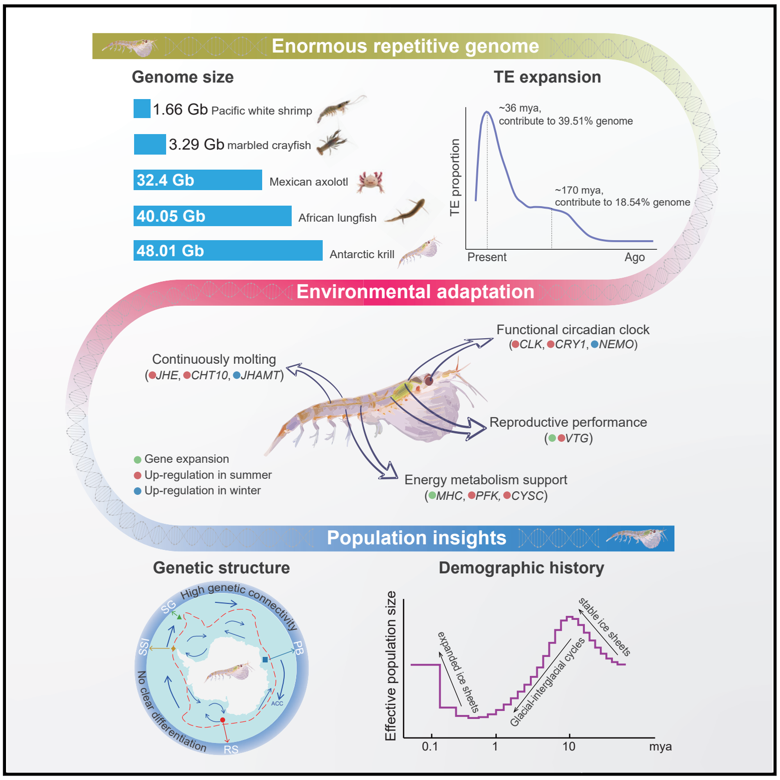
Cell

Resource

The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights

Graphical abstract



Highlights

- Assembly of the 48.01 Gb chromosome-level Antarctic krill genome
- Extensive repeat expansions contributed to the giant Antarctic krill genome
- Genetic adaptations to extreme variability of the Antarctic environment
- Population analysis reveals no clear geographic differentiation in Antarctic krill

Authors

Changwei Shao, Shuai Sun, Kaiqiang Liu, ..., Xianyong Zhao, Bettina Meyer, Guangyi Fan

Correspondence

shaocw@ysfri.ac.cn (C.S.),
bettina.meyer@awi.de (B.M.),
fanguangyi@genomics.cn (G.F.)

In brief

The giant and highly repetitive Antarctic krill genome reveals environmental adaptations and population dynamics of Earth's most abundant wild animal.

Shao et al., 2023, Cell *186*, 1–16
March 16, 2023 © 2023 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2023.02.005

CellPress

Cell

CellPress
OPEN ACCESS

Resource

The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights

Changwei Shao,[1,2,27,28,*] Shuai Sun,[3,4,5,27] Kaiqiang Liu,[1,2,27] Jiahao Wang,[3,27] Shuo Li,[1,2,27] Qun Liu,[3,6,27] Bruce E. Deagle,[7,8] Inge Seim,[9] Alberto Biscontin,[10] Qian Wang,[1,2] Xin Liu,[4,11,12,13] So Kawaguchi,[8] Yalin Liu,[3] Simon Jarman,[14] Yue Wang,[4,15] Hong-Yan Wang,[1,2] Guodong Huang,[4] Jiang Hu,[16] Bo Feng,[1,2] Cristiano De Pittà,[10] Shanshan Liu,[3] Rui Wang,[1,2] Kailong Ma,[4,17] Yiping Ying,[18] Gabrielle Sales,[10] Tao Sun,[3] Xinliang Wang,[18] Yaolei Zhang,[3,4]

(Author list continued on next page)

[1]National Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, Shandong 266071, China
[2]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao, Shandong 266237, China
[3]BGI-Qingdao, BGI-Shenzhen, Qingdao, Shandong 266555, China
[4]BGI-Shenzhen, Shenzhen, Guangdong 518083, China
[5]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[6]Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
[7]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australian National Fish Collection, National Research Collections Australia, Hobart, TAS 7000, Australia
[8]Australian Antarctic Division, Channel Highway, Kingston, TAS 7050, Australia
[9]Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal University, Nanjing, Jiangsu 210023, China
[10]Department of Biology, University of Padova, Padova 35121, Italy
[11]BGI-Beijing, Beijing 102601, China
[12]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
[13]State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Murdoch University, Murdoch, WA 6150, Australia
[14]School of Molecular and Life Sciences, Curtin University, Perth, WA 6009, Australia
[15]State Key Laboratory of Quality Research in Chinese Medicine and Institute of Chinese Medical Sciences, University of Macau, Macao 999078, China
[16]Nextomics Biosciences Institute, Wuhan, Hubei 430073, China
[17]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

(Affiliations continued on next page)

SUMMARY

Antarctic krill (*Euphausia superba*) is Earth's most abundant wild animal, and its enormous biomass is vital to the Southern Ocean ecosystem. Here, we report a 48.01-Gb chromosome-level Antarctic krill genome, whose large genome size appears to have resulted from inter-genic transposable element expansions. Our assembly reveals the molecular architecture of the Antarctic krill circadian clock and uncovers expanded gene families associated with molting and energy metabolism, providing insights into adaptations to the cold and highly seasonal Antarctic environment. Population-level genome re-sequencing from four geographical sites around the Antarctic continent reveals no clear population structure but highlights natural selection associated with environmental variables. An apparent drastic reduction in krill population size 10 mya and a subsequent rebound 100 thousand years ago coincides with climate change events. Our findings uncover the genomic basis of Antarctic krill adaptations to the Southern Ocean and provide valuable resources for future Antarctic research.

INTRODUCTION

Krill, malacostracan crustaceans of the order Euphausiacea, are abundant components of the pelagic ecosystem of all oceans. The biomass of Antarctic krill (*Euphausia superba*) (Figure 1A) is 300–500 million tons, the largest of any wild animal species on the planet.[1] The highly abundant species is a cornerstone of the Antarctic marine ecosystem, forming an ecological link

Yunxia Zhao,[18] Shanshan Pan,[3] Xiancai Hao,[1,2] Yang Wang,[4] Jiakun Xu,[1,18] Bowen Yue,[1,2] Yanxu Sun,[1,2] He Zhang,[4] Mengyang Xu,[3,4] Yuyan Liu,[1,2] Xiaodong Jia,[19] Jiancheng Zhu,[18] Shufang Liu,[1,2] Jue Ruan,[20] Guojie Zhang,[4,21] Huanming Yang,[4,22] Xun Xu,[3,4] Jun Wang,[3] Xianyong Zhao,[1,18] Bettina Meyer,[23,24,25,*] and Guangyi Fan[3,4,20,26,*]

[18]Key Lab of Sustainable Development of Polar Fisheries, Ministry of Agriculture and Rural Affairs, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, Shandong 266071, China
[19]Joint Laboratory for Translational Medicine Research, Liaocheng People's Hospital, Liaocheng, Shandong 252000, China
[20]Agricultural Genomics Institute, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China
[21]Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark
[22]James D. Watson Institute of Genome Science, Hangzhou 310058, China
[23]Section Polar Biological Oceanography, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
[24]Institute for Chemistry and Biology of the Marine Environment, Carlvon Ossietzky University of Oldenburg, 26111 Oldenburg, Germany
[25]Helmholtz Institute for Functional Marine Biodiversity (HIFMB), University of Oldenburg, 26129 Oldenburg, Germany
[26]Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Qingdao, BGI-Shenzhen 518120, China
[27]These authors contributed equally
[28]Lead contact
*Correspondence: shaocw@ysfri.ac.cn (C.S.), bettina.meyer@awi.de (B.M.), fanguangyi@genomics.cn (G.F.)

https://doi.org/10.1016/j.cell.2023.02.005

between primary producers and higher trophic levels—from ice algae to fish, birds, and marine mammals.[2] Antarctic krill plays a vital role in the biogeochemical cycling of carbon and recycles the trace element iron that fosters phytoplankton growth in the Southern Ocean.[3,4] Recent investigations support the existence of an endogenous timing system in Antarctic krill, enabling important life-cycle events to synchronize with Antarctica's seasonal polar environment.[5] However, knowledge of the molecular mechanisms underlying Antarctic krill adaptations to an environment characterized by high seasonality in day length, food availability, and sea ice extent is limited.[5–9] It is also not firmly established whether demographically separate populations exist.[10,11] The krill genome has been estimated at 42–48 gigabases (Gb).[12,13] Its large genome size and complexity have so far prevented its assembly and hindered research on the genetic underpinnings of Antarctic krill adaptations.[13,14] However, recent studies on lungfishes[15,16] and the Mexican axolotl[17,18] demonstrate that daunting technical challenges inherent in the assembly of large animal genomes can be overcome. Here, we present the sequencing, assembly, genome features analysis, and population genetic analysis of the Antarctic krill.

RESULTS

Chromosome-level genome assembly and evaluation

To assemble the Antarctic krill genome, we generated 3.06 terabases (Tb) PacBio continuous long reads (CLR), 734.99 Gb PacBio high-fidelity circular consensus sequencing (HiFi-CCS) reads, 4.01 Tb short reads, and 11.38 Tb Hi-C reads (Table S1). A genome assembly spanning 48.01 Gb was generated (Table S1), the largest animal assembly reported to date. It is about 50% larger than the Mexican axolotl[17,18] and 20%–30% larger than two lungfish species.[15,16] The assembly has a longer contig N50 (178.99 kilobases [kb]) than 120 of 154 available invertebrate genome assemblies (Figure 1B; Table S1). It also has a scaffold N50 of 1.08 Gb with 66.01% of contigs anchored to 17 chromosomes,[19,20] while the unanchored contigs were shorter with a low gene density (Figure S1A; Table S1). A comparison with other malacostracan crustacean assemblies, using evaluations based on genomic short reads, transcriptome data, and non-exonic ultra-conserved elements (UCEs) of eukaryotes, revealed that the Antarctic krill assembly is of comparative quality and completeness despite its much larger size (Table S1; STAR Methods).

Repetitive sequences, particularly long nearly identical repeats, can greatly impact genome assembly.[21] Invertebrate genomes often have a larger proportion and unit length of tandem repeats (TRs) than vertebrate genomes, likely contributing to reported difficulties in their assembly.[22] Repetitive DNA in the Antarctic krill genome is exceptionally abundant, making genome assembly particularly challenging. The most common satellite repeat sequences were of longer unit length than in the most closely related invertebrate genome available, the crustaceans *Procambarus virginalis* (t test, $p < 2.2 \times 10^{-16}$) and *Litopenaeus vannamei* (t test, $p < 2.2 \times 10^{-16}$) (Figure 1C; Table S2). We found that the genome assembly harbors a large proportion of TRs (25.77%), which was still underestimated because TRs are difficult to assemble, especially for TRs with long unit length (>50 base pairs [bp]) and high abundance (Figure 1C; Table S2). The high TR proportion affected the genome assembly, reflected by a negative correlation between contig length and TRs observed (Pearson's $r = -0.14$, $p < 10^{-4}$) (Figure S1B). The Antarctic krill genome possesses higher density of repeat regions (in 10 kb windows) than the Mexican axolotl, lungfish, and two malacostracan crustaceans (*t*-test, $p < 2.2 \times 10^{-16}$) (Figure 1D). We also found that 93.43% of contigs ended in repetitive sequences, and neighboring TEs with high sequence similarities (identity >98%) cluster at short intervals in the assembly, forming extended stretches of repeats (Figure S1C).

Attributes of a giant invertebrate genome

Giant genome sizes appear common in crustaceans of the Southern Ocean and North Atlantic, but there is no evidence of polyploidy (whole-genome duplication) in Antarctic krill.[12] Our genome assembly reveals that the huge Antarctic krill genome can be ascribed to repetitive sequence expansions. 72.15% of

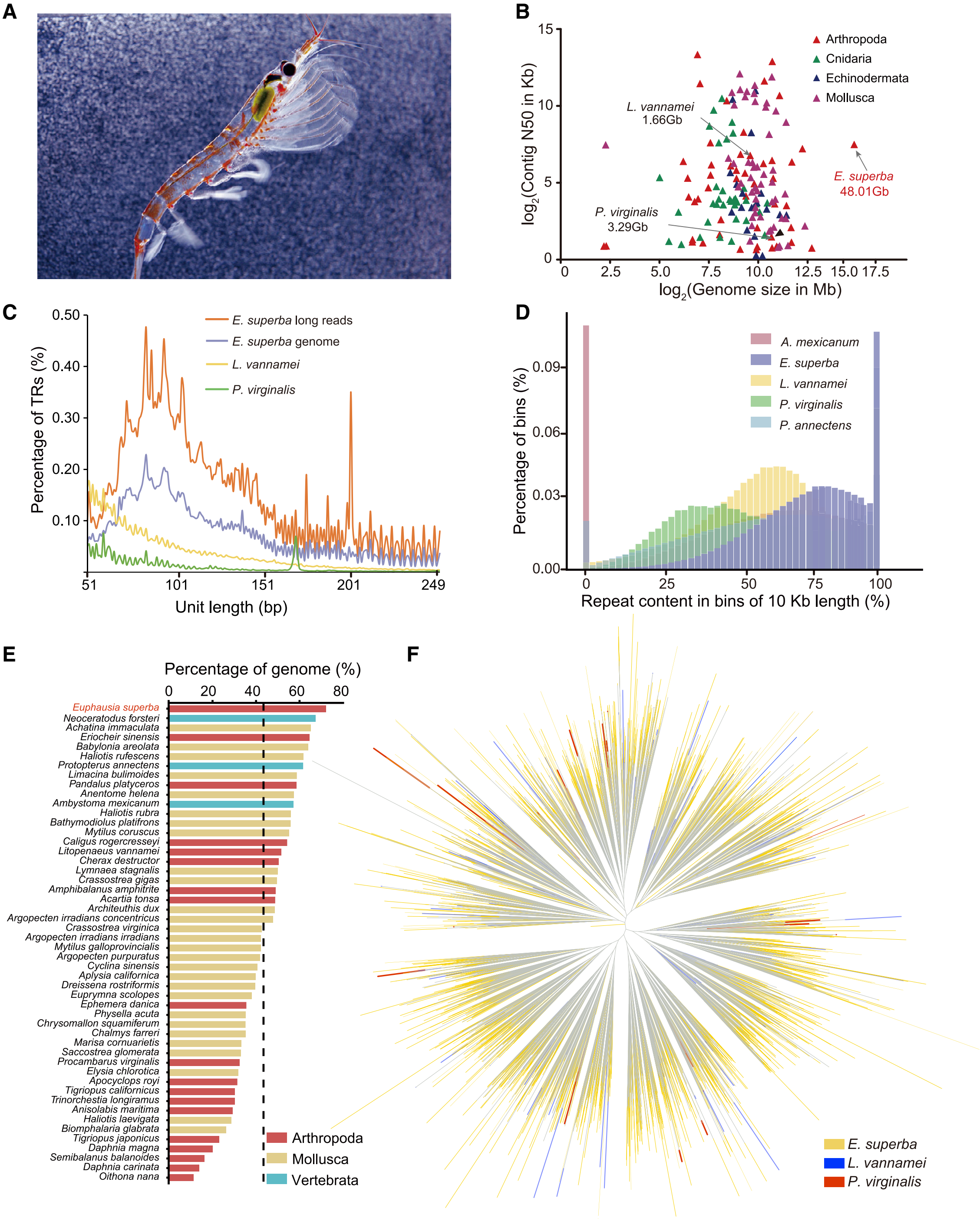


Figure 1. The Antarctic krill genome and its repetitive sequence landscape

(A) Image of an Antarctic krill (*Euphausia superba*) (photo credit Simon Payne, Australian Antarctic Division).

(B) The relationship between genome size and contig N50 of 154 invertebrate genome assemblies.

(C) Comparison of 51–249 bp TRs annotated in long reads and genome assembly of Antarctic krill, *L. vannamei*, and *P. virginalis*.

(D) The distribution of repeat content in the bins of non-overlapping 10 kb windows.

(E) Composition of the first-round repetitive sequences in invertebrate and vertebrate genomes. The dashed line represents the average of repetitive sequences across these species.

(F) Phylogenetic tree of the DNA/CMC-EnSpm TE superfamily in Antarctic krill, *L. vannamei*, and *P. virginalis*.

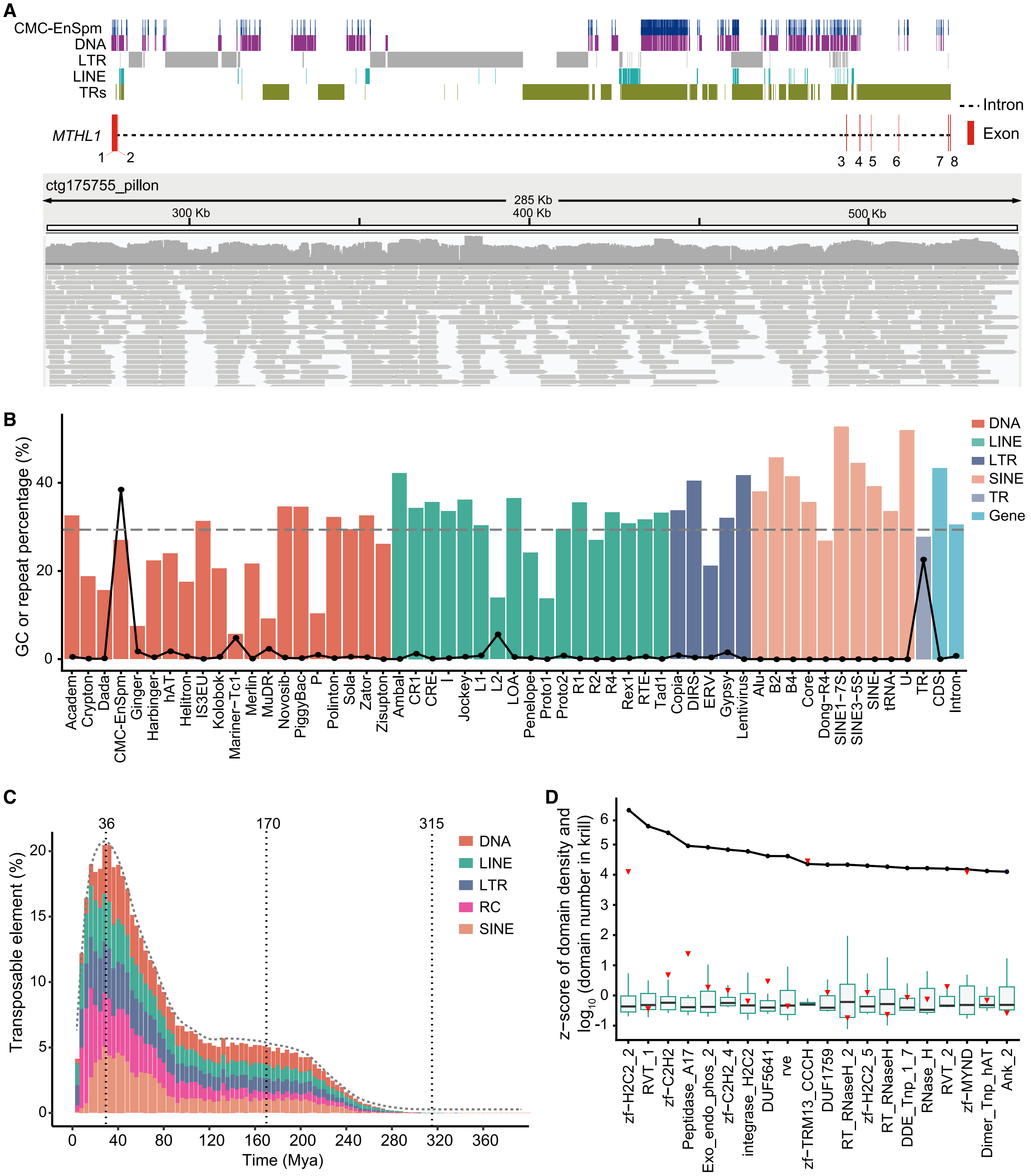


Figure 2. Effect of abundant repetitive sequences on the Antarctic krill genome

(A) Example of an Antarctic krill gene locus (*MTHL1*) with a high abundance of repetitive sequences in introns. This region was shown to be free of assembly errors by aligned PacBio CCS reads (visualized using Integrative Genomics Viewer [IGV] below).

(B) Histogram of GC content and genome proportion for each subtype of repetitive sequences and genes. The column height of the histogram indicates the GC content, the polyline indicates genome proportion, and the dashed line indicates average genome-wide GC content, DNA transposon (DNA), long interspersed nuclear element (LINE), long terminal repeat (LTR), short interspersed nuclear element (SINE), and tandem repeat (TR).

(legend continued on next page)

the genome was identified as repetitive sequence using standard repeat-masking procedures and reached up to 92.45% after additional repeat annotation, slightly higher than that reported for the Australian lungfish[15] (90.00%) (Figure 1E; Table S2). Transposable elements (TEs) constitute 78.22% of the Antarctic krill genome, and DNA TEs make up the largest proportion (Table S2). Notably, DNA/CMC-EnSpm accounted for 91.91% of DNA TEs, which formed 42.02% of the genome (Table S2). A phylogenetic tree of DNA/CMC-EnSpm among Antarctic krill, *L. vannamei*, and *P. virginalis* revealed no specific clades with dramatic expansion in Antarctic krill (Figure 1F).

We annotated 28,834 protein-coding genes in the Antarctic krill genome, which had gene models similar to genes of other related species, and coding sequence length comparable to the length of full-length transcripts (Figures S1D and S1E; Table S1).The gene and intron lengths of Antarctic krill are notably shorter than those of lungfishes and Mexican axolotl (Figures S1D and S1F; Table S1), suggesting that repetitive sequence expansions inserted in genic regions in Antarctic krill are limited compared to vertebrates with comparable genome sizes. However, compared with 46 other published marine invertebrates, the prevalence of TE insertions significantly increases intron length in Antarctic krill (Figures 2A and S1F). A previous study reported that the insertion of TEs into genes does not greatly impact gene regulation in African lungfish.[16] Similarly, we also observed that the gene length was independent of gene expression levels between Antarctic krill and other invertebrate species as well as between tissues of Antarctic krill (Figures S1G–S1J).

Dynamics of repetitive sequence expansion and its genetic mechanisms

Compared to 46 other invertebrate genomes, the repeat subtypes of Antarctic krill are more prevalent but show a similar composition and expansion pattern (Figure S2A; Table S2). A positive correlation between the proportion of TRs and TEs in invertebrates was observed (Figure S2B). In the Antarctic krill, most (96.39%) TRs overlapped with TEs (DNA transposon, long terminal repeat [LTR], and long interspersed nuclear element [LINE]) (STAR Methods). The high proportion of TRs may result from TE expansions coupled with slippage mutations of associated TRs, as recently reported for two penaeid shrimps (*L. vannamei* and *Fenneropenaeus chinensis*).[23]

The GC content of Antarctic krill is 29.36%, lower than 140 of 154 (90.91%) published invertebrate genome assemblies (Figures S2C and S2D; Table S1). This low GC content reflects a large proportion of GC-poor DNA transposons (Figure 2B). It has been argued that CpG dinucleotide loss follows genome size expansions by TEs, which limits the deleterious effects of TEs insertion.[24] We observed two putative TEs expansion events in the Antarctic krill, ∼36 and ∼170 mya (Figure 2C). The most recent event contributed to 39.51% of their genome expansion and is close to the emergence time of *Euphausia*, a krill genus with consistently large genome size,[12,25] while the proportion attributable to the other expansion is 18.54% (Figure 2C; Table S2).

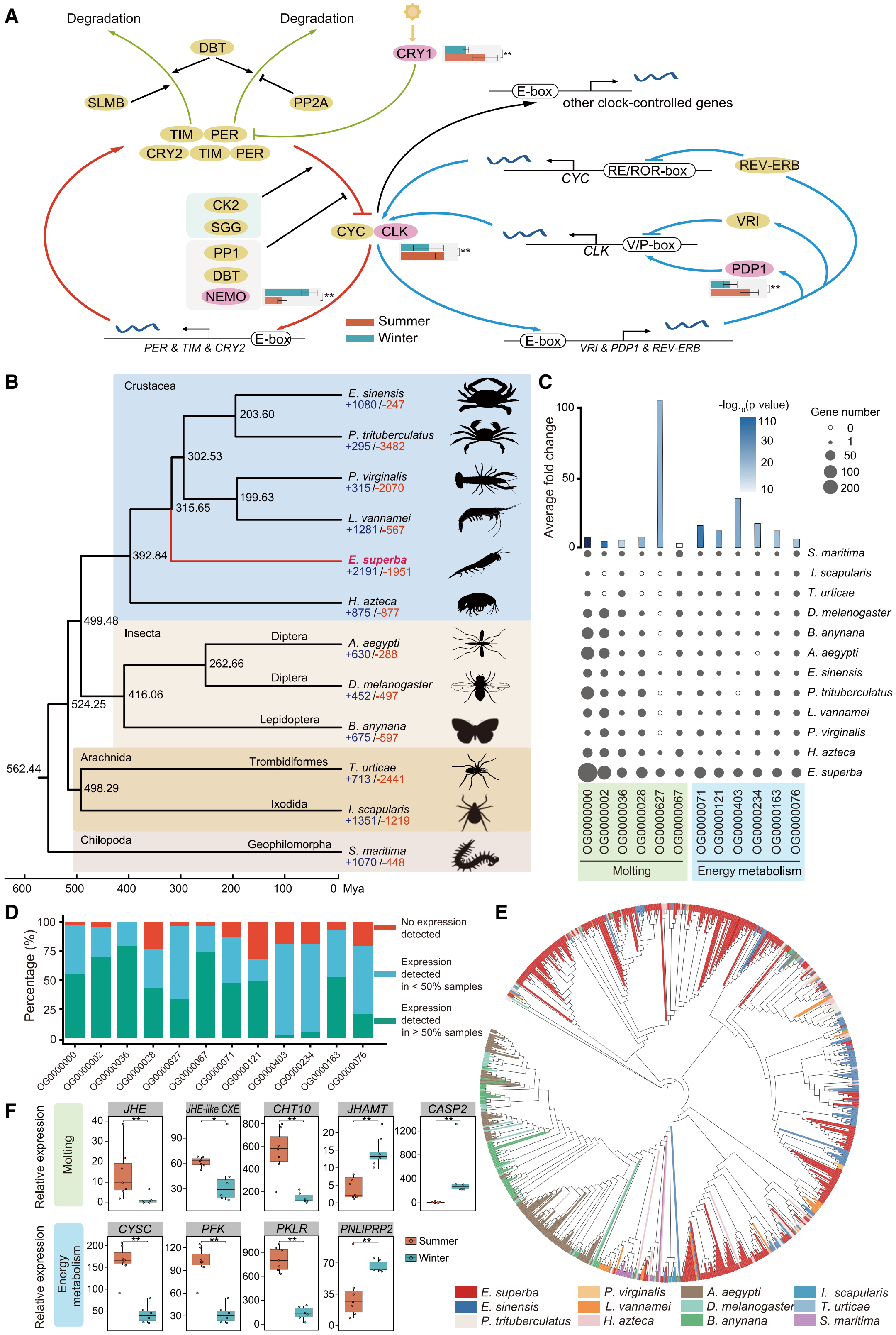
Levels of accumulation of TEs in host genomes result from the balance of TE activity and repression over long timescales.[24,26,27] To investigate this in Antarctic krill, we interrogated the genome using the protein domain Pfam database.[28] The top 20 domains account for 55.91% of detected Pfam domains, and 11 of the 20 domains play roles in transposable element (TE) activity (Figure 2D), such as reverse transcriptase (RVT_1) and integrase (integrase_H2C2). Notably, among the 20 top domains, we found three domains (zf-H2C2_2, zf-TRM13_CCCH, and zf-MYND) with a density higher than 46 other invertebrate species (*Z* score >3) (Figure 2D). The domain zf-TRM13_CCCH is found at the N terminus of TRM13 methyltransferase proteins, while zf-MYND is found in SET and MYND domain-containing (SMYD) methyltransferase proteins.[29] DNA methylation may enable TE-driven genome expansion.[24,30] We speculate that these protein domains in particular underly Antarctic krill genome size dynamics.

The genomic basis of environmental adaptations of Antarctic krill

Antarctic krill are able to maintain great abundance in the Southern Ocean because they have evolved seasonal synchronization strategies.[7] This makes adaptations for living with the highly variable levels of light, temperature, and sea ice the key to understanding the seasonal life cycle of Antarctic krill.[5–9] Light and temperature changes can entrain and reset the circadian system.[31,32] Antarctic krill are exposed to a cold environment with dramatic light changes caused by seasonal changes and have evolved genetic adaptations in circadian rhythm. It has been proposed that in Antarctic krill,[5] as in other eukaryotes,[33] the transcription factors CLOCK (CLK) and CYCLE (CYC) bind E-box elements upstream of the genes encoding their inhibitors, CRYPTOCHROME2 (CRY2), PERIOD (PER), and TIMELESS (TIM), to generate a self-sustained circadian rhythm (feedback loop). We found that 625 genes in the Antarctic krill genome contain at least one E-box (Antarctic-krill-specific consensus sequence CA[AT/TA]TG) within their promoter region (Table S3). These include the main clock inhibitors PER, TIM, and CRY2 and the three key circadian transcription factors VRI, PDP1, and REV-ERB that directly regulate CLK and CYC expression. Our findings provide a model of the molecular architecture of the krill circadian clock (Figure 3A), confirming that a dual feedback loop mechanism likely exists.[5] Many of the putative clock-controlled genes (58.3%) showed a daily oscillatory

(C) Insertion time of transposable elements (TEs) in Antarctic krill, the vertical dot-line from left to right, indicate two burst peaks of TEs and the divergence time of Antarctic krill, respectively. The percentages of each type of transposons are calculated separately. RC denotes rolling circle repeat.

(D) The distribution of number and density of the top 20 domains in the Antarctic krill genome. Bold line at the top of figure represents the number of domains, transformed by $\log_{10}$, boxplot on the bottom of the figure represents the distribution of domain density across the 47 invertebrate genomes. The domain density was calculated as domain number divided by genome size and *Z* score normalized. The median (Q2) is shown as a horizontal black line within the box, and the lower and upper ends of a box represent the first (Q1) and third quartile (Q3). The whiskers are defined by the inter-quartile range (IQR = Q3–Q1), extending no further than 1.5 × IQR. The domain density of Antarctic krill is highlighted with a red inverted triangle.



(legend on next page)

Please cite this article in press as: Shao et al., The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights, Cell (2023), https://doi.org/10.1016/j.cell.2023.02.005

expression profile in a previous study[6] (Table S3). We further assessed the seasonal differences in expression of genes in the biorhythm feedback loops, revealing that four circadian genes (*CLK*, *CRY1*, *NEMO*, and *PDP1*) show differential expression between summer and winter (Figure 3A). *CLK*, *CRY1*, and *PDP1* were upregulated during the summer, while *NEMO* was upregulated during the winter (fold change [FC] > 2, Benjamini-Hochberg corrected p value [p-*adj*] < 0.01) (Figure 3A). In *Drosophila*, *NEMO* is a serine/threonine kinase that regulates the speed of the circadian clock.[34] The increased *NEMO* expression in winter might suggest an involvement in the complex transition to the quiescent state, which was previously reported to be influenced by the circadian clock in Antarctic krill,[6] leading to sexual regression and decreased activity, growth, and metabolic rates.[35]

Antarctic krill have evolved physical adaptations and patterns of behavior governed by the circadian rhythm system that help them conserve energy and survive under low temperature and dramatically changing light conditions.[36] They can molt continuously throughout their life cycle, but their growth rates vary seasonally.[37] The intermolt period of Antarctic krill in winter is generally double that of summer and autumn (every 26–29 days),[37] with almost no relation to feeding regimes.[4] We identified 25 significantly expanded gene families in the Antarctic krill genome (p-*adj* < 0.05) (Figure 3B; Table S3). Twelve are directly involved in the molt cycle (six families) and energy metabolism (six families) (Figure 3C). Most genes in these families are expressed, indicating that the additional gene copies are functional (Figure 3D; Table S3).

Chitin is an essential building block of the crustacean cuticle.[38] The expansion of genes encoding proteins with chitin-binding domains in Antarctic krill (Figure 3E) may reflect finely regulated cuticle formation and resorption during the molt cycle.[38] Six expanded gene families associated with energy metabolism may reduce the maintenance costs of continuous molting (Figure 3C; Table S3). These families comprise genes encoding proteins with ATP binding domains, including *MHC*, *DDX5*, and *DDR2* (Table S3). In particular, we noted 69 myosin genes in Antarctic krill (on average 16-fold more than other species) (Table S3). Myosin gene expansions may have a range of functions relating to the unique life cycle of Antarctic krill, such as muscle contractions associated with body shrinkage during winter.[8]

We identified several genes that are differentially expressed between the summer and winter. All six copies of the gene encoding vitellogenin (*VTG*) were upregulated in the summer season (FC > 2, p-*adj* < 0.05) (Table S3). VTG is an essential egg yolk protein in invertebrates that provides a nutrient reservoir during the energetically demanding spawning season.[39] Additional energy metabolism-related genes—including *CYSC*, *PFK*, and *PKLR*—also showed upregulation during the summer (Figure 3F) and may support increased vitellogenesis and frequent molting at this time. One of two homologs of *PNLIPRP2*, a digestive lipase gene, was upregulated during the winter season and may aid survival during food shortages[40,41] (Figure 3F; Table S3). In addition, genes promoting molting and hence growth (*JHE*, *JHE-like CXE*, and *CHT10*) were upregulated in the summer when food availability is high, while genes inhibiting molting (*JHAMT* and *CASP2*) were upregulated in winter (Figure 3F). This finding agrees with a previous study of Antarctic krill molting during a time of relatively high temperatures, a long light regime, and increased food availability.[37] These results suggest that genomic innovations in the molt cycle and reproduction are adaptations to the extreme seasonal food availability in the Southern Ocean.

Antarctic krill population dynamics

There has been a long-standing debate about whether Antarctic krill represent a single genetically homogeneous population with panmixia in the Southern Ocean.[10,11] To investigate the population structure of Antarctic krill, we collected 75 individuals from four Southern Ocean regions with high biomass: South Georgia (SG) and South Shetland Island (SSI) in the Atlantic Ocean sector, Prydz Bay (PB) in the Indian Ocean sector, and Ross Sea (RS) in the Pacific Ocean Sector and carried out genome sequencing to an average depth of 17.72× (Figure 4A; Table S4).

We applied multiple quality control steps and obtained 364.57 million SNPs with an average density of one SNP per 37 bp (Table S4; STAR Methods), allowing for population genomic analysis of Antarctic krill. The mean nucleotide diversity ($\theta\pi$) and observed autosomal heterozygosity are 2.25×10^{-3} and 2.08×10^{-3}, respectively, with similar genetic diversity in the four geographical groups (Table S4). We also observed low pairwise F_{ST} values between Antarctic krill geographical groups, with a maximum F_{ST} of 1.92×10^{-3} (Figure 4B) and only 0.052%

Figure 3. Candidate genomic changes underlying adaptations to the Antarctic marine environment

(A) Connection of the circadian dual-feedback loop of Antarctic krill. Genes shaded in pink (*CRY1*, *CLK*, *PDP1*, and *NEMO*) showed significantly different expression between summer and winter (FC > 2, p-*adj* < 0.01) indicated by a bar graph with asterisks. While other genes shaded by yellow showed no expression difference between summer and winter. E-box denotes a promoter element found upstream of clock-controlled genes regulated by the transcription factors CLOCK (CLK) and CYCLE (CYC).

(B) Loss and gain of gene families mapped onto the phylogeny of 12 invertebrate species. Blue and red numbers indicate number of gene families with gained and lost on each branch, respectively (contains the significant and non-significant gene families). Divergence time estimates (million years ago, mya) is shown at each node and the red branch indicates Antarctic krill.

(C) Gene families significantly expanded in Antarctic krill are associated with molting and energy metabolism (family-wise p value < 0.05 based on a Monte–Carlo re-sampling procedure). The number of genes in each gene family represented by bubble size. The average FC of Antarctic krill gene number in each family compared to other species is shown in the upper histogram, with the shading indicating $\log_{10}$ p value.

(D) Expression of 12 significantly expanded gene family members in 55 Antarctic krill transcriptome samples. For each gene family, the frequency (percentage of expressed samples) is shown.

(E) Phylogenetic tree of genes in the molting associated gene family OG0000000 in 12 species.

(F) Boxplots of differentially expressed genes in a summer-winter comparison of Antarctic krill sampled from the Lazarev Sea. The y axis represents the relative expression, * and ** indicate Benjamini-Hochberg corrected p value below 0.05 and 0.01, respectively.

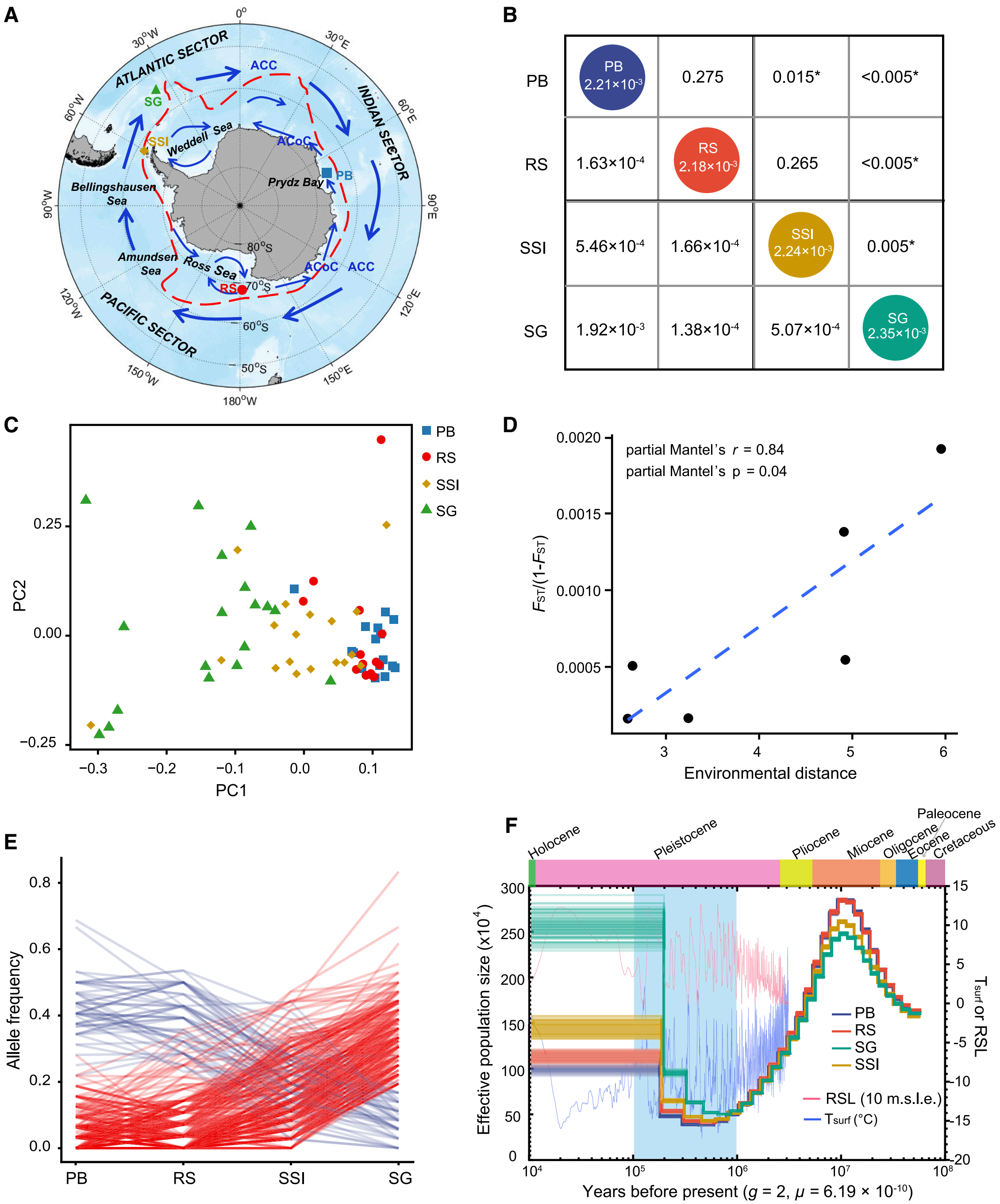


Figure 4. Antarctic krill population dynamics

(A) Antarctic krill samples were collected from four distinct geographical regions around Antarctica, including the Prydz Bay (PB, square), the Ross Sea (RS, dot), the South Shetland Islands (SSI, diamond) and the South Georgia (SG, triangle). The red dashed-lines represent the Antarctic Divergence separating the prevailing Antarctic Circumpolar Current (ACC, thick blue arrowed-lines) with the Antarctic Coastal Current (ACoC, thin blue arrowed-lines) and the gyres in the Ross Sea and the Weddell Sea.

(B) The genetic diversity ($\theta\pi$) of four geographical groups is shown in the diagonal circles. The population differentiation (F_{ST}) between pairwise Antarctic krill groups is on the lower triangle. The significance of F_{ST} by permutation test is shown on the upper triangle, and the p values less than 0.05 are marked by asterisks.

(C) Principal component analysis (PCA) of 66 individuals (depth >10×) from four geographical groups based on 47,555,257 SNPs after linkage disequilibrium (LD) pruning.

(legend continued on next page)

SNPs with $F_{ST} > 0.15$ (Table S4). The very low average F_{ST} and small percentage of differentiated SNPs indicate that there is no substantial differentiation within and between geographical groups. These results—together with the inferred gene flow and genetic connectivity between PB, RS, SSI, and SG using ABBA-BABA statistics[42] (Figure S3A), TreeMix inference[43] (Figures S3B and S3C), and BA3-SNPs[44] (Figure S3D)—indicate extensive mixing and connectivity on a large geographical scale between Antarctic krill geographical groups.

However, principal component analysis (PCA) (Figures 4C, S3E, and S3F), multidimensional scaling analysis (MDS) (Figure S3G), and neighbor-joining (NJ) tree (Figure S3H) suggest that the genetic structure is identifiable in Antarctic krill, especially between SG and PB-RS. The population structure analysis by STRUCTURE[45] and NGSadmix[46] also revealed the different proportions of ancestry components for four geographical groups (Figures S3I–S3P). This result was also supported by the permutation test that the F_{ST} values from random selected individuals were significantly lower than that between different geographical groups (permutation test, $p < 0.05$) (Table S4). This pattern of differentiation mainly depends on a small number of SNPs. When outlier SNPs with $F_{ST} > 0.1$ are removed (0.36% of SNPs), individuals from the four geographical groups were mixed[43] (Figures S3Q–S3T). Overall, these results indicate the minor geographic structure that is detectable is due to a small percentage of differentiated SNPs, which reflects the power of our very large SNP dataset.

In a large population, genetic drift is limited, and this slows differentiation of neutral markers. However, natural selection is more effective at fixing beneficial mutations, even if selection is weak.[47,48] We examined associations between SNPs and ten environmental variables at four geographical locations as an indication of potential environmental selection[49] (Figure S4A; Table S4). The isolation-by-distance (IBD) test was not significant (Figure S4B), but the isolation-by-environment (IBE) analysis revealed that the genetic differentiation is significantly correlated with environmental distance (partial Mantel test, $p = 0.04$, $r = 0.84$, two-sided) (Figure 4D). Furthermore, we detected 387 potentially adaptive SNPs dispersed across the genome associated with environmental variables using the latent factor mixed model[50] (Figure S4C). The allele frequency of these 387 adaptive SNPs in four groups revealed the distinct genetic patterns between SG-SSI and PB-RS groups (Figures 4E and S4D). Our results suggest that environmental selection may play an important role in driving genetic structure in different groups of Antarctic krill.

To uncover the demographic history of Antarctic krill, we employed the pairwise sequentially Markovian coalescent (PSMC) method[51] and PopSizeABC[52] inference to estimate past effective population sizes (*Ne*). We found a drastic reduction in *Ne* from approximately ten mya (Figures 4F and S4E–S4L), coinciding with extensive amplitudes of glacial-interglacial variations during the Pleistocene Epoch and an overall decrease in Southern Ocean temperature.[53,54] The overall peak in population size around ten mya is associated with the formation time of a stable Antarctic Ice Sheet and the presence of a consistent Antarctic circumpolar current (ACC) (Figure 4F).[55] We also observed a subsequent expansion in the Antarctic krill group from more than 100 thousand years ago, which has previously been linked with a cooler climate during the late Pleistocene age and a larger area of beneficial sea ice habitat (Figure 4F).[56] The cooler climate in the last ice age expanded Antarctic ice sheets, which would have provided an expanded habitat and ecological release for krill.[56] This population bottleneck and subsequent expansion are supported by an excess of rare alleles (Tajima's *D* ranging from −1.35 to −1.31) (Table S4). These inferences from genomic data correspond with historical and recent temperature changes;[57] however, the impact of temperature on Antarctic krill over long time frames is complex since the species continuously evolves to live in different conditions and the ecological context changes over time. The impact of rapid climate change on Antarctic krill will be hard to predict. Antarctic krill is a key species of important Southern Ocean food webs, yet how changes in ocean temperature and primary production may impact their habitat quality remain poorly understood. The habitat of krill will likely shift to higher latitudes in these areas, but how climate change will impact krill population size, and consequently the Antarctic ecosystem that depends on krill, are critical questions that need to be addressed urgently.[58]

DISCUSSION

Antarctic krill form a critical link in the food web in the Southern Ocean and influences ecosystem functionality because of their enormous biomass. Here we present an Antarctic krill genome, the largest animal assembly to date. The only assembled animal genomes of comparable size are the recently released, slightly smaller, Mexican axolotl, Australian lungfish, and African lungfish assemblies.[15–18] Although animals with huge genomes often have a high proportion of repetitive sequences, their TE expansions show different patterns. That is, DNA transposons, especially DNA/CMC-EnSpm, are predominant in Antarctic krill, while the LINE and LTR are the most dominant TEs of the lungfish genomes[15,16] and LTRs are dominant for the Mexican axolotl. Furthermore, we identified an ancient accumulation of TEs with the two recent bursts, and the most recent event is close to the emergence time of *Euphausia*, which may partly explain the large genome size of this krill genus.[12,25] A large genome size appears common in polar crustaceans.[12] As a result, we demonstrated that the Antarctic krill assembly will provide an incentive for sequencing efforts that can answer why and how a huge genome size is typical in polar crustaceans.

(D) IBE analyses (partial Mantel test, two-sided, controlling for the effect of geographic distance) for four geographical groups to reveal the correlation between environmental distances and the genetic distances ($F_{ST}/(1\text{-}F_{ST})$).

(E) Distribution of the allele frequency of 387 adaptive SNPs in four geographical groups, with differences of allele frequency in distribution pattern between PB/RS and SSI/SG.

(F) Estimation of historical effective population sizes (*Ne*) of Antarctic krill using PSMC. A generation time (*g*) of two years and mutation rate (μ) of 6.19×10^{-10} substitutions per site per generation was employed. Blue lines represent surface temperatures, red lines for relative sea level (RSL). The light blue shading indicates a period of expansion after a population bottleneck of Antarctic krill.

Circadian rhythm is controlled by molecular clock genes that cooperate to generate cyclic changes in their abundance and activity in response to environmental cues.[59] The extreme seasonal changes in terms of the day length of the Southern Ocean may fundamentally alter the circadian system. Here, we connected the dual-feedback loop system of the Antarctic krill circadian clock, the fundamental genetic architecture of circadian oscillations. Comparing the circadian components with other organisms (mammals and *Drosophila*), we found that the main framework of the circadian system has not changed, but the gene expression (*CRY1*, *CLK*, *NEMO*, and *PDP1*) of the feedback pathway may show a different expression pattern. It is clear that more studies are needed in order to reveal the concrete functional roles of these genes, which are major drivers of Antarctic krill adaptations to the Antarctic environment.

Previous Antarctic krill population genetic studies have relied on inferences from a limited number of markers from mtDNA[10,56] and low-coverage restriction-site-associated DNA sequencing.[11] Our assembly and hundreds of millions of SNPs across 75 individuals greatly expand population genetic insights. Given the large population size of Antarctic krill, the genome-wide estimation of genetic diversity is relatively low[60] (Table S4). The low genetic diversity in large populations, known as "Lewontin's paradox," could be due to a low mutation rate of Antarctic krill[61,62] (Table S4). This low mutation rate may be a by-product of the low GC content of the krill genome and/or the increase power of selection in large populations.[63] We also found that the ratio of the effective population size (*N*e) to census population size (*Ne*/*N*c) was 5 × 10^{-9} (STAR Methods), providing one of the most extreme examples of this difference found to date.[64] Thus, a mixture of evolutionary force including relatively small *Ne*/*N*c, low mutation rate, and natural selection likely shape the low genetic diversity in Antarctic krill.

Our results indicate that the Antarctic krill population is essentially panmictic, with high levels of connectivity on a large geographical scale around the entire Antarctic continent. The species range overlaps extensively with what is considered to be the strongest ocean current (ACC) in the world accompanying by the Antarctic Coastal Current (ACoC), and movement of krill in the currents is likely to account for the overall genetic homogenetity[65–69] (Figure 4A). The large statistical power of our dataset did detect extremely low levels of genetic differentiation. Our analysis indicates that this very subtle signal is dependent on a small number of loci that are subject to selection. Krill from SG, the only population we analyzed located north of the Antarctic Divergence, were the most genetically differentiated, presumably due to selective forces acting on a few parts of the genome in this distinct environment. Generally, in groups comprised of large numbers of individuals, as we see in Antarctic krill, genetic drift will have very limited effects on SNP frequency, allowing loci with minor selective advantages to become established.[70] This also indicates that changes in gene frequencies between these geographical groups are most likely the result of natural selection related to local adaptation. Confirmation of any weak geographic structuring at neutral markers would require more widespread sampling and ideally collection of krill over a time series to assess temporal stability.[71] Our results suggest that area-specific fishery conservation measures may still be warranted to maintain krill functional genetic diversity.

The major technological highlight of the study is assembly of the largest animal genome ever sequenced. This technical challenge was exacerbated by the hyper-abundant TR DNA in the genome, which became one of the major biological findings of our work. We carefully analyzed the repeat sequences that contribute to the enormous genome size, and this provides one of the best examples of genome size expansion caused by repeat element activity. The assembled genome allowed us to comprehensively analyze genes involved in photoperiodicity throughout the whole genome. Physiological responses to the highly variable light conditions of the Antarctic are central to krill biology, and studying this adaptation in detail was greatly improved by the genome resource we have generated. Finally, genome-wide SNPs are used to address the long-standing question of population genetic differentiation in Antarctic krill. Population structure was very limited, with neutral SNPs showing no major genetic differentiation but with some evidence that local conditions may be having selective effects on a subset of the SNPs in the genome. In summary, the krill genome and detailed population genetic analysis will undoubtedly aid future research relevant for management on this keystone Antarctic species.

Limitations of study

Due to the large genome size and TR abundance of Antarctic krill, the continuity of genome assembly is not as long as the vertebrate genomes of lungfishes and Mexican axolotl. Additionally, it would have been valuable to have more widespread sampling and ideally collection of krill over a time series to assess temporal stability and reliability of the inference on population dynamics and natural selection.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Source organisms
- METHOD DETAILS
  - ○ Sampling and sequencing
  - ○ Genome assembly and evaluation
  - ○ Genome annotation
  - ○ Repetitive sequences analysis
  - ○ Comparative genomics related to adaptation
  - ○ SNP detection
  - ○ Population structure and genetic diversity
  - ○ Gene flow and population connectivity
  - ○ Detecting SNPs related to natural selection
  - ○ Demographic history inference
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.02.005.

ACKNOWLEDGMENTS

This study was supported by the National Key R&D Program of China (2022YFD2400100, 2022YFC2807500); the Marine S&T Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (No. 2022QNLM030004); the Central Public-interest Scientific Institution Basal Research Fund, CAFS (NO.2019QY01); the Major Science and Technology Innovation Project of Shandong Province (2018SDKJ0302); the AoShan Talents Cultivation Program Supported by Qingdao National Laboratory for Marine Science and Technology (2017ASTCP-ES06); the Taishan Scholars Program (NO.tstp20221149 to C.S.); the National Ten-Thousands Talents Special Support Program to C.S.; the Central Public-interest Scientific Institution Basal Research Fund, CAFS (No.2020TD19); the Specially-Appointed Professor Program of Jiangsu Province to I.S.; the Jiangsu Foreign Expert Bureau to I.S.; and the Helmholtz Earth and Environment Research Program "Changing Earth - Sustaining our Future", Topic 6, Subtopic 6.2 to B.M. We appreciate the support from the supercomputing center of Qingdao National Laboratory for Marine Science and Technology. We also appreciate the technical support from China National GeneBank for the library construction and sequencing in Hi-C and whole-genome short reads.

AUTHOR CONTRIBUTIONS

C.S. initiated the Antarctic krill genome project. C.S., G.F., and B.M. conceived the study and coordinated the work. C.S., G.F., S.S., K.L., Jiahao Wang, Shuo Li, I.S., X.L., S.J., B.E.D., G.Z. and H.Y. wrote the manuscript with contributions from all other authors. C.S., X.Z., G.F., B.E.D., Yang Wang, S.K., Y.Y., and Shanshan Liu. coordinated and performed sample collection and sequencing. K.L., J.H., Shuo Li, Q.L., Jiahao Wang, T.S., and J.R. performed genome assembly into contigs and genome correction and transcript alignment. Jiahao Wang, K.L., S.S., S.P., and H.Z. performed repeat analysis. Shuo Li, K.L., G.H., B.F., R.W., K.M., X.J., and X.H. undertook transcriptome and genome annotation analysis. S.L., A.B., C.D.P., G.S., Yaolei Zhang, K.L., B.F., J.X., X.Z., Jun Wang, and B.M. conducted comparative genomics analysis. Y.S., B.Y., Q.W., and H.-Y.W. performed the isolation of RNA and DNA and the validation of SNPs by PCR and Sanger sequencing. S.S., B.E.D., S.J., Yaolei Zhang, M.X., Yalin Liu, Yuyan Liu, J.Z., X.W., Shufang Liu; Yue Wang, and X.X. conducted population analysis.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 10, 2022
Revised: December 11, 2022
Accepted: February 2, 2023
Published: March 2, 2023

REFERENCES

1. Bar-On, Y.M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. Proc. Natl. Acad. Sci. USA *115*, 6506–6511. https://doi.org/10.1073/pnas.1711842115.
2. Hill, S.L., Murphy, E.J., Reid, K., Trathan, P.N., and Constable, A.J. (2006). Modelling Southern Ocean ecosystems: krill, the food-web, and the impacts of harvesting. Biol. Rev. Camb. Philos. Soc. *81*, 581–608. https://doi.org/10.1017/s1464793106007123.
3. Cavan, E.L., Belcher, A., Atkinson, A., Hill, S.L., Kawaguchi, S., McCormack, S., Meyer, B., Nicol, S., Ratnarajah, L., Schmidt, K., et al. (2019). The importance of Antarctic krill in biogeochemical cycles. Nat. Commun. *10*, 4742. https://doi.org/10.1038/s41467-019-12668-7.
4. Manno, C., Fielding, S., Stowasser, G., Murphy, E.J., Thorpe, S.E., and Tarling, G.A. (2020). Continuous moulting by Antarctic krill drives major pulses of carbon export in the north Scotia Sea, Southern Ocean. Nat. Commun. *11*, 6051. https://doi.org/10.1038/s41467-020-19956-7.
5. Biscontin, A., Wallach, T., Sales, G., Grudziecki, A., Janke, L., Sartori, E., Bertolucci, C., Mazzotta, G., De Pittà, C., Meyer, B., et al. (2017). Functional characterization of the circadian clock in the Antarctic krill, *Euphausia superba*. Sci. Rep. *7*, 17742. https://doi.org/10.1038/s41598-017-18009-2.
6. Biscontin, A., Martini, P., Costa, R., Kramer, A., Meyer, B., Kawaguchi, S., Teschke, M., and De Pittà, C. (2019). Analysis of the circadian transcriptome of the Antarctic krill *Euphausia superba*. Sci. Rep. *9*, 13894. https://doi.org/10.1038/s41598-019-50282-1.
7. Ducklow, H.W., Baker, K., Martinson, D.G., Quetin, L.B., Ross, R.M., Smith, R.C., Stammerjohn, S.E., Vernet, M., and Fraser, W. (2007). Marine pelagic ecosystems: the West Antarctic Peninsula. Philos. Trans. R. Soc. Lond. B Biol. Sci. *362*, 67–94. https://doi.org/10.1098/rstb.2006.1955.
8. Seear, P.J., Goodall-Copestake, W.P., Fleming, A.H., Rosato, E., and Tarling, G.A. (2012). Seasonal and spatial influences on gene expression in Antarctic krill *Euphausia superba*. Mar. Ecol. Prog. Ser. *467*, 61–75. https://doi.org/10.3354/meps09947.
9. Urso, I., Biscontin, A., Corso, D., Bertolucci, C., Romualdi, C., De Pittà, C., Meyer, B., and Sales, G. (2022). A thorough annotation of the krill transcriptome offers new insights for the study of physiological processes. Sci. Rep. *12*, 1–15. https://doi.org/10.1038/s41598-022-15320-5.
10. Bortolotto, E., Bucklin, A., Mezzavilla, M., Zane, L., and Patarnello, T. (2011). Gone with the currents: lack of genetic differentiation at the circum-continental scale in the Antarctic krill *Euphausia superba*. BMC Genet. *12*, 32. https://doi.org/10.1186/1471-2156-12-32.
11. Deagle, B.E., Faux, C., Kawaguchi, S., Meyer, B., and Jarman, S.N. (2015). Antarctic krill population genomics: apparent panmixia, but genome complexity and large population size muddy the water. Mol. Ecol. *24*, 4943–4959. https://doi.org/10.1111/mec.13370.
12. 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 Biol. *35*, 959–962. https://doi.org/10.1007/s00300-011-1137-4.
13. Huang, Y., Bian, C., Liu, Z., Wang, L., Xue, C., Huang, H., Yi, Y., You, X., Song, W., Mao, X., et al. (2020). The First Genome Survey of the Antarctic Krill (*Euphausia superba*) Provides a Valuable Genetic Resource for Polar Biomedical Research. Mar. Drugs *18*, 185. https://doi.org/10.3390/md18040185.
14. Jarman, S.N., and Deagle, B.E. (2016). Genetics of Antarctic Krill. In Biology and Ecology of Antarctic Krill Advances in Polar Ecology, V. Siegel, ed. (Springer International Publishing), pp. 247–277. https://doi.org/10.1007/978-3-319-29279-3_7.
15. Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J.M., Irisarri, I., Wong, W.Y., Nowoshilow, S., Kneitz, S., Kawaguchi, A., et al. (2021). Giant lungfish genome elucidates the conquest of land by vertebrates. Nature *590*, 284–289. https://doi.org/10.1038/s41586-021-03198-8.
16. Wang, K., Wang, J., Zhu, C., Yang, L., Ren, Y., Ruan, J., Fan, G., Hu, J., Xu, W., Bi, X., et al. (2021). African lungfish genome sheds light on the vertebrate water-to-land transition. Cell *184*, 1362–1376.e18. https://doi.org/10.1016/j.cell.2021.01.047.
17. Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. Nature *554*, 50–55. https://doi.org/10.1038/nature25458.
18. Schloissnig, S., Kawaguchi, A., Nowoshilow, S., Falcon, F., Otsuki, L., Tardivo, P., Timoshevskaya, N., Keinath, M.C., Smith, J.J., Voss, S.R., and Tanaka, E.M. (2021). The giant axolotl genome uncovers the evolution, scaling, and transcriptional control of complex gene loci. Proc.

Natl. Acad. Sci. USA *118*. e2017176118. https://doi.org/10.1073/pnas.2017176118.

19. Catherine, T.-Q., Leitão, A., and Cuzin-Roudy, J. (1998). Chromosome Diversity in Mediterranean and Antarctic Euphausiid Species (Euphausiacea). J. Crustac Biol. *18*, 290–297. https://doi.org/10.2307/1549322.
20. Van Ngan, P., Gomes, V., Suzuki, H., and Passos, M.J.A.C.R. (1989). Preliminary study on chromosomes of Antarctic krill, *Euphausia superba* Dana. Polar Biol. *10*, 149–150. https://doi.org/10.1007/BF00239161.
21. Tørresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A.V., Promponas, V.J., et al. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. *47*, 10994–11006. https://doi.org/10.1093/nar/gkz841.
22. Zhang, X., Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, X., et al. (2019). Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat. Commun. *10*, 356. https://doi.org/10.1038/s41467-018-08197-4.
23. Yuan, J., Zhang, X., Wang, M., Sun, Y., Liu, C., Li, S., Yu, Y., Gao, Y., Liu, F., Zhang, X., et al. (2021). Simple sequence repeats drive genome plasticity and promote adaptive evolution in penaeid shrimp. Commun. Biol. *4*, 186–214. https://doi.org/10.1038/s42003-021-01716-y.
24. Zhou, W., Liang, G., Molloy, P.L., and Jones, P.A. (2020). DNA methylation enables transposable element-driven genome expansion. Proc. Natl. Acad. Sci. USA *117*, 19359–19366. https://doi.org/10.1073/pnas.1921719117.
25. Patarnello, T., Bargelloni, L., Varotto, V., and Battaglia, B. (1996). Krill evolution and the Antarctic ocean currents: evidence of vicariant speciation as inferred by molecular data. Mar. Biol. *126*, 603–608. https://doi.org/10.1007/BF00351327.
26. Bruno, M., Mahgoub, M., and Macfarlan, T.S. (2019). The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. Annu. Rev. Genet. *53*, 393–416. https://doi.org/10.1146/annurev-genet-112618-043717.
27. Fedoroff, N.V. (2012). Transposable elements, epigenetics, and genome evolution. Science *338*, 758–767. https://doi.org/10.1126/science.338.6108.758.
28. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.
29. Spellmon, N., Holcomb, J., Trescott, L., Sirinupong, N., and Yang, Z. (2015). Structure and Function of SET and MYND Domain-Containing Proteins. Int. J. Mol. Sci. *16*, 1406–1428. https://doi.org/10.3390/ijms16011406.
30. Niu, S., Li, J., Bo, W., Yang, W., Zuccolo, A., Giacomello, S., Chen, X., Han, F., Yang, J., Song, Y., et al. (2022). The Chinese pine genome and methylome unveil key features of conifer evolution. Cell *185*, 204–217.e14. https://doi.org/10.1016/j.cell.2021.12.006.
31. Duffy, J.F., and Wright, K.P. (2005). Entrainment of the Human Circadian System by Light. J. Biol. Rhythms *20*, 326–338. https://doi.org/10.1177/0748730405277983.
32. Liu, Y., Merrow, M., Loros, J.J., and Dunlap, J.C. (1998). How Temperature Changes Reset a Circadian Oscillator. Science *281*, 825–829. https://doi.org/10.1126/science.281.5378.825.
33. Seitz, S.B., Voytsekh, O., Mohan, K.M., and Mittag, M. (2010). The role of an E-box element. Plant Signal. Behav. *5*, 1077–1080. https://doi.org/10.4161/psb.5.9.12564.
34. Yu, W., Houl, J.H., and Hardin, P.E. (2011). NEMO kinase contributes to core period determination by slowing the pace of the *Drosophila* circadian oscillator. Curr. Biol. *21*, 756–761. https://doi.org/10.1016/j.cub.2011.02.037.
35. Meyer, B., Auerswald, L., Siegel, V., Spahic, c., Pape, C., Fach, B., Teschke, M., Lopata, A.L., and Fuentes, V. (2010). Seasonal variation in body composition, metabolic activity, feeding, and growth of adult krill *Euphausia superba* in the Lazarev Sea. Mar. Ecol. Prog. Ser. *398*, 1–18. https://doi.org/10.3354/meps08371.
36. Ikeda, T., and Dixon, P. (1982). Body shrinkage as a possible overwintering mechanism of the Antarctic krill, *Euphausia superba* Dana. J. Exp. Mar. Biol. Ecol. *62*, 143–151. https://doi.org/10.1016/0022-0981(82)90088-0.
37. Meyer, B. (2012). The overwintering of Antarctic krill, *Euphausia superba*, from an ecophysiological perspective. Polar Biol. *35*, 15–37. https://doi.org/10.1007/s00300-011-1120-0.
38. Seear, P.J., Tarling, G.A., Burns, G., Goodall-Copestake, W.P., Gaten, E., Özkaya, Ö., and Rosato, E. (2010). Differential gene expression during the moult cycle of Antarctic krill (*Euphausia superba*). BMC Genom. *11*, 582. https://doi.org/10.1186/1471-2164-11-582.
39. Robinson, R. (2008). For Mammals, loss of yolk and gain of milk went hand in hand. PLoS Biol. *6*, e77. https://doi.org/10.1371/journal.pbio.0060077.
40. Berton, A., Sebban-Kreuzer, C., Rouvellac, S., Lopez, C., and Crenon, I. (2009). Individual and combined action of pancreatic lipase and pancreatic lipase-related proteins 1 and 2 on native versus homogenized milk fat globules. Mol. Nutr. Food Res. *53*, 1592–1602. https://doi.org/10.1002/mnfr.200800563.
41. Wente, W., Brenner, M.B., Zitzer, H., Gromada, J., and Efanov, A.M. (2007). Activation of liver X receptors and retinoid X receptors induces growth arrest and apoptosis in insulin-secreting cells. Endocrinology *148*, 1843.
42. Malinsky, M., Matschiner, M., and Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. Mol. Ecol. Resour. *21*, 584–595. https://doi.org/10.1111/1755-0998.13265.
43. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. *8*, e1002967. https://doi.org/10.1371/journal.pgen.1002967.
44. Mussmann, S.M., Douglas, M.R., Chafin, T.K., and Douglas, M.E. (2019). BA3-SNPs: Contemporary migration reconfigured in BayesAss for next-generation sequence data. Methods Ecol. Evol. *10*, 1808–1813. https://doi.org/10.1111/2041-210X.13252.
45. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. Genetics *155*, 945–959. https://doi.org/10.1093/genetics/155.2.945.
46. Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. Genetics *195*, 693–702. https://doi.org/10.1534/genetics.113.154138.
47. Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. *10*, 195–205. https://doi.org/10.1038/nrg2526.
48. Gravel, S. (2016). When is selection effective? Genetics *203*, 451–462. https://doi.org/10.1534/genetics.115.184630.
49. Hauser, L., and Carvalho, G.R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. Fish Fish. *9*, 333–362. https://doi.org/10.1111/j.1467-2979.2008.00299.x.
50. Caye, K., Jumentier, B., Lepeule, J., and François, O. (2019). LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. Mol. Biol. Evol. *36*, 852–860. https://doi.org/10.1093/molbev/msz008.
51. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496. https://doi.org/10.1038/nature10231.
52. Boitard, S., Rodríguez, W., Jay, F., Mona, S., and Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. PLoS Genet. *12*, e1005877. https://doi.org/10.1371/journal.pgen.1005877.
53. Elderfield, H., Ferretti, P., Greaves, M., Crowhurst, S., McCave, I.N., Hodell, D., and Piotrowski, A.M. (2012). Evolution of ocean temperature and

ice volume through the mid-pleistocene climate transition. Science *337*, 704–709. https://doi.org/10.1126/science.1221294.

54. Miller, K.G., Browning, J.V., Schmelz, W.J., Kopp, R.E., Mountain, G.S., and Wright, J.D. (2020). Cenozoic sea-level and cryospheric evolution from deep-sea geochemical and continental margin records. Sci. Adv. *6*, eaaz1346. https://doi.org/10.1126/sciadv.aaz1346.

55. Young, D.A., Wright, A.P., Roberts, J.L., Warner, R.C., Young, N.W., Greenbaum, J.S., Schroeder, D.M., Holt, J.W., Sugden, D.E., Blankenship, D.D., et al. (2011). A dynamic early East Antarctic Ice Sheet suggested by ice-covered fjord landscapes. Nature *474*, 72–75. https://doi.org/10.1038/nature10114.

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

57. Buizert, C., Fudge, T.J., Roberts, W.H.G., Steig, E.J., Sherriff-Tadano, S., Ritz, C., Lefebvre, E., Edwards, J., Kawamura, K., Oyabu, I., et al. (2021). Antarctic surface temperature and elevation during the Last Glacial Maximum. Science *372*, 1097–1101. https://doi.org/10.1126/science.abd2897.

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

59. Patke, A., Young, M.W., and Axelrod, S. (2020). Molecular mechanisms and physiological importance of circadian rhythms. Nat. Rev. Mol. Cell Biol. *21*, 67–84. https://doi.org/10.1038/s41580-019-0179-2.

60. Robinson, J.A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B.Y., vonHoldt, B.M., Marsden, C.D., Lohmueller, K.E., and Wayne, R.K. (2016). Genomic flatlining in the endangered Island Fox. Curr. Biol. *26*, 1183–1189. https://doi.org/10.1016/j.cub.2016.02.062.

61. Xu, S., Stapley, J., Gablenz, S., Boyer, J., Appenroth, K.J., Sree, K.S., Gershenzon, J., Widmer, A., and Huber, M. (2019). Low genetic variation is associated with low mutation rate in the giant duckweed. Nat. Commun. *10*, 1243. https://doi.org/10.1038/s41467-019-09235-5.

62. Krasovec, M., Rickaby, R.E.M., and Filatov, D.A. (2020). Evolution of mutation rate in astronomically large phytoplankton populations. Genome Biol. Evol. *12*, 1051–1059. https://doi.org/10.1093/gbe/evaa131.

63. Kiktev, D.A., Sheng, Z., Lobachev, K.S., and Petes, T.D. (2018). GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *115*, E7109–E7118. https://doi.org/10.1073/pnas.1807334115.

64. Hedgecock, D., and Pudovkin, A.I. (2011). Sweepstakes reproductive success in sighly fecund marine fish and shellfish: a review and commentary. Bull. Mar. Sci. *87*, 971–1002. https://doi.org/10.5343/bms.2010.1051.

65. Piñones, A., Hofmann, E.E., Daly, K.L., Dinniman, M.S., and Klinck, J.M. (2013). Modeling the remote and local connectivity of Antarctic krill populations along the western Antarctic Peninsula. Mar. Ecol. Prog. Ser. *481*, 69–92. https://doi.org/10.3354/meps10256.

66. Piñones, A., Hofmann, E.E., Dinniman, M.S., and Klinck, J.M. (2011). Lagrangian simulation of transport pathways and residence times along the western Antarctic Peninsula. Deep Sea Res. Part II Top. Stud. Oceanogr. *58*, 1524–1539. https://doi.org/10.1016/j.dsr2.2010.07.001.

67. Atkinson, A., Siegel, V., Pakhomov, E.A., Rothery, P., Loeb, V., Ross, R.M., Quetin, L.B., Schmidt, K., Fretwell, P., Murphy, E.J., et al. (2008). Oceanic circumpolar habitats of Antarctic krill. Mar. Ecol. Prog. Ser. *362*, 1–23. https://doi.org/10.3354/meps07498.

68. Thorpe, S.E., Murphy, E.J., and Watkins, J.L. (2007). Circumpolar connections between Antarctic krill (*Euphausia superba* Dana) populations: Investigating the roles of ocean and sea ice transport. Deep Sea Res. Oceanogr. Res. Pap. *54*, 792–810. https://doi.org/10.1016/j.dsr.2007.01.008.

69. Youngs, M.K., Thompson, A.F., Flexas, M.M., and Heywood, K.J. (2015). Weddell Sea export pathways from surface frifters. J. Phys. Oceanogr. *45*, 1068–1085. https://doi.org/10.1175/JPO-D-14-0103.1.

70. Waples, R.S. (1998). Separating the wheat from the Chaff: Patterns of genetic differentiation in high gene flow species. J. Hered. *89*, 438–450. https://doi.org/10.1093/jhered/89.5.438.

71. Knutsen, H., Olsen, E.M., Jorde, P.E., Espeland, S.H., André, C., and Stenseth, N.C. (2011). Are low but statistically significant levels of genetic differentiation in marine fishes 'biologically meaningful'? A case study of coastal Atlantic cod. Mol. Ecol. *20*, 768–783. https://doi.org/10.1111/j.1365-294X.2010.04979.x.

72. Matthews, B.J., Dudchenko, O., Kingan, S.B., Koren, S., Antoshechkin, I., Crawford, J.E., Glassford, W.J., Herre, M., Redmond, S.N., Rose, N.H., et al. (2018). Improved reference genome of *Aedes aegypti* informs arbovirus vector control. Nature *563*, 501–507. https://doi.org/10.1038/s41586-018-0692-z.

73. Nowell, R.W., Elsworth, B., Oostra, V., Zwaan, B.J., Wheat, C.W., Saastamoinen, M., Saccheri, I.J., van't Hof, A.E., Wasik, B.R., Connahs, H., et al. (2017). A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*. GigaScience *6*, 1–7. https://doi.org/10.1093/gigascience/gix035.

74. Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W., Pfeiffer, B.D., George, R.A., Svirskas, R., et al. (2015). The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. *25*, 445–458. https://doi.org/10.1101/gr.185579.114.

75. Tang, B., Wang, Z., Liu, Q., Zhang, H., Jiang, S., Li, X., Wang, Z., Sun, Y., Sha, Z., Jiang, H., et al. (2019). High-Quality Genome Assembly of Eriocheir japonica sinensis Reveals Its Unique Genome Evolution. Front. Genet. *10*, 1340. https://doi.org/10.3389/fgene.2019.01340.

76. Poynton, H.C., Hasenbein, S., Benoit, J.B., Sepulveda, M.S., Poelchau, M.F., Hughes, D.S.T., Murali, S.C., Chen, S., Glastad, K.M., Goodisman, M.A.D., et al. (2018). The Toxicogenome of *Hyalella azteca*: A Model for Sediment Ecotoxicology and Evolutionary Toxicology. Environ. Sci. Technol. *52*, 6009–6022. https://doi.org/10.1021/acs.est.8b00837.

77. Miller, J.R., Koren, S., Dilley, K.A., Harkins, D.M., Stockwell, T.B., Shabman, R.S., and Sutton, G.G. (2018). A draft genome sequence for the *Ixodes scapularis* cell line. F1000Res. *7*, 297. https://doi.org/10.12688/f1000research.13635.1.

78. Tang, B., Zhang, D., Li, H., Jiang, S., Zhang, H., Xuan, F., Ge, B., Wang, Z., Liu, Y., Sha, Z., et al. (2020). Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (*Portunus trituberculatus*). GigaScience *9*, giz161. https://doi.org/10.1093/gigascience/giz161.

79. Gutekunst, J., Andriantsoa, R., Falckenhayn, C., Hanna, K., Stein, W., Rasamy, J., and Lyko, F. (2018). Clonal genome evolution and rapid invasive spread of the marbled crayfish. Nat. Ecol. Evol. *2*, 567–573. https://doi.org/10.1038/s41559-018-0467-9.

80. Grbić, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouzé, P., Grbić, V., Osborne, E.J., Dermauw, W., Ngoc, P.C.T., Ortego, F., et al. (2011). The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature *479*, 487–492. https://doi.org/10.1038/nature10640.

81. Chipman, A.D., Ferrier, D.E.K., Brena, C., Qu, J., Hughes, D.S.T., Schröder, R., Torres-Oliva, M., Znassi, N., Jiang, H., Almeida, F.C., et al. (2014). The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. PLoS Biol. *12*, e1002005. https://doi.org/10.1371/journal.pbio.1002005.

82. Sales, G., Deagle, B.E., Calura, E., Martini, P., Biscontin, A., De Pittà, C., Kawaguchi, S., Romualdi, C., Meyer, B., Costa, R., and Jarman, S. (2017). KrillDB: A *de novo* transcriptome database for the Antarctic krill (*Euphausia superba*). PLoS One *12*, e0171908. https://doi.org/10.1371/journal.pone.0171908.

83. Höring, F., Biscontin, A., Harms, L., Sales, G., Reiss, C.S., De Pittà, C., and Meyer, B. (2021). Seasonal gene expression profiling of Antarctic krill

in three different latitudinal regions. Mar. Genomics *56*, 100806. https://doi.org/10.1016/j.margen.2020.100806.

84. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

85. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

86. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963. https://doi.org/10.1371/journal.pone.0112963.

87. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat. Biotechnol. *31*, 1119–1125. https://doi.org/10.1038/nbt.2727.

88. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. *3*, 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

89. Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310.

90. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

91. Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. Genome Res. *12*, 656–664. https://doi.org/10.1101/gr.229202.

92. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

93. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *Chapter 4*, 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

94. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11. https://doi.org/10.1186/s13100-015-0041-9.

95. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580. https://doi.org/10.1093/nar/27.2.573.

96. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. USA *117*, 9451–9457. https://doi.org/10.1073/pnas.1921046117.

97. Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. *35*, W265–W268. https://doi.org/10.1093/nar/gkm286.

98. Katoh, K., Misawa, K., Kuma, K.i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.

99. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

100. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. https://doi.org/10.1093/nar/gkab301.

101. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience *7*, 1–6. https://doi.org/10.1093/gigascience/gix120.

102. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

103. Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinf. *11*, 431. https://doi.org/10.1186/1471-2105-11-431.

104. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. *29*, 644–652. https://doi.org/10.1038/nbt.1883.

105. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

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

107. Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988–995. https://doi.org/10.1101/gr.1865504.

108. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295. https://doi.org/10.1038/nbt.3122.

109. Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with. BMC Genom. *7*, 327. https://doi.org/10.1186/1471-2164-7-327.

110. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

111. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157. https://doi.org/10.1186/s13059-015-0721-2.

112. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

113. Emms, D.M., and Kelly, S. (2018). STAG: Species Tree Inference from All Genes. Preprint at bioRxiv. https://doi.org/10.1101/267914.

114. Emms, D.M., and Kelly, S. (2017). STRIDE: species tree root inference from gene duplication events. Mol. Biol. Evol. *34*, 3267–3278. https://doi.org/10.1093/molbev/msx259.

115. Sanderson, M.J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics *19*, 301–302. https://doi.org/10.1093/bioinformatics/19.2.301.

116. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

117. De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269–1271. https://doi.org/10.1093/bioinformatics/btl097.

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

119. Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. *45*, W12–W16. https://doi.org/10.1093/nar/gkx428.

120. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

121. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

122. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

123. Kendig, K.I., Baheti, S., Bockol, M.A., Drucker, T.M., Hart, S.N., Heldenbrand, J.R., Hernaez, M., Hudson, M.E., Kalmbach, M.T., Klee, E.W., et al. (2019). Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. Front. Genet. *10*, 736. https://doi.org/10.3389/fgene.2019.00736.

124. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

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

126. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms. Fly *6*, 80–92. https://doi.org/10.4161/fly.19695.

127. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

128. Baum, B.R. (1989). PHYLIP: phylogeny inference package. Version 3.2. Q. Rev. Biol. *64*, 539–541. https://doi.org/10.1086/416571.

129. Meisner, J., and Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. Genetics *210*, 719–731. https://doi.org/10.1534/genetics.118.301336.

130. Fumagalli, M., Vieira, F.G., Korneliussen, T.S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., and Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. Genetics *195*, 979–992. https://doi.org/10.1534/genetics.113.154740.

131. Luu, K., Bazin, E., and Blum, M.G.B. (2017). *pcadapt*: an R package to perform genome scans for selection based on principal component analysis. Mol. Ecol. Resour. *17*, 67–77. https://doi.org/10.1111/1755-0998.12592.

132. Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. *15*, 1179–1191. https://doi.org/10.1111/1755-0998.12387.

133. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.

134. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

135. Fitak, R.R. (2021). *OptM*: estimating the optimal number of migration edges on population trees using *Treemix*. Biol. Methods Protoc. *6*, bpab017. https://doi.org/10.1093/biomethods/bpab017.

136. Wilson, G.A., and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. Genetics *163*, 1177–1191. https://doi.org/10.1093/genetics/163.3.1177.

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

138. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.

139. Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. (2013). Package 'vegan.'. Package Version 2. Community Ecol., 1–295.

140. Flanagan, S.P., and Jones, A.G. (2017). Constraints on the F_{ST}–Heterozygosity Outlier Approach. J. Hered. *108*, 561–573. https://doi.org/10.1093/jhered/esx048.

141. Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920. https://doi.org/10.1093/bioinformatics/bts277.

142. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

143. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208. https://doi.org/10.1093/nar/gkp335.

144. Smith, B.T., Boyle, J.M., Garbow, B.S., Ikebe, Y., Klema, V.C., and Moler, C.B. (1974). How to Use EISPACK. In Matrix Eigensystem Routines — EISPACK Guide Lecture Notes in Computer Science, B.T. Smith, J.M. Boyle, B.S. Garbow, Y. Ikebe, V.C. Klema, and C.B. Moler, eds. (Springer), pp. 5–106. https://doi.org/10.1007/978-3-540-38004-7_2.

145. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag) https://doi.org/10.1007/978-0-387-98141-3.

146. Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., and Liu, X. (2021). rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Dev. Reprod. Biol. *19*, 619–628. https://doi.org/10.1016/j.gpb.2020.10.007.

147. Ryu, T., Seridi, L., and Ravasi, T. (2012). The evolution of ultraconserved elements with different phylogenetic origins. BMC Evol. Biol. *12*, 236–311. https://doi.org/10.1186/1471-2148-12-236.

148. Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. (2000). InterPro–an integrated documentation resource for protein families, domains and functional sites. Bioinformatics *16*, 1145–1150. https://doi.org/10.1093/bioinformatics/16.12.1145.

149. Attwood, T.K. (2002). The PRINTS database: A resource for identification of protein families. Brief. Bioinform. *3*, 252–263. https://doi.org/10.1093/bib/3.3.252.

150. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M., and Sigrist, C.J.A. (2006). The PROSITE database. Nucleic Acids Res. *34*, D227–D230. https://doi.org/10.1093/nar/gkj063.

151. Corpet, F., Gouzy, J., and Kahn, D. (1998). The ProDom database of protein domain families. Nucleic Acids Res. *26*, 323–326. https://doi.org/10.1093/nar/26.1.323.

152. Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., et al. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. *33*, D284–D288. https://doi.org/10.1093/nar/gki078.

153. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

154. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

155. Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. *28*, 33–36. https://doi.org/10.1093/nar/28.1.33.

156. Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *33*, D501–D504. https://doi.org/10.1093/nar/gki025.

157. Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. *28*, 45–48. https://doi.org/10.1093/nar/28.1.45.

158. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. *34*, 1812–1819. https://doi.org/10.1093/molbev/msx116.

159. Yella, V.R., Kumar, A., and Bansal, M. (2018). Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. Sci. Rep. *8*, 4520. https://doi.org/10.1038/s41598-018-22129-8.

160. Orsi, A.H., Whitworth, T., and Nowlin, W.D. (1995). On the meridional extent and fronts of the Antarctic Circumpolar Current. Deep Sea Res. Oceanogr. Res. Pap. *42*, 641–673. https://doi.org/10.1016/0967-0637(95)00021-W.

161. Hunt, G.L., Drinkwater, K.F., Arrigo, K., Berge, J., Daly, K.L., Danielson, S., Daase, M., Hop, H., Isla, E., Karnovsky, N., et al. (2016). Advection in polar and sub-polar environments: Impacts on high latitude marine ecosystems. Prog. Oceanogr. *149*, 40–81. https://doi.org/10.1016/j.pocean.2016.10.004.

162. Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for Ancient Admixture between closely related populations. Mol. Biol. Evol. *28*, 2239–2252. https://doi.org/10.1093/molbev/msr048.

163. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the neandertal genome. Science *328*, 710–722. https://doi.org/10.1126/science.1188021.

164. Copernicus Climate Change Service (2019). ERA5 monthly averaged data on single levels from 1979 to present https://doi.org/10.24381/CDS.F17050D7.

165. Siegel, V. (1987). Age and growth of Antarctic Euphausiacea (Crustacea) under natural conditions. Mar. Biol. *96*, 483–495. https://doi.org/10.1007/BF00397966.

166. Baer, C.F., Miyamoto, M.M., and Denver, D.R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat. Rev. Genet. *8*, 619–631. https://doi.org/10.1038/nrg2158.

Please cite this article in press as: Shao et al., The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights, Cell (2023), https://doi.org/10.1016/j.cell.2023.02.005

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Euphausia superba (Antarctic krill)	This study	N/A
Euphausia pacifica (North Pacific krill)	This study	N/A
Critical commercial assays		
NucleoBond HMW DNA KIT	MACHEREY-NAGEL	N/A
AMPure XP beads	Beckman Coulter Agencourt	B37419AA
T4 DNA polymerase	Enzymatics	P708L
Tris pH 8.0	Thermo Fisher Scientific	N/A
BluePippin	Sage Science	N/A
Dynabeads MyOne Streptavidin T1	Thermo Fisher Scientific	N/A
QIAsymphony RNA Kit	Qiagen	931,636
DNA Polymerase Binding Kit	Pacific Biosciences	101-046-400
VAHTS mRNA-seq v2 Library Prep Kit	Vazyme	NR611-02
TruSeq Stranded mRNA LT Sample Prep Kit	Illumina	RS-122-9004DOC
SMARTer PCR cDNA Synthesis Kit	Takara Biotechnology	634,925
SMRTbell Template Prep Kit 1.0	Pacific Biosciences	100-259-100
Deposited data		
Genome sequencing data for *Euphausia superba*	This study	China National GeneBank DataBase (CNGB): CNP0001930
Genome assembly of *Euphausia superba*	This study	China National GeneBank DataBase (CNGB): CNP0001930
Transcriptome data for gene structure annotation and gene expression analysis of *Euphausia superba*	This study	China National GeneBank DataBase (CNGB): CNP0001930
Re-sequencing data of 75 *Euphausia superba* individuals	This study	China National GeneBank DataBase (CNGB): CNP0001930
Genome and annotation of *Aedes aegypti*	Matthews et al.[72]	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/204/515/GCF_002204515.2_AaegL5.0/
Genome and annotation of *Bicyclus anynana*	Nowell et al.[73]	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/900/239/965/GCF_900239965.1_Bicyclus_anynana_v1.2/
Genome and annotation of *Drosophila melanogaster*	Hoskins et al.[74]	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/215/GCF_000001215.4_Release_6_plus_ISO1_MT/
Genome and annotation of *Eriocheir sinensis*	Tang et al.[75]	https://ftp.cngb.org/pub/Assembly/GCA/013/436/485/GCA_013436485.1_ASM1343648v1/
Genome and annotation of *Hyalella azteca*	Poynton et al.[76]	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/764/305/GCF_000764305.2_Hazt_2.0.2/
Genome and annotation of *Ixodes scapularis*	Miller et al.[77]	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/920/785/GCF_016920785.2_ASM1692078v2/
Genome and annotation of *Portunus trituberculatus*	Tang et al.[78]	http://gigadb.org/dataset/100678
Genome and annotation of *Litopenaeus vannamei*	Zhang et al.[22]	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/789/085/GCF_003789085.1_ASM378908v1/
Genome and annotation of *Procambarus virginalis*	Gutekunst et al.[79]	http://marmorkrebs.dkfz.de/downloads/genome/pvirGEN-0.4/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Genome and annotation of *Tetranychus urticae*	Grbić et al.[80]	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/239/435/GCF_000239435.1_ASM23943v1/
Genome and annotation of *Strigamia maritima*	Chipman et al.[81]	http://metazoa.ensembl.org/Strigamia_maritima/Info/Index
RNA-seq data of CO_2 treatment of *Euphausia superba*	Sales et al.[82]	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA362526
RNA-seq data of different seasons and regions of *Euphausia superba*	Höring et al.[83]	https://www.ncbi.nlm.nih.gov/bioproject/PRJEB30084/
RNA-seq data of temperature treatment of *Euphausia superba*	East China Sea Fisheries Research Institute	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA640244/
Software and algorithms		
NextDenovo v2.30	Nextomics	https://github.com/Nextomics/NextDenovo
SAMtools v1.7	Li et al.[84]	https://github.com/samtools/samtools
BWA v0.7.12	Li and Durbin[85]	https://github.com/lh3/bwa/releases/
HiFi-CCS v4.0.0	Pacific Biosciences	https://www.pacb.com/support/software-downloads/
Pilon v1.23	Walker et al.[86]	https://github.com/broadinstitute/pilon/releases
Lachesis v201701	Burton et al.[87]	https://github.com/shendurelab/LACHESIS
Juicer-box v1.91	Durand et al.[88]	https://github.com/aidenlab/juicer/releases
GMAP v2021-12-12	Wu et al.[89]	http://research-pub.gene.com/gmap/
HISAT2 v2.1.0	Kim et al.[90]	https://github.com/DaehwanKimLab/hisat2/releases
BLAT v319	Kent et al.[91]	https://github.com/djhshih/blat/releases
BEDTools v2.29	Quinlan and Hall[92]	https://github.com/arq5x/bedtools2/releases/
RepeatMasker v4.0.6	Tarailo-Graovac and Chen[93]	http://repeatmasker.org/RepeatMasker/
RepBase v202101	Bao et al.[94]	https://www.girinst.org/server/RepBase/
Tandem Repeats Finder (TRF) v4.07	Benson et al.[95]	https://github.com/Benson-Genomics-Lab/TRF/releases
RepeatModeler v1.0.4	Flynn et al.[96]	https://github.com/Dfam-consortium/RepeatModeler/releases
LTR-Finder v1.06	Xu et al.[97]	https://github.com/xzhub/LTR_Finder
MAFFT v6.864b	Katoh et al.[98]	https://github.com/GSLBiotech/mafft/releases
FastTree v2.1.10	Price et al.[99]	http://www.microbesonline.org/fasttree/
TreeBeST v1.9.2	Heng Li	http://treesoft.sourceforge.net/treebest.shtml
FigTree v1.4.3	Andrew Rambaut	https://github.com/rambaut/figtree
iTols v6	Letunic and Bork[100]	https://itol.embl.de/
parseRM.pl	Aurelie Kapusta	https://github.com/4ureliek/Parsing-RepeatMasker-Outputs
SOAPnuke v2.1.5	Chen et al.[101]	https://github.com/berry08/SOAPnuke2
EMBOSS v6.5.7	Rice et al.[102]	https://ftp.ebi.ac.uk/pub/software/unix/EMBOSS/
HMMER v3.1b2	Johnson et al.[103]	http://hmmer.org/download.html
Iso-Seq v3.2.2	Pacific Biosciences	https://github.com/PacificBiosciences/IsoSeq
Trinity v2.5.1	Grabherr et al.[104]	https://github.com/trinityrnaseq/trinityrnaseq/releases
CD-HIT-EST v4.5.4	Li and Godzik[105]	http://www.bioinformatics.org/cd-hit/
BLAST v2.8.1	Camacho et al.[106]	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.8.1/
TransDecoder v5.0.2	Brian Haas	https://github.com/TransDecoder/TransDecoder
Gene-Wise v2.4.1	Birney et al.[107]	https://www.ebi.ac.uk/~birney/wise2/
StringTie v1.3.5	Pertea et al.[108]	https://github.com/gpertea/stringtie/releases
PASApipeline v2.0.2	Campbell et al.[109]	https://github.com/PASApipeline/PASApipeline

(*Continued on next page*)

Please cite this article in press as: Shao et al., The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights, Cell (2023), https://doi.org/10.1016/j.cell.2023.02.005

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
EvidenceModeler v1.1.1	Haas et al.[110]	https://github.com/EVidenceModeler/EVidenceModeler/releases
OrthoFinder v2.4.0	Emms et al.[111]	https://github.com/davidemms/OrthoFinder/releases
Diamond v2.0.4.142	Buchfink et al.[112]	https://github.com/bbuchfink/diamond/releases/
STAG v1.0.0	Emms et al.[113]	https://github.com/davidemms/STAG/releases
STRIDE v1	Emms et al.[114]	https://github.com/davidemms/STRIDE
r8s v1.71	Sanderson et al.[115]	http://loco.biosci.arizona.edu/r8s/
PAML v4.5	Yang et al.[116]	http://abacus.gene.ucl.ac.uk/software/paml.html
CAFE v5.0	De Bie et al.[117]	https://github.com/hahnlab/CAFE
Fastp v0.20.0	Chen et al.[118]	https://github.com/OpenGene/fastp/releases
FastQC v0.11.9	Simon Andrews	https://github.com/s-andrews/FastQC
CPC2 v1.0.1	Kang et al.[119]	https://github.com/gao-lab/CPC2_standalone
Bowtie2 v2.3.4.1	Langmead et al.[120]	http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
RSEM v1.2.12	Li et al.[121]	https://github.com/deweylab/RSEM/releases
DEseq2 v1.14.1	Love et al.[122]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
preprocessCore v1.44.0	Ben Bolstad	https://github.com/bmbolstad/preprocessCore
Sentieon Genomics Tools	Kendig et al.[123]	https://www.sentieon.com/
GATK v3.8.1	McKenna et al.[124]	https://github.com/broadinstitute/gatk/releases
GenMap v1.3.0	Pockrandt et al.[125]	https://github.com/cpockrandt/genmap/releases
SnpEff v5.1	Cingolani et al.[126]	http://pcingola.github.io/SnpEff/
PLINK v1.90b6.6	Purcell et al.[127]	https://zzz.bwh.harvard.edu/plink/download.shtml
PHYLIP v3.69	Baum et al.[128]	https://evolution.gs.washington.edu/phylip/
PCAngsd v1.10	Meisner and Albrechtsen[129]; Fumagalli et al.[130]	http://www.popgen.dk/software/index.php/PCAngsd
NGSadmix v3.2	Skotte[46]	http://www.popgen.dk/software/index.php/NgsAdmix
STRUCTURE v2.3.4	Pritchard et al.[45]	https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/release/
PCAdapt v4.3.2	Luu et al.[131]	https://github.com/bcm-uga/pcadapt
CLUMPAK (last accessed on 2022.11.20)	Kopelman et al.[132]	http://clumpak.tau.ac.il/
EIGENSOFT v7.2.1	Price et al.[133]	https://github.com/DReichLab/EIG
VCFtools v0.1.17	Danecek et al.[134]	https://sourceforge.net/projects/vcftools/files/
Dsuite v0.4	Malinsky et al.[42]	https://github.com/millanek/Dsuite
TreeMix v1.13	Pickrell et al.[43]	https://bitbucket.org/nygcresearch/treemix/downloads/
OptM v0.1.6	Fitak[135]	https://cran.r-project.org/web/packages/OptM/index.html
BA3-SNPs v3.0.4	Wilson and Rannala[136]; Mussmann et al.[44]	https://github.com/stevemussmann/BayesAss3-SNPs
PSMC v0.6.5-r67	Li et al.[51]	https://github.com/lh3/psmc
BCFtools v1.4	Danecek et al.[134]	http://www.htslib.org/download/
PopSizeABC v1	Boitard et al.[52]	https://github.com/stsmall/popsizeabc
minimap v2-2.17	Li et al.[137]	https://github.com/lh3/minimap2
IGV v2.4.14	Robinson et al.[138]	https://software.broadinstitute.org/software/igv/download
geosphere v1.5-18	Robert J. Hijmans	https://github.com/rspatial/geosphere
vegan v2.6-4	Oksanen[139]	https://cran.rstudio.com/
LEA v3.10.0	Caye et al.[50]	https://cran.rstudio.com/

(*Continued on next page*)

Please cite this article in press as: Shao et al., The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights, Cell (2023), https://doi.org/10.1016/j.cell.2023.02.005

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FSThet v1.0.1	Flanaga and Jones[140]	https://cran.rstudio.com/
PerformanceAnalytics v1.5.3	Justin M. Shea	https://github.com/braverock/Performance Analytics/releases
pheatmap v1.0.12	Raivo Kolde	https://github.com/raivokolde/pheatmap
BEDOPS v2.4.35	Shane et al.[141]	https://github.com/bedops/bedops/releases
seqkit v0.10.2	Shen et al.[142]	https://github.com/shenwei356/seqkit/releases
MEME v5.3.3	Grant et al.[143]	https://meme-suite.org/meme/doc/download.html
Eigen v3.6.2	Smith et al.[144]	https://cran.rstudio.com/
ggplot2 v3.4.0	Hadley Wickham et al.[145]	https://github.com/tidyverse/ggplot2
CMplot v4.2.0	Yin et al.[146]	https://github.com/YinLiLin/CMplot
scatterpie v0.1.8	Guangchuang Yu	https://github.com/GuangchuangYu/scatterpie

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Changwei Shao (shaocw@ysfri.ac.cn).

Materials availability

This study did not generate any new unique reagents.

Data and code availability

- The genome sequences, genome assembly, transcriptome sequencing data and population re-sequencing data have been deposited at CNGB under accession code CNP0001930.
- All software and packages used was publicly accessible, and this study does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Source organisms

The experimental procedures were in accordance with the guidelines approved by the institutional review board on bioethics and biosafety of BGI (IRB-BGI). The experiment was authorized by IRB-BGI (under NO. BGI-IRB A20007), and the procedures in IRB-BGI meet good clinical practice (GCP) principles. For genome sequencing, the muscle tissue of a female Antarctic krill was used for PacBio CCS sequencing, while the other female for PacBio CLR and WGS short-reads sequencing. Additionally, the muscle tissues from twelve Antarctic krill were used for Hi-C sequencing. For transcriptome sequencing, the muscle tissue and whole body from two female Antarctic krill were used for Iso-seq long reads sequencing, respectively; the short-reads RNA sequencing data was obtained from the tissues (head, eye, appendage, gill, and muscle) of another seven female Antarctic krill individually. For whole genome re-sequencing, the muscles of adult Antarctic krill from four geographical groups, Prydz Bay (18 samples, including 9 female and 9 male), Ross Sea (20 samples, including 14 female and 6 male), South Georgia (18 samples, including 8 female and 10 male), and South Shetland Islands (19 samples, including 9 female and 10 male) were collected. In addition, to explore the studies on population genetics, and evolutionary history of Antarctic krill, three North Pacific krill samples as outgroups were collected from the Yellow Sea (123°58.36′E 34°59.35′N).

METHOD DETAILS

Sampling and sequencing

Whole-genome sequencing of short reads

Female adult of Antarctic krill for genome sequencing was collected from South Shetland Island. Total genomic DNA was extracted using NucleoBond HMW DNA KIT (MACHEREY-NAGEL, Germany), then were stored in TE buffer. To construct whole genome sequencing short read libraries, six major steps were performed: genomic DNA interruption, fragment selection, end-repair, adding adapter, PCR amplification, and library purification. In details, the extracted DNA was firstly sheared into 50–800 bp fragments with treating time of 20 s, acoustic duty factor of 25%, peak incident power of 500 W, cycles per burst of 500 and for 24 cycles. Fragments ranged from 150 bp to 500 bp were selected and treated with T4 DNA polymerase (Enzymatics, Beverly, USA) for 30 min at 20°C to obtain blunt ends. The DNA fragments were next ligated to T-tailed adapters and amplified. The temperature profile was 3 min at 95°C followed by 8 cycles of 20 s at 98°C, 15 s at 60°C, 30 s at 72°C, and 10 min at 72°C for further elongation. AMPure XP beads (Agencourt, Beverly, USA) were used to purify the PCR reaction. The whole-genome sequencing (WGS) short reads libraries were subsequently sequenced on DNBSEQ-T1 and BGI-SEQ500 platforms in BGI-Qingdao. A total of 4.01 Tb WGS short reads was generated (Table S1).

Hi-C sequencing

To construct Hi-C library, the formaldehyde-fixed muscular tissue was placed into a pre-chilled dounce homogeniser with nuclei isolation buffer consisted of 10 mM Tris-HCl, 2 mM EDTA, 250 mM sucrose, 3 mM $CaCl_2$, 2 mM MgAC2, 1x protease inhibitor cocktail, 1 mM DTT, 0.2% NP-40, 0.4 U/μL RNase inhibitor and 1% BSA, then homogenized and strained through a 100 μm cell strainer, centrifuged at 300 x g for 5 min at 4°C to pellet nuclei. Resuspend the nuclei and then strained through a 40 μm cell strainer (Falcon). The extracted nuclei were resuspended with 250 μL 0.5% SDS at 62°C for 10 min, adding 725 μL water and 125 μL 10% Triton X-100 to quench the SDS, 37°C for 15 min. Next the chromatin was digested by restriction enzyme (DpnII/MboI/Sau3AI) (NEB) at 37°C overnight, labeled with biotin-14-dCTP and end-repaired. The resulting blunt-ended fragments were ligated *in situ* at 23°C for 4h and then treated with ExoI at 37°C for 45 min. Collect the nuclei pellet and isolated DNA, by adding 10 mg/mL proteinase K and 1% SDS 65°C overnight, the formaldehyde cross-link was reversed. In order to remove the biotin-14-dCTP from the end of unligated DNA, 10 μg of DNA were incubated with T4 DNA polymerase, dTTP and dATP, 20°C for 1h. Shearing the DNA with LE220, the biotin-containing fragments were captured by Dynabeads MyOne Streptavidin T1 (Invitrogen). Following by end-repaired, adaptor ligation, PCR enrichment and cyclization, the Hi-C libraries were created and sequenced on DNBSEQ-T1 in BGI-Qingdao. A total of 11.38 Tb Hi-C data was generated (Table S1).

PacBio CLR sequencing

The same female adult of Antarctic krill as above whole genome sequencing of short reads was used to do the long read sequencing. Total genomic DNA was extracted using NucleoBond HMW DNA KIT (MACHEREY-NAGEL, Germany), and then stored in TE buffer. To construct library for PacBio CLR sequencing, the Antarctic krill genomic DNA was sheared to ~20 kb, and short fragments shorter than 7 kb were filtered out using BluePippin (Sage Science, MA, USA) and converted into the proprietary SMRTbell library using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, CA, USA). The SMRT bell libraries were subsequently sequenced using a PacBio Sequel instrument with the V3.0 sequencing reagent and SMRT Cell (1M, V3) by Tianjin Biochip Corporation. A total of 3.06 Tb CLR subreads was generated (Table S1).

PacBio HiFi-CCS sequencing

The genomic DNA was extracted from muscle of another female Antarctic krill using NucleoBond HMW DNA KIT (MACHEREY-NAGEL, Germany). The integrity of the DNA was determined using the Agilent 4200 Bioanalyzer (Agilent Technologies, Palo Alto, California). To construct library for HiFi-CCS sequencing, eight micrograms of genomic DNA were sheared and concentrated with AMPure PB magnetic beads. Each SMRT bell library was constructed using Pacific Biosciences SMRTbell Template Prep Kit v1.0. The constructed library was selected by Sage ELF for molecules 11–15 kb in size, followed by primer annealing and binding of SMRT bell templates to polymerases with the DNA Polymerase Binding Kit (Pacific Biosciences, CA, USA). Sequencing was performed on the Pacific Bioscience Sequel II for 30 h by Annoroad Gene Technology company. A total of 8.54 Tb CCS subreads was generated (Table S1).

RNA sequencing of short reads

Five tissues were obtained, including three head, one eye, one appendage, one gill and one muscle (each tissue from one individual) for total RNA extraction of Antarctic krill, VAHTS mRNA-seq v2 Library Prep Kit for Illumina (Vazyme NR611-02) was used following the manufacturer's instructions. Libraries were constructed using the TruSeq Stranded mRNA LT Sample Prep Kit (Illumina, San Diego, CA, USA) and sequenced on the Illumina sequencing platform (Illumina HiSeq 2500) to generate 150 bp paired end reads by Berry Genomics. A total of 111.45 Gb short RNA reads generated (Table S1).

RNA sequencing of long reads

We collected muscle from one individual and whole body from another individual for full length transcriptome sequencing. Total RNA was extracted using QIAsymphony RNA Kit (Qiagen, Germany), and then stored in TE buffer. For long-read RNA-seq (PacBio Iso-Seq), RNA samples were assessed by measuring the RNA Integrity Number (RIN) and concentration using an Agilent 2100 instrument. RNA purity and contamination can be assessed through UV-spectrophotometry using a Nanodrop spectrophotometer. A total amount of 2 μg RNA was required for the RNA sample preparation. The first-strand cDNA synthesis and PCR amplification was performed with NEBNextSingle Cell/Low Input cDNA Synthesis & Amplification Module and PacBio Iso-Seq Express Oligo Kit. SMRTbell Template Prep Kit was used to generate SMRTbell libraries with processes including DNA damage repair, end-repair, polyA-tailing and adaptor ligation. Library quality was assessed by the Agilent Bioanalyzer 2100 system. Finally, the primers and enzyme were binding to the SMRT template to construct a complete SMRT bell library. The SMRTbell libraries were prepared according to the Isoform Sequencing Protocol (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit Sequencing was carried out on the Pacific Bioscience Sequel II in Berry Genomics. A total of 267.45 Gb Iso-seq subreads was generated, from which 15.00 Gb CCS reads was produced (Table S1).

Genome assembly and evaluation

Assembly of Antarctic krill genome

Five major steps were conducted to assemble the Antarctic krill genome: 1) correct CLR (Continuous Long Reads) and generate the HiFi-CCS reads based on HiFi-CCS subreads; 2) all-to-all alignment with corrected CLR reads and HiFi-CCS reads; 3) build a preliminary genome assembly; 4) error correction (‘polishing’) using WGS short reads; 5) anchor the preliminary genome to chromosome-level with Hi-C data. The details were as follows.

In the first step, CLR were self-corrected by NextDenovo v2.30 (https://github.com/Nextomics/NextDenovo) to reduce the error ratio of reads with the parameters ‘read_cuoff = 1k, seed_cutoff = 10k, seed_cutfiles = 148, blocksize = 10g, pa_raw_align = 1, pa_correction = 1, minimap2_options_raw = -x ava-pb -t 14, correction_options = -p 28 -dbuf, sort_options = -m 80g -t 28 -k 60’. At this stage, 3059.79 Gb CLR reads were fed into the NextDenovo and 743.72 Gb corrected CLR were generated (Table S1). To evaluate the quality of corrected CLR data, we randomly select 1,000 CLR reads and mapped WGS short reads to them using BWA[85] v0.7.12. The resulting mapping error rate were calculated using SAMtools[84] v1.7, which suggested that the corrected CLR reads were adequate for downstream genome assembly. For the HiFi-CCS reads, we first determined the predicted accuracy of HiFi-CCS reads and pass number (one can presume that a higher number of passes can produce more multiple alignment information resulting in better-quality HiFi-CCS reads). HiFi-CCS reads with a predicted accuracy of at least Q20 (99%) were retained by the HiFi-CCS v4.0.0 (https://www.pacb.com/support/software-downloads/) with the parameter ‘–min-passes 5’. Finally, a total of 734.99 Gb (8.61% of HiFi-CCS subreads) HiFi-CCS reads with an N50 length of 11,446 bp were generated, accounting for 15.31-fold genome coverage.

In the second step, the all-to-all alignment proceeded by combining corrected CLR and HiFi-CCS reads using minimp2[137] v2-2.17 with the parameters ‘-x ava-pb -t 28 -w17 –mode 1 –kn 18’. In the third step, Nextgraph in NextDenovo v2.40 was used to cope with the all-to-all alignment and generate a primary genome assembly. In the fourth step, WGS short reads were aligned to primary assembly using BWA and SAMtools. The alignment was then fed into Pilon[86] v1.23 to lower base call errors and improve genome accuracy (‘polishing’). Thus, we obtained a draft assembly that consisted of 298,755 contigs with an N50 of 178.99 kb and a total length of 47.97 Gb (Table S1). In the fifth step, we used Lachesis[87] v201701 to cluster, order and orient the contigs with following parameters ‘CLUSTER_MIN_RE_SITES = 48, CLUSTER_MAX_LINK_DENSITY = 1, CLUSTER_NONINFORMATIVE_RATIO = 3, ORDER_MIN_N_RES_IN_TRUNK = 50, ORDER_MIN_N_RES_IN_SHREDS = 10’. Then Juicer-box[88] v1.91 was used to generate Hi-C contact matrix and visualize it. Next, the top 17 longest super-scaffolds based on the karyotype of Antarctic krill were selected as chromosomes,[19,20] which consists of 182,702 contigs with a length of 31.67 Gb (66.01% of contigs in length) anchored onto chromosomes (Table S1).

Evaluation of the assembled genome

We compared the Antarctic krill genome assembly with 154 published invertebrate genomes, especially malacostracan crustaceans that much smaller genomes, the marbled crayfish[79] (*P. virginalis*) (3.29 Gb, contig N50 1.19 kb) and Pacific white shrimp[22] (*L. vannamei*) (1.66 Gb, contig N50 86.86 kb), the Antarctic krill assembly shows much longer contig N50 values (Figure 1B; Table S1).

We further evaluated the completeness and accuracy of the assembled genome in three methods. In the first method, the general completeness was examined by mapping WGS short reads to genome. The WGS short reads were filtered using SOAPnuke[101] v2.1.5 with the parameters ‘-n 0.1 -q 0.5 -L 12’, and then were mapped to the assembled genome using BWA to evaluate the completeness of the whole genome. A total of 7,868,405,470 bp (14.70×) WGS short clean reads were mapped to the genome with a mapping rate of 96.42% (Table S1). In the second method, RNA sequencing reads were mapped to the reference assembly to further validate the coding regions of the genome. A total of 10.30 Gb of HiFi-CCS reads (called from 267.45 Gb subreads) (N50: 3,905 bp) were produced by the single-molecule real-time (SMRT) isoform sequencing (Iso-Seq) v3.3.3 (https://github.com/PacificBiosciences/IsoSeq) using PacBio platform, and then were mapped to genome using GMAP[89] v2021-12-12, and achieving a mapping rate of 79.84%. Meanwhile, 419.06 Gb of short RNA-seq reads produced by Illumina HiSeq 2500 were aligned to genome using HISAT2[90] v2.1.0 with a mapping rate of 78.68% (Table S1). In the third method, UCEs were used to assess the completeness of the non-coding region of the genome. A self-identified UCEs were constructed based on distantly genetic-related species included a demosponge (*Amphimedon queenslandica*) from the phylum Porifera, a hydra (*Hydra magnipapillata*), a sea anemone (*Nematostella vectensis*) from the phylum Cnidaria, a sea urchin (*Strongylocentrotus purpuratus*) from the phylum Echinodermata, a fruit fly (*Drosophila melanogaster*) from the phylum Arthropoda, and human (*Homo sapiens*) from the phylum Chordata.[147] Briefly, pairwise whole-genome alignments were performed using BLAT[91] v319 with default parameters. Only the regions with a minimum coverage of 75%, a minimum identity of 75%, and more than 50 bp in length in all species were retained, and finally 81 UCEs were yielded. The sequences of these 81 UCEs were subsequently aligned to the genomes of Antarctic krill using BLAT. All BLAT alignments were filtered for a minimum of 75% identity and coverage. We identified 55 of 81 non-exonic UCEs in the Antarctic krill genome. The same evaluation was performed for *Litopenaeus vanname* and *P. virginalis*. We could detect 50 of 81 UCEs in the genome of *L. vannamei*, and 50 of 81 UCEs in *P. virginalis*, revealing that the Antarctic krill assembly is of similar quality and completeness despite its much larger size.

Genome annotation

Identification of repetitive sequences

We used two approaches to identify repeat elements in the genome: homolog-based prediction and *de novo* prediction. We used RepeatMasker[93] v4.0.6 and RepeatProteinMask[96] v4.0.6 to perform homolog prediction based on the RepBase library[94] v202101

Please cite this article in press as: Shao et al., The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights, Cell (2023), https://doi.org/10.1016/j.cell.2023.02.005

and used Tandem Repeats Finder[95] v4.07 to find tandem repeats. RepeatModeler[96] v1.0.4 and LTR-Finder[97] v1.06 were used to perform *de novo* prediction of repeat sequences and the results were combined as the library for RepeatMasker to identify and classify repeat elements. We identified 72.15% repetitive sequences in the assembly (Table S2). To confirm the tandem repeats in genome, we also predicted the tandem repeat on long reads using the same pipeline, and 35.25% tandem repeats in whole genome were annotated (Figure 1C). We next downloaded the genomes of *A. mexicanum, L. vannamei*, *P. virginalis*, and *P. annectens*. The same method was used to get the repeat information of the species and identified 57.13%, 51.64%, 32.42%, and 61.70% repetitive sequences, respectively of first round identification (Table S2). Two rounds of repetitive sequences in Australian lungfish genome were formerly predicted.[15] Referring to the method used in the repeat annotation of Australian lungfish, we masked the genome with repeat sequence of the first round and performed additional repeat annotation in Antarctic krill and other genomes. The repetitive sequences reached up to 92.45% of Antarctic krill genome (Table S2). We also re-annotated the first round of the other 46 invertebrate genomes using the same pipeline, which combined *de novo* and homolog-based methods (Figure 1E; Table S2).

To investigate factors that could influence the krill genome assembly, we assessed the following features in each contig: length, GC content, percentage of tandem repeat (TR) regions, percentage of TE regions (excluding nested TRs), and percentage of genic regions. We obtained a pairwise correlation matrix using the R package 'PerformanceAnalytics' with default parameters. Furthermore, 96.39% of TRs regions were calculated to be overlapped with TEs using BEDTools[92] v2.29.

Gene structure prediction

We used both homology-based and RNA-seq-based methods to predict the gene models in the Antarctic krill genome. For homology-based prediction, we mapped the protein sequences of six published *Arthropoda* genomes (*L. vannamei*,[22] *P. virginalis*,[79] *Portunus trituberculatus*,[78] *Eriocheir sinensis*,[75] *Hyalella azteca*,[76] and *D. melanogaster*[74]) onto the Antarctic krill genome using BLAT and then used Gene-Wise[107] v2.4.1 to predict gene structures. For next-generation RNA-sequencing annotation, we aligned the RNA-sequencing data[82,83] to the genome using HISAT2 and used the alignments as input for StringTie[108] v1.3.5. TransDecoder v5.0.2 (https://github.com/TransDecoder/TransDecoder) was used to predict ORFs and identify gene structure. For the full-length transcript annotation, long-read RNA-seq transcripts were obtained by removing the redundant sequences using CD-HIT-EST[105] v.4.5.4 with the parameters '-aL 0.90 -AL 100 -aS 0.99 -AS 30', followed mapping of the non-redundant transcripts to genome by BLAT and GMAP. Genes predicted from the above methods were then merged to a consensus gene set using EvidenceModeler[110] v.1.1.1 (EVM). The genes with intron size >130 kb (~1% of whole introns) was manually checked using the alignment results of full-length transcripts and modified according to the evidence within gene loci. The genes shorter than 150 bp were removed. Finally, PASApipeline[109] v2.0.2 was developed to improve the quality of genome annotation. The final non-redundant gene set contains 28,834 genes (Table S1).

Gene function annotation

The web resource Inter-Pro[148] was used to correlate protein domains and motifs with the publicly available databases Pfam,[28] PRINTS,[149] PROSITE,[150] Pro-Dom,[151] and PANTHER.[152] Gene Ontology[153] (GO) classifications of genes were extracted from Inter-Pro. We also mapped the Antarctic krill genes to KEGG[154] (Kyoto Encyclopedia of Genes and Genomes) database, and employed BLAST[106] v.2.8.1 to search the public protein databases COG[155] (Cluster of Orthologous Groups), NR[156] (non-redundant protein sequences in NCBI), SwissProt,[157] and TrEMBL[157] (TRanslation of EMBL (nucleotide sequences that are not in Swiss-Prot). The final functional annotation results showed that 92.35% of genes have homologous genes in public databases.

Repetitive sequences analysis

Comparison of repetitive sequences

We collected the genome-wide percentages of five repeat types, including tandem repeats (TRs), DNA transposons, long terminal repeats (LTRs), long interspersed nuclear elements (LINEs), and short interspersed nuclear elements (SINEs) for the 46 invertebrates, and we obtained a correlation matrix using the R package 'PerformanceAnalytics' v1.5.3 (https://github.com/braverock/PerformanceAnalytics) with default parameters. To further depict the relationship among the subtypes of repeats, we counted the percentages of top 54 different repeat subtypes, using the R package 'pheatmap' v1.0.12 (https://rdrr.io/cran/pheatmap/man/pheatmap.html), and performed clustering analysis by both species and repeat subtypes.

We counted the tandem repeat length of each unit length in long reads and the genome assembly of Antarctic krill, *L. vannamei*, and *P. virginalis*. We divided the tandem repeats into three groups according to the repetitive unit length of 2-6 bp (micro satellites), 7–50 bp (small satellites), and 51–249 bp (satellites) (Table S2) and plotted the percentage of each unit length (Figure 1C). The Fisher exact test was used to compare the TRs between Antarctic krill and other species. Satellite sequences with unit length between 51 and 249 bp, which were abundant in Antarctic krill, has a relatively higher frequency and longer average length in long reads than genome assembly, with the average length of 1507.35 bp and 1,076.06 bp, respectively (Table S2). Among these TRs, for the unit length of 200 bp, the TR frequency of long reads is quite more than that of genome assembly (Figure 1C).

The phylogenetic tree of DNA/CMC-EnSpm

In Antarctic krill genome, we identified DNA transposons made up the largest portion of transposable elements (45.72%) and the dominant subtype DNA/CMC-EnSpm (42.02%) (Table S2). We randomly extracted a total of 10,000 DNA/CMC-EnSpm sequences, 389 and 21, and 9,590 for *L. vannamei*, *P. virginalis*, and Antarctic krill, respectively (the number was calculated by the proportion of DNA/CMC-EnSpm of each species). And we aligned these sequences using MAFFT[98] v6.864b to build the phylogenetic tree using FastTree[99] v2.1.10, rooted by TreeBeST v1.9.2 (http://treesoft.sourceforge.net/treebest.shtml) and visualized with iTols[100] v6 (Figure 1F).

Transposable elements activity

To explore the expansion history of transposable elements of Antarctic krill, we further used a custom Perl script parseRM.pl (https://github.com/4ureliek/Parsing-RepeatMasker-Outputs) to estimate the TE activities in Antarctic krill based on alignment outputs from RepeatMasker. The substitution rate was set as 6.19 × 10^{-10} per site per generation. The analysis result was packed into bin per 2 mya.

Calculation of GC content

We mapped WGS short reads to the corrected genome using BWA, using the 'makewindows' function to divide the genome into 1 kb non-overlapping windows. Next, the GC content was obtained using the 'nuc' function in BEDTools[92] v2.29, while the reads depth of each window was obtained using the 'bedcov' function in SAMtools. We also directly compared the GC content of sequencing reads. Firstly, we randomly selected the WGS short (NGS) reads, CLR, and HiFi-CCS reads, then the GC content of each type reads were calculated, revealing a similar GC content (Figure S2C). Then the GC content and genome size of published invertebrate assemblies were calculated (Table S1). Finally, the relationship between the GC content against genome size among these species was showed by a scatterplot (Figure S2D).

After repeat annotation, we converted the annotation file to BED format using gff2bed in BEDOPS[141] v2.4.35, then the GC content of each type of repetitive sequences and gene regions were calculated using the BEDTools 'nuc' function.

Genome-wide protein domains annotation

To depict the genetic basis for the TE expansion, we annotated all Pfam-A protein domains from the Pfam database[28] of 18 arthropods and 28 molluscan genomes. All these genomes were collected from open access databases, such as NCBI and Ensembl. We translated each genome with 6-phase model using EMBOSS[102] v6.5.7 and searched domains across the genomes using HMMER[103] v3.1b2 with default parameters. Furthermore, we analyzed the top 20 domains detected in Antarctic krill, and calculated their density in all 47 invertebrate species (Figure 2D).

Comparative genomics related to adaptation

Phylogeny, gene family, and divergence time

To determine the phylogenetic relationship of the Antarctic krill and other species (*Aedes aegypti,*[72] *Bicyclus anynana,*[73] *D. melanogaster,*[74] *E. sinensis,*[75] *H. azteca,*[76] *Ixodes scapularis,*[77] *P. trituberculatus,*[78] *L. vannamei,*[22] *P. virginalis,*[79] *Tetranychus urticae,*[80] *and Strigamia maritima*[81]) were used. The longest transcripts of each gene were selected. We also filtered the genes with shorter than 30 amino acids. OrthoFinder[111] v2.4.0 was used to identify single-copy genes families based on the alignments generated using Diamond[112] v2.0.4.142. In 12 invertebrates, we identified 95 single-copy gene families. With these single copy genes, a phylogenetic tree was inferred, using the OrthoFinder hybrid species-overlap/duplication-loss coalescent model in STAG[113] v1.0.0, and rooted using STRIDE[114] v1 (Figure 3B).

For divergence time calculation, the tree of 12 species was changed to an ultrametric tree using r8s[115] v1.71 with calibration data form TimeTree website.[158] MCMCTREE analysis in the PAML[116] v4.5 package was employed to estimate divergence times with the nucleotide substitution model set as JC69 and other parameters set as default. CAFE[117] v5.0 was used to identify gene-family expansion with '–error_model' parameter as the statistical foundation. There were 2,191 expanded families (25 significant expansion after p value correction; Benjamini-Hochberg p value <0.05) and 1,951 contracted families (30 significant contraction after p value correction; Benjamini-Hochberg p value <0.05) in Antarctic krill (Figure 3B; Table S3).

Transcriptome analysis

Fastp[118] v0.20.0 was used for the removal of adapter sequences and for quality trimming of reads with default parameters. FastQC v0.11.9 (https://github.com/s-andrews/FastQC) was used to check the quality of reads. HISAT2 was used to align paired end reads of each sample to the Antarctic krill reference genome. Novel transcripts were detected using StringTie on HISAT2 alignment results. CPC2[119] v1.0.1 was used to predict potential coding transcripts. Bowtie2[120] v2.3.4.1 was used to re-align, with novel potential coding transcripts and reference transcripts as queries.

The expression level of genes was calculated using RSEM[121] v1.2.12. The expression of genes in expanded families is shown in Figure 3D and Table S3. For seasonal transcriptome, samples in Lazarev Sea were used to analyze different expression because this place has extreme day and night phenomenon. Genes differentially expressed were identified by DEseq2[122] v1.14.1. Differentially expressed genes (DEGs) were identified using a Benjamini-Hochberg corrected p value (p-*adj*) cutoff value of 0.05 and a minimum absolute 2-fold change ($\log_2$FC > 1). DEGs related with molting and energy metabolism were shown (Figure 3F; Table S3).

To compare the relationship between gene length and expression level among krill tissues, we performed quantile normalization on the expression matrix using the ''preprocessCore'' v1.44.0 in R (https://github.com/bmbolstad/preprocessCore). As for the expression of long genes (>180 kb) in krill, we identified the homolog genes from the long genes among *E. superba*, *Homarus americanus*, *Procambarus clarkii*, *Penaeus indicus*, and *Penaeus monodon* using the RBH method. Taking krill as a reference, the homologous gene pairs were identified as homologous gene sets to calculate the quantile normalized gene expression matrix. Nonparametric tests and Benjamini-Hochberg corrected p value were used to compare the distribution difference between different groups (Figures S1G–S1J). The long gene was visualized using IGV[138] v2.4.14 (Figure 2A).

Circadian E-box identification

We took advantage of the long RNA-seq to estimate the position of transcription start sites (TSSs) of each gene. In eukaryotes, core promoter regions – including initiator (Inr), TATA-box, downstream promoter element, and CpG islands – are usually found from

Please cite this article in press as: Shao et al., The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights, Cell (2023), https://doi.org/10.1016/j.cell.2023.02.005

500 bp upstream to 500 bp downstream the TSS. Since other promoter elements, such as ATG deserts and E-boxes, may be found at a greater distance,[159] we extended the search for the E-box element up to 2,000 bp upstream the TSS. Next, we analyzed the putative promoter region of the main clock genes that, according to literature, should be under the CLK/CYC control looking for the consensus E-box motif 'CA[AT/TA]TG'. Briefly, a frequency matrix of E-box (transcription factor binding sites), MA0249.1 and MA0249.2, were downloaded from the JASPAR CORE database (http://jaspar.genereg.net). We extracted the gene sequences from the genome according to the gene structure annotation, then mapped the gene sequence back to the PacBio HiFi-CCS reads using BLAT with default parameters. We only kept the hits where the start 100 bp of the gene sequences were located within the HiFi-CCS reads for the downstream analysis. Next, the 2,000 bp 5′ flanking sequences were extracted from the HiFi-CCS reads by 'subseq' function of seqkit[142] v0.10.2. The E-box frequency matrices, MA0249.1.meme and MA0249.2.meme, were used to scan the flanking sequence from the previous step by applying the 'fimo' function of MEME[143] v5.3.3 to locate the position of E-box sequences. Finally, we sorted the candidate regions according to mapping identity and filtered out candidate regions with less than three reads.

SNP detection

Follow the same method in WGS short reads sequencing, we produced the whole genome sequencing reads of Antarctic krill (75 individuals) and North Pacific krill (three individuals used as an outgroup) on DNBSEQ-T1 platform (a total of 64.60 Tb data), and then filtered the low-quality reads using SOAPnuke with the parameters '-n 0.1 -q 0.5 -L 12'. To accelerate the alignment and SNP detection for our huge genome, Sentieon Genomics Tools v202010.02 (https://www.sentieon.com/) was used for alignment, sorting, duplicate removal, re-alignment, haplotype calling for each sample, and joint calling. Serious steps of filtering steps were performed to obtain a high-quality SNP dataset. Firstly, we used GATK[124] v3.8.1 VariantFiltration to perform hard filtering of SNPs with the criterion of '$QD < 2.0 \| MQ < 40.0 \| FS > 60.0 \| ReadPosRankSum < -8.0 \| MQRankSum < -12.5 \| SOR > 3.0$' as recommended by GATK. Secondly, only biallelic SNPs were used in the downstream analysis. Thirdly, we used the program GenMap[125] v1.3.0 to calculate the mappability across the whole genome, for the uniqueness of 35-mers with no more than one mismatch. The length of genome with mappability ≥ 0.1 is 13,621,110,494 (Table S4). After the filtration, a total of 364,568,426 SNPs was used to do the population genetic analyses. Random selected SNPs located on the genic and inter-genic regions were verified by PCR amplification and sanger sequencing.

Population structure and genetic diversity

Population structure analysis

In the Antarctic map, the Antarctic Divergence separating the prevailing Antarctic Circumpolar Current (ACC) with the Antarctic Coastal Current (ACoC), the gyres in the Ross Sea, and the Weddell Sea were schematic illustration based on previous study.[160,161] We firstly excluded the individuals with sequencing depth less than 10× to avoid bias caused by low sequencing depth, and thus a total of 66 samples were used in population structure analysis. To avoid potential bias caused by linkage disequilibrium (LD), LD pruning were performed using PLINK[127] v1.90b6.6 with the parameters '–indep-pairwise 2,000,100 0.2'. We finally produced a prune SNP dataset containing 47,555,257 SNPs from whole genome-wide 364,568,426 SNPs after LD pruning. The 1-IBS (identity by state) genetic distance matrix was calculated using PLINK with parameter '–distance 1-ibs', and a neighbor-joining phylogenetic tree was constructed using the PHYLIP[128] package v3.69. The Newick tree was visualized in Figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). PCA (principal component analysis) was performed using EIGENSOFT[133] v7.2.1 with parameters 'numoutevec: 20 numoutlieriter: 0 qtmode: 0'. Meanwhile, we also performed the PCA analysis based on removing high divergent SNPs (using three different F_{ST} cutoffs including 0.05, 0.10 and 0.15, separately) to detect the population structure with same parameters (Figures S3Q–S3T).

STRUCTURE[45] v2.3.4 was used to perform ancestry inference. Considering the limitation in computational speed and computer memory of this software to handle large-scale SNP datasets, 20 small datasets with each containing 100,000 random SNPs selected from genome-wide LD pruned SNPs were produced. The ancestry inference for 20 small datasets was run with 10,000 burn-in periods with 20,000 Markov chain Monte Carlo (MCMC) steps using an admixture model and correlated allele frequencies among groups for each value K (number of assumed ancestral components) ranging from 1 to 6, separately. To calculate the value of K with the best model, we used the ΔK method which was achieved in CLUMPAK.[132]

Another two tools were also used to discover the population structure, including PCAngsd[129,130] v1.10 and NGSadmix v3.2. For PCAngsd analysis, the input BEAGLE files based on genotype likelihoods were firstly conversed by VCFtools[134] v0.1.17, and then PCAngsd was run with no iteration for estimation of individual allele frequencies and none iteration for minor allele frequencies estimation in BEAGLE files, as well as 100 iterations for estimation of individual allele frequencies and 200 iterations for minor allele frequencies estimation in PLINK files. The output covariance matrix was performed the PCA using R function "eigen" v3.6.2[144], and package ggplot2[145] v3.4.0 for plotting (https://github.com/tidyverse/ggplot2). NGSadmix was run 5 replications with K (number of assumed ancestral components) ranging from 1 to 6. To detect the true value of K, we used the ΔK method which achieved in CLUMPAK.

In order to detect whether there are some SNP subsets, which influence whether the structuration pattern can be detected, we constructed the filtered SNP sets by gradually removing the SNPs according to their F_{ST} values. We observed that the number of SNPs with $F_{ST} > 0.15$ only accounts for 0.052% of whole SNPs, and the percentage of SNPs with $F_{ST} > 0.05$ is 3.32%. We found these four groups were 'mixed' when the SNPs with high F_{ST} values were removed. In details, when the SNPs with the $F_{ST} > 0.1$, the individuals from four geographical groups are completely mixed, although 99.9% SNPs were retained (Figures S3Q–S3T). We observed the SNPs of $F_{ST} > 0.05$ are distributed in 76.19% of assembled contigs, which implies that these high divergent SNPs are scattered distributed across the genome, rather than concentrated in several genomic regions. The very small percentage of SNPs ($F_{ST} > 0.1$) revealed that the Atlantic krill is no strong population structure.

Population genetic diversity

The genetic diversity ($\theta\pi$) was firstly calculated using VCFtools for each site with parameter "–site-pi" for each group, separately. Then, the $\theta\pi$ of each group was calculated as that the sum of $\theta\pi$ on each site was divided by the length of effective genome length (that is, the remain genome length by estimation of genome mappability using GenMap). Tajima's *D* of each group was calculated using VCFtools with non-overlapping window of 100 kb length with parameter "–TajimaD 100,000". The relatedness between each pairwise of two individuals was calculated based on the unadjusted Ajk statistic using VCFtools with parameter "–relatedness".

F_{ST} of each group was calculated using with non-overlapping window of 100 kb length with parameter "–fst-window-size 100,000 –fst-window-step 100,000". We used permutation test to examine the significance of observed F_{ST}. In details, we randomly shuffled the individuals to construct 'mock groups' and calculate the pairwise F_{ST} for the mock group pair, and we kept shuffling and calculated F_{ST} for 200 times using VCFtools with same parameters. Then we compared the observed F_{ST} value of real geographical groups to those of mock groups, one-side of more than 5% was used to determine the significance of F_{ST}. From the testing of F_{ST} values between geographical groups and 200 mock groups replicates, we observed the signal of genetic structuration among SG, SSI, and PB geographical groups.

Gene flow and population connectivity

D-statistic,[162,163] also known as the ABBA-BABA statistic, was performed to assess the evidence of gene flow between groups. All SNPs were used to assess the gene flow among four geographic groups, with three North Pacific krill samples specified to be the outgroup, using 'Dtrios' command in Dsuite[42] v0.4 with default parameters. TreeMix[43] v1.13 was used to infer the migration events among these four groups with three North Pacific krill individuals as outgroup. We used 4,073,894 SNPs with inferred ancestral alleles (ancestral alleles were defined as at the status of homologous genotype and at least two outgroup individuals have called genotypes) for this analysis. TreeMix was applied to this dataset to generate maximum likelihood trees and rooted by outgroup (North Pacific krill). Linkage disequilibrium was accounted by grouping SNPs into blocks of 2000 (–k 2000). Standard errors (–se) and bootstrap replicates (–bootstrap) were used to evaluate the confidence in the inferred tree topology and the weight of migration events. Migration events (ranging from 0 to 8) were estimated, and ten independent runs were conducted. The optimal number of migration edges was estimated by OptM[135] v0.1.6. The result suggests the optimal number of migration edges is 1, and the highest likelihood was chosen to graph this given migration scenario. The contemporary migration rates among sampling locations were evaluated using the Bayesian inference approach BayesAss[136] as implemented in the BA3-SNPs[44] program, with 1.0×10^6 iterations, a burn-in of 1.0×10^5 steps and a sampling frequency of 100 (with parameters "-i 1,000,000 –sampling 100 –burnin 100,000") during the Markov chain Monte Carlo (MCMC) sampling. Ten independent runs initialized with different seeds were conducted to examine convergence by comparing the posterior mean parameter estimates for concordance.

Detecting SNPs related to natural selection

To investigate and compare the role of geography and environment in shaping spatial genetic variation. The genetic distance is based on linearized F_{ST}, which is calculated as $F_{ST}/(1 - F_{ST})$. The pairwise geographic distance was calculated using the R package 'geosphere' v1.5-18 (https://github.com/rspatial/geosphere) based on the information of longitude and latitude. For each sampling location, environmental data was collected from publicly available databases[164] (ERA5 monthly averaged data on single levels from 1959 to present: https://doi.org/10.24381/cds.f17050d7; Copernicus Marine Environment Monitoring Service-Global ocean biogeochemistry hindcast: https://doi.org/10.48670/moi-00019; Copernicus Marine Environment Monitoring Service-Global Ocean Physics Reanalysis: https://doi.org/10.48670/moi-00021; Moderate Resolution Imaging Spectroradiometer (MODIS) : https://doi.org/10.5067/TERRA/MODIS/L3M/CHL/2018, https://doi.org/10.5067/AQUA/MODIS/L3M/CHL/2022). Specifically, we chose ten environmental variables with potential impacts on krill physiology and ecology from 2000 to 2020. The environmental variables were firstly scaled using the function "scale" in R package "base" v4.0.4 with parameter "center = TRUE, scale = TRUE", and then the pairwise Euclidean distance between each pair of sampling location was calculated using the function "vegdist" in R package "vegan" v2.6-4[139] with parameter "method = euclidean, binary = TRUE, diag = TRUE, upper = TRUE". To test the significance between genetic and geographic distance, patterns of isolation-by-distance (IBD) were examined using Mantel test by function "mantel" in R package "vegan" Mantel tests with 999 permutations with parameter "method = pearson, permutations = 999". The isolation-by-environment (IBE) was tested using partial Mantel test by function "mantel" in R package "vegan" Mantel tests with parameter "method = pearson, permutations = 999" while controlling for the effect of geographic distance.

To detect the potential loci and genes under local adaptation, we first extracted the SNPs in gene body and flanking 1 kb of the upstream and downstream on genes. Meanwhile, the minor allele frequency less than 0.05 were removed, and finally 1,143,372 SNPs were remained to perform this analysis. We next used two tools including PCAdapt[131] v4.3.2 and FSThet[140] v1.0.1, which were based on different approaches to reduce the false positive discovery. The first method inferred outliers based on principal component analysis (PCA), which assumes that candidate markers are outliers with respect to how they are related to population structure. We used PCAdapt with parameter "$K = 2$", and the p values were adjusted using Benjamini-Hochberg (BH) method using the function "p.adjust" in R v4.0.4. We detected 2,314 putative adaptive SNPs under an expected false discovery rate of $\alpha = 0.1$. The second method corresponds to FSThet, which identifies candidate loci by calculating smoothed quantiles between loci with strong differentiation F_{ST} relative to their expected heterozygosity. We used FSThet with default parameters and detected 71,060 SNPs. We obtained 1,028 SNPs potentially associated with natural selection by the intersect of the two tools.

Based on the environmental factors collected and genotypes called in this study, we detected whether genetic variation exhibit association with ecological variables. We used a univariate latent-factor linear mixed model (LFMM) implemented in the function "lfmm2" of R package LEA[50] v3.10.0 with parameter "-K 2" to search identified the associations between allele frequencies and ten environmental variables, separately. Among the 1,028 SNPs detected by both PCAdapt and FSThet, 387 SNPs were also support by at least one environmental factor, which located on 228 genes involved in 25 pathways. We also plot the genome-wide distribution using R package 'CMplot' v4.2.0[146] and the allele frequency of 387 adaptive SNPs in four groups using R package 'scatterpie' v0.1.8 (https://github.com/GuangchuangYu/scatterpie).

Demographic history inference

The generation time and mutation rate were essential for demographic history inference. The generation time was previously reported to be 2 years of Antarctic krill from egg to adult.[165] We then estimated the mutation rate following the previously described method following the formula: $\mu = D \times g/2T$, where D is the sequence divergence, T is the estimated divergence time, and g is the generation time.[166] The pairwise alignment of between Antarctic krill and closely related species *E. sinensis* on whole-genome level showed the observed sequence divergence of 17.2%, and further corrected to actual sequence divergence of 19.5% using the Jukes-Centor model. The divergence time was estimated to be 315.65 mya in the comparative genome analysis above, and the generation time is 2 years. Thus, the mutation per site per generation was estimated to be 6.19×10^{-10}, and at a relatively low level for Antarctic krill compared to other animals (Table S4).

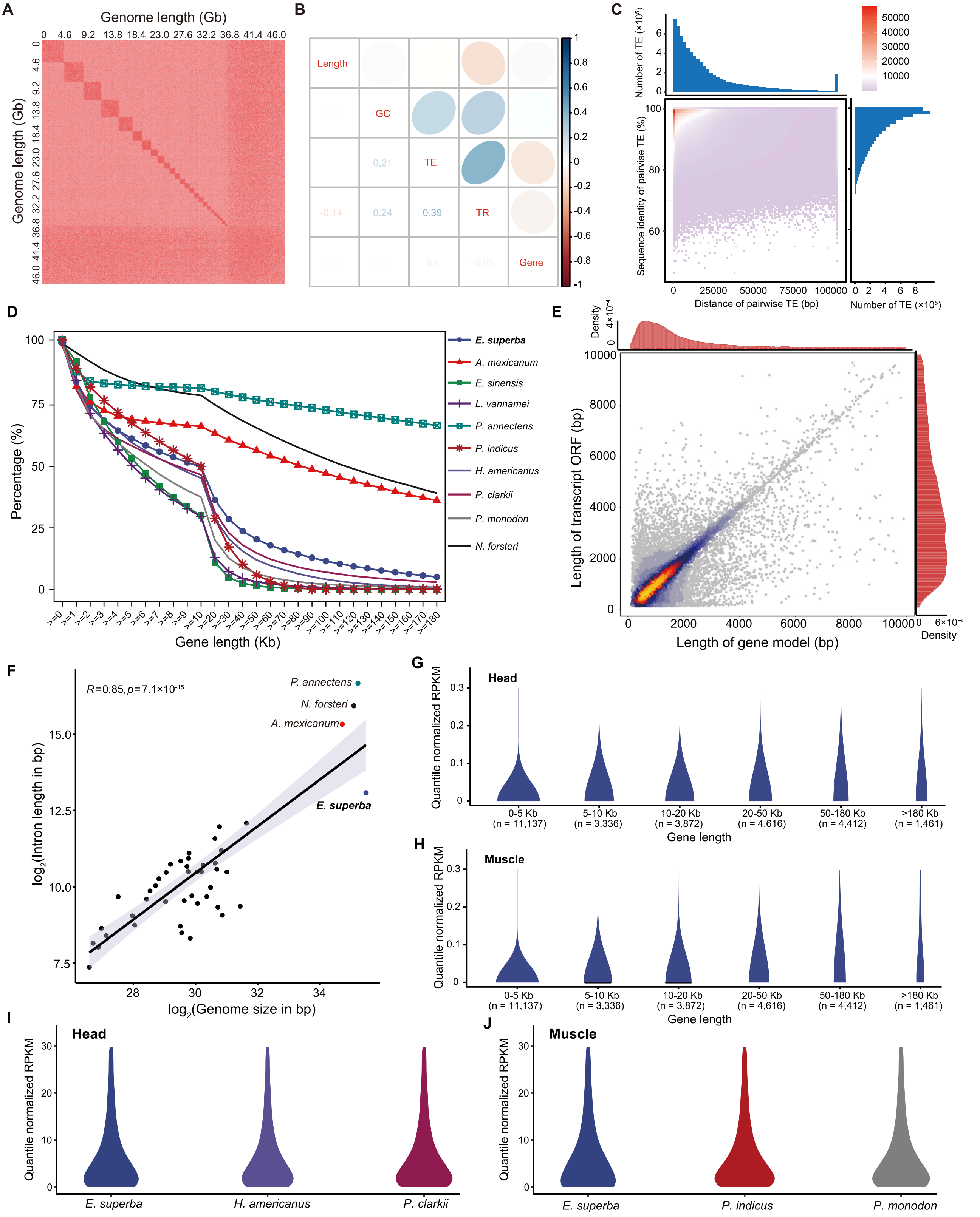
PSMC[51] v0.6.5-r67 was used to infer the history of effective population size (*N*e). We first constructed a high-quality of diploid consensus sequence for each sample using SAMtools mpileup and BCFtools[134] v1.4 call with the parameters '-C50' and '-d 3 -D 100', separately. Then, the consensus sequence was transformed to FASTA-like format using fq2psmcfa with parameter '-q20'. Finally, PSMC was used to infer the history of *N*e parameter '-N25 -t15 -r5 -p 4 + 25*2 + 4+6' with 100 rounds of bootstrapping. The PSMC figure was drawn using an estimated Antarctic krill generation time (g) of 2 and a mutation rate per generation per site (μ) of 6.19×10^{-10}. An approximate Bayesian computation method PopSizeABC[52] v1 was used to estimate demographic history. First the simulate summary statistics were carried out 5 times by the script 'simul_data.py'. In each time, 100 simulated datasets and 100 independent segments in each dataset were used. Then the script 'stat_from_vcf' was used to compute observed summary statistics. Finally, the all simulate summary statistics and observed summary statistics were used to perform ABC (approximate bayesian computation) estimation by the script 'abc.R'. The minor allele frequency threshold was 0.2 for the allele frequency spectrum (AFS) and IBS (identity by state) statistics computation and LD (linkage disequilibrium) statistics computation in all steps.

The recent effective population size (*N*e) of Antarctic krill was around 2×10^6 (estimated by PSMC in recent 10 thousand years), and the census population size (*N*c) of Antarctic krill was estimated to be about 4×10^{14} (in total 300–500 million tones, and a weight of about 1 g of each individual). Thus, the ratio of the effective population size (*N*e) to census population size (*N*e/*N*c) is extremely low to 5×10^{-9} ($2 \times 10^6/4 \times 10^{14}$).

QUANTIFICATION AND STATISTICAL ANALYSIS

All quantitative and statistical analyses were performed using the R computational environment and packages described above. The significance of difference of means between two data groups were conducted using the unpaired two-tailed Student's *t* test in Figure 1C. Differential gene expression was assessed by Wald test, and then was adjusted for multiple testing using the Benjamini-Hochberg procedure as implemented in DESeq2[122] in Figures 3A and 3F. Gene family expansion analysis was estimated based on a Monte–Carlo re-sampling procedure in Figure 3C. The permutation test was used in Figure 4B. The significance of patterns of isolation-by-distance (IBD) and isolation-by-environment (IBE) were examined using Mantel test by function "mantel" in R package "vegan"[139] in Figures 4D and S4B. Pearson's correlation coefficient and the significance was calculated using R package 'PerformanceAnalytics' v1.5.3 in Figures S1B, S1F, and S2B. The significance of *D*-statistics was calculated using Dsuite[42] v0.4 in Figure S3A. n in the Figure S1 represents number of genes. Quantification approaches and statistical analyses used in this can be also found in the relevant sections of the method details.

Supplemental figures



(legend on next page)

Figure S1. Genome sequencing, assembly, and annotation of Antarctic krill, related to Figure 1

(A) Hi-C contact heatmap of the chromosomes and scaffolds of the Antarctic krill genome assembly, ordered by scaffold length.

(B) Pearson's correlation coefficient (*r*) between contig length and various genome features. GC, percentage of GC; TE, percentage of transposable element; TR, percentage of tandem repeat; Gene, protein-coding gene regions.

(C) Distribution of TE identity and distance in the Antarctic krill genome.

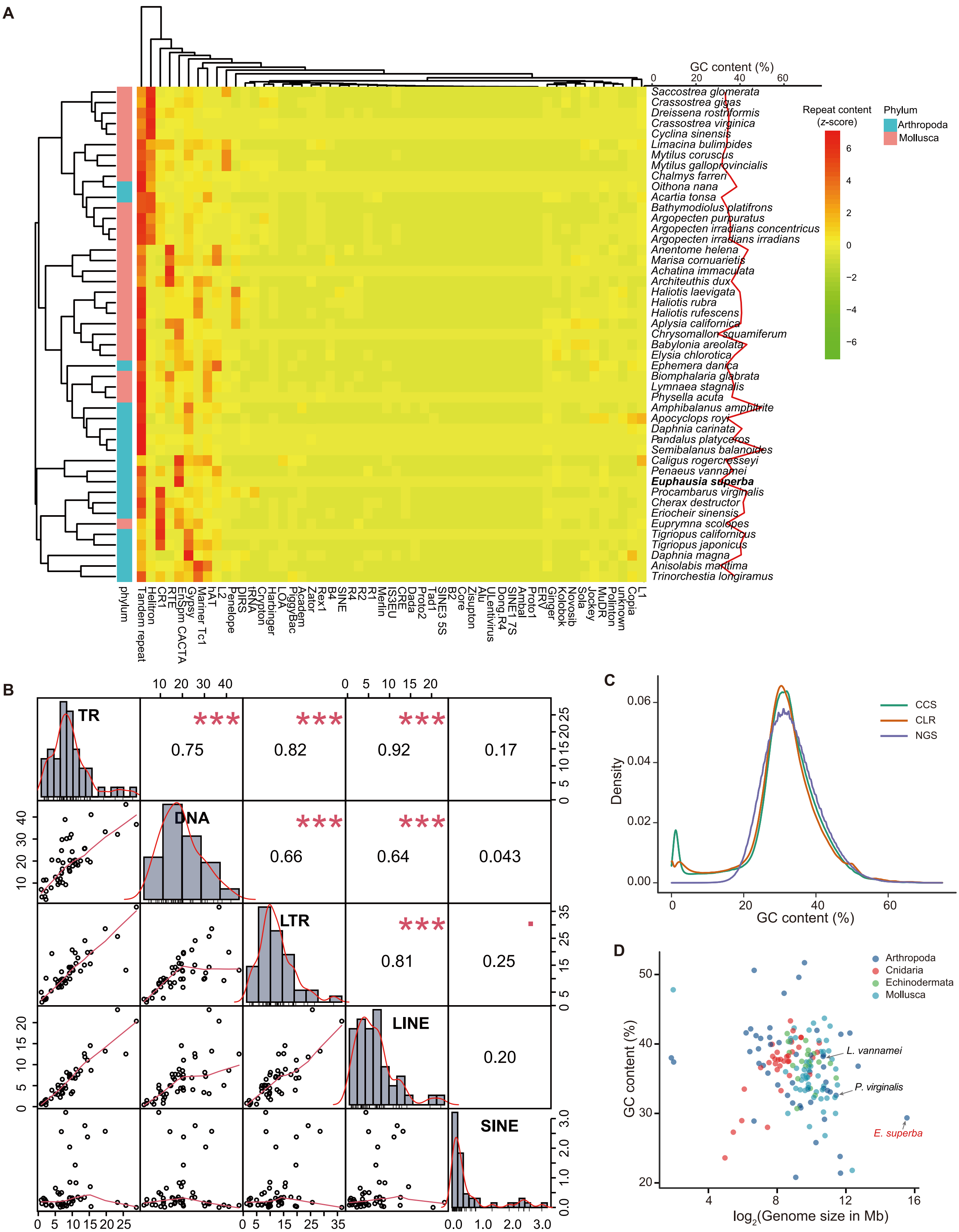
(D) Comparison of gene length among crustacean species and three species with large genomes.

(E) Pairwise comparison of the length between gene models and full-length transcripts.

(F) Comparison of the relationship between intron size and genome size among Antarctic krill, crustacean species and three species with large genomes.

(G and H) The relationship of gene expression level and gene length of Antarctic krill in two tissues.

(I and J) Comparison of the gene expression level and gene length of long genes in Antarctic krill and their homologous genes in other small genome crustaceans.



(legend on next page)

Figure S2. Comparison of genome repetitive elements and GC content of Antarctic krill and other invertebrates, related to Figure 2

(A) Heatmap and hierarchical clustering of repetitive sequence subtypes in invertebrate genomes, the green-red bar represents *Z* score percentage of repeat.

(B) Pearson's correlation analysis of repetitive sequence subtypes in invertebrates. A line of best fit is shown. Gray histograms with distribution lines (in red) of the subtypes are also shown. Three asterisks (***) denote significant correlation at $P < 0.001$. TR denotes tandem repeat; DNA, DNA transposon; LTR, long terminal repeat; LINE, long interspersed nuclear element; SINE, short interspersed nuclear element.

(C) Comparison of Antarctic krill GC content of sequencing reads. CCS, circular consensus sequencing reads; CLR, continuous long reads; NGS, next-generation sequencing reads (i.e., short-insert library by BGISEQ-500). Calculated as GC% per 500 bp non-overlapping sliding window for CCS and CLR, GC content per read for NGS.

(D) The relationship of genome size and GC content of 154 invertebrate genome assemblies.

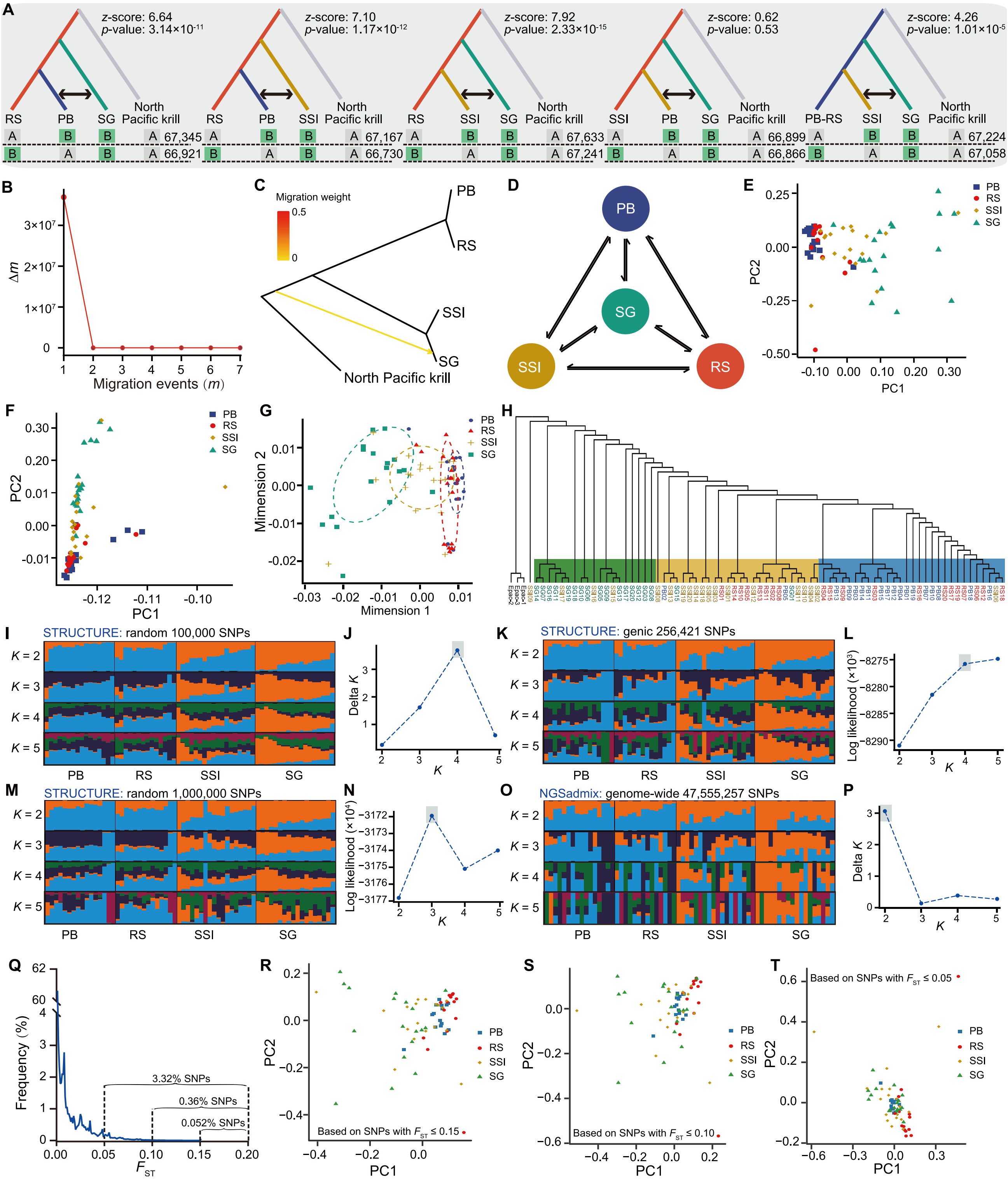


Figure S3. Population connectivity and differentiation between four groups and population structure, related to Figure 4

(A) Four ABBA-BABA tests among four geographical groups with North Pacific krill as outgroup. The numbers of ABBA and BABA sites, *Z* score and *p* value are also shown.

(B) The optimal value of migration event (*m*) is inferred from the second-order rate of change in likelihood (Δm) across incremental values of *m* using OptM, which shows the optimal value of migration event (*m*) = 1.

(legend continued on next page)

(C) TreeMix analysis of Antarctic krill samples from four geographical groups. The arrow corresponds to the direction of migration, and the migration weight is given according to the color of the arrows.

(D) The contemporary migration rates among four geographical groups inferred by BA3-SNPs. The arrow shows the migration direction.

(E and F) Population structure detected by PCAngsd based on genotype (GT) and genotype likelihood (GL) tags in VCF file, separately.

(G) Population structures identified by multidimensional scaling (MDS) analysis, and first two MDS dimensions (C1 and C2) are plotted. The points represent samples and the colors of points corresponding to the sampling location.

(H) The neighbor-joining tree constructed for the samples and the different shade colors were used to display the main groups. The green was mainly for SG group, the yellow mainly for SSI and the blue mainly for PB and RS groups.

(I) The STRUCTURE result shows the estimation of the proportion of ancestry under model K from 2 to 5 based on 100,000 SNPs randomly from genome-wide SNPs.

(J) The best model with $K = 4$ was suggested by Delta K (ΔK, Evanno Method), and was shaded in gray.

(K) The STRUCTURE result shows the estimation of the proportion of ancestry under model K from 2 to 5 based on SNPs in genic regions (256,421 SNPs after LD pruning).

(L) The best model with $K = 4$ was suggested by maximum of natural logarithm of the likelihood and was shaded in gray.

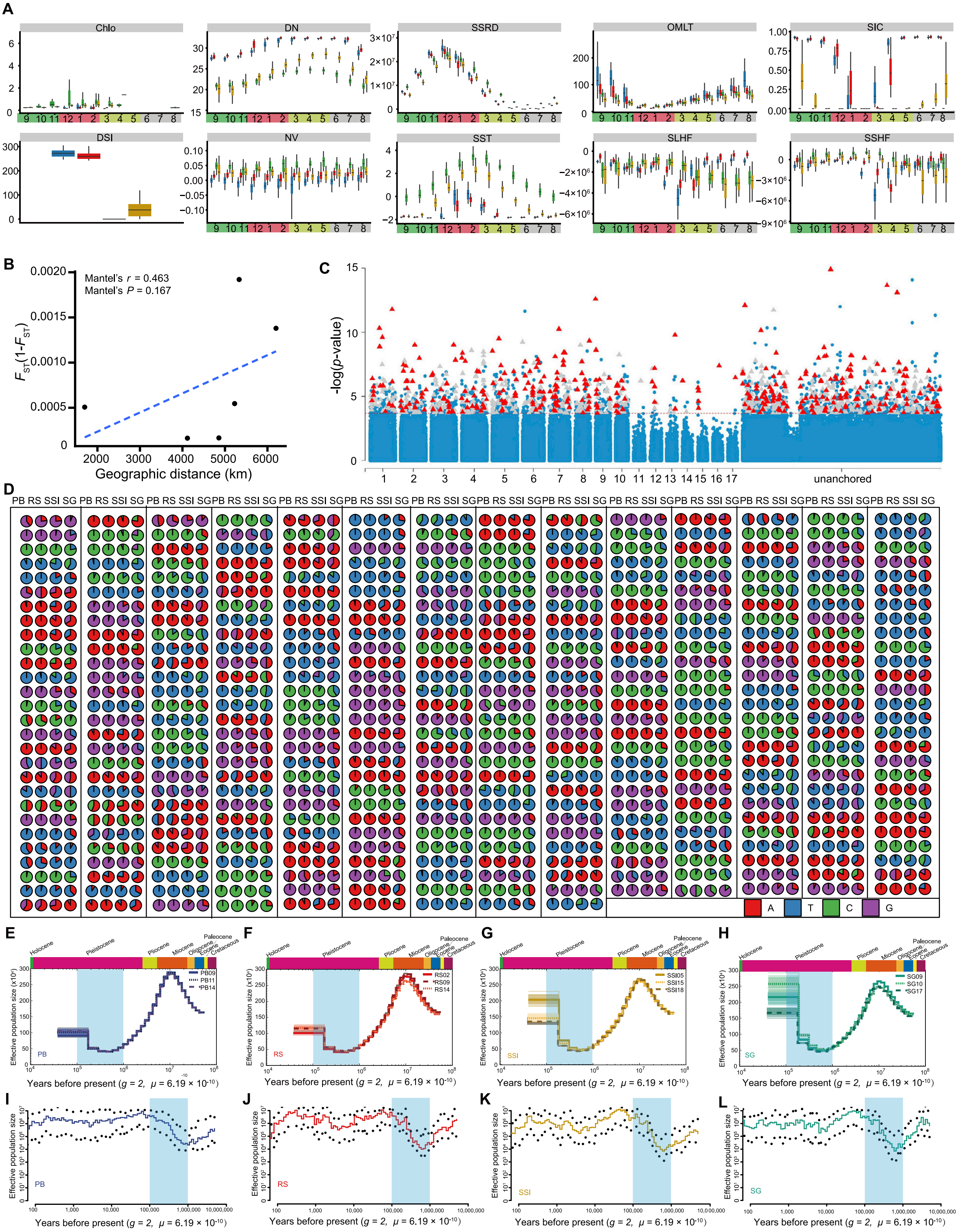
(M) The STRUCTURE result shows the estimation of the proportion of ancestry under model K from 2 to 5 based on 1,000,000 SNPs randomly from genome-wide SNPs.

(N) The best model with $K = 3$ was suggested by maximum of natural logarithm of the likelihood and was shaded in gray.

(O) The population structure of Antarctic krill was detected based on genome-wide SNPs using NGSadmix K from 2 to 5.

(P) The best model with $K = 2$ was suggested by Delta K (ΔK, Evanno Method), and was shaded in gray. PB denotes Prydz Bay; SG, South Georgia; RS, Ross Sea, SSI, South Shetland Island.

(Q) The F_{ST} value distribution of genome-wide SNPs. The percentage of SNPs with $F_{ST} > 0.15$ is 0.052%, the percentage of SNPs with $F_{ST} > 0.10$ is 0.36%; the percentage of SNPs with $F_{ST} > 0.05$ is 3.32%. (R-T) PCA analysis was conducted based on the SNPs with $F_{ST} \leq 0.15$, $F_{ST} \leq 0.1$ and $F_{ST} \leq 0.05$, separately.



(legend on next page)

Figure S4. The detection of SNPs associated with environment variables and inference of population demographic history, related to Figure 4

(A) The boxplot of ten environmental variables across 12 months from 2000 to 2020, four seasons are shaded by different colors, spring from 9 to 11, summer from 12 to 2, autumn from 3 to 5, and winter from 6 to 8. SSRD denotes mean surface solar radiation downwards of summer (J/m^2); SSHF denotes mean surface sensible heat flux of winter (J/m^2); SST denotes mean sea surface temperature of winter ($^\circ$C); NV denotes mean northward velocity of winter (m/s); SLHF denotes mean surface latent heat flux of summer (J/m^2); OMLT denotes mean ocean mixed layer thickness of summer (m); SIC denotes mean sea ice concentration of summer (%); DSI denotes mean duration of sea ice (retreat - advance); Chloe denotes mean chlorophyll of winter (mg/m^3). For each box of the boxplots, the center line represents the median, the bottom line represents the 25th percentiles and the top line represents the 75th percentiles. The whiskers of the boxplots show 1.5 inter-quartile range (IQR) below the 25th percentiles and 1.5 IQR above the 75th percentiles.

(B) The isolation-by-distance (IBD) test for relationship between demographic distance and genetic distance (F_{ST}/(1-F_{ST})).

(C) The distribution of SNPs associated with environmental variables. The blue points (including dots and triangles) donate all genic SNPs tested in this analysis. The triangles (both in gray and red) indicting the SNPs only supported by natural selection, and the triangles in red indicting the SNPs supported by natural selection and environmental associations.

(D) The allele frequency of SNPs associated with natural selection and environmental variables in four geographical groups. Different colors represent different nucleotides in the pie chart.

(E–H) Inference of population demographic history using PSMC with randomly selected high-sequencing individuals (>20×) with 100 PSMC bootstraps (in light color).

(I–L) Inference of population demographic history using popSizeABC, with 90% confidence interval of bootstrapping shown by the black dotted points. For all figure panels in this figure, the light blue shading indicates a period of expansion after population bottleneck of Antarctic krill. A generation time (g) of two years and mutation rate (μ) of 6.19×10^{-10} substitutions per site per generation was employed. PB denotes Prydz Bay; RS, Ross Sea; SSI, South Shetland Island; SG, South Georgia.