nature portfolio

Corresponding author(s): Juliana A. Vianna

Last updated by author(s): Mar 24, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection. Genomic data were obtained through standard sequencing protocols (Illumina platform) and public repositories. Morphological data were collected manually from museum specimens using calipers following standardized protocols. Geographic coordinates for species distribution modeling were compiled from the Global Biodiversity Information Facility (GBIF) and published literature.
Data analysis	We used the following tools and software for data processing, analysis, and visualization: Genome processing and variant calling: BWA-mem2 v2.2.1 for read mapping Samtools v1.15.1 for file manipulation Picard MarkDuplicates v2.23.8 for duplicate removal GATK v3.8 for local realignment around indels bcftools v1.12 for variant calling bedtools v2.30.0 for coverage masking and BED file generation SNP filtering and population structure: PLINK v1.9 for SNP filtering, LD pruning, and PCA VCFtools v0.1.16 for MAF filtering and heterozygosity estimation ADMIXTURE v1.3.0 for ancestry estimation Dsuite v0.4 for D-statistics and introgression analysis TreeMix v1.13 for modeling ancestral gene flow

Selection and adaptation analysis:
RAiSD v2.9 and nSL for selective sweep detection
bedtools for intersecting SNPs with outlier windows
Redundancy Analysis (RDA) using vegan and stats packages in R
Phylogenomics and divergence time estimation:
MAFFT v7.505 for sequence alignment
IQ-TREE v2.2.0 for maximum likelihood gene trees
ASTRAL v5.7.3 for species tree inference
StarBEAST3 v1.0.0 (BEAST2.6.6) for divergence time estimation using multispecies coalescent models

Ecological niche modeling and statistics:
R (v4.2.0) with biomod2 for species distribution modeling
Python (v3.9) with Jupyter Notebooks for statistical analysis and figure generation

Figure editing:
Adobe Illustrator 2023 for final figure layout and graphical editing

All software used is publicly available, except for Adobe Illustrator, which was used exclusively for figure preparation. All custom scripts and input files are available in the GitHub repository: https://github.com/dalynoll/Gentoo-penguin-adaptive-divergence.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

GenBank accession of genomes, scripts and pipelines for data analysis were stored in https://github.com/dalynoll/Gentoo-penguin-adaptive-divergence

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study investigates adaptive divergence and species delimitation in the Gentoo penguin complex, historically considered a single species. Using whole-genome resequencing, morphometric analyses, genotype-environment associations, and ecological niche modeling, we tested whether divergent lineages represent distinct evolutionary units. Our integrative approach identified signatures of positive selection, genomic discontinuities, ecological differentiation, and morphological divergence aligned with oceanographic and climatic gradients across the Southern Ocean. Based on these findings, we propose a taxonomic revision recognizing four species of Gentoo penguins.
Research sample	All genomic samples used in this study were previously collected and published in earlier works (e.g., Vianna et al. 2017; Pertierra et al. 2020), with the exception of a small number of newly collected samples from Macquarie Island. These were included to complete the geographic representation of Gentoo penguin lineages. The same blood or DNA samples were re-used to perform whole-genome resequencing. Samples were selected based on DNA integrity and geographic representativeness across historically sampled breeding colonies. While DNA quality and coverage varied among samples, particularly in some localities such as Macquarie Island, sequencing results were sufficient to include these individuals in population-level genomic and phylogenomic analyses. Morphological data were partially obtained from previously published studies (e.g., Tyler et al. 2020), and complemented with new measurements taken from additional museum specimens representing under-sampled regions (e.g., Macquarie, Crozet, and South Orkney Islands). All measurements were collected using standardized protocols consistent with those applied in previous work to ensure comparability.
Sampling strategy	No formal sample size calculation was performed. Instead, we included all available individuals with sufficient DNA quality from previously collected samples. Individuals and localities were selected based on their representation of the full geographic and ecological diversity of the Gentoo penguin complex, and to build upon prior phylogeographic studies (e.g., Vianna et al. 2017; Pertierra et al. 2020). Our goal was to ensure lineage representation across the Southern Ocean. Some regions, such as the South Sandwich and Heard/McDonald Islands, were not included due to lack of available samples. Samples from certain localities, such as Macquarie and South Georgia Islands, were included in phylogenomic analyses but excluded from some population-level analyses due to the low number of individuals available. This decision was made to ensure statistical robustness and avoid overinterpretation of underrepresented populations. However, genomic data from these samples were retained for broader analyses of lineage divergence, species tree inference, and niche modeling. Two individuals from Macquarie Island were newly sampled to complement the geographical representation of gentoo penguin lineages. These samples were obtained from naturally deceased individuals in the context of population monitoring from archived collections, collected under Department of Primary Industries, Parks, Water and Environment Tasmania staff permits. No live animals were captured or handled as part of this study.
Data collection	Most DNA samples used for genome sequencing were originally collected during previous field campaigns and published in earlier studies (e.g., Vianna et al. 2017; Pertierra et al. 2020). These archived samples were re-used in this study for whole-genome resequencing. However, a small number of new samples were obtained from Macquarie Island to complete the geographic representation of Gentoo penguin lineages, which were obtained from naturally deceased individuals in the context of population monitoring from archived collections. Morphological data were compiled from both published measurements (Tyler et al. 2020) and additional measurements taken by JY from museum specimens housed in international collections. All geographic and taxonomic metadata were validated by the authors based on field records, sample provenance, and expert knowledge of regional Gentoo penguin populations.
Timing and spatial scale	Samples were collected over a span of several years between 2015 and 2021, as part of different expeditions and collaborative projects. The earliest samples were obtained in 2015, with most fieldwork conducted prior to the publication of Vianna et al. (2017). Two additional samples from Macquarie Island were collected more recently to complete the geographic representation. Despite the temporal dispersion, all samples represent adult individuals during the breeding season and were selected to provide a representative snapshot of the main breeding colonies. The study covers ten distinct localities across the Southern Ocean, encompassing both Subantarctic and Antarctic environments, with sampling sites separated by thousands of kilometers.
Data exclusions	No individuals were excluded from the genomic dataset based on sequencing quality. However, samples from certain localities (e.g., Macquarie and South Georgia Islands) were excluded from some population-level analyses due to the low number of individuals available per site, which limited statistical power. For South Georgia, a single high-quality genome assembly available in GenBank (Cole et al. 2022) was used in phylogenomic analyses to represent this lineage. All available samples were retained in broader analyses of phylogeny, divergence time estimation, and ecological niche modeling to ensure lineage-level representation.
Reproducibility	Key analyses such as population structure (PCA and ADMIXTURE), selective sweep detection (RAiSD and nSL), and phylogenomic inference (IQ-TREE and ASTRAL) were re-run using filtered datasets to confirm consistency of results. The code and scripts used for data filtering, analysis, and figure generation are available in a public GitHub repository: https://github.com/dalynoll/Gentoo-penguin-adaptive-divergence, allowing full reproducibility of the computational workflow.
Randomization	Randomization was not applicable to this study. Sample inclusion was based on DNA integrity, availability, and geographic representation rather than random assignment. This is consistent with standard practices in evolutionary genomics, where the aim is to characterize natural population structure and divergence across predefined geographic or ecological units.
Blinding	Blinding was not applicable to this study. Analyses were based on genomic and ecological data from non-experimental samples.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Most samples used in this study were collected prior to 2020 and reused for whole-genome sequencing. Field work was conducted during austral summer breeding seasons, under typical Subantarctic and Antarctic conditions including low temperatures, high humidity, and logistical constraints associated with remote field sites. According to field records and prior publications, sampling followed ethical guidelines and was carried out during brief handling procedures to minimize disturbance to the animals.
Location	Samples were collected across ten Gentoo penguin colonies throughout the Southern Ocean, including sites in the Falkland/Malvinas Islands, Martillo Island (Argentina), Antarctic Peninsula, South Georgia, Kerguelen, Crozet, Marion, and Macquarie Islands. With the exception of Macquarie Island, all samples were originally collected and published in previous studies (e.g., Vianna et al. 2017; Pertierra et al. 2020) and were reused in the present study. The two samples from Macquarie Island were newly collected to complete geographic coverage. Exact coordinates and site-level metadata are provided in Supplementary Table S1.
Access & import/export	All samples were collected and transported under appropriate national and international permits, as documented in previous publications and institutional records. Newly collected samples from Macquarie Island were obtained with explicit collection and export permits from the Australian government. The Acknowledgments section of the manuscript includes details on institutional support and collaborators who facilitated sample access, collection, and handling across all sites.
Disturbance	All samples were obtained following established wildlife handling protocols to minimize disturbance. Blood collection was performed quickly and with minimal stress to the animals, typically during the breeding season. No long-term behavioral effects or environmental disruptions were reported during or after sampling. For the newly collected samples from Macquarie Island, similar low-impact protocols were followed in accordance with local regulations and permit requirements.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☐ Clinical data
☐	☐ Dual use research of concern
☐	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	The study did not involve laboratory animals.
Wild animals	The study involved wild Gentoo penguins sampled across multiple colonies in the Southern Ocean. Blood samples were collected by trained personnel during the austral summer breeding season, following ethical handling protocols. Most samples were collected and reported in previous studies (e.g., Vianna et al. 2017; Pertierra et al. 2020), with the exception of two new samples obtained from Macquarie Island. All individuals were released immediately after sample collection. No animals were held in captivity.
Reporting on sex	Sex was identified bioinformatically based on read depth patterns in BAM files, distinguishing between male (ZZ) and female (ZW) individuals using scaffolds associated with sex chromosomes. While sex was determined for all individuals, it was not a primary factor in the study design and was not used to stratify analyses. The study focused on geographic lineage divergence rather than sex-based variation.
Field-collected samples	The study involved field-collected blood or DNA samples obtained from wild penguins during their breeding season. No laboratory housing, maintenance, or experimental intervention was involved beyond DNA extraction and genomic analysis.
Ethics oversight	All samples were collected under institutional, national, and international permits and protocols, as described in the Supplementary Materials and cited publications. Ethical approval and guidance were obtained from relevant authorities, including INACH (Chile), IPEV (France), TAAF, PROANTAR (Brazil), and the Australian Antarctic Division for Macquarie Island. The handling of animals followed approved procedures to minimize stress and comply with wildlife welfare standards.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*