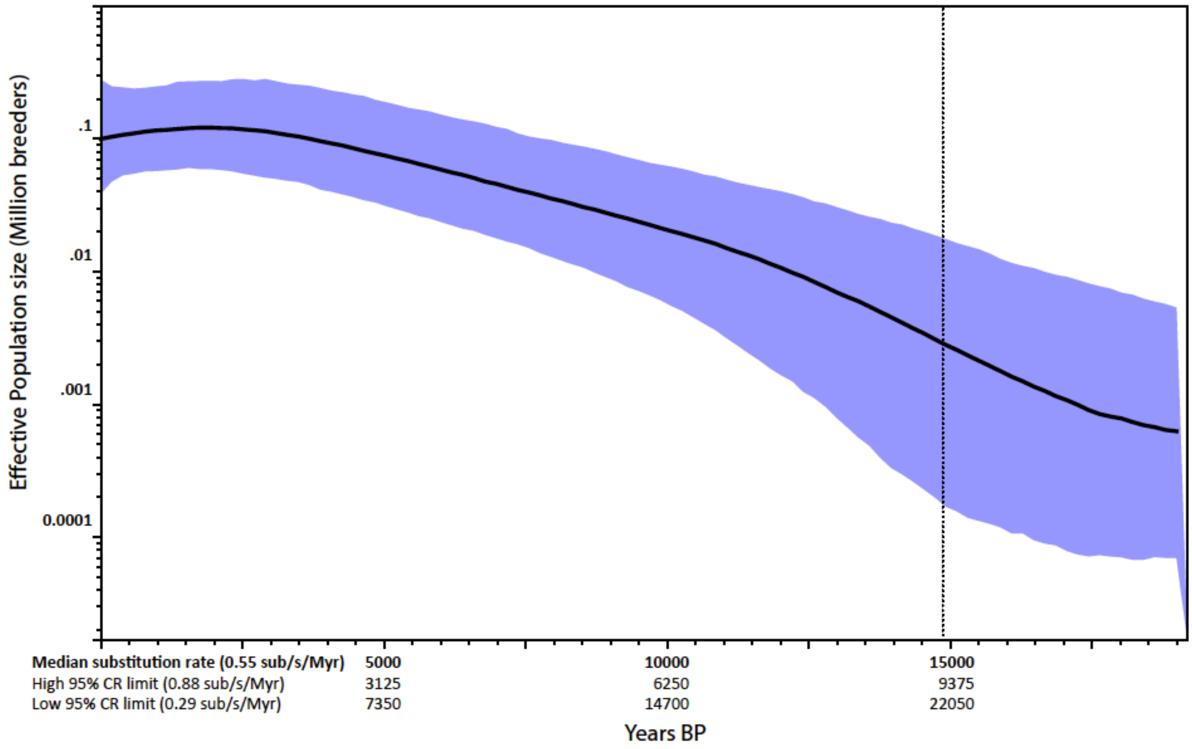


Figure S3. Bayesian Skyride Plot inferred from mitochondrial Control Region data describing population trend through time. Time is scaled according to median (solid black line), 95% upper and lower credibility region (filled blue area), as estimated in Millar *et al* [6].


Effective Population size (Million breeders)
Median substitution rate (0.55 sub/s/Myr)
High 95% CR limit (0.88 sub/s/Myr)
Low 95% CR limit (0.29 sub/s/Myr)
Years BP