Received: 29 May 2024 | Revised: 12 August 2024 | Accepted: 2 September 2024

DOI: 10.1111/1755-0998.14022

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Exome capture of Antarctic krill (*Euphausia superba*) for cost effective genotyping and population genetics with historical collections

Oliver W. White[1,2,3] | Sarah Walkington[1] | Hugh Carter[1] | Lauren Hughes[1] | Melody Clark[4] | Thomas Mock[5] | Geraint A. Tarling[4] | Matthew D. Clark[1]

[1]The Natural History Museum, London, UK

[2]NERC Environmental Omics Facility (NEOF), NEOF Visitor Facility, School of Biosciences, University of Sheffield, Sheffield, UK

[3]NERC Environmental Omics Facility, Centre for Genomic Research, University of Liverpool, Liverpool, UK

[4]British Antarctic Survey, Cambridge, UK

[5]School of Environmental Sciences, University of East Anglia, Norwich, UK

Correspondence
Matt Clark, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.
Email: matt.clark@nhm.ac.uk

Funding information
NERC Environmental Omics Facility (NEOF), Grant/Award Number: NEOF1430; Internal Natural History Museum London Science Investment Fund; Natural Environmental Research Council

Handling Editor: Simon Creer

Abstract

Antarctic krill (*Euphausia superba* Dana) is a keystone species in the Southern Ocean ecosystem, with ecological and commercial significance. However, its vulnerability to climate change requires an urgent investigation of its adaptive potential to future environmental conditions. Historical museum collections of krill from the early 20th century represent an ideal opportunity to investigate how krill have changed over time due to predation, fishing and climate change. However, there is currently no cost-effective method for implementing population scale collection genomics for krill given its genome size (48 Gbp). Here, we assessed the utility of two inexpensive methods for population genetics using historical krill samples, specifically low-coverage shotgun sequencing (i.e. 'genome-skimming') and exome capture. Two full-length transcriptomes were generated and used to identify 166 putative gene targets for exome capture bait design. A total of 20 historical krill samples were sequenced using shotgun and exome capture. Mitochondrial and nuclear ribosomal sequences were assembled from both low-coverage shotgun and off-target of exome capture data demonstrating that endogenous DNA sequences could be assembled from historical collections. Although, mitochondrial and ribosomal sequences are variable across individuals from different populations, phylogenetic analysis does not identify any population structure. We find exome capture provides approximately 4500-fold enrichment of sequencing targeted genes, suggesting this approach can generate the sequencing depth required to call identify a significant number of variants. Unlocking historical collections for genomic analyses using exome capture, will provide valuable insights into past and present biodiversity, resilience and adaptability of krill populations to climate change.

KEYWORDS
bioinformatics, exome capture, genome skimming, museomics

1 | INTRODUCTION

Antarctic krill (*Euphausia superba*, hereafter krill) is the most successful wild living animal species on the planet in terms of population biomass (300–500 million tonnes; Atkinson et al., 2009) and a keystone species in the Southern Ocean ecosystem. Krill has a circumpolar distribution covering the entire Southern Ocean, with krill swarms forming some of the largest aggregations of animal life recorded to date. Krill's significance extends far beyond its abundance. It plays a pivotal role in the food web of the Southern Ocean, as a consumer of phytoplankton and prey for charismatic megafauna including penguins, seals and whales. The Southern Ocean is one of the largest carbon sinks in the world, and krill plays a key role in the carbon cycle, removing up to 40 million tonnes annually (Cavan et al., 2019). In addition, it has become increasingly commercially important, supporting the krill industry's catch of >$200M annually for omega-3 dietary supplements and aquaculture feed (Tou et al., 2007).

The polar regions are thought to be most at risk from climatic warming, with Southern Ocean temperature increases of more than double the global average (Meredith et al., 2019), especially around the sub-Antarctic Islands and along the Antarctic Peninsula. Rapid warming is likely to have profound implications for marine species distributions and ocean ecosystem functioning, especially polar species, which are cold-adapted and have a low tolerance to fluctuating temperatures (Peck et al., 2004; Portner, 2002). Critically, krill is a stenothermic species, adapted to a narrow temperature range between −2 and 5°C (McBride et al., 2021), making it particularly sensitive to climate change. Indeed, over the past 90 years, the range of Antarctic krill has contracted southward towards colder waters (Atkinson et al., 2019). Recent models suggest a temporal shift in habitat quality for krill in the Antarctic peninsula, with habitat quality improving in spring and declining in summer and autumn (Veytia et al., 2020). Such temporal shifts in habitat quality may affect krill population dynamics, by creating a disjunction between the annual cycle of the Antarctic environment and current krill life cycles, with knock on effects in other species (Veytia et al., 2020).

Thus, determining the population genetics of krill, and how this species will respond to future climate change, is crucial for both ecosystem functioning and fisheries management. For example, there may be distinct sub-populations of krill, with varying susceptibility or adaptation to rapidly changing environmental conditions (Tarling, 2020), such as the retreat of sea ice. An understanding of population structure would aid stock management through rotating fishing quotas across populations, allowing genetic diversity and resilience to be maintained, and avoiding overfishing of subpopulations. Indeed, a key aim of the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) and the SCAR Krill Expert Group (SKEG) is developing a 'krill stock hypothesis'. For example, fisheries may be targeting populations of krill living away from ice in warmer waters that are more easily fished, but these could be the populations with the greatest potential to adapt to climate change. However, to date, we do not fully understand krill population structure, to what degree sub-populations are interconnected and able to replenish spatial regions adversely affected by climatic anomalies such as Southern Annular Modes and El Nino/La Nina. Without this understanding, we cannot fully determine how krill will cope with environmental change in combination with fishing and other human impacts on the Southern Ocean.

Previous population genetics studies of krill have largely relied upon a small selection of markers from allozyme variation (Fevolden & Schneppenheim, 1989) and mitochondrial DNA (Goodall-Copestake et al., 2010; Zane et al., 1998), but these only consider single mitochondrial genes and were not informative of variation at the genomic level. A reduced representation approach (RAD-seq) has been employed to investigate population structure of krill (Meyer et al., 2015), but this study has significant limitations because most genetic data were from multicopy genomic regions. However, recently, a chromosome level assembly of the krill genome has been generated (Shao et al., 2023), and resequencing of 75 individuals from multiple populations confirmed limited population structure but identified 387 sites associated with environmental variables.

There are limited molecular markers available for krill, likely due to the lack of genomic resources until recently (Shao et al., 2023). Genome resequencing for population genetics studies is becoming increasingly plausible as the cost of sequencing continues to fall. However, for species with large genomes such as krill (48 Gbp; Jeffery, 2012), resequencing is cost-prohibitive due to the amount of sequence data required to identify variants confidently. Thus, there is a need for cost-effective sequencing options to study populations genetics such as (1) low coverage shotgun sequencing that is "genomes skims" or (2) exome capture. Low coverage genome skims are used to assemble multicopy components of the genome including organelle genomes or ribosomal genes which are sequenced at a higher coverage (Straub et al., 2012). Low coverage genome skims have been widely employed to assemble organelle and/or ribosomal genes from museum collections, including marine Solariellid gastropods (White et al., 2024) and Eurepini crickets (He et al., 2024). Exome capture uses probe sequences to enrich for specific targets which are typically expressed genes that is exons (Mascher et al., 2013). Exome capture has been applied successfully to other large repetitive genomes, notably barley (Mascher et al., 2013; Hordeum vulgare 5Gb) and conifers (Suren et al., 2016; Pinus; 18–35 Gb). However, to date, there is no cost-effective method for implementing population genetics at the scale required for krill.

Natural history collections offer an invaluable resource to understand how krill will respond to climate change. Historical collections allow research to look back in time, investigating population change and adaptation, which is invaluable for forecasting future change. For example, the Natural History Museum London is home to c. 20,000 krill spirit preserved (ethanol) accessions, with trawl-net samples suitable for population studies covering the last 130 years. The collections are also home to samples of great historical importance, including those collected during Scott's Discovery expedition in 1901–1904. These represent potential flagship collections with which to investigate the impact of climate change in a

keystone species in the Southern Ocean ecosystem. Working with high throughput sequencing data from museum specimens has until recently been challenging since DNA is typically fragmented and contaminated with non-endogenous sequences, restricting their use in genomics studies. However, advances in DNA isolation, a reduction in cost of DNA sequencing and the availability of novel bioinformatics tools means it is increasingly possible to use museum samples for genomic analyses (Burrell et al., 2015).

This study investigates the relative utility of (1) low coverage shotgun sequencing and (2) exome capture for population genetic analyses with historical collections of krill. Target sequences for exome capture are identified using two full-length transcriptomes for krill, generated using PacBio IsoSeq data from recently collected samples. We hypothesize (1) shotgun sequencing will allow the assembly of mitochondrial genomes, which are present in high copy number, and (2) exome capture will enable nuclear gene assembly with sufficient depth for population genetic analyses.

2 | METHODS

2.1 | PacBio IsoSeq sequencing

Samples were collected using an RMT8 net (Baker et al., 1973) in the vicinity of South Georgia on 11th December 2019 (Cruise JR19001, Event 65 Net 1, maximum depth) on board the RRS James Clark Ross. Samples were immediately flash frozen within liquid nitrogen and stored at −80°C. Two samples, one male (29659_1) and one female (29659_4) were shipped to NERC Environmental Omics Facility (Sheffield, UK). Samples were dissected for muscle tissue before isolating RNA using a QIAGEN RNeasy Mini kit (QIAGEN, Manchester, UK). RNA quality and integrity was checked before library construction and sequencing using four Sequel II SMRT Cells. High quality transcripts were generated from raw data for each sample using the manufacturer's (PacBio, California, United States) SMRTlink pipeline.

2.2 | Target selection and bait design

The high-quality transcripts generated from PacBio IsoSeq data were used to identify putative target sequences. For high-quality transcripts recovered from each sample, the following analyses were performed: (1) BUSCO search against core Eukaryotic genes (eukaryota_odb10; 10/09/2020 using the transcriptome mode), (2) BLASTN (Camacho et al., 2009) search against annotated transcript sequences downloaded from KrillDB[2] (Urso et al., 2022) and (3) Blobtools (Laetsch et al., 2017) analysis to define transcript taxonomy based on a BLASTN search against the NCBI nucleotide database (nt, downloaded June 2022) and the taxrule "bestsumorder". Transcripts were prioritized for exome capture if they had a (1) BUSCO hit to a core eukaryotic gene, (2) blast hit description from KrillDB[2] suggesting a role in environmental responses (i.e. heat, cold, temperature sensitive) or core homeobox genes, and finally (3) a sequence identified as being from *Euphausia superba* based on the blobtools output. Putative targets from each sample were combined and redundant sequences were identified and removed using cd-hit-est with a similarity threshold of 0.8 (Fu et al., 2012). Note we also avoided targeting sequences with annotations suggesting they originated from the mitochondrial genome or genes with putative repetitive elements (e.g. microsatellites) as these are expected to occur at a higher frequency and will bias the sequencing of these targets.

The putative target sequences were shared with Daciel Arbor Biosciences and tested for bait design suitability. Specifically, targets were softmasked for simple and low complexity repeats and baits were designed using 80 nucleotide (nt) probes and 4× tiling (i.e. one probe every ~20 nt). Note that the krill reference genome (Shao et al., 2023; CNGB CNP0001930; accessed March 2023) was published shortly after our baits were designed. Therefore, our baits design was from cDNA sequences and could not account for intron-exon boundaries and genome mappability (i.e. genome uniqueness and the likelihood of mapping short 80 nt baits to genomic regions). All target sequences and bait sequences were later mapped to the *E. superba* reference genome using minimap and bwa mem respectively. To investigate the impact of mappability for bait design, genome mappability was quantified using GenMap (Pockrandt et al., 2020) with a kmer size of 36 and no sequence errors for the krill genome and a pre-indexed human genome (GRCh38) provided with GenMap for comparison.

2.3 | Shotgun and exome capture sequencing

A total of 27 samples were collected from the NHM spirit-preserved collection with the aim of sampling different station numbers from the early historical samples collected from 1926 to 1979 (Table 1; Figure 1). DNA was isolated in a BSC hood cleaned with 5% bleach (w/v) and UV sterilization. Krill were rinsed in sterile molecular grade water for 12–24 h, and the lower abdominal segments were dissected using a sterile scalpel and prepared for overnight lysis (Ruane & Austin, 2017). After lysis, any remaining tissues were ground with a sterile micro pestle. DNA was isolated using a modified method from Ruane and Austin (2017) with double quantities of binding buffer. DNA samples were shipped to Daicel Arbor BioSciences (Ann Arbor, United States) where total DNA was re-quantified via a spectrofluorimetric assay and visualized using the TapeStation 4200 (Agilent) platform with a High Sensitivity D1000 tape. Samples containing high molecular weight DNA or no visible DNA were sonicated to generate an average insert size of approximately 300 nt before taking up to 80% of the available mass (up to 5 ng) to a single-stranded library preparation protocol that produces dual-indexed Illumina-compatible libraries. A total of 20 samples were selected for downstream shotgun and exome capture sequencing based on total genomic DNA mass (Table 1) and TapeStation plots were visualized. For the shotgun sequence data, 20 libraries (Table 1) were pooled in equimolar

TABLE 1 Historical samples from the NHM spirit-preserved collection used for shotgun and exome capture with information for collection station number, source (Discovery/William Scoresby), year, net type, depth (start-end depth), location, total genomic DNA mass and if the sample was used for sequencing.

Ref.	Station	Collection	Year	Net	Depth (m)	Location	Latitude	Longitude	gDNA (ng)	Sequenced
K1	1311	Discovery	1934	N100H	5-0	Bellingshausen sea	−67.313	−82.9	13.5	Y
K2	539	Discovery	1930	n70B	137-0	South Shetland Islands	−61.8	−54.858	18.5	Y
K3	2365	Discovery	1938	TYFB	350-200	South Atlantic	−53.39	4.842	38	Y
K4	125	Discovery	1926	N100H	70	South Georgia	−53.475	−36.342	26.1	Y
K5	9968	Discovery	1979	RMT8M-3	980-1990	NA	−58.872	20.347	177.1	Y
K6	2139	Discovery	1937	N100H	5-0	Prydz bay	−56.58	95.057	501.8	Y
K7	539	WS	1931	N100B	100-0	South Georgia	−57.692	−23.2	111.8	Y
K8	1404	Discovery	1934	N100B	117-0	South Georgia	−53.997	−39.465	37.1	Y
K9	538	WS	1931	N100B	97-0	South Georgia	−57.058	−24.533	100.1	Y
K10	1405	Discovery	1934	N100B	110-0	South Georgia	−54.002	−40.123	99.2	Y
K11	542	WS	1931	N100B	77-0	South Georgia	−58.65	−18.217	115.4	Y
K12	1297	Discovery	1934	N100H	5-0	Amunden sea	−70.42	−129.262	87.1	Y
K13	42	Discovery	1926	N7-T	120-204	South Georgia	NA	NA	28.1	Y
K14	622	Discovery	1931	N100-B	155-0	South Georgia	−59.092	−36.417	45.9	Y
K15	643	Discovery	1931	N100B	93-0	South Shetland Islands	−61.742	−56.117	44.6	Y
K16	NA	Discovery	1926	HN	0-1	NA	NA	NA	171	Y
K17	568	WS	1931	N100B	106-0	South Georgia	−53.621	−37.3	25	Y
K18	53	WS	1927	N100H	0-5	South Georgia	NA	NA	14.8	N
K19	39	WS	1926	N100H	87	South Georgia	−54.133	−35.717	6.2	N
K20	30	WS	1926	100H	0-5	South Georgia	−53.571	−38.604	5.3	N
K21	36	WS	1926	N100H	77	South Georgia	−55.338	−34.775	49.3	Y
K22	542	WS	1927	N100H	70	South Georgia	−58.65	−18.217	11	N
K23	43	WS	1927	N100H	70	South Georgia	−54.9	−36.833	11	N
K24	1309	Discovery	1934	N100B	106-0	Bellingshausen sea	−67.765	−89.392	14.6	Y
K25	53H	WS	1927	N100H	0-5	South Georgia	NA	NA	9.9	N
K26	53O	WS	1927	N100H	0-5	South Georgia	NA	NA	8	N
K27	658	Discovery	1931	N100B	120-0	South Georgia	−53.638	−40.454	39.2	Y

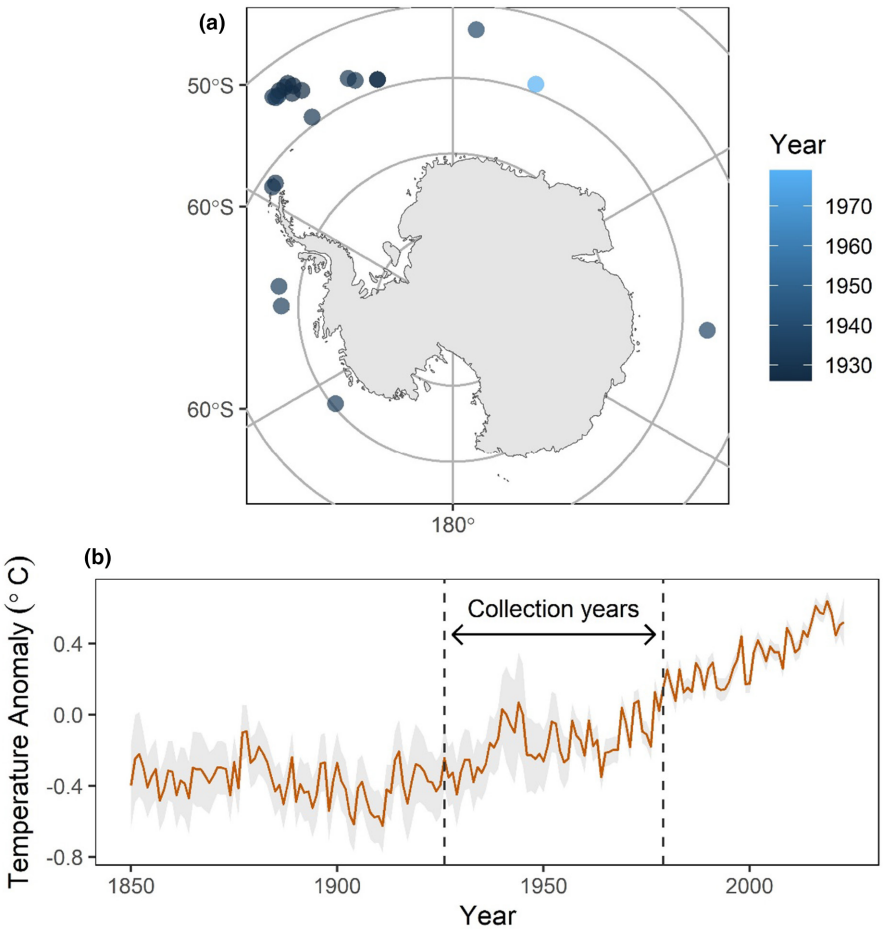

FIGURE 1 (a) Location of krill (*Euphausia superba*) sampling locations with points coloured by collection year and (b) the Southern Ocean surface temperature relative to 1961–1990 mean, annotated with the minimum and maximum collection years of samples used in this study (1926–1979). The sea surface temperature data were downloaded from the Met Office Hadley Centre observations datasets (accessed April 2023).

ratios and sequenced on a partial NovaSeq 6000 S4 PE150 lane, targeting approximately 1M read pairs per sample. Capture pools were prepared from up to 250 ng of 6 libraries per reaction for the historical samples. Each capture pool was dried down to 7 μL by vacuum centrifugation. Captures were performed following the myBaits v5.02 protocol using myBaits custom design (myBaits design ID: D10573KRILL) with an overnight hybridization and washes at 62°C. The captures were pooled in approximately equimolar ratios. For the exome capture sequence data, 20 samples (Table 1) were sequenced on the Illumina NovaSeq 6000 platform on a partial S4 PE150 lane to approximately 2M read pairs per library.

2.4 | Assembly of mitochondrial and ribosomal sequences

To evaluate the utility of shotgun and exome capture data for the assembly multi-copy sequences, custom bioinformatic pipelines skim2mito 0.0.1 and skim2rrna 0.0.1 were used to assemble and annotate mitochondrial genomes and ribosomal genes respectively (White et al., 2024). In addition to the 20 samples sequenced for this study, we also analysed a subset (1M reads per FTP file; Table S1) of publicly available data for 78 individuals sampled by Shao et al. (2023). This additional data includes 75 *E. superba* and three outgroup samples *E. pacifica*.

The skim2mito and skim2rrna pipelines each follow a similar methodology, with adapter removal and quality control with fastp 3.3.6 (Chen et al., 2018), target sequence assembly using GetOrganelle 1.7.7.0 (Jin et al., 2020) with custom seed and gene reference databases downloaded and formatted using go_fetch.py (https://github.com/o-william-white/go_fetch). Assembly quality was evaluated by a BLASTN search (2.13.0 Camacho et al., 2009) against custom databases. For mitochondrial sequences, a custom blast database was generated from the NCBI mitochondrion RefSeq database (https://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/), and for ribosomal sequences, the blast database was generated from the SILVA 138 database (Quast et al., 2013). Assembly quality was also evaluated by mapping quality filtered reads to the assembled sequence using minimap 2.24 (Li, 2018). Blast hits and mapped reads were summarized with blobtools (Laetsch et al., 2017) with the most likely taxonomy of the assembled sequence estimated using the taxrule "bestsumorder". Assembled mitochondrial sequences were annotated using MITOS2 2.1.0 (Bernt et al., 2013), and assembled ribosomal sequences were annotated using barrnap 0.9 (https://github.com/tseemann/barrnap). Annotated genes were extracted and aligned with mafft 7.508 (Katoh & Standley, 2013). Note that only protein coding genes mitochondrial genes were

extracted by skim2mito for downstream analyses. Individual sequences with ≥50% missing data in gene alignments were removed, and poorly aligned regions were trimmed with gblocks 0.91b (Castresana, 2000). Phylogenetic analysis was implemented for each gene using IQTREE 2 2.2.0.3 (Minh et al., 2020) with 1000 ultrafast bootstrap replicates. Note that IQTREE 2 uses ModelFinder (Kalyaanamoorthy et al., 2017) by default to optimize model selection on individual genes. Phylogenetic trees were visualized with the R package ggtree (R Core team, 2020; Yu et al., 2017).

The outputs of skim2mito and skim2rrna were evaluated manually to remove samples and/or genes with excessive missing data and evidence of contamination from non-target species. Specifically, samples with more than 50% missing data across all annotated genes were identified and removed. In addition, any individual gene alignments with more than 50% of missing data across samples were removed. Finally, assembled sequences from non-target contaminant species were identified based on visual inspection of gene trees and blobtools taxonomy identification. With the final selection of annotated gene sequences, gene2phylo 0.0.1 (White et al., 2024) was implemented for phylogenetic analysis with all annotated mitochondrial and ribosomal genes using a partitioned analysis by individual genes with IQTREE 2 with 1000 ultrafast bootstrap replicates and an analysis of gene trees using ASTRAL-III (Zhang et al., 2018).

Using the final set of genes, the number and percentage of parsimony informative sites was calculated with phykit (Steenwyk et al., 2021). The similarity of assembled shotgun and exome data from the same samples was measured to ensure that the results were consistent, using percentage similarity of the best blast hit (Camacho et al., 2009). Mitochondrial assembles were visualized using Circos 0.69–8 and a custom python script (https://github.com/o-william-white/circos_plot_organelle) to check gene order, read coverage, GC content and repeat content.

2.5 | Targeted gene sequencing

The utility of shotgun and exome capture for targeted gene sequencing was also compared using the sequence data generated by the present study. Specifically, raw sequence data generated from shotgun and exome capture were initially processed with adapterremoval2 2.3.3, to remove adapter sequences, trim low-quality bases and merge overlapping reads. Trimmed pairs of reads which were not collapsed, trimmed singleton reads where one mate was discarded, merged reads and merged reads that have been trimmed were concatenated into single file. Concatenated reads were mapped to the *E. superba* reference genome (CNGB: CNP0001930) using BWA-mem 0.7.17-r1188. SAMtools 1.14 (Danecek et al., 2021) was used to filter primary mapped reads (-F 2308). Duplicate reads were tagged using picard MarkDuplicates 3.1.0 before collecting hybrid-selection metrics using picard CollectHsMetrics (Broad Institute, 2019). The output of CollectHsMetrics was used to compare fold enrichment and mean coverage of targeted sequences. The number of reads that mapped to on or off-target regions were quantified using bedtools mutlicov 2.30.0. SNP calling was implemented using bcftools mpileup 1.16 (Danecek et al., 2021) for shotgun and exome capture bam files, before filtering for sites with a minimum quality of 30 and depth of 5. Note that SNP calling was implemented using all shotgun or all exome capture BAM files, rather than per individual BAM file.

To estimate insert size across sampled collections, adapterremoval2 was repeated without collapsing reads, and paired reads were mapped with BWA-mem as above, before estimating mean insert size with SAMtools. To investigate the extent and composition of any contaminants, kraken2 2.1.2 was used to identify unmapped reads using the NCBI non-redundant nucleotide database (accessed 14/11/2022). To understand where sequence data were retained or lost throughout each step of the pipeline, Sankey plots were generated for each sample using the R package networkD3.

To investigate the sequencing effort required for accurate variant calling with exome capture data, a single pool of four samples (Table 1; K8, K10, K21 and K24) was sequenced again to approximately 160M read pairs. Note that sample K21 was poorly represented within this pool with a low proportion of reads and was not presented in the results. The sequence data from each sample were randomly subsampled at regular proportions (1%, 5%, 10%–100%) to investigate the effect of sequencing effort and processed as above.

3 | RESULTS

3.1 | PacBio IsoSeq sequencing

The number of high-quality isoforms generated using the SMRTlink pipeline were 115,591 and 248,047 for the male (29659_1) and female (29659_4) sample respectively. The number of low-quality isoforms was 18 and 40 for the male (29659_1) and female (29659_4) sample respectively. The transcript N50 ranged from 2115 (29659_1) to 2018 (29659_4). Each sample exhibited a similar proportion of BUSCO gene categories (Figure 2a), with 60.4% and 70.9% of complete BUSCO genes recovered for the male and female sample respectively. The majority of BUSCO genes identified as "complete and single copy" or "complete and duplicated" (61%) that were unique to the female sample was 61% (29659_4; Figure 2b), with 36% shared and 3% unique to the male sample (29659_1; Figure 2b). The sequence taxonomy as defined by blobtools was most frequently identified as the class Malacostraca (Figure 2c).

3.2 | Target selection and bait design

A total of 175 transcripts were initially selected as putative target sequences, identified as core BUSCO eukaryotic genes (64), having a blast description suggesting a role in environmental responses or core homeobox genes (65) or being identified as a gene from *Euphuasia superba* (46). Nine transcripts were removed for having a high similarity to other putative target sequences or being a putative repetitive sequence. A final set of 166 target sequences with a

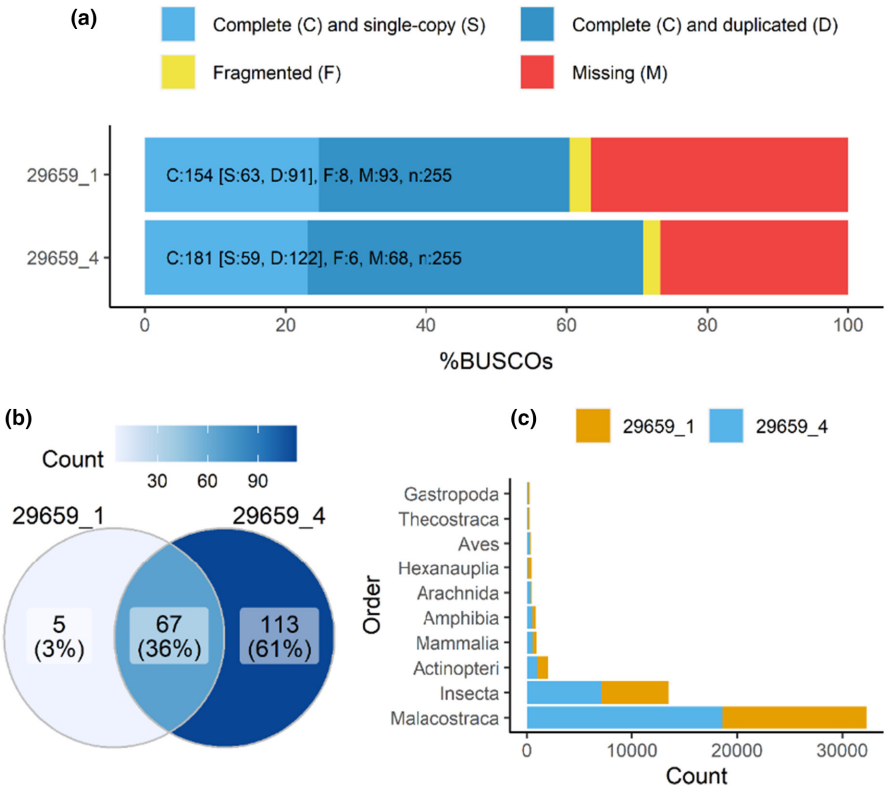

FIGURE 2 Summary of high-quality isoforms generated for the male (29659_1) and female (29659_4) krill samples. (a) Proportion of BUSCO categories identified, (b) Venn diagram of complete or duplicated BUSCO genes identified across samples and (c) number transcripts identified across the ten most frequent classes.

total target size of 395,924 nt were selected and tested for bait deign suitability.

A total of 18,468 baits were designed using 80 nt probes and a maximum of 4× tiling. Of these, 8228 baits passed filters based on softmasking for repeats. Following bait design and filtering, 43 target sequences were completely covered by baits, 120 targets were covered by baits with gaps up to 100 nt and three targets had no baits. With this bait design, 301,711 nt from a total of 395,924 nt (~75%) in the target sequences could be targeted.

All 166 target sequences mapped to the *E. superba* reference genome (CNGB: CNP0001930). Of the 8228 of filtered baits, a total of 6411 baits mapped to the reference genome. Of these, 3241 mapped to a target sequence. Mean target coverage (i.e. proportion of bases covered by a least one bait) was 38.8%.

Genome mappability was quantified for the krill genome (CNGB: CNP0001930) and a pre-indexed human genome (GRCh38) provided with GenMap to investigate the impact of mapping short DNA fragments. A mappability value of 1 indicates that the k-mer sequence occurs only once whereas a low mappability value indicates that this k-mer belongs to a repetitive region. Our analysis highlighted that 57.61% of the krill genome has a mappability ≤0.5, for comparison only 14.79% of the human genome has a mappability ≤0.5 (Figure S1).

3.3 | Assembly of mitochondrial and ribosomal sequences

Of the 20 samples newly sequenced for this study, GetOrganelle assembled mitochondrial sequences from 12 (60%) and 11 (55%) samples for shotgun and exome capture data respectively. Of these, six assemblies (30%; Figure S2; Table S2) were circular for the shotgun data, whereas only one sample was circular for the exome capture data (5%). GetOrganelle assembled ribosomal sequences from 14 (70%) and 10 (50%) samples for shotgun and exome capture data respectively (Table S2). Of the 78 samples previously sequenced by Shao et al. (2023), we were able to assemble mitochondrial and ribosomal sequences from all samples.

Following manual inspection of the outputs of skim2mito and skim2rrna, a total of 10 shotgun and 13 exome capture samples were removed for having more 50% missing data across annotated mitochondrial and ribosomal genes. In addition, the nuclear 5.8S gene was removed as it had more than 50% missing sequences across samples. Two contigs assembled from the sample "SSI03_South_Shetland_Islands" with *cox1* annotations were identified as likely human contaminants and removed. In addition, eight nuclear contigs with ribosomal 18S and 28S annotations were removed due to likely contaminations from non-target species. This resulted in a final set of 17 genes including 13 mitochondrial protein-coding genes, two mitochondrial ribosomal genes and two nuclear ribosomal.

Analysis of number and percentage of parsimony informative sites (Table S3) shows that most genes show some degree of variation. For example, *cox1* had 261 (17%) parsimony informative sites. However, the partitioned phylogenetic analysis using IQ-TREE2 (Figure S3) and an analysis of gene trees using ASTRAL-III (Figure S4) suggested there was little evidence of population structure based on these genes.

The shotgun and exome capture data from the same samples generated assemblies with high BLASTN sequence similarity (Table S2;

>98%). The high similarity of assemblies generated by shotgun and or exome capture data from the same samples is confirmed by our phylogenetic analyses, where samples sequenced using shotgun data and exome capture were identified as sister species with maximum bootstrap support values (Figures S3 and S4).

3.4 | Targeted gene sequencing

A total of 154M and 58M raw paired reads were generated for shotgun and exome capture sequencing respectively (Table S4). After processing raw reads with adapterremoval2 to remove adapter sequences, trim low-quality bases and merge overlapping reads, a total of 146M and 63M reads were retained for downstream analysis of shotgun and exome capture sequencing respectively. Note that trimmed pairs of reads which were not collapsed, trimmed singleton reads where one mate was discarded, merged reads and merged reads that have been trimmed were concatenated into single file. As a result, the number of quality filtered reads produced by adapterremoval2 can be higher than the number of raw paired reads which is the case for exome capture. The mean number of quality filtered reads per sample was 7.3M and 3.2M for shotgun and exome capture respectively. A mean of 2.6M (35.64%) and 1.23M (38.91%) of quality filter reads mapped accurately to the krill reference genome for the shotgun and exome capture data respectively.

Of the unmapped reads, the most frequent (top five) contaminants as identified by kraken2 were the families Nitrobacteraceae (39/44 samples), Burkholderiaceae (38/44), Sphingomonadaceae (33/44), Hominidae (31/44) and Suidae (23/44) (Data Dryad https://doi.org/10.5061/dryad.v6wwpzh4p). Of the mapped reads, 0.43M (16.52%) and 0.60M (48.81%) were identified as duplicates for shotgun and exome capture data respectively.

The mean insert size of mapped reads was 41.51 and 48.81 bp for shotgun and exome capture respectively. Insert size exhibited a positive correlation with the year of sample collection (Figure 3a) for both shotgun and exome capture sequence data, with the exception of a single sample (K5) which was collected more recently in 1979. The mean fold enrichment was 0.76 and 3463 for shotgun and exome capture sequencing respectively, representing mean fold change of ~4500 for exome capture (Figure 3b). For the shotgun sequence data, increasing sequencing effort (i.e. total reads) did not appreciably increase mean target coverage, whereas for the exome capture data, mean target coverage increased with increasing sequencing effort (Figure 3c). However, sample age did not appear to be correlated with mean target coverage. Sankey plots to visualize where sequence data were retained or lost throughout each step of the pipeline are presented on Data Drayd (https://doi.org/10.5061/dryad.v6wwpzh4p).

No SNPs were identified for the target sites using the shotgun sequencing. For the exome data, an average of 34 SNPs were identified across samples (Figure 3d). However, the proportion of missing data was relatively high with an average of 66% of missing sites across samples.

To investigate the sequencing effort required to call SNPs accurately across samples, a single pool of three samples (Table 1; K8, K10 and K24) was sequenced again and mapped reads were subsampled at regular intervals (1%, 5%, 10%–100%) using SAMtools view --subsample. Increasing the sequencing effort increased the number of on target SNPs, although the number of SNPs discovered starts to plateau with increased sequencing effort (Figure 3e). It is notable that mean target coverage was not uniform across samples, even with all the additional sequence data generated (Figure S5). Indeed, mean target coverage shows a strong correlation with bait coverage (Figure 3f).

4 | DISCUSSION

Despite the ecological and commercial significance of Antarctic krill, there are few genetic resources available for this species, and therefore there is only limited knowledge about krill population structure. The availability of historical museum specimens offers the unique opportunity to link changes in krill population structure with environmental change, which may allow researchers to assess the fate of this keystone species in the future. To meet this aim, our study assessed the relative utility of two cost-effective methods for population genetic analyses frequently adopted for historical museum collections: low coverage genome skimming and exome capture. To design our exome capture bait sequences, we generated full-length transcriptomes sequenced from recently collected samples.

Mitochondrial and ribosomal sequences were assembled using shotgun and exome capture. However, shotgun data were more effective at recovering complete mitochondrial and ribosomal sequences, which is to be expected since targeted exome capture has reduced the off-target sequencing. In addition, the number of reads generated by shotgun sequencing (mean 7.3M per sample) was higher than exome capture sequencing (mean 3.2M per sample). Multicopy parts of the genome such as organelle genomes and ribosomal tandem repeats are common targets for "genome skimming" studies because these regions are sequenced at a higher depth compared to the rest of the genome (Straub et al., 2012). Although genes were variable based on the proportion of parsimoniously informative sites, there was little evidence of population structure, corroborating previous work based on the *cox1* gene (Goodall-Copestake et al., 2010).

Shao et al. (2023) identified some population structure based on a large whole genome SNP dataset, suggesting that the identification of many nuclear genomic variants holds promise for future studies. In addition, the identification of nuclear genomic variants from museum collections may identify alleles associated with selection across time for environmental change. Although low-coverage shotgun data may not hold utility for large scale population genomic studies in this system, it is likely that shotgun sequencing will be useful for pre-screening historical samples to confirm the presence of endogenous sequence data from the target organism and levels of DNA degradation.

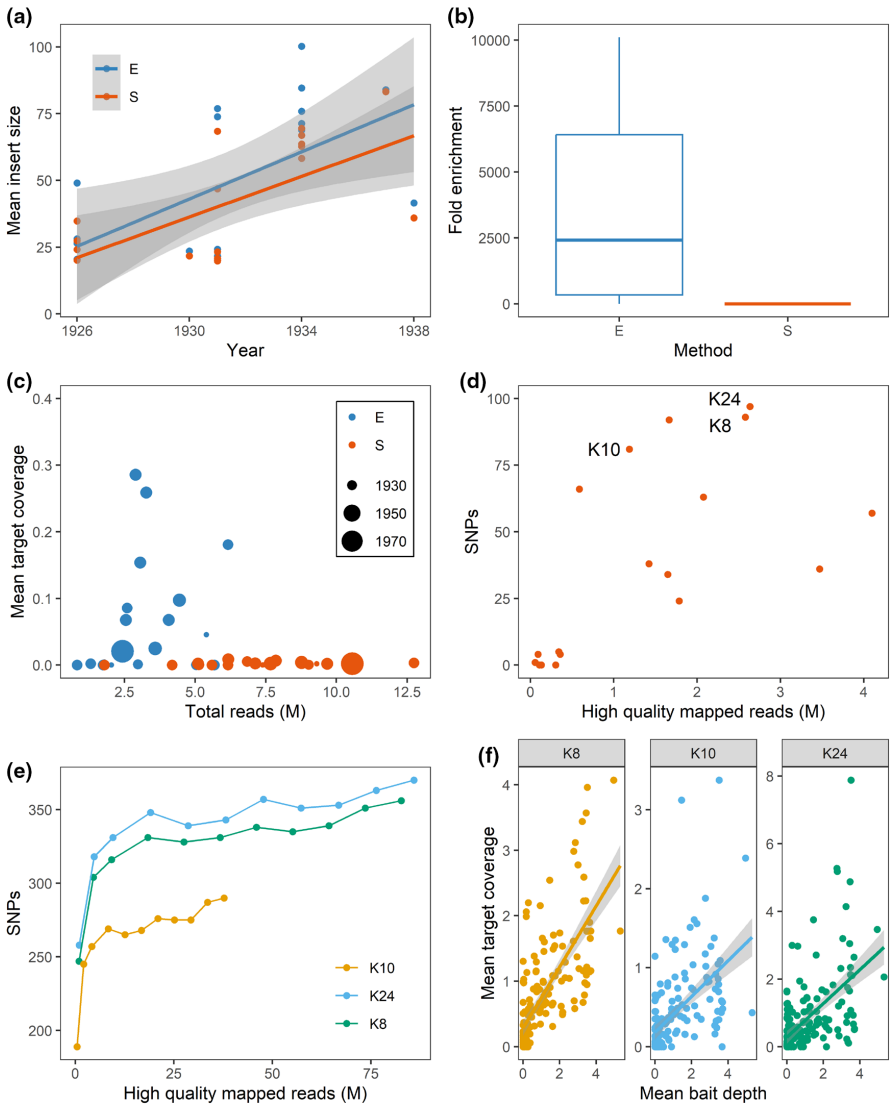

FIGURE 3 Summary of shotgun and exome capture sequence data for targeted regions from ethanol fixed samples. (a) Relationship between sample age and mean insert size for shotgun [S] and exome capture [E]. Note that sample K5 [Year 1979 and mean insert size 37.8–41.5] was excluded as this sample was preserved using formalin and had a shorter mean insert size relative to the collection year. (b) Boxplot of fold enrichment for shotgun and exome capture. (c) Relationship between total sequencing reads and mean target coverage, with point size showing sample age. (d) Relationship between the number of SNPs and high-quality mapped reads for exome capture data. The position of three pooled samples sequenced at higher coverage are annotated. (e) Relationship between high-quality mapped reads and the number of SNPs for three additional sequence datasets sub-sampled to measure the effect of increased sequencing effort. (f) Relationship between mean bait depth and mean target sequence coverage for three samples sequenced at a higher coverage.

This study generated full length transcriptome sequences for two recently collected samples, which were instrumental in the selection of target sequences and subsequent bait design. The transcriptome sequences were not complete, with only 60.4% and 70.9% of complete BUSCO genes recovered, which may be explained by our use of dissected muscle tissue for RNA sampling.

Exome capture sequencing resulted in 4500-fold increase of target sequence coverage relative to shotgun sequence data (Figure 3a). As a result, it was possible to identify on-target variant SNPs using exome capture data (Figure 3d), while shotgun data yielded no on target SNPs. Although it was possible to call SNPs with the exome capture data, there was a high proportion of missing data in the SNPs initially identified. Additional sequencing of a single pool of samples highlighted that increasing the sequencing effort increases the number of SNPs identified, although this starts to plateau with increasing sequencing effort (Figure 3e). A plateau in sequence complexity and subsequent SNPs is to be expected (Daley & Smith, 2013), especially for historical sequence datasets.

Given the highly repetitive nature of the krill genome (Shao et al., 2023), it may not be possible to design short bait sequence that map specifically to all protein coding genes. For example, we were able to design baits for 76% of nucleotides in the original gene target list. In addition, the mappability of short sequences in the krill genome was found to be much lower than the human genome. Specifically, our analysis highlighted that 57.61% of the krill genome has a mappability ≤0.5, whereas only 14.79% of the human genome has a mappability ≤0.5 (Figure S1).

Although our study highlights that exome capture can be successfully applied to historical museum collections, there are clear differences in DNA sample and sequence quality. Spirit (or wet) collections are typically stored in liquid preservatives including ethanol and may have been fixed with formalin prior to storage which may damage DNA (Ruiz-Gartzia et al., 2022). Recent studies investigating the utility of spirit-preserved collections for genomic studies have highlighted that overall specimen condition had the greatest impact on recovering high quality genomic DNA (Hahn et al., 2022; O'Connell et al., 2022; Straube et al., 2021). In our study, it is notable that insert size exhibited a positive correlation with the year of sample collection (Figure 3a) for both shotgun and exome capture sequence data, with the exception of a single

sample (K5) collected in 1979. This sample had a shorter insert size than expected (37.8 shotgun, 41.5 exome), which may have been due to this sample being fixed in formalin instead of ethanol. Although formalin has been used in museum specimen preservation shortly after it became commercially available in the late 1890's/early 1900's, its use was primarily restricted to soft bodied taxa until ~1950's when bulk formalin fixing of samples became more frequently employed. The early Discovery Investigation collections (1926–1939) making up the bulk of the material tested in this study were entirely preserved in 75% ethanol (Kemp et al., 1929). Therefore, it is plausible that the older historical collections will have greater utility for the future population genetic studies because more recent material is likely formalin fixed, requiring more complex laboratory techniques due to DNA cross-linking and damage (Hykin et al., 2015; Ruane & Austin, 2017).

4.1 | Future work

Future studies attempting to design bait sequences for exome capture and use for population genetics could build on the lessons learned from this study. The most important developments would be to improve the bait design, use knowledge from recently published genome sequences, and to pre-screen samples with shotgun sequencing prior to exome capture to identify libraries which are most likely to be successful. Target sequence coverage was not uniform across samples suggesting some targets were preferentially sequenced over others (Figure S5). The variation in target sequence coverage may be explained by differences in bait coverage. Indeed, there was a positive association between target coverage and bait coverage (Figure 3f), suggesting that changing bait design only to include targets with a minimum threshold of bait depth (e.g. ≥4) would increase target recovery. At the time of bait design for the present study, the version of the krill genome assembly available was a contig level assembly with 298,755 contigs, rather than the 17-chromosome reference level assembly described by Shao et al. (2023). Therefore, our baits design was from cDNA sequences and could not account for intron-exon boundaries, which may have negatively impacted bait specificity at intron-exon boundaries. In addition, our investigation of genome mappability, highlighted that a large proportion of the krill genome has low mappability due to repetitive sequences (Figure S1). Where possible, target sequences with low mappability of short bait sequences should be avoided. Pre-screening historical samples with low-coverage genome-skims will also allow researchers to identify which samples are most likely to yield sequence libraires with sufficient complexity to call SNPs effectively. For example, there was a strong correlation between the number of uniquely mapped reads from shotgun and exome capture libraries (Figure S6). In addition, mitochondrial assembly status also appeared to be associated the number of uniquely mapped exome capture reads, except for a single sample. Specifically, samples with contig and circular mitochondrial genomes, typically had a higher number of uniquely mapped reads.

5 | CONCLUSION AND OUTLOOK

This study provides first evidence that it is possible to isolate DNA from historical krill collections and that these samples can be used for genomic and genetic analyses. In addition, lessons learn from our analyses will improve the efficiency of future work through the understanding of bait design and sample quality requirements. This study has important implications for the utility of historical spirit-preserved collections, including the Discovery collections of the 1920s and 30s. For example, future work could investigate (1) what historical krill diversity looked like prior to the onset of anthropogenic climate change and widespread fishing, (2) how contemporary krill diversity is being impacted by climate change and predation pressure, (3) the relationship between the biological characteristics of krill and temperature and (4) how adaptable krill are to climatic change. Considering the central importance of krill in the food web of the Southern Ocean and as the most dominant animal on Earth considering its biomass, the implications for conservation and fisheries management are potentially profound. Although, our study has focused on krill, the approach could also be applied to a broader range of taxa.

AUTHOR CONTRIBUTIONS

OW and MC designed the study. GT provided the samples required for full length transcriptomes and LH and HC provided access the historical samples for sampling, shotgun and exome capture sequencing. SW performed the wet lab work. OW wrote the first draft of the manuscript; all co-authors contributed to the preparation of the final manuscript.

ACKNOWLEDGEMENTS

The lab work was performed by staff of the NERC Environmental Omics Facility (NEOF) funded by the UK Natural Environmental Research Council (NERC). The authors acknowledge Gavin Horsburgh, Lucy Knowles (NEOF Visitor Facility, University of Sheffield, UK) for the RNA extraction and preparation of the sample for sequencing; and Xuan Liu (NEOF, University of Liverpool, UK) for PacBio sequencing and bioinformatic processing of the PacBio RNAseq data. Probe design, ssDNA library builds, hybridization capture and sequencing were organized by Daicel Arbor Biosciences (Ann Arbor, United States). The authors also acknowledge Jennifer Klunk from Daicel Arbor Biosciences for technical support and advice regarding Daicel Arbor Biosciences services. We would also like to acknowledge project collaborator Simeon Hill (British Antarctic Survey, Cambridge, UK) for comments on the manuscript.

FUNDING INFORMATION

This research was funded by the NERC Environmental Omics Facility (NEOF) Early Career Researchers Pilot project competition (NEOF1430) and an Internal Natural History Museum London Science Investment Fund. The NERC Environmental Omics Facility (NEOF) is funded by the UK Natural Environmental Research Council (NERC).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Raw sequence data and assembled mitochondrial and ribosomal sequences are accessible from the European Nucleotide Archive (ENA) under project accession PRJEB77367. Note that assembled mitochondrial and ribosomal sequences from shotgun and exome capture libraries for each sample were identical where they overlapped. However, the assembled sequences from shotgun data were typically longer and more complete. Therefore, only the assembled sequences from the shotgun data were uploaded to ENA. Configuration files, reference data used for the Snakemake pipelines skim2mito, skim2rrna and gene2phylo are available on Data Dryad (https://doi.org/10.5061/dryad.v6wwpzh4p). Assembled sequences from previously published data generated by Shao et al. (2023) are also available on Data Dryad. The exome capture bait design is available from Daicel Arbor BioSciences as a myBaits custom design kit (myBaits design ID: D10573KRILL).

BENEFITS SHARING STATEMENT

Benefits generated: The benefits from this research accrue from the sharing of our code, data and results on public databases as described above.

ORCID

Oliver W. White https://orcid.org/0000-0001-6444-0310
Hugh Carter https://orcid.org/0009-0004-7642-347X
Lauren Hughes https://orcid.org/0000-0002-5679-1732
Melody Clark https://orcid.org/0000-0002-3442-3824
Thomas Mock https://orcid.org/0000-0001-9604-0362
Geraint A. Tarling https://orcid.org/0000-0002-3753-5899
Matthew D https://orcid.org/0000-0002-8049-5423

REFERENCES

Atkinson, A., Hill, S. L., Pakhomov, E. A., Siegel, V., Reiss, C. S., Loeb, V. J., Steinberg, D. K., Schmidt, K., Tarling, G. A., Gerrish, L., & Sailley, S. F. (2019). Krill (*Euphausia superba*) distribution contracts southward during rapid regional warming. *Nature Climate Change*, *9*(2), 142–147. https://doi.org/10.1038/s41558-018-0370-z

Atkinson, A., Siegel, V., Pakhomov, E. A., Jessopp, M. J., & Loeb, V. (2009). A re-appraisal of the total biomass and annual production of Antarctic krill. *Deep-Sea Research Part I: Oceanographic Research Papers*, *56*(5), 727–740. https://doi.org/10.1016/j.dsr.2008.12.007

Baker, A. d. C., Clarke, M. R., & Harris, M. J. (1973). The N.I.O. Combination net (RMT 1 + 8) and further developments of rectangular midwater trawls. *Journal of the Marine Biological Association of the United Kingdom*, *53*(1), 167–184. https://doi.org/10.1017/S0025315400056708

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., & Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Broad Institute. (2019). *Picard toolkit*. Broad Institute.

Burrell, A. S., Disotell, T. R., & Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution*, *79*, 35–44. https://doi.org/10.1016/j.jhevol.2014.10.015

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. https://academic.oup.com/mbe/article/17/4/540/1127654, https://doi.org/10.1093/oxfordjournals.molbev.a026334

Cavan, E. L., Belcher, A., Atkinson, A., Hill, S. L., Kawaguchi, S., McCormack, S., Meyer, B., Nicol, S., Ratnarajah, L., Schmidt, K., Steinberg, D. K., Tarling, G. A., & Boyd, P. W. (2019). The importance of Antarctic krill in biogeochemical cycles. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-12668-7

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, *10*(4), 325–327. https://doi.org/10.1038/nmeth.2375

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Fevolden, S. E., & Schneppenheim, R. (1989). Genetic homogeneity of krill (*Euphausia superba* Dana) in the Southern Ocean. *Polar Biology*, *9*(8), 533–539. https://doi.org/10.1007/BF00261038

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Goodall-Copestake, W. P., Pérez-Espona, S., Clark, M. S., Murphy, E. J., Seear, P. J., & Tarling, G. A. (2010). Swarms of diversity at the gene cox1 in Antarctic krill. *Heredity*, *104*(5), 513–518. https://doi.org/10.1038/hdy.2009.188

Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., & Holleley, C. E. (2022). Unlocking inaccessible historical genomes preserved in formalin. *Molecular Ecology Resources*, *22*(6), 2130–2147. https://doi.org/10.1111/1755-0998.13505

He, S., Su, Y. N., Tan, M. K., Zwick, A., Warren, B. H., & Robillard, T. (2024). Museomics, molecular phylogeny and systematic revision of the Eurepini crickets (orthoptera: Gryllidae: Eneopterinae), with description of two new genera. *Systematic Entomology*, *49*, 389–411. https://doi.org/10.1111/syen.12622

Hykin, S. M., Bi, K., & McGuire, J. A. (2015). Fixing formalin: A method to recover genomic-scale DNA sequence data from formalin-fixed museum specimens using high-throughput sequencing. *PLoS One*, *10*(10), 1–16. https://doi.org/10.1371/journal.pone.0141579

Jeffery, N. W. (2012). The first genome size estimates for six species of krill (malacostraca, Euphausiidae): Large genomes at the north and south poles. *Polar Biology*, *35*(6), 959–962. https://doi.org/10.1007/s00300-011-1137-4

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(241), 241. https://doi.org/10.1101/256479

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.

Molecular Biology and Evolution, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kemp, S., Hardy, C., & Mackintosh, N. A. (1929). Discovery report – Discovery investigations, objects, equipment and methods. *Progress of the Discovery Investigations*, *124*, 483–486.

Laetsch, D. R., Blaxter, M. L., & Leggett, R. M. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*(1287), 1–16.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Mascher, M., Richmond, T. A., Gerhardt, D. J., Himmelbach, A., Clissold, L., Sampath, D., Ayling, S., Steuernagel, B., Pfeifer, M., D'Ascenzo, M., Akhunov, E. D., Hedley, P. E., Gonzales, A. M., Morrell, P. L., Kilian, B., Blattner, F. R., Scholz, U., Mayer, K. F. X., Flavell, A. J., … Stein, N. (2013). Barley whole exome capture: A tool for genomic research in the genus Hordeum and beyond. *Plant Journal*, *76*(3), 494–505. https://doi.org/10.1111/tpj.12294

McBride, M., Schram Stokke, O., Renner, A., Krafft, B., Bergstad, O., Biuw, M., Lowther, A., & Stiansen, J. (2021). Antarctic krill *Euphausia superba*: Spatial distribution, abundance, and management of fisheries in a changing climate. *Marine Ecology Progress Series*, *668*, 185–214. https://doi.org/10.3354/meps13705

Meredith, M., Sommerkorn, M., Cassotta, S., Derksen, C., Ekaykin, A., Hollowed, A., Kofinas, G., Mackintosh, A., Melbourne-Thomas, J., Muelbert, M., & Ottersen, G. (2019). Polar regions. In *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*.

Meyer, B., Martini, P., Biscontin, A., De Pittà, C., Romualdi, C., Teschke, M., Frickenhaus, S., Harms, L., Freier, U., Jarman, S., & Kawaguchi, S. (2015). Pyrosequencing and de novo assembly of Antarctic krill (*Euphausia superba*) transcriptome to study the adaptability of krill to climate-induced environmental changes. *Molecular Ecology Resources*, *15*(6), 1460–1471. https://doi.org/10.1111/1755-0998.12408

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

O'Connell, K. A., Mulder, K. P., Wynn, A., de Queiroz, K., & Bell, R. C. (2022). Genomic library preparation and hybridization capture of formalin-fixed tissues and allozyme supernatant for population genomics and considerations for combining capture- and RADseq-based single nucleotide polymorphism data sets. *Molecular Ecology Resources*, *22*(2), 487–502. https://doi.org/10.1111/1755-0998.13481

Peck, L. S., Webb, K. E., & Bailey, D. M. (2004). Extreme sensitivity of biological function to temperature in Antarctic marine species. *Functional Ecology*, *18*(5), 625–630. https://doi.org/10.1111/j.0269-8463.2004.00903.x

Pockrandt, C., Alzamel, M., Iliopoulos, C. S., & Reinert, K. (2020). GenMap: Ultra-fast computation of genome mappability. *Bioinformatics*, *36*(12), 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222

Portner, H. O. (2002). Climate variations and the physiological basis of temperature dependent biogeography: Systemic to molecular hierarchy of thermal tolerance in animals. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *132*, 739–761. https://doi.org/10.1016/S1095-6433(02)00045-4

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

R Core team. (2020). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing* (3.6.3). R Foundation for Statistical Computing.

Ruane, S., & Austin, C. C. (2017). Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. *Molecular Ecology Resources*, *17*(5), 1003–1008. https://doi.org/10.1111/1755-0998.12655

Ruiz-Gartzia, I., Lizano, E., Marques-Bonet, T., & Kelley, J. L. (2022). Recovering the genomes hidden in museum wet collections. *Molecular Ecology Resources*, *22*(6), 2127–2129. https://doi.org/10.1111/1755-0998.13631

Shao, C., Sun, S., Liu, K., Wang, J., Li, S., Liu, Q., Deagle, B. E., Seim, I., Biscontin, A., Wang, Q., Liu, X., Kawaguchi, S., Liu, Y., Jarman, S., Wang, Y., Wang, H.-Y., Huang, G., Hu, J., Feng, B., … Fan, G. (2023). The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. *Cell*, *186*, 1279–1294. e19. https://doi.org/10.1016/j.cell.2023.02.005

Steenwyk, J. L., Buida, T. J., Labella, A. L., Li, Y., Shen, X. X., & Rokas, A. (2021). PhyKIT: A broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics*, *37*(16), 2325–2331. https://doi.org/10.1093/bioinformatics/btab096

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, *99*(2), 349–364. https://doi.org/10.3732/ajb.1100335

Straube, N., Lyra, M. L., Paijmans, J. L. A., Preick, M., Basler, N., Penner, J., Rödel, M. O., Westbury, M. V., Haddad, C. F. B., Barlow, A., & Hofreiter, M. (2021). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. *Molecular Ecology Resources*, *21*(7), 2299–2315. https://doi.org/10.1111/1755-0998.13433

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., Aitken, S. N., & Holliday, J. A. (2016). Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, *16*(5), 1136–1146. https://doi.org/10.1111/1755-0998.12570

Tarling, G. A. (2020). Routine metabolism of Antarctic krill (*Euphausia superba*) in South Georgia waters: Absence of metabolic compensation at its range edge. *Marine Biology*, *167*(108). https://doi.org/10.1007/s00227-020-03714-w

Tou, J. C., Jaczynski, J., & Chen, Y. C. (2007). Krill for human consumption: Nutritional value and potential health benefits. *Nutrition Reviews*, *65*(2), 63–77. https://doi.org/10.1111/j.1753-4887.2007.tb00283.x

Urso, I., Biscontin, A., Corso, D., Bertolucci, C., Romualdi, C., De Pittà, C., Meyer, B., & Sales, G. (2022). A thorough annotation of the krill transcriptome offers new insights for the study of physiological processes. *Scientific Reports*, *12*(1), 11415. https://doi.org/10.1038/s41598-022-15320-5

Veytia, D., Corney, S., Meiners, K. M., Kawaguchi, S., Murphy, E. J., & Bestley, S. (2020). Circumpolar projections of Antarctic krill growth potential. *Nature Climate Change*, *10*(6), 568–575. https://doi.org/10.1038/s41558-020-0758-4

White, O., Hall, A., Price, B. W., Williams, S. T., & Clark, M. (2024). A snakemake toolkit for the batch assembly, annotation, and phylogenetic analysis of mitochondrial genomes and ribosomal genes from genome skims of museum collections. *BioRxiv*, 2023.08.11.552985. https://doi.org/10.1101/2023.08.11.552985

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zane, L., Ostellari, L., Maccatrozzo, L., Bargelloni, L., Battaglia, B., & Patarnello, T. (1998). Molecular evidence for genetic subdivision

of Antarctic krill (*Euphausia superba* Dana) populations. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1413), 2387–2391. https://doi.org/10.1098/rspb.1998.0588

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*, 153. https://doi.org/10.1186/s12859-018-2129-y

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: White, O. W., Walkington, S., Carter, H., Hughes, L., Clark, M., Mock, T., Tarling, G. A., & Clark, M. D. (2024). Exome capture of Antarctic krill (*Euphausia superba*) for cost effective genotyping and population genetics with historical collections. *Molecular Ecology Resources*, *00*, e14022. https://doi.org/10.1111/1755-0998.14022