MOLECULAR ECOLOGY

Molecular Ecology (2015) doi: 10.1111/mec.13370

Antarctic krill population genomics: apparent panmixia, but genome complexity and large population size muddy the water

BRUCE E. DEAGLE,*† CASSANDRA FAUX,* SO KAWAGUCHI,*† BETTINA MEYER‡§ and SIMON N. JARMAN*

*Australian Antarctic Division, Kingston, Tasmania, Australia, †Antarctic Climate and Ecosystems Cooperative Research Centre, Hobart, Tasmania, Australia, ‡Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany, §Institute for Chemistry and Biology of the Marine Environment, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

Abstract

Antarctic krill (*Euphausia superba*; hereafter krill) are an incredibly abundant pelagic crustacean which has a wide, but patchy, distribution in the Southern Ocean. Several studies have examined the potential for population genetic structuring in krill, but DNA-based analyses have focused on a limited number of markers and have covered only part of their circum-Antarctic range. We used mitochondrial DNA and restriction site-associated DNA sequencing (RAD-seq) to investigate genetic differences between krill from five sites, including two from East Antarctica. Our mtDNA results show no discernible genetic structuring between sites separated by thousands of kilometres, which is consistent with previous studies. Using standard RAD-seq methodology, we obtained over a billion sequences from >140 krill, and thousands of variable nucleotides were identified at hundreds of loci. However, downstream analysis found that markers with sufficient coverage were primarily from multicopy genomic regions. Careful examination of these data highlights the complexity of the RAD-seq approach in organisms with very large genomes. To characterize the multicopy markers, we recorded sequence counts from variable nucleotide sites rather than the derived genotypes; we also examined a small number of manually curated genotypes. Although these analyses effectively fingerprinted individuals, and uncovered a minor laboratory batch effect, no population structuring was observed. Overall, our results are consistent with panmixia of krill throughout their distribution. This result may indicate ongoing gene flow. However, krill's enormous population size creates substantial panmictic inertia, so genetic differentiation may not occur on an ecologically relevant timescale even if demographically separate populations exist.

Keywords: genotyping by sequencing, RAD sequencing, repetitive DNA, zooplankton

Received 22 May 2015; revision received 23 August 2015; accepted 1 September 2015

Introduction

Antarctic krill (*E. superba*) are a species of pelagic crustacean found in Southern Ocean waters surrounding Antarctica. They are one of the world's most abundant animals with a total biomass estimated to be between 100 and 500 million tonnes (Nicol & Endo 1997). The species plays a critical ecological role in the Southern Ocean by linking photosynthetic phytoplankton and small zooplankton at the bottom of the food web with animals at higher trophic levels (Marchant & Murphy 1994). There is also a substantial commercial fishery for this krill species and the catch has been expanding in recent years (Nicol *et al.* 2012).

Despite having a circumpolar distribution, the density of Antarctic krill in different areas is far from uniform (Marr 1962; Siegel 2005; Atkinson *et al.* 2008). The high-

Correspondence: Bruce Deagle, Fax: +61 3 6232 3288; E-mail: Bruce.Deagle@aad.gov.au

est abundances occur in the South Atlantic basin (location of current krill fishing activities) and adjoining waters around the Antarctic Peninsula. There are also 'krill-rich' areas around the Ross Sea and off East Antarctica (Indian Ocean sector), but other parts of the range have low abundance (Atkinson *et al.* 2008). This patchy distribution is surprising given the dominant oceanographic feature of the Southern Ocean is the Antarctic Circumpolar Current. This current transports surface water eastward around the entire continent, connecting each ocean basin, and has the potential to transport krill thousands of kilometres during their lifespan (Hofmann & Murphy 2004). However, on a finer scale, the oceanography is more complex. Closer to the continent, where high densities of krill are found (Nicol *et al.* 2000), there is a continental countercurrent which travels in a westward direction. This westward flow may limit the long distance movement of krill, especially when coupled to ontogenetic onshore–offshore krill migrations (see Nicol 2006). There are also several large gyres (e.g. in the Ross Sea and Weddell Sea) which partially isolate these water masses from surrounding regions (Hofmann & Murphy 2004). Vertical migration by krill also makes it hard to predict the impact of the dominant surface currents on their distribution (Hofmann & Murphy 2004). Finally, the distribution of krill may be influenced by their ability to actively swim rather than simply being passive drifters (Trathan *et al.* 1993).

There has been long-standing interest in the ecological genetics of Antarctic krill to investigate potential population structuring (Valentine & Ayala 1976; Bortolotto *et al.* 2011). A number of detailed studies documenting allozyme variation reached the overarching conclusion that the species represent a single genetically homogeneous population (summarized in Fevolden & Schneppenheim 1989). Since this time, a number of DNA-based studies have been carried out primarily looking at mtDNA variation (Zane *et al.* 1998; Goodall-Copestake *et al.* 2010; Batta-Lona *et al.* 2011; Bortolotto *et al.* 2011). All mtDNA data sets show a high diversity of haplotypes, but very low levels of genetic structuring. Despite limited genetic divergence between sampled sites, there have been cases where significant genetic differences have been reported. Zane *et al.* (1998) found differentiation between collections at two sites in the South Atlantic region (Weddell Sea vs. South Georgia; φ_{ST} = 0.0213 based on 154 bp of mtDNA sequence data). An extension of this study found similar φ_{ST} values between two samples collected at one location in different years; however, incorporation of a larger number of sampling sites meant none of the results were statistically significant after correction for multiple comparisons (Bortolotto *et al.* 2011). To examine apparent sample-to-sample variation, two studies have looked at fine-scale genetic structuring. One looked at mtDNA differentiation between krill swarms in the Scotia Sea near South Georgia and failed to detect swarm-level structuring (Goodall-Copestake *et al.* 2010). The other found weak local temporal structuring in mtDNA haplotypes at sites off the Western Antarctic Peninsula and interpreted the findings as evidence for multiple sources of recruitment in this region (Batta-Lona *et al.* 2011). Based on these studies, krill is still considered to be panmictic across its range; however, the analysis of more powerful molecular data sets may provide a different view. Recent work has revealed more complex patterns of dispersal and connectivity in other open ocean zooplankton species (discussed in Peijnenburg & Goetze 2013).

Given the central role of krill in the Antarctic ecosystem, the genetic resources and population genetics data sets available for this species are relatively modest. There are two main explanations for this. First, collecting specimens from the Southern Ocean is difficult; in fact, there have been no DNA-based population genetics studies that include samples from Eastern Antarctica between 0° and 180° longitude (a distance of >8000 km at the Antarctic circle). Second, the number of genetic markers employed has been very limited. Microsatellite markers are often used in high-resolution population genetics analyses, but in Antarctic krill, their application has been restricted. In the only Antarctic krill study that has applied microsatellites, Bortolotto *et al.* (2011) tested several markers but most were discarded due to their unusual structure (interruptions, variable repeat motifs) and the occurrence of more than two alleles per individual. In their population genetics data set, variation in only three microsatellite markers was characterized and no genetic differentiation was found. The complexity of microsatellites may be related to the exceptionally large genome size of Antarctic krill; at *c.* 47 gigabases (Gbp), it is more than 15 times larger than the human genome (Jeffery 2012). With such a large genome, it is unlikely that a genome sequencing and assembly project will be a source of new population genetic markers in near future.

Even in nonmodel species without a reference genome, advances in high-throughput sequencing (HTS) have enabled identification of single nucleotide polymorphisms (SNPs) from markers distributed throughout the genome (Narum *et al.* 2013). This is accomplished by focused sequencing of specific parts of the genome, which allows enough read coverage to be obtained from each locus to document allelic variation within and between individuals. Often regions adjacent to restriction enzyme sites are characterized, commonly using a method called restriction site-associated DNA

sequencing (RAD-seq) (Baird *et al.* 2008; Davey *et al.* 2011). It is possible to use RAD-seq to identify SNPs simultaneously at thousands of loci. For population genomic studies, approaches involve either (i) identifying SNPs in representative individuals, then developing assays to carry out population scale genotyping (e.g. Larson *et al.* 2014), or (ii) obtaining sequences from many individuals from the study populations and using these sequences directly to obtain a population genetics data set (e.g. Hohenlohe *et al.* 2010). While most detailed RAD-seq studies come from organisms with a sequenced genome, it is possible to assemble a library of marker sequences (i.e. the reduced genome) to use for identification of variants. There have now been several nonmodel marine species where this approach has lead to significant insight into population structure (e.g. Reitzel *et al.* 2013). One major benefit of obtaining genomewide markers is the potential to detect markers involved in local adaptation by identification of loci that are highly differentiated relative to neutral markers (e.g. Hess *et al.* 2013; Roda *et al.* 2013). This new insight into the distribution of adaptive genetic variation within populations may be particularly informative in marine species where high levels of gene flow and limited divergence in neutral markers is a common feature (Nielsen *et al.* 2009; Limborg *et al.* 2012; Hess *et al.* 2013; Milano *et al.* 2014).

The explosion of interest in using RAD-seq and related techniques has been accompanied by many studies examining various technical aspects of the methodology [e.g. optimization of laboratory protocols and bioinformatic pipelines (Arnold *et al.* 2013; Davey *et al.* 2013; Gautier *et al.* 2013; Puritz *et al.* 2014)]. To make these experiments tractable, they are often carried out on species with relatively well-characterized and/or small genomes (e.g. Arnold *et al.* 2013), or focus on a specific issue (e.g. Mastretta-Yanes *et al.* 2015). Despite some technical complications being pointed out, the approach is being enthusiastically adopted by researchers studying a wide range of species. As the methodology matures, studies will understandably focus less on the genotyping process and more on the biological questions being answered. In fact, several commercial companies now provide services to perform RAD-seq (or similar) genotyping and initial data processing. While this development has many positive aspects (e.g. support from specialist scientists allows wider adoption of the methods), it does disconnect the end-user from many of the technical challenges.

In this study, we investigated population structure of Antarctic krill by examining genetic variation in samples collected from five sites across the species circum-Antarctic distribution. This includes two sites from East Antarctica – a vast geographic region not included in previous population genetics studies. Here, we sequenced two mtDNA gene regions examined in previous Antarctic krill studies. We also obtained RAD-seq data from >140 individual krill using a commercial service provider with the goal of obtaining a comparable nuclear genotype data set. The RAD-seq data provide new insight into the krill genome and the genetic structuring of this key Antarctic species. Our analysis also provides a case study for the use of standard RAD-seq protocols in nonmodel organisms with complex uncharacterized genomes.

Methods

Sampling and DNA extraction

Adult krill samples were collected at five areas around the Antarctic continent spanning the species distribution (Fig. 1). Sampling was conducted by plankton trawling on Australian, German and American research voyages between 2005 and 2013 (Table 1). To limit possible effects of swarm-specific genetic signatures, krill were taken from spatially or temporally distinct sampling events within these areas when possible. Our goal was to look for signatures of overarching genetic structuring between geographic regions rather than ephemeral fine-scale genetic patterns. Specimens were stored in 95% ethanol or frozen at −80 °C. DNA was extracted using a Qiagen DNeasy Tissue Kit. For further sampling details, see Appendix S1 (Supporting Information).

MtDNA sequencing and data analysis

Two mtDNA fragments were PCR amplified from 140 individual krill: 655 bp from the cytochrome c oxidase subunit I gene (COI) and 569 bp from NADH dehydrogenase subunit 1 gene (ND1). Purified amplicons were sequenced in both directions using the PCR primers (Appendix S2, Supporting Information) and the BIGDYE TERMINATOR KIT (v3.1; Applied Biosystems). Capillary separation was carried out at the Australian Genome Research Facility. Sequences from these gene regions collected in previous population genetics studies [COI (Goodall-Copestake *et al.* 2010); ND1 (Bortolotto *et al.* 2011)] were downloaded from GenBank to allow direct comparisons between data sets.

Genetic diversity indices (haplotype number, segregating sites, mean number of pairwise differences π, Tajima's D) were calculated using MEGA (Tamura *et al.* 2011). Analysis of molecular variance (AMOVA) was used to investigate the partitioning of variance within and among sample sites using the software GENALEX (version 6.5) (Peakall & Smouse 2012). Genetic differentiation

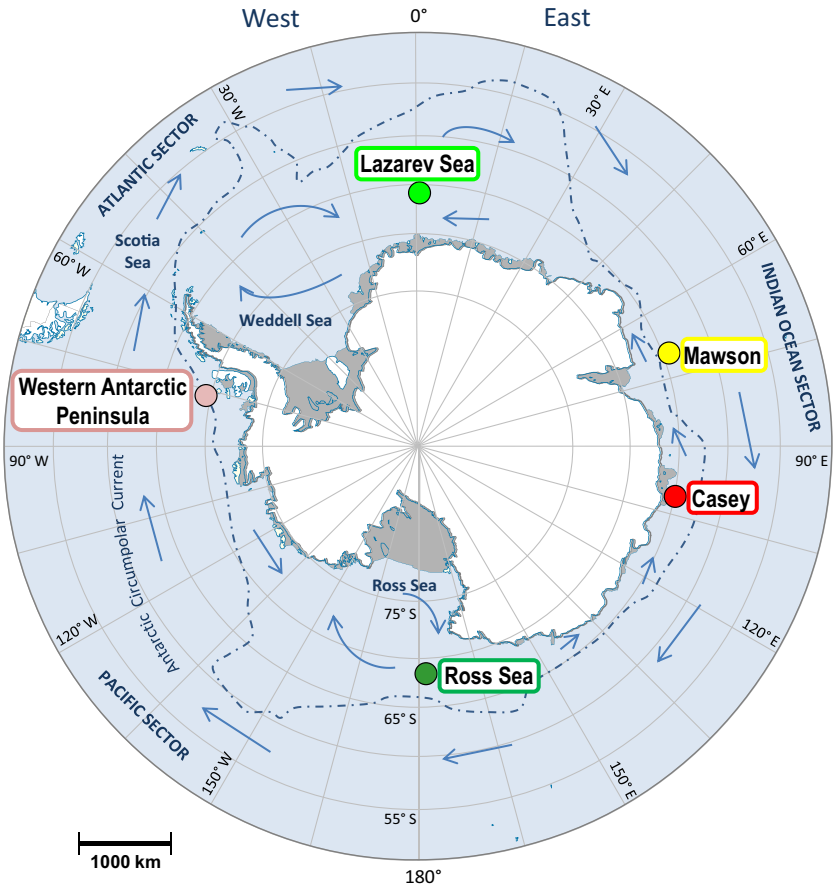

Fig. 1 Krill sample collection sites in the Southern Ocean. Arrows illustrate general surface water circulation patterns; the dotted line shows the southern boundary of the Antarctic Circumpolar Current.

Table 1 Sample site information and pairwise φ_{ST} estimates based on mtDNA sequences

Location	Sample size		Austral summer	Lat	Long	ID	Pairwise φ_{ST} and *P*-values†				
	MtDNA*	RAD*					Cas	Maw	Laz	WAP	Ross
East Antarctica (Casey)	26	21	2010/2011	64° S	100° E	Cas	—	0.325	0.346	0.357	0.106
East Antarctica (Mawson)	30	22	2011/2012	66° S	70° E	Maw	−0.018	—	0.383	0.355	0.068
Lazarev Sea	30	38	2004/2005 and 2007/2008	66° S	0°	Laz	−0.006	−0.001	—	0.337	0.337
Western Antarctic Peninsula	24	16	2010/2011	69° S	76° W	WAP	−0.020	−0.021	−0.012	—	0.218
Ross Sea	30	23	2012/2013	68° S	178° E	Ross	0.025	0.031	−0.012	0.013	—

*Number in final data set; for mtDNA, this includes krill with data from ND1 or COI; for RAD samples, this is the number of krill in the filtered data set.
†Based on combined COI and ND1 sequences. The φ_{ST} values are below diagonal and above diagonal are *P*-values derived from comparison with 9999 from random permutations.

between sites was accessed by calculating pairwise φ_{ST} values. Significance of the resulting F-statistics was determined by comparison with 9999 random permutations. A nonhierarchical statistical parsimony network was constructed to explore genealogical relationships between haplotypes and their geographic distribution (using TempNet, a freely available R script) (Prost & Anderson 2011). MtDNA sequences from previous studies were also incorporated in haplotype networks.

RAD sequencing

RAD-seq was carried out on 148 krill samples, including four replicates of one individual krill used to monitor genotyping error rates (DNA obtained from separate extractions for replicates). Samples came from the same collections used for mtDNA sequencing, but in a few cases DNA extracts from different krill were used due to the requirement for high-quality template

for RAD-seq analysis. Library preparation was carried out in two separate batches (processed several months apart) by Floragenex (Eugene, Oregon, USA) following the protocol of Etter *et al.* (2011). Briefly, genomic DNA was digested with *Sbf*I (recognition sequence: CCTGGA*GG; New England Biolabs) and libraries from individual krill were barcoded with six base tags differing by >2 nucleotides. After random shearing with a Bioruptor (Diagenode), DNA 250 bp to 500 bp in size was isolated and RAD fragment libraries were sequenced on an Illumina HiSeq 2000 using single-end 100 bp chemistry. FASTQ sequence data were demultiplexed and trimmed to 90 bp.

RAD reference sequence assembly, SNP calling and initial data filtering

As there is no reference genome for Antarctic krill, a set of unique 90-bp sequences (RAD tags) was assembled from 17.3 million single-end reads from an individual krill. The following parameters were applied to cluster sequences from this krill into RAD tags using a proprietary bioinformatics pipeline (Floragenex): minimum sequence coverage of 5 and maximum of 500, maximum number of two haplotypes per cluster and a maximum of three mismatches allowed per cluster. Complete analysis was also carried out on data derived using reference RAD loci assembled from a different krill; results for each were congruent so only one analysis is presented.

To facilitate SNP calling, sequence reads from remaining krill samples were aligned to the reference RAD tags using BOWTIE (version 0.11.3; Langmead *et al.* 2009). Reads mapping to more than one reference sequence were discarded, and the maximum number of mismatches allowed was three. SNPs were called using SAMtools (0.0.12a; Li *et al.* 2009) under the following parameters: minor variant frequency of 0.075, minimum 6× coverage, minimum phred genotype quality score of 15 and minimum per cent of the samples genotyped of 80%. SNP variants for all individuals were tabulated (using the 'pileup' module) and 'core data set' exported in variant call format 4.1. Several different parameter sets were trialled but intermediate stringency was ultimately chosen as a compromise between SNP numbers vs. coverage. Throughout the study, 'RAD tags' refer to the 90-bp DNA sequences produced by clustering of closely related RAD haplotypes. RAD haplotypes differ from one another by a small number of SNPs (following Mastretta-Yanes *et al.* 2015).

The SNPs on RAD tags from the core data set were further processed to produce a 'filtered data set' [carried out in R (R_Core_Team 2013)]. Krill with <4 million reads in total were removed, as were any SNPs with total read coverage of <4000 or >80 000 sequences. In the remaining markers, we observed that occasional SNPs (0.3%) had erroneous three-variant calls within individual krill. Some of these triallelic calls may have resulted from sequencing error, but they tended to be concentrated within particular RAD tags, and most were likely caused by clustering of sequences from multiple genetic loci. We excluded any RAD tag containing a SNP with more than three triallelic calls in the 148 genotyped krill. Remaining triallelic calls were coded as missing data in the krill in which they occurred. We removed rare genetic variants (uninformative for population structure analysis) by excluding SNPs fixed for the most common allele in >95% of the krill. Hardy–Weinberg equilibrium tests (genetics R package) were carried out to check for major deviation in genotypes observed at each SNP loci in krill from within each collection site. SNPs with a *P*-value <0.0001 in any site, or <0.001 in multiple collection sites, were discarded (*P*-value computed using 20000 simulations).

Alternative RAD data processing steps

Despite the initial data filtering steps, there was strong evidence that in many cases sequences grouped to single RAD tags were not from a single genetic locus (see Results for further details). This meant that these data could not be analysed using conventional population genetics methods to draw inferences about population structure. We therefore carried out two alternative sets of analyses.

First, we directly analysed the raw count data of different SNPs at variable sites (i.e. rather than the derived genotype). Sequence counts showed consistency in our replicated sample indicating that counts characterize the prevalence of nucleotide polymorphisms on a particular RAD tag (see Results). This makes no assumption that sequences were derived from a single genetic locus. In a diploid individual, a variable single-copy SNP would be expected to have a minor allele frequency *c.* 50% in recovered sequences. If the RAD tag was duplicated, minor allele frequency would be expected to be *c.* 25% or *c.* 50%. Regardless of copy number, on average the number of recovered sequences should reflect the dose of a particular SNP.

Our second analysis involved carrying out a further very conservative data filtering step and using only the remaining scored genotypes. Here, we only accepted RAD tags containing multiple SNPs that were variable within several individual krill, and the count data from these SNPs had to be consistent with a maximum of two haplotypes in individual krill (see Fig. 2). Specifically, we determined which RAD tags had multiple

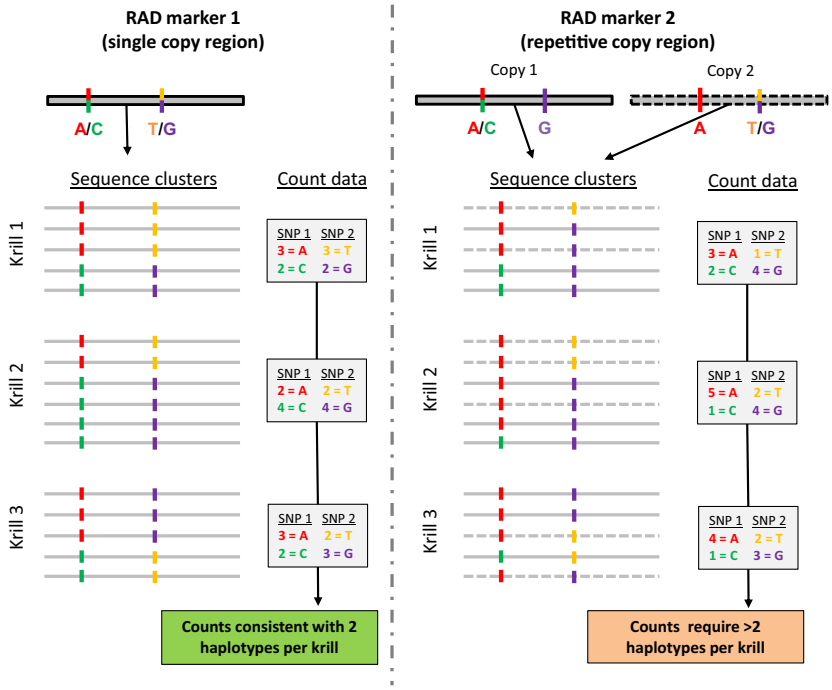

Fig. 2 Schematic illustrating a single-copy and a repetitive RAD tag marker with multiple heterozygous SNPs. The nucleotide count data can be used to determine whether variation is consistent with the marker being present as a single copy (i.e. only two haplotypes in individual krill).

heterozygous SNPs in >5 individual krill, then examined sequence count data at these RAD tags within each krill. RAD tags were only considered if their count data were consistent with one or two haplotypes in >95% of the sequenced individuals (referred to as 'haplotype consistent' genotypes). It should be noted that many markers could be single copy but not meet these criteria (e.g. those on less variable RAD markers).

Analysis of RAD-seq data

To investigate population structure, we used individual-based multivariate methods. For the sequence count data, we performed principal component analysis (PCA) using counts of all variable nucleotides in the core (n = 12 114) and filtered (n = 2197) data sets. PCA is commonly used in the analysis of SNP data as an unsupervised clustering method to discern underlying population structure. It summarizes highly multivariate genetic data into a few synthetic variables which capture variation observed across the data set. For the haplotype consistent genotypes, we used PCA and discriminant analysis of principle components (DAPC) implemented in the adegenet R package (Jombart *et al.* 2010). DAPC is a supervised method that produces synthetic variables maximizing differences between predefined sample groups (i.e. sampling locations) while minimizing variation within groups.

Results

mtDNA

From 140 krill included in mtDNA analysis, 136 COI and 139 ND1 sequences were obtained. Sequences were trimmed to 593 bp (COI) and 494 bp (ND1) to standardize read length. The full 1087 bp of sequence was obtained in 135 krill. Nucleotide diversity was high for both genes (COI: $\pi = 0.0106 \pm 0.0022$ SE; ND1 $\pi = 0.0132 \pm 0.0026$ SE). There was no evidence for the presence of pseudo-gene sequences (i.e. no low-quality sequences, stop codons or outlier sequences). All COI substitutions (80 positions) were synonymous; three nonsynonymous changes occurred in ND1. Consistent with previous studies, Tajima's D (Tajima 1989) was strongly negative (−1.86), indicating an excess of rare mtDNA haplotypes. The overall mtDNA sequence mismatch distribution was bimodal (Appendix S2, Supporting Information), reflecting the presence of two divergent lineages. However, AMOVA results from the analysis of combined COI and ND1 sequences showed that variation among populations was encompassed by that found within populations (100% of variation within populations; global $\varphi_{ST} = -0.002$; $P = 0.447$). Pairwise comparison of φ_{ST} values between sampling sites confirmed the lack of genetic structuring (P-values ranged between 0.068 and 0.383; Table 1).

Haplotype diversity was very high for both mtDNA markers, with most haplotypes being found only in single individuals. For COI, there were 93 haplotypes in 136 sequences. The two most common COI haplotypes made up 16% and 5% of sequences, respectively. These two sequences differed by eight substitutions (1.36% divergence), and each of these formed clusters with several closely related sequences (Fig. 3). The statistical parsimony network of COI haplotypes shows that sequences from these clusters were found across all collection sites and there are only minor frequency differences on a circum-continental scale (Fig. 3). Inclusion of representative COI sequences from the Scotia Sea (Goodall-Copestake *et al.* 2010) highlights the similar diversity of sequences collected in the two studies and extensive haplotype sharing (Fig. 3).

For ND1, there were 95 haplotypes in 139 sequences and, again, two main clusters consistent with variation within COI were observed (see Appendix S2, Supporting Information). The three most common ND1 haplotypes in our data set matched those found previously by Bortolotto *et al.* (2011) and were present at remarkably similar frequency: haplotype #8 (34% vs. 37%), #12 (12% vs. 9%) and #58 (10% vs. 9%). These data emphasize mtDNA admixture around the continent, including our new sites from East Antarctica.

Initial RAD-seq filtering and batch effect

We obtained over a billion reads from the 148 krill in our study (a mean of 6.8 million reads per sample). The reference assembly contained 239 441 distinct RAD tags [based on *in silico* estimates, we expected *c.* 185 000 RAD tags given the krill genome size (Jeffery 2012) and a GC content of 32% (Jarman *et al.* 1999)]. When reads from each krill sample were compared against the reference, a total of 1 800 000 putative SNPs were identified. However, most SNPs only had sufficient sequencing coverage to be called in a small number of krill. The core data set exported for downstream data filtering included just those SNPs with genotype calls in at least 80% of the krill samples and contained 12 114 SNPs on 816 RAD tags (mean of 14.8 SNPs per RAD tag).

Further data filtering steps are detailed in a flowchart provided in Appendix S3 (Supporting Information). From a total of 148 samples sequenced, 24 krill had fewer than 4 million reads or other data quality limitations, and were removed from the data set. Many SNPs were removed because of our strict filtering of all SNPs on RAD tags with low-level triallelic calling errors. More than 70% of remaining SNPs were excluded because they were fixed for the most common allele in >95% of genotype calls. During initial population structure analyses, some separation between samples processed in two different laboratory batches was noted. This was problematic as Ross Sea krill were only included in the second batch. To mitigate this confounding effect, we used DAPC to differentiate between krill SNP data sets from separate laboratory batches (using only data from sites included in both batches, i.e. no data from Ross Sea krill). We then removed the SNPs with highest loadings (top 5%) (Appendix S3, Supporting Information). After these steps, the filtered data set contained 2197 SNPs on 512 RAD tags from 124 samples including four replicates from one individual krill.

Evidence that most krill RAD tags were derived from multiple loci

We initially carried out population genetic analysis on the filtered genotype data set; however, it became apparent that sequence reads used to call SNPs on many RAD tags were not derived from a single locus. Instead, sequence reads aligned to individual RAD tags were often composite clusters containing sequence variants derived from distinct genomic locations (i.e. repetitive regions). There are several lines of evidence which led to this conclusion.

First, sequence count data for many heterozygous loci from individual krill showed strong directional bias away from the expected 50:50 ratio (e.g. Fig. 4a). We originally examined these count data to investigate inconsistencies between genotype calls in the four replicate samples from an individual krill. Genotype errors, measured as percentage of SNP genotype call mismatches between replicates, ranged between 12% within batches and 20% between batches (filtered data set; excluding comparisons with missing data). The replicates each had a large number of reads (between 11.6 and 18.6 million), so low sequencing coverage of RAD tags was not an issue. Instead, most genotype inconsistencies resulted from loci where count proportions consistently fell on the threshold between heterozygotes and homozygotes regardless of sequence coverage (Fig. 4a). It is unlikely these counts result from sequencing errors because if we consider only those SNP loci called as homozygous in all replicates, the percentage minor allele sequences was low (0.41%; Fig. 4b). This indicates these 'true homozygotes' rarely have errors. In contrast, in SNP loci called as homozygous in some but not all replicates, the minor allele sequences made up 3.75% of sequences in homozygotes (Fig. 4c). This shows that some loci in this krill have consistently low minor SNP allele sequence counts, an expected feature of RAD-seq data derived from repetitive loci.

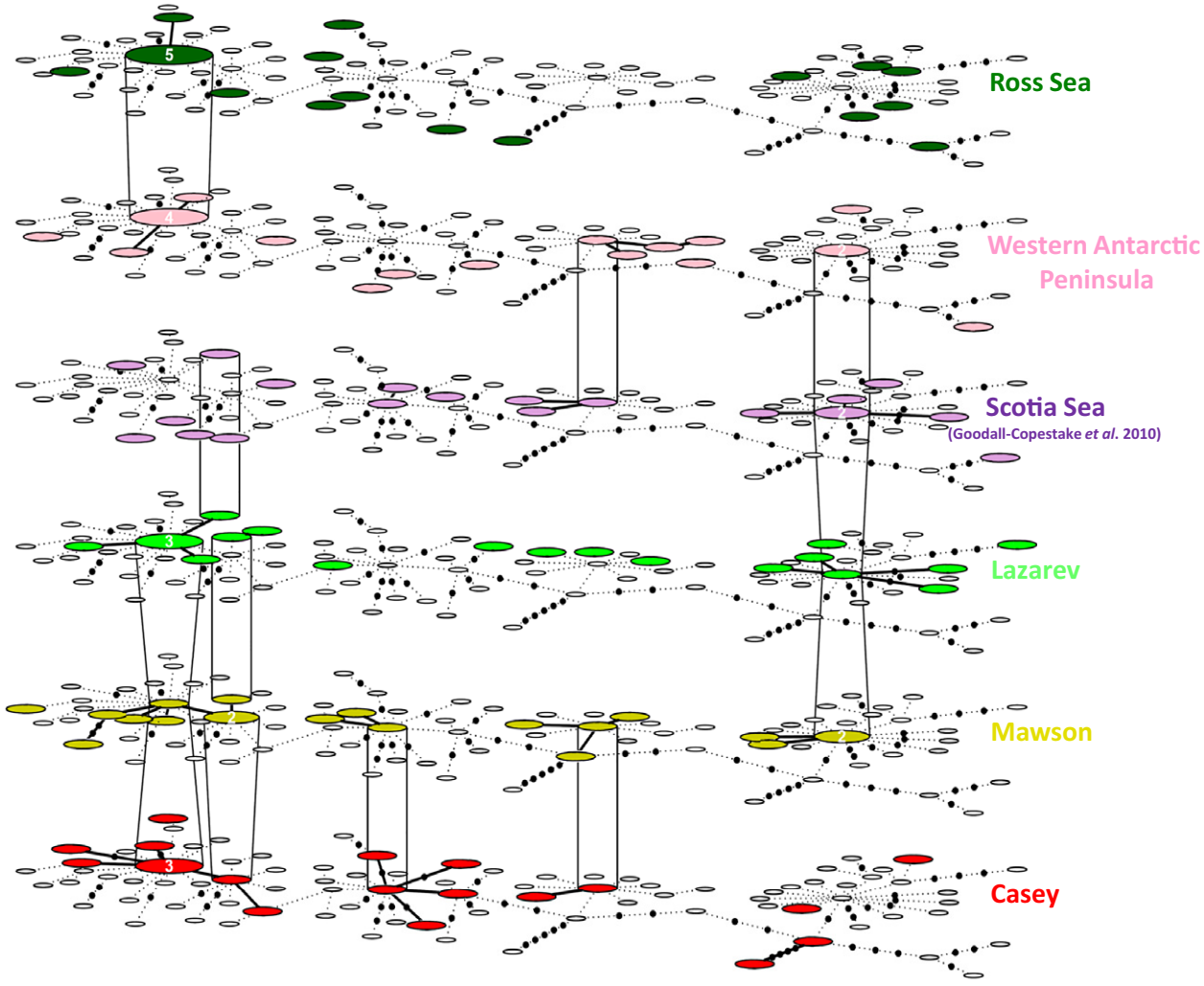

Fig. 3 Relationships between mtDNA COI sequences collected from six circum-Antarctic sampling locations illustrated using a three-dimensional statistical parsimony network (20 individuals per site to avoid cluttering; Scotia Sea sequences from Goodall-Copestake *et al.* (2010)). Unique haplotypes are represented by circles; colours indicate the presence of a haplotype in sample from a particular location. Numbers in circles show haplotype frequency if greater than one. Within each layer, haplotypes are connected by a line if they are separated by one mutation; each additional mutation is indicated by a small black dot. Shared haplotypes between adjacent layers are joined by vertical lines.

Second, we carried out clustering of RAD tags to see whether they were distinctive sequences, or whether they formed groups of related sequences. Clustering of all 239 441 RAD tags with a 10% similarity threshold grouped 35% of these 'unique' tags into groups (performed using USEARCH Edgar 2010). The number of members within a cluster decreased exponentially as cluster size increased, showing a diverse group of repetitive regions exist in the krill genome (Appendix S4, Supporting Information). The largest cluster grouped only 127 RAD tags, but the analysis of raw sequence reads from one krill indicated *c.* 5% of total reads were closely related to this cluster. Of most consequence for our analysis, RAD tags in the core data set (i.e. the population genomic data set) were highly enriched for sequences closely related to other RAD tags; 81% of these were within the 10% similarity threshold to another RAD tag (Fig. 5). A BLAST search of the reference RAD tags in the core data set against the NCBI nucleotide collection did not match any known mobile elements (consistent with previous findings; Leese *et al.* 2012). However, there were matches to apparent repetitive regions, including one adjacent to a previously described krill microsatellite sequence (Candeias *et al.* 2014). There were also matches to two arthropod protein-coding sequences (see Appendix S4, Supporting Information).

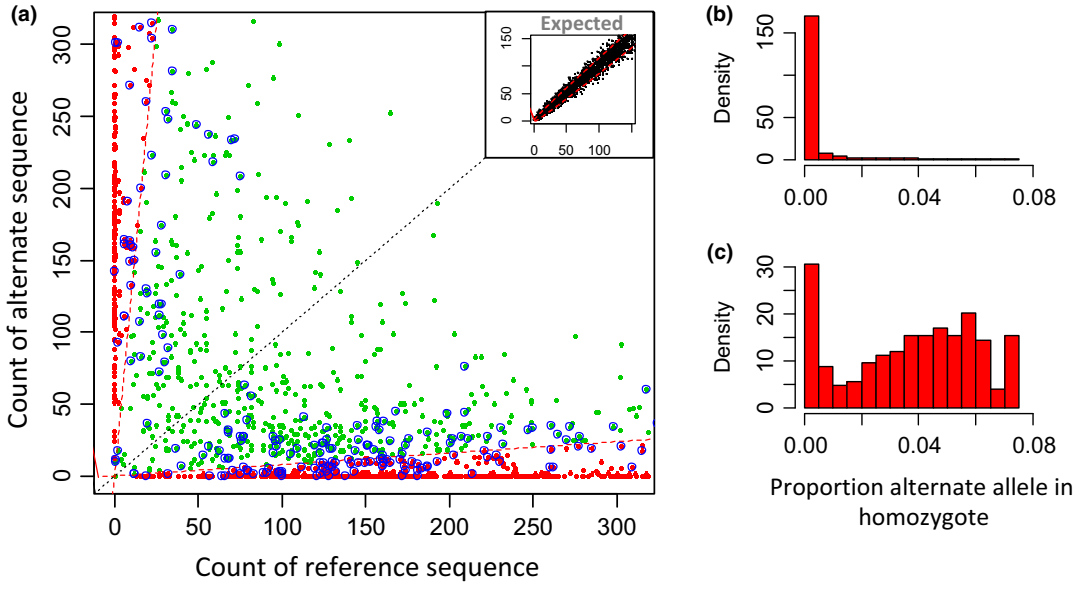

Fig. 4 (a) Plot of reference vs. alternate nucleotide counts in sequences from a single krill. Each point represents a proposed biallelic SNP locus from the filtered data set (>2000 SNPs); green points show loci scored as heterozygous, and red points show those scored as homozygous. The points circled in blue show loci with genotype calls that changed in a replicated sample processed with DNA from the same krill. Inset shows the expected distribution of counts from random sampling of binomial distribution with $P = 0.5$. (b) Density histogram showing proportion of sequences from the minor SNP allele at loci scored as homozygotes all replicates from an individual krill. (c) Density histogram showing proportion of sequences from minor allele for the homozygote genotypes in cases where one replicate was scored as heterozygous.

In our third analysis, we examined RAD tags at the haplotype level to see whether counts of physically linked SNPs from an individual were consistent with the presence of two haplotypes (see Fig. 2). In SNPs that were informative, only a small fraction was consistent with expectations of data derived from a single locus (details in section below).

Population structure analysis based on RAD count data

The count data showed consistency between replicates from an individual krill (Fig. 4c), indicating this could be used to characterize variation at multicopy loci. We carried out our PCA on count data from SNPs included in the filtered data set ($n = 2197$) and also on the core data set ($n = 12\ 114$); results for each data set were very similar. When analyses were carried out on raw count data, separation on the first axis was based purely on sequencing depth of each sample (Fig. 6a). This separation was driven by rare sequences only detected when sequencing depth was high. The second axis showed separation of sequences based on laboratory batch (Fig. 6a). Again, this was driven by rare sequences picked up in differing frequencies in the separate sequencing runs. To remove these effects, we resampled the count data so the maximum coverage that a RAD tag could have within an individual was 25. This standardized count data set removed both the depth and batch effects (although batch effect remained in the core data set; see Appendix S5, Supporting Information).

PCA of standardized count data separated the individual krill processed as four replicates vs. all remaining individuals on the first principle component axis (Fig. 6b). Less than 2% of the variation is explained by this axis, but this result indicates the primary source of variation within these data is between individuals (i.e. any population signal is overpowered by this replicated sample). The scree plot shows this component contains considerably more information than the remaining eigenvalues, which are all about the same size. When the replicated sample is removed, PCA does not separate any clusters of krill on the first two principal component axes (Fig. 6c). Instead, individual krill tend to be separated on eigenvectors and the bulk of krill from different populations overlie each other without any clear pattern reflecting geographic origin. Analysis of the core data set sequences including only samples from the first laboratory batch (i.e. removing the confounding batch effect) produced a similar result (Fig. 6d). Therefore, despite being able to effectively fingerprint individual krill and uncover a very minor batch effect, multivariate analysis of sequence count data failed to uncover any population-related structure.

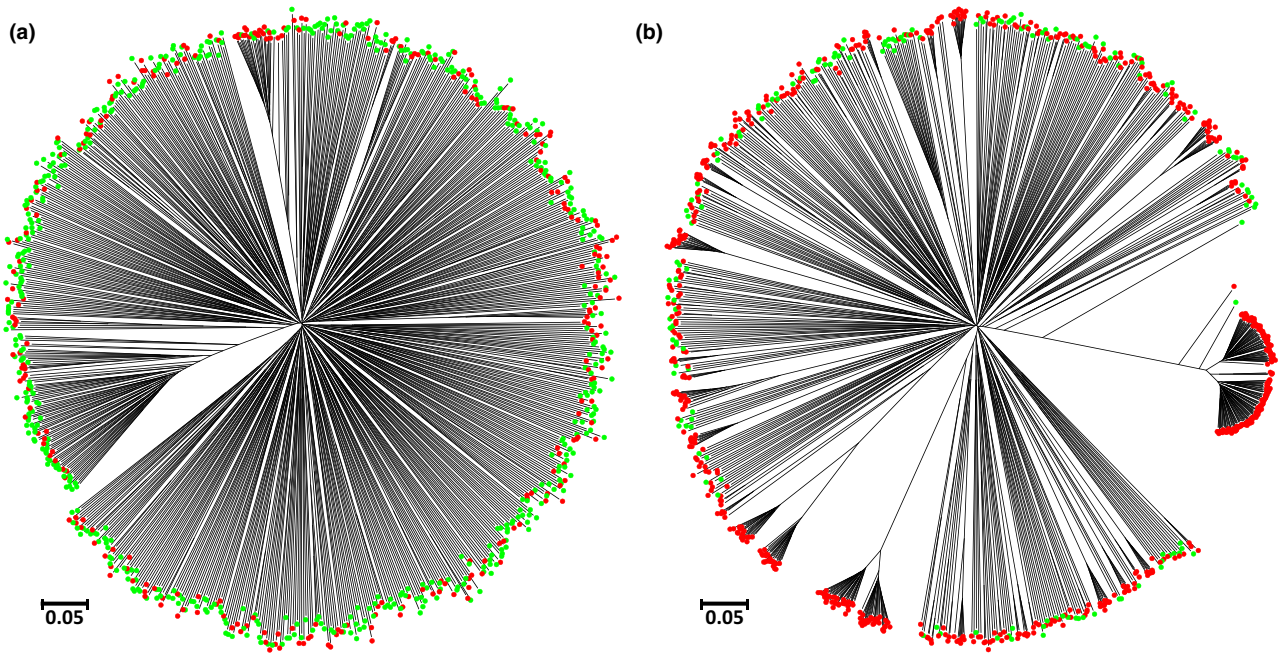

Fig. 5 Distance-based tree showing relatedness among RAD tags from a single krill, labelled to highlight closely related repetitive regions. RAD tags with green markers are >10% different from all others in the complete set of >200 000 RAD tag reference sequences; those coloured red are ≥90% identical to at least one other reference sequence. (a) 816 RAD tags randomly selected from the complete set of RAD tag references; (b) the 816 RAD tags in the core data set. The comparison indicates the core markers selected for use in our population genomic analysis are enriched for repetitive DNA regions. Units are the number of base differences per site.

Population structure analysis based on haplotype consistent genotypes

In the filtered genotype data set, only 66 SNPs on 23 different RAD tags had informative sequence counts consistent with our criteria for having a maximum of two haplotypes in >95% of the samples (Appendix S6, Supporting Information). Most RAD tags were discarded as uninformative, but >30% were ruled out because of the presence of more than three RAD haplotypes in multiple krill. PCA of the 66 haplotype consistent genotypes did not separate krill by sampling locations. Using DAPC, where variation is maximized between sampling locations, the krill still fail to form distinct clusters (see Appendix S6, Supporting Information for additional details).

Discussion

Our investigation into the population genetic structure of Antarctic krill was carried out using a combination of mtDNA sequencing and RAD-seq. The mtDNA showed a lack of population structure across the species' range coupled with a high degree of genetic diversity within each sampled site. Examination of the RAD-seq data indicated that most markers identified for population genomic analysis were present in multiple genomic copies. Using read counts as a proxy for copy number variation in the SNP markers, it was possible to clearly discriminate individual krill, but no population genetic structure was discernible. Analysis of a small number of stringently selected RAD-seq markers also found no genetic structuring.

mtDNA

The current mtDNA sequence data set includes samples from throughout the species' circum-Antarctic range, extending previous sampling to include sites from Eastern Antarctica. The overall lack of mtDNA genetic structuring that we observed is consistent with past findings (Goodall-Copestake *et al.* 2010; Bortolotto *et al.* 2011). Our COI sequences from sites separated by several thousand kilometres closely mirror the haplotypes diversity previously identified from krill swarms in a small geographic area in the Scotia Sea (Goodall-Copestake *et al.* 2010). Similarly, comparison of our new ND1 sequences with results from Bortolotto *et al.* (2011) shows there is extensive sharing of ND1 haplotypes between studies, indicating mixing of mtDNA around the entire continent. Combining sequence data from both mtDNA genes reveals a substantial division between two mtDNA sequence clusters, resulting in a bimodal mismatch distribution (the

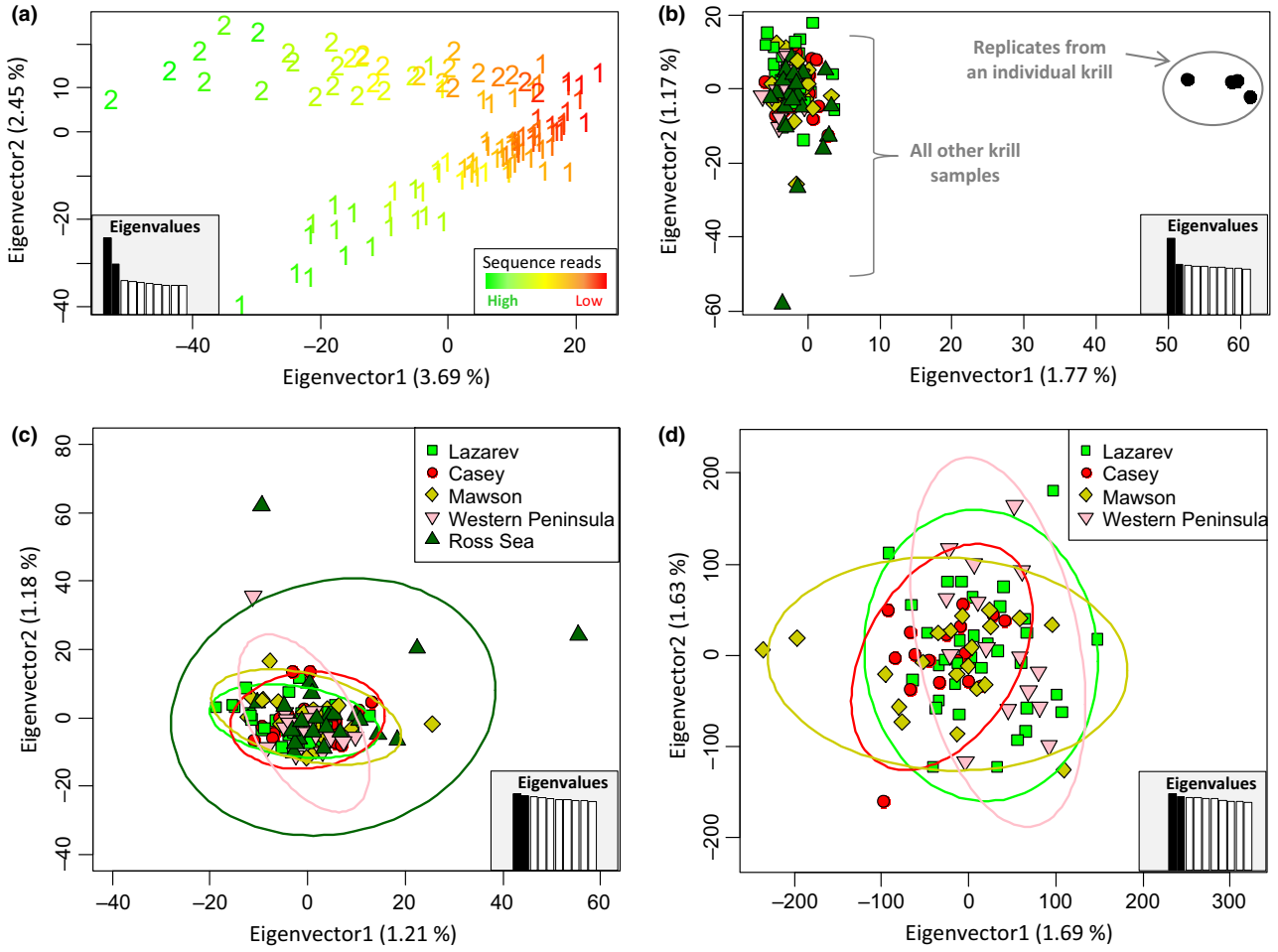

Fig. 6 Principal component analysis of SNP marker nucleotide counts from Antarctic krill samples. In all graphs, plotting characters represent individual krill. (a) Analysis using raw count data in the filtered data set (n = 2197 markers). Counts were scaled to have unit variance. Colour indicates mean sequencing coverage for a given sample. Plotting symbol indicates which laboratory batch a sample was processed in. (b) Analysis using count data in the filtered data set resampled to retain a maximum of 25 sequences per marker in each krill. This data set includes replicates from one krill which was RAD-sequenced four times independently as well as data from 121 other krill from five sites (differentiated by plotting character). Most variation in this data set separates the replicates from remaining samples. This indicates variation between individuals is present and can be repeatedly measured, but there is no overarching population structure. The scree plot indicates only one eigenvector contains much information with the eigenvalues for the rest being about the same size. (c) Analysis of same data set as in previous panel, but replicate samples removed. Inertia ellipses (95%) are shown around the population centroid. (d) Analysis using count data from markers in the core data set (n = 12 114) resampled to retain a maximum of 25 sequences per marker in each krill. Replicate samples were excluded and only samples processed in the first laboratory batch were considered because a batch effect is present (see Appendix S4, Supporting Information).

distribution of number of differences between pairs of haplotypes). The mismatch distribution was previously reported to be unimodal (Zane *et al.* 1998; Bortolotto *et al.* 2011) because the short ND1 region sequenced did not allow two peaks to be discriminated. Our finding indicates that contemporary krill mtDNA has a deep history, with different mtDNA molecules not having shared a common ancestor for at least several hundred thousand years [assuming arthropod mtDNA substitution rate of *c.* 2–3% per million years (Papadopoulou *et al.* 2010)]. The presence of these divergent mtDNA haplogroups probably reflects stochastic retention of mtDNA polymorphisms; this is not unexpected in species with a very large effective population size (see section on population size below). This conclusion is supported by the presence of some intermediate haplotypes. Still, a possible ancient divide between krill populations that have subsequently mixed cannot be discounted. The excess of rare alleles (reflected in the strongly negative Tajima's D) indicates an expanding population or selective sweeps relative to a null model (this issue is discussed in detail in

Bortolotto *et al.* 2011; Goodall-Copestake *et al.* 2010; Zane *et al.* 1998).

RAD data from large uncharacterized genomes

The development of high-throughput reduced representation genomic sequencing approaches has allowed population genomic studies to be carried out on a wide range of taxa including many nonmodel organisms (Narum *et al.* 2013). However, most evaluations of genotype accuracy examine data sets from model species. For example, Davey *et al.* (2013) examined RAD-seq data sets from the very high-quality genome of *Caenorhabditis elegans* and from manually curated loci from *Heliconius melpomene,* to document read coverage bias correlated with restriction fragment size. Similarly, much of the work involving development of algorithms for data analysis also makes use of model systems to evaluate the methodology (e.g. Nielsen *et al.* 2012; Arnold *et al.* 2013). There have only been a few cases where the quality of RAD-seq data sets from species with large complex genomes has been examined in detail to investigate the additional challenges these genomes impose (e.g. Pan *et al.* 2015). Our Antarctic krill RAD-seq data set further illustrates the difficulties that might be expected.

In our analysis, many markers met criteria for SNP calling in a standard RAD-seq population genomic bioinformatic pipeline. However, sequencing depth was low for the vast majority of markers and these were excluded from the final data set. This is a common theme in many NGS applications: large amounts of data are collected, these are highly filtered, and analysis is performed on a small fraction of the 'best' sequences (see DeWoody *et al.* 2013). This filtering can introduce strong biases. In our case, SNPs having the required read coverage and meeting other standard RAD-seq SNP calling parameters were primarily multicopy markers (i.e. only over-represented markers would have enough coverage to make it into the final data set). Larger genomes are expected to contain an increased proportion of repetitive DNA sequences. These will consist of a diverse array of repetitive elements present at a broad range of frequencies (i.e. many paralogous regions present at levels from duplicates to thousands of copies) (e.g. Kovach *et al.* 2010). In our data, the most highly repetitive sequences could be filtered out based on excessive read counts (e.g. >5% of reads came from one cluster of closely related sequences). However, when these high-copy-number markers are removed, markers present at intermediate copy number may be misidentified as being single copy when overall sequencing coverage is low.

There are several features of our RAD-seq data set that alerted us to the likelihood that many SNP loci were likely from repetitive regions. These features are likely to be present in similar RAD-seq data sets for other organisms with complex, uncharacterized genomes and are worth identifying early in RAD-seq projects. The fact that <1% of the identified SNPs were included in the population data set is a strong indication that only markers with unusually high coverage were being selected. Furthermore, plotting of sequence counts in heterozygotes from these markers shows a strong allele bias (alternate counts should have close to 50% representation in a heterozygote). Finally, the analysis of RAD tag haplotypes provides clear evidence that there were large Mendelian inconsistencies. Identifying problems in data is one thing, providing a remedy is another. There are alternative RAD-seq pipelines for calling genotypes which may be better at dealing with repetitive regions (e.g. Dou *et al.* 2012). However, reanalysis of this krill RAD-seq data is unlikely to provide information on single-copy sequences allowing population structure analysis simply because the read coverage for these markers is too low to call genotypes in enough samples. The high depth coverage obtained in a few krill probably contains reliable data on single-copy SNPs in these individuals, and these data could potentially be useful for the development of targeted krill SNP genotyping assays.

Rather than focusing on genotypes from single-copy homologous genetic regions, it is also possible to carry out more inclusive analyses of RAD-seq data sets (e.g. Gouin *et al.* 2015; Waples *et al.* 2015). Here, we used sequence depth as a proxy for copy number of the variant sequences in individual krill and looked for patterns of population structure using multivariate methods. For organisms with diverse complex genomes, this type of analysis has the benefit of allowing a much higher proportion of the data to be used by simultaneously incorporating allelic variation and copy number polymorphisms. Using sequence count data is only possible if count numbers are standardized across samples; otherwise, rare sequences at RAD tags with low coverage are consistently missed (i.e. having a minor allele count of <10% is not unexpected from a multicopy marker, but this variant is detected more often in samples with higher coverage). By directly analysing standardized sequence counts from our krill RAD-seq data, we reliably identified an individual krill processed as replicate samples. We also uncovered count signatures indicative of which laboratory batch samples were run in (this accounted for <2% of the variance; discussed below). Despite consistently measuring these low-level sources of variation, no population-specific sequence counts were identified. This strongly indicates a lack of

population-related structure in these data. Further application of this count-based approach in study systems where the biological signal is stronger will help establish its broad utility.

Another approach to deal with the presence of multicopy regions in a RAD-seq data set is to use additional post hoc filtering to identify single-copy homologous genetic regions. We took advantage of sequence variability in our markers to screen for RAD tags that contained multiple heterozygous SNPs and retained only those markers with two haplotypes in individual krill. While this step presumably enriches data sets for single-copy markers, in our case it also substantially reduced the amount of useable data. This excessive filtering may simply be masking the problem by finding multicopy SNPs that follow the pattern expected for single-copy regions. We also attempted several other filtering approaches such as including only RAD tags which were at least 10% divergent from all other RAD tags; including only loci with low coverage; or including only those with limited bias in allele counts for heterozygotes. No markers in our data set met all these criteria, suggesting that the vast majority of markers are in fact multicopy. This type of post hoc filtering would be more successfully applied in data sets with a higher proportion of single-copy markers.

Despite some commercial companies offering standard services to provide RAD-seq population genotyping from species with large complex genomes, detailed preliminary studies are required to ensure success. One such pilot study was recently carried out for pine trees (Pan *et al.* 2015; genome size 22–32 Gbp vs. 47 Gbp for krill). Analysis was simplified in this study because it focused on haploid tissue isolated from pine seeds; therefore, all heterozygous loci in an individual seed must originate from repetitive genomic regions or sequencing error. Several libraries were created and showed large variation in repetitive DNA content depending on restriction enzyme used. The sequence depth required for saturation varied between $3.5e^5$ and $1.4e^7$ and was not easily predictable based on *in silico* digestion of the pine genome (our mean coverage for each krill was $6.8e^6$). Beyond choice of appropriate restriction enzyme and coverage, other taxon-specific genome features need to be evaluated. For example, in diploid samples, the difficulty in identifying repetitive markers will be magnified in genetically diverse species such as krill because the divergence between alleles will encompass that likely to be seen in many paralogs. This diversity will also result in increased sequence variation in restriction sites resulting in null alleles (i.e. heterozygous restriction sites), and these impacts need to be evaluated (Gautier *et al.* 2013).

The lack of population genetic structuring in our RAD-seq data set highlighted a laboratory batch effect. This effect was relatively minor, but some of the recovered sequences did differentiate samples from different runs. It is not clear whether this resulted from slight changes in methodology between batches, or low-level contamination. This technical issue is problematic when it confounds differences between populations (i.e. when sequences from a new site are added to the analysis in a separate batch). We removed the effect using multivariate methods to identify batch-related sequences in samples from populations included in both batches. It is not clear whether this batch effect would impact a more standard RAD-seq data set with cleaner genotype calls, but randomization of populations across batches would still be prudent experimental design. Including replicate samples across batches would also provide increased confidence in any RAD-seq study.

Large population size and lack of genetic structuring in Antarctic krill

Despite ongoing DNA-based examination of population structuring in Antarctic krill, the statement made more than 25 years ago based on allozymes data sets still seems valid: 'the genetic data obtained to date have substantiated the hypothesis of a single genetically homogeneous breeding population of *E. superba* in Antarctic waters' (Fevolden & Schneppenheim 1989). However, the fact that Antarctic krill is a hyperabundant species should affect the way that we view this conclusion. The census population size of krill is exceptionally large (i.e. in the order of several hundred trillion), and the effective population size estimates (Ne) range from hundreds of thousands to millions (see Zane *et al.* 1998; Goodall-Copestake *et al.* 2010). While Ne has a major impact on many population genetic parameters, its influence is arguably still underappreciated in the analysis of population structure (Waples 1998; Cano *et al.* 2008; Cutter *et al.* 2013). Population genetic structuring in neutral genetic markers results primarily from genetic drift (i.e. stochastic sampling of alleles between generations), and the effect of drift is inversely related to population size. A very large meta-population will have an extremely slow rate of genetic differentiation between large subpopulations, even in the absence of any homogenizing gene flow (e.g. Dey *et al.* 2013). Because of the muted impact of drift, even an extremely low relative rate of migration will prevent differentiation (Waples & Gaggiotti 2006). The panmictic inertia exhibited in large populations is even more pronounced in expanding populations, such as krill, because the sampling effect between generations is less potent. For these reasons, lack of genetic

differentiation measured with neutral markers does not provide solid information on the demographic connectivity in Antarctic krill (i.e. the extent of demographic linkages between regions; see Lowe & Allendorf (2010) for discussion).

There are a few ways that demographic connectivity could be further evaluated using genetics. Because large populations differentiate slowly, finding stable genetic differences between krill from different locations would indicate that migration is very low and the sites could be considered to be demographically independent (Waples & Gaggiotti 2006). Continued evaluation of highly variable neutral genetic markers in many krill from different sites would provide higher power compared to available studies (Peijnenburg & Goetze 2013), and this could lead to further insight if subtle population structure is present. Rather than focusing on genetic markers influenced only by demographically driven selectively neutral processes, it may be possible to find genomic regions shaped by natural selection. The efficiency of selection is expected to be highest in large populations (Cutter *et al.* 2013; Peijnenburg & Goetze 2013), and markers under divergent selection will change in frequency much faster than neutral markers. Genome scans for outlier loci have been used in other species to identify genomic regions under different selective forces in different geographic locations. A limitation of this approach in species with very large populations is that linkage disequilibrium is expected to be low (Reich *et al.* 2001). This means regions of the genome selected for together will be small (unless a selective sweep is recent), and therefore, nontargeted genome scans (such as RAD-seq) are less likely to uncover genomic regions under selection. The current study clearly illustrates that a very large sequencing effort would be required to obtain reliable single-copy genetic markers from even a small portion of the krill genome. Focusing on functional genetic variation in cDNA markers (RNA-seq) or candidate genes will increase the chances of uncovering markers under selection (Davey *et al.* 2011).

Conclusions

A primary goal of the current study was to obtain a population genomic data set with many nuclear markers from Antarctic krill using standard RAD-seq methodology. However, careful examination of the RAD-seq genotype calls we obtained from Floragenex, including comparison of data derived from replicate samples, showed that most of the newly discovered markers were from multicopy genomic regions. Using methods of data analysis appropriate for multicopy variants, we detected genetic structure caused by individual and technical variability, but no population-related structure. This conclusion was supported by a small number of higher quality RAD-seq loci and parallel analysis of mtDNA variability. While our data lend further support to the hypothesis of panmixia in this key Antarctic species, the goal of obtaining genotypes at many single-copy nuclear loci in a krill population genetics study remains elusive.

Rather than providing a blueprint for future population genomic studies on nonmodel organisms with large genomes, the current study illustrates the many challenges that exist. As a positive outcome, we also outline alternative methods of analysis that can be applied to get the most out of an imperfect data set. We would strongly recommend a pilot study before attempting to obtain RAD-seq population genomic data sets from species with similarly large genomes due to the disproportionate amount of repetitive DNA and the unpredictable sequence composition of the repetitive regions. The inclusion of replicate samples can be very useful to uncover difficulties within the data set (e.g. Mastretta-Yanes *et al.* 2015). Depending on the question that is being addressed, it may also be wise to choose a genotyping approach that targets fewer loci to ensure sufficient coverage of the single-copy markers.

Continued creative scrutiny of demographic connectivity in krill will be important to provide an accurate picture for ongoing management of the expanding krill fishery (Nicol *et al.* 2012). It will also be important for understanding how the species might respond to changing environmental conditions (Kawaguchi *et al.* 2013). If krill are truly panmictic and genetically homogeneous on a broad scale, then adaptation to local conditions would be limited. In this case, adaptive genetic diversity may not be present and this would not bode well for the future of krill (as interpreted in Flores *et al.* 2012; Kawaguchi *et al.* 2013). Alternatively, if there is some population structure and local adaptation, krill could be well poised for adaptive evolutionary responses due to their high intraspecific diversity (Peijnenburg & Goetze 2013). Finding an approach to illuminate further details of the species population dynamics and evolutionary potential remains an important goal in Southern Ocean ecosystem research.

Acknowledgements

We thank the many people who helped by providing krill samples, in particular D. K. Steinberg (WAP, Palmer LTER sampling grid); Rob King and Natasha Waller (East Antarctic samples); Margaret Lindsay (Ross Sea); and the crews on each of the Antarctic voyages. We thank Jason Boone and Jenna Donovan for facilitating RAD-seq at Floragenex, and Susanne Saphic for

extracting DNA from Lazarev Sea krill. Thanks also to Paige Eveson for help with some of the trickier R coding and to Steve Nicol for his enthusiasm in all things krill. This project is a contribution to the research program PACES II (topic 1, work-package 5) of the Alfred Wegener Institute. B.E.D. was supported by the R. J. L. Hawke Post Doctoral Fellowship in Antarctic Environmental Science. The research was funded by the Australian Antarctic Science Program (AAS Project 4015).

References

Arnold B, Corbett-Detig R, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Atkinson A, Siegel V, Pakhomov EA *et al.* (2008) Oceanic circumpolar habitats of Antarctic krill. *Marine Ecology Progress Series*, **362**, 1–23.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Batta-Lona PG, Bucklin A, Wiebe PH, Patarnello T, Copley NJ (2011) Population genetic variation of the Southern Ocean krill, *Euphausia superba*, in the Western Antarctic Peninsula region based on mitochondrial single nucleotide polymorphisms (SNPs). *Deep-Sea Research Part II-Topical Studies in Oceanography*, **58**, 1652–1661.

Bortolotto E, Bucklin A, Mezzavilla M, Zane L, Patarnello T (2011) Gone with the currents: lack of genetic differentiation at the circum-continental scale in the Antarctic krill *Euphausia superba*. *BMC Genetics*, **12**, 32.

Candeias R, Teixeira S, Duarte CM, Pearson GA (2014) Characterization of polymorphic microsatellite loci in the Antarctic krill *Euphausia superba*. *BMC Research Notes*, **7**, 73.

Cano JM, Shikano T, Kuparinen A, Merila J (2008) Genetic differentiation, effective population size and gene flow in marine fishes: implications for stock management. *Journal of Integrated Field Science*, **5**, 1–10.

Cutter AD, Jovelin R, Dey A (2013) Molecular hyperdiversity and evolution in very large populations. *Molecular Ecology*, **22**, 2074–2095.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

DeWoody JA, Abts KC, Fahey AL *et al.* (2013) Of contigs and quagmires: next-generation sequencing pitfalls associated with transcriptomic studies. *Molecular Ecology Resources*, **13**, 551–558.

Dey A, Chan CKW, Thomas CG, Cutter AD (2013) Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. *Proceedings of the National Academy of Sciences*, **110**, 11056–11060.

Dou J, Zhao X, Fu X *et al.* (2012) Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology Direct*, **7**, 17.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS One*, **6**, e18561.

Fevolden SE, Schneppenheim R (1989) Genetic homogeneity of krill (*Euphausia superba* Dana) in the Southern Ocean. *Polar Biology*, **9**, 533–539.

Flores H, Atkinson A, Kawaguchi S *et al.* (2012) Impact of climate change on Antarctic krill. *Marine Ecology Progress Series*, **458**, 1–19.

Gautier M, Gharbi K, Cezard T *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Goodall-Copestake WP, Perez-Espona S, Clark MS *et al.* (2010) Swarms of diversity at the gene cox1 in Antarctic krill. *Heredity*, **104**, 513–518.

Gouin A, Legeai F, Nouhaud P *et al.* (2015) Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads. *Heredity*, **114**, 494–501.

Hess JE, Campbell NR, Close DA, Docker MF, Narum SR (2013) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology*, **22**, 2898–2916.

Hofmann EE, Murphy EJ (2004) Advection, krill, and Antarctic marine ecosystems. *Antarctic Science*, **16**, 487–499.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genetics*, **6**, e1000862.

Jarman S, Elliott N, Nicol S, McMinn A, Newman S (1999) The base composition of the krill genome and its potential susceptibility to damage by UV-B. *Antarctic Science*, **11**, 23–26.

Jeffery NW (2012) The first genome size estimates for six species of krill (Malacostraca, Euphausiidae): large genomes at the north and south poles. *Polar Biology*, **35**, 959–962.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Kawaguchi S, Ishida A, King R *et al.* (2013) Risk maps for Antarctic krill under projected Southern Ocean acidification. *Nature Climate Change*, **3**, 843–847.

Kovach A, Wegrzyn J, Parra G *et al.* (2010) The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics*, **11**, 420.

Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Larson WA, Seeb JE, Pascal CE, Templin WD, Seeb LW (2014) Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, **71**, 698–708.

Leese F, Brand P, Rozenberg A *et al.* (2012) Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. *PLoS One*, **7**, e49202.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Limborg MT, Helyar SJ, De Bruyn M *et al.* (2012) Environmental selection on transcriptome-derived SNPs in a high gene

flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, **21**, 3686–3703.
Lowe WH, Allendorf FW (2010) What can genetics tell us about population connectivity? *Molecular Ecology*, **19**, 3038–3051.
Marchant H, Murphy EJ (1994) Interactions at the base of the Antarctic marine food web. In: *Southern Ocean Ecology: The BIOMASS Perspective* (ed El-Sayed SZ), pp. 267–285. Cambridge University Press, Cambridge.
Marr JWS (1962) The natural history and geography of the Antarctic krill (*Euphausia superba* Dana). *Discovery Reports*, **32**, 33–464.
Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.
Milano I, Babbucci M, Cariani A *et al.* (2014) Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). *Molecular Ecology*, **23**, 118–135.
Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.
Nicol S (2006) Krill, currents, and sea Ice: *Euphausia superba* and its changing environment. *BioScience*, **56**, 111–120.
Nicol S, Endo Y (1997) Krill fisheries of the world. In: *Fisheries Technical Paper*. FAO, Rome.
Nicol S, Pauly T, Bindoff NL *et al.* (2000) Ocean circulation off east Antarctica affects ecosystem structure and sea-ice extent. *Nature*, **406**, 504–507.
Nicol S, Foster J, Kawaguchi S (2012) The fishery for Antarctic krill - recent developments. *Fish and Fisheries*, **13**, 30–40.
Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276.
Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One*, **7**, e37558.
Pan J, Wang B, Pei ZY *et al.* (2015) Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. *Molecular Ecology Resources*, **15**, 711–722.
Papadopoulou A, Anastasiou I, Vogler AP (2010) Revisiting the insect mitochondrial molecular clock: the mid-Aegean trench calibration. *Molecular Biology and Evolution*, **27**, 1659–1672.
Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, **28**, 2537–2539.
Peijnenburg KTCA, Goetze E (2013) High evolutionary potential of marine zooplankton. *Ecology and Evolution*, **3**, 2765–2781.
Prost S, Anderson CNK (2011) TempNet: a method to display statistical parsimony networks for heterochronous DNA sequence data. *Methods in Ecology and Evolution*, **2**, 663–667.
Puritz JB, Matz MV, Toonen RJ *et al.* (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.
R_Core_Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Reich DE, Cargill M, Bolk S *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.
Roda F, Liu H, Wilkinson MJ *et al.* (2013) Convergence and divergence during the adaptation to similar environments by an Australian groundsel. *Evolution*, **67**, 2515–2529.
Siegel V (2005) Distribution and population dynamics of *Euphausia superba*: summary of recent findings. *Polar Biology*, **29**, 1–22.
Tajima F (1989) Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
Tamura K, Peterson D, Peterson N *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **10**, 2731–2739.
Trathan P, Priddle J, Watkins J, Miller D, Murray A (1993) Spatial variability of Antarctic krill in relation to mesoscale hydrography. *Marine Ecology Progress Series*, **98**, 61–71.
Valentine JW, Ayala FJ (1976) Genetic variability in krill. *Proceedings of the National Academy of Sciences*, **73**, 658–660.
Waples RS (1998) Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, **89**, 438–450.
Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.
Waples RK, Seeb LW, Seeb JE (2015) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources*, [Epub ahead of print]. DOI: 10.1111/1755-0998.12394.
Zane L, Ostellari L, Maccatrozzo L *et al.* (1998) Molecular evidence for genetic subdivision of Antarctic krill (*Euphausia superba* Dana) populations. *Proceedings of the Royal Society B-Biological Sciences*, **265**, 2387–2391.

All authors contributed to aspects of study conception and design. B.M. and S.K. provided krill samples. B.E.D. and C.F. did the laboratory work. B.E.D. analysed the data and wrote the manuscript. All authors edited and approved the final manuscript.

Data accessibility

Data derived from RAD-seq analysis (RAD tag sequences, genotype calls, count data, etc.) and all mtDNA sequences have been deposited in Dryad (doi:10.5061/dryad.3023m). Raw Illumina sequence data (FASTQ) along with output files from data processing steps have been deposited in the Australian Antarctic Data Centre (http://dx.doi.org/10.4225/15/556FAB354BE19).

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Additional sample collection information.

Appendix S2 Supplementary mtDNA information: primer sequences, sequence mismatch distribution plots and ND1 haplotype networks.

Appendix S3 Schematic of filtering steps performed on RAD-seq genotype data; includes outline of steps used to remove a batch effect.

Appendix S4 Supplementary analysis of RAD tag clusters including BLAST search with core dataset RAD tags.

Appendix S5 Additional analysis of population structure using RAD-seq sequence count data.

Appendix S6 Analysis of haplotype consistent genotypes.