

RESEARCH

Open Access



A new phased assembly of the Antarctic spiny plunderfish provides novel insights into the evolution of the notothenioid radiation

Jacopo Martelossi¹, Ksenia Krasheninnikova², Amy Denton², Jonathan M. D. Wood², Thomas C. Mathers², Richard Durbin^{2,3}, Nova Fong⁴, David L. Bentley⁴, Melody S. Clark⁵ and Iliana Bista^{1,2*}

*Correspondence:
ilianabista@gmail.com

¹ Senckenberg Research Institute and Natural History Museum, Frankfurt Am Main, Germany

² Wellcome Sanger Institute, Tree of Life, Wellcome Genome Campus, Hinxton, UK

³ Department of Genetics, University of Cambridge, Cambridge, UK

⁴ Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, USA

⁵ British Antarctic Survey, High Cross, Madingley Road, Cambridge, UK

Abstract

Background: Notothenioids are a well characterised species flock endemic to the Antarctic and an important model group for studying genome adaptation to extreme cold. A new reference assembly and clade-wide comparative genomic analysis allow for investigation of cryonotothenioid evolution and the appearance of novel functionalities linked to cold adaptation.

Results: A new phased assembly of a model notothenioid, *Harpagifer antarcticus*, demonstrated low levels of haplotypic variability across the genome. Nevertheless, numerous insertions from multiple LINE-L2 clades were found, suggesting ongoing transposition with potential contribution to speciation. Contrary to expectations the *afgp* locus was highly similar between haplotypes, except for large length allelic variants of *afgp* genes. Analysis suggests a model for the *afgp* locus expansion in *H. antarcticus* through segmental tandem duplications involving two pairs of *afgp* genes at time. Syntenic reconstruction of genomes from across the clade demonstrates conserved macrosyntenic relationships and group specific chromosomal fusions of notothenioids. Quantification of genome gain and transposition rates during cryonotothenioid diversification showed a first ancestral slow genome expansion concurrent with historic temperature drops. This was followed by lineage-specific massive peaks of genomic gain and transposition activity. Finally, we identified a set of genes that underwent ancestral diversifying selection and acquired novel conserved non-coding elements during the cryonotothenioid emergence. These were related to antioxidants and proteostasis, which may have facilitated the notothenioid Antarctic radiation.

Conclusions: Diversifying selection and genomic gain linked to transposon activity are primary contributors to lineage-specific evolutionary dynamics through the clade which facilitated adaptation to life in the cold.

Keywords: Cold adaptation, Transposons, LINE-L2, *Afgp*, Haplotype diversity, Genome expansion, Proteostasis, Diversifying selection, Antioxidant activity



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Background

The notothenioids are one of the most diverse and dominant fish groups in the Southern Ocean and a rare example of a well characterised marine flock [1]. The clade is formed by approximately 140 species grouped in eight families, including three non-Antarctic (basal groups) and five mostly high specialised Antarctic families which are also known as cryonotothenioids [2]. They are characterised by genomic adaptations to extreme cold and they are considered to be an important model group for the study of cold adaptation [3].

The appearance of the antifreeze glycoproteins (AFGPs) is one of the most characteristic innovations enabling freeze avoidance in this group, and are present in members of the cryonotothenioid clade [4]. Also Zona Pellucida (ZP) genes were found to contribute to freeze avoidance, especially during early developmental times [5]. However, multiple other adaptations on the molecular and cellular level have also been considered to play important roles in the survival and establishment of notothenioids (e.g., cold stable enzyme production, lipid composition of cellular membranes, modified mechanisms of protein homeostasis) (summarised in [6]). The mechanisms through which such adaptations may have arisen, or the timing of their appearance along the ancestral cryonotothenioid clade remain poorly understood.

Assembly of notothenioid genomes has been confounded by the high number of repeat elements present in these genomes [7]. This is exemplified by the structure of the antifreeze gene locus, which is characterised by tandemly arranged genes [4, 8], and high levels of repeat content making it very challenging to assemble and analyse accurately [7]. For the *afgp* gene locus in particular it has been suggested that high levels of structural variation on the haplotype level may be present. These estimates were based on BAC library sequencing reconstruction of the *afgp* locus of the Antarctic toothfish species *Dissostichus mawsoni* [9]. In that case it was shown that the two haplotypes of the *afgp* locus were characterised by extensive structural variation, with a distinct number of *afgp* gene copies, with a difference in overall size and mapping within each haplotype of the locus. Other reconstructions of the *afgp* locus based on whole genome sequencing data were only able to map the consensus structure of one assembled haplotype [7, 10, 11] though haplotig data have also been analysed [7]. To be able to test this hypothesis on other species, haplotype resolved (phased) genome assemblies are required.

Using Oxford Nanopore Technologies (ONT) and Hi-C data we assembled a new haplotype resolved chromosomal reference genome assembly for *Harpagifer antarcticus* (Antarctic spiny plunderfish). This is the first phased assembly for any notothenioid species. In addition, *H. antarcticus*, as a small, shallow-water and experimentally tractable member of the cryonotothenioids, is also an important model species for the study of cold adaptation [12–14].

Previous work on notothenioid genomes revealed the significant contributions of transposable elements (TEs) in the evolution of the clade. For the cryonotothenioid clade in particular it has been shown that a recent massive burst of transposons was linked to a significant genome expansion compared to the sub-Antarctic group [7, 15, 16], while the evolution of various other genomic features has also been linked to transposon activity such as expansion of the antifreeze locus, and the loss of haemoglobins in the Channichthyidae family [7]. Furthermore, it has been suggested that TE accumulation in

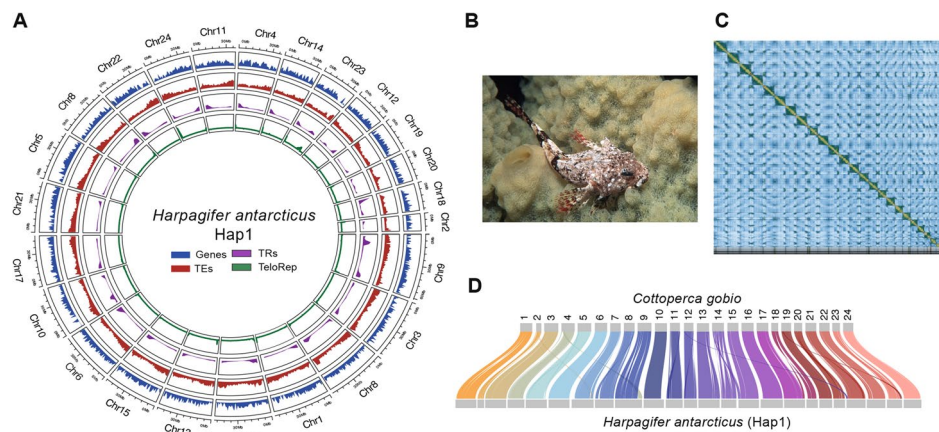


Fig. 1 Genome assembly and genomic features. **A** Circos plot showing the 24 chromosomes of the fHarAnt1.2 genome assembly (Hap1) and the distribution of protein coding genes (Genes), transposable elements (TEs), tandem repeats (TR), and canonical telomeric repeats (TeloRep). Chromosomes are ordered by length and named based on syntenic positions with *Cottoperca gobio* reference genome. Synteny between Hap1 – Hap2 and genomic features for both haplotypes are shown in Additional file 1: Fig. S4. [24]. **B** Photograph of a *H. antarcticus* specimen from Ryder Bay, Antarctic Peninsula (Photo credit: BAS Photo Library. Photographer: Simon Brockington). **C** Hi-C-contact map of fHarAnt1.2 Hap1. **D** Syntenic relationships between *H. antarcticus* Hap1 and *C. gobio* genomes

some notothenioid groups, may have promoted karyotypic reorganisations likely leading to species diversification [17].

In general, TEs are known to be a major source of genomic variation across eukaryotes by disrupting gene structures, reshaping regulatory networks, and co-opting novel regulatory elements [18]. The extensive presence of TEs throughout the genome might amplify the occurrence of recombination events thus promoting deletions, inversions, and translocations [19], while the presence of environmental stressors may also promote TE activity [20]. Using whole genome comparisons can be a powerful method for understanding the mechanisms and patterns of transposon activity at a large scale, while it could also enable highly accurate quantification of genomic losses and gains occurring during diversification events [21–23].

In the present study we employed large scale genome wide analyses of chromosomal references across the notothenioid radiation along with novel genomic resources, to understand the 1) overall contribution of transposable elements in the cryonotothenioid diversification, through timed and quantified analysis of genomic gains and losses, 2) the evolutionary dynamics of the antifreeze gene locus across the cryonotothenioids, and models of expansion of *afgps* and haplotypic diversity in the *H. antarcticus* phased assembly, 3) impact of other genomic features that may have enabled genomic adaptation to extreme conditions.

Results

Genome assembly and gene annotation

The assembled genome for *H. antarcticus* using ONT long reads and Hi-C data resulted in a fully phased assembly with two high-quality haplotypes (Fig. 1, Additional file 1: Fig. S1). The genome was assembled at a total length of 1,069 Mb (Hap1) and 1,222 Mb

(Hap2), and contig N50 of 17 Mb (Hap1) and 12.8 Mb (Hap 2). Each haplotype was assigned to 24 chromosomes (Fig. 1A, C) with a scaffold N50 of 44.8 Mb (Hap1) and 41 Mb (Hap2) (Additional file 2: Table S1). Both haplotypes were highly complete, with 99% BUSCO completeness for both haplotypes, and QV scores of QV 40 (Hap1) and QV 40.4 (Hap2) (Additional file 2: Table S1). Chromosomes were named based on syntenic relationships with *Cottoperca gobio* genome assembly (fCotGob3.1) [24] (Fig. 1C; Additional file 2: Table S2). Completeness of the assembly was confirmed by a clear enrichment of telomeric repeats on at least one end of all chromosomes on both haplotypes (Fig. 1A, Additional file 1: Fig. S2), confirming previous higher level localisation data using in situ hybridization data of the (TTAGGG)_n repeat motif [25]. Based on *k*-mer analysis, the heterozygosity rate of the *H. antarcticus* genome was estimated at 0.37%.

De-novo gene annotation (BRAKER) for Hap1 identified 23,057 protein-coding genes, with an average length of 14,904 bp, an average of 8.3 introns per gene, and mean intron and exon lengths of 1,555 bp and 174 bp, respectively (Additional file 2: Table S3). By lifting Hap1 annotation onto Hap2 we identified 22,463 (Additional file 2: Table S3) protein coding genes with 696 flagged as pseudogenes (i.e., containing premature stop codons). Functional annotation of these pseudogenes revealed multiple members of high copy number gene families, such as Zinc-finger proteins, leucine rich repeat containing proteins, immunoglobulins, and TRIM proteins (Additional file 2: Table S4) suggesting that haplotypic pseudogenization might mainly involve redundant genes subjected to relaxed purifying selection. 19 GO terms were enriched for these genes related to various biological processes, including immunity (GO:0002292), brain morphogenesis (GO:0048854, GO:0021987), chromatin remodelling (GO:0006338) and spermatogenesis (GO:0007283) (Additional file 2: Table S5). BUSCO completeness of the gene annotation was high for both haplotypes with 97.7% for Hap1 and 95.6% for Hap2 (Additional file 2: Table S3).

Genomic landscape of repetitive elements

Annotation of transposable elements on the fHarAnt1.2 genome revealed 40% coverage (Total repeats 43%) (Additional file 2: Table S6) on the assembled chromosomes of both haplotypes, with a clear enrichment towards chromosome ends (Fig. 1A). Each of the three major TE classes (LINEs, LTRs, DNA), were almost equally represented in Hap1, covering between 11–14% of the chromosomes. Unplaced scaffolds are dominated by a single satellite family, which we named HarAnt-Sat1, with a monomer length of 230 bp and occupying ~30% of their base pairs, whereas other repetitive sequences are much more depleted compared to the assembled chromosomes (Additional file 2: Table S6).

Classification at the RepeatMasker superfamily level was achieved for just 38.6% of LTR and 40.7% of DNA elements, in contrast to 97.5% of LINE elements. These predominantly consisted of LINE-L2 which accounted for 9.7% (Hap1) and 8.5% (Hap2) of the genome and correspond to 278,816 (Hap1) and 278,567 (Hap2) insertions (Additional file 2: Table S6). Tandem repeats, excluding HarAnt-Sat1, and simple repeats accounted for <1–2% of the genome on the haplotypes, respectively. Repeat landscape profiles describing activity of TEs through absolute time, were obtained by translating CpG-corrected Kimura divergence into absolute time using an approximated neutral substitution rate estimated from four-fold degenerate sites of BUSCO genes. These revealed that

the TE content of the Antarctic spiny plunderfish genome mainly derived from a recent burst of activity of all three major TE groups that occurred after the diversification of cryonotothenioids, between 4 and 7 Million Years Ago (MYA) (Fig. 2A).

To further characterize the evolutionary history of the highly abundant LINE-L2 elements we first employed a phylogenetic approach to identify evolutionarily distinct LINE-L2 families (Fig. 2B). Among the 33,670 insertions harbouring a reverse transcriptase segment, 86% clustered with reference L2 elements, whereas others are part of other L2-related clades (Crack, Daphne L2A and L2B) [26]. Eight different high copy-number LINE-L2 families were identified, which were named HarAnt_LINE-L2_I – VIII. These families accounted for 60% of the total LINE-L2 content, and 5.7% of the genome (both haplotypes). The two most abundant of these were L2_IV (27.4% LINE-L2, 1.5% of the genome), and L2_II (22.9% LINE-L2, 1.3% of the genome) (Fig. 2C). Dated repeat landscape plots suggest that all families were already active before the extant cryonotothenioid radiation, with a first peak concurrent with their emergence time ~ 10 MYA and a second more recent activity peak between 4 MYA and 500 thousand years ago (Fig. 2D). The two most abundant appear to have decreased their transposition rate over the last 1 MYA, whereas the less abundant ones LINE-L2_VI, LINE-L2_VII, and LINE L2_VIII, have become predominant.

Subsequent analysis of LINE-L2 families across 12 notothenioid genomes (Additional file 2: Table S7), identified all LINE-L2 families across all cryonotothenioid species (Fig. 2E). The three most abundant LINE-L2 families in the *H. antarcticus* genome were also the most abundant ones in the other species (Fig. 2E). In the sub-Antarctic species *E. maclovinus* and *C. gobio*, some low-abundance families, II and III respectively, were missing and therefore likely have originated along the cryonotothenioid stem branch (Fig. 2E).

A total of 170 (18S), 534 (5S), 166 (5.8S), and 181 (28S) ribosomal RNA (rDNA) genes were annotated in the *H. antarcticus* genome. Among the 18S, 5.8S, and 28S genes, multiple copies were found to be dispersed throughout the genome, however, the majority, 97.1% of 18S, 97.6% of 5.8S, and 97.2% of 28S, were located in a single main 45S rDNA locus at the end of chromosome 18 (Additional file 1: Fig. S3A, B). The 5S genes occur in three distinct contexts: (i) within the 45S locus forming 18S–5.8S–28S–5S arrays (155 copies), (ii) as 20 interspersed single copies scattered across the genome, and (iii) in two large tandem arrays on chromosomes 13 and 21 (Additional file 1: Fig. S3A, C), which contain 151 and 208 paralogues, respectively. These data expand our knowledge of 45S and 5S rDNAs in *H. antarcticus*, which were previously thought to be co-localised to a single locus [27]. However, the dispersed nature of the 5S locus is consistent with reports for other notothenioids and fish in general [27].

Haplotypic variation in the *H. antarcticus* genome

Alignment of the two haplotypes of the fHarAnt.1.2 assembly showed heterozygosity rate of 0.15%, consistent with the low heterozygosity estimated based on *k*-mers, and a generally low number of structural variants (Additional file 2: Table S8; Additional file 1: Fig. S4). More specifically 1.68% of the genome was present in a hemizygous state (i.e., composed of heterozygous insertions and deletions), and 8.9 Mb and 9.1 Mb was affected by heterozygous inversions on Hap1 and Hap2, respectively. Hemizygous

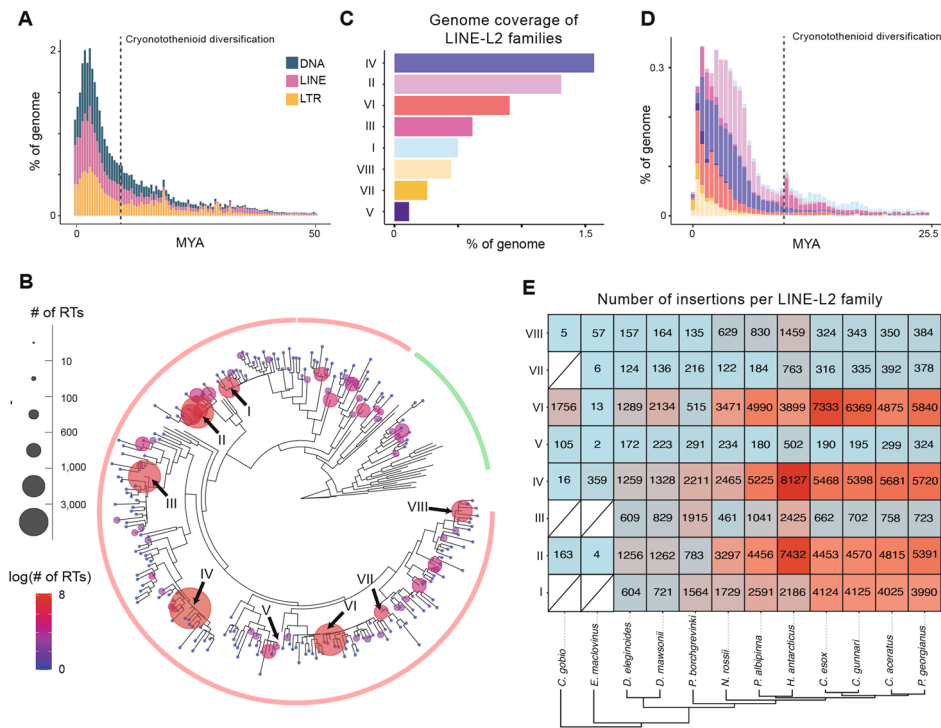
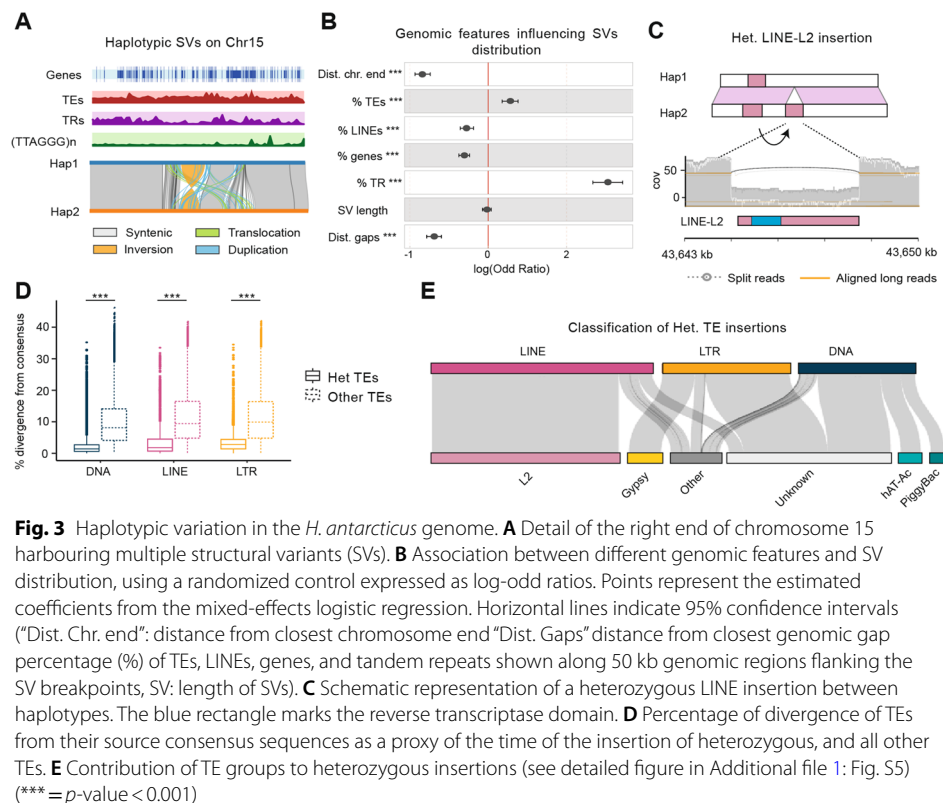


Fig. 2 LINE-L2 elements in the *H. antarcticus* genome. **A** Repeat landscape profile showing the activity of the three main TE groups through absolute time. Each bin represents one million years (MYA = million years ago). **B** Maximum likelihood phylogenetic tree of clustered LINE-L2 reverse-transcriptase sequences (fHarAnt1.2 genome). Tips represent the longest element for each cluster. The size and colour of bubbles is proportional to the number of elements contained in each cluster. Arrows highlight the eight identified high-copy number families. Tips without bubbles represent reference sequences extracted from [26]. **C** Genome coverage per LINE-L2 family, and **(D)** activity through absolute time of the eight LINE-L2 families highlighted in panel B. Each bin represents 500 thousand years. **E** Number of insertions across notothenioid genomes for each LINE-L2 family identified in *H. antarcticus*

deletions relative to Hap1 affect a total of 849 exons, and 136 genes are completely deleted on Hap2 and thus might be subjected to presence-absence variation at the population level. Importantly, 66.6% of all SVs were also detected in a heterozygous state when performing a read-based SV call, while only 1.7% were genotyped as homozygous for the alternate allele, indicating low levels of false positives due to assembly errors. The remaining 31.7% were not detected using ONT reads. Using generalized linear mixed models and chromosome identity as random effect, a significant positive association was observed between SV occurrence (considering inversions, duplications, and translocations) and repetitive content within 50 kb flanking regions of the breakpoints. The strongest predictor of SV occurrence was the density of tandem repeats, with an odd ratio of 2.7. Additionally, we found a negative association of SV occurrence with gene content, distance from both chromosome ends, and assembly gaps as well as LINE-L2 content (Fig. 3A, B).

The phased assembly allowed the inspection of the genome for the detection of potentially recently active TEs, as recent insertions are expected to be found mostly in a heterozygous state and thus correspond to insertion/deletion events between the two haplotypes (Fig. 3C). Overall, 20% (16,657) of the insertions/deletions have a reciprocal overlap of at least 75% with annotated TEs and might therefore correspond to



heterozygous TE insertions. Accordingly, these TEs have a significantly lower divergence from their consensus sequences than other TEs, which is used as a proxy for insertion time (One-tail Wilcoxon rank sum test, p -value < 0.001) (Fig. 3D). A 39.1% of all heterozygous TE insertions correspond to LINE-L2 elements followed by unknown LTRs with 17.2%, and unknown DNA transposons with 12.4% (Fig. 3E; Additional file 1: Fig. S5). Among all heterozygous TE insertions, most are located in intergenic regions (12,187), followed by introns (4,642) and 2,500 bp gene flanking regions (2,578). Only 23 insertions affect coding exons, leading to allelic pseudogenization of the impacted genes.

When evaluating the insertion patterns of heterozygous LINE-L2 elements relative to all other TEs, we observed a minor yet statistically significant underrepresentation of LINE-L2 within 2,500 bp gene-flanking regions (12.2% and 13.9%; chi-square post-hoc test with Bonferroni correction, p -value = 0.004). Conversely, when comparing heterozygous LINE-L2 insertions to those shared between both haplotypes, heterozygous elements were markedly overrepresented in intergenic regions (63.66% and 49.93%; chi-square post-hoc test with Bonferroni correction, p -value < 0.001) and significantly depleted within intronic regions (24.09% and 38.58%; chi-square post-hoc test with Bonferroni correction, p -value < 0.001).

Annotation of the ZP genes was done based on classification into 11 vertebrate types as described by [5]. In total 97 and 103 ZP genes were annotated on Hap1 and Hap2, respectively. Of these, five on Hap1 and three on Hap2 are likely pseudogenes as they possess multiple premature stop codons, whereas 48 and 52 putatively functional copies could be classified into one of the 11 types using phylogenetic analyses. Genes ZPAX2

and ZPAC1 showed the highest copy number in both haplotypes (Additional file 1: Fig. S6). For ZPAX1 and ZPC5, we found seven functional copies on Hap1 and eight on Hap2 for the former, and six functional copies on both haplotypes for the latter (Additional file 1: Fig. S6).

Genome evolution during notothenioid diversification

Synteny analyses of 12 notothenioid chromosome-scale genomes (Fig. 4A; Additional file 2: Table S7) revealed strong conservation of macrosyntenic relationships during their diversification, together with lineage-specific Robertsonian chromosomal fusions (Fig. 4B), in line with previous karyotype analyses [28]. Most assemblies, including *H. antarcticus*, show the predominant notothenioid karyotype of $n=24$, which largely comprises acrocentric chromosomes with the exception of *Notothenia rossii* (Nototheniinae) ($n=12$), and *Pagothenia borchgrevinki* (Trematominae) and *Pogonophryne albipinna*

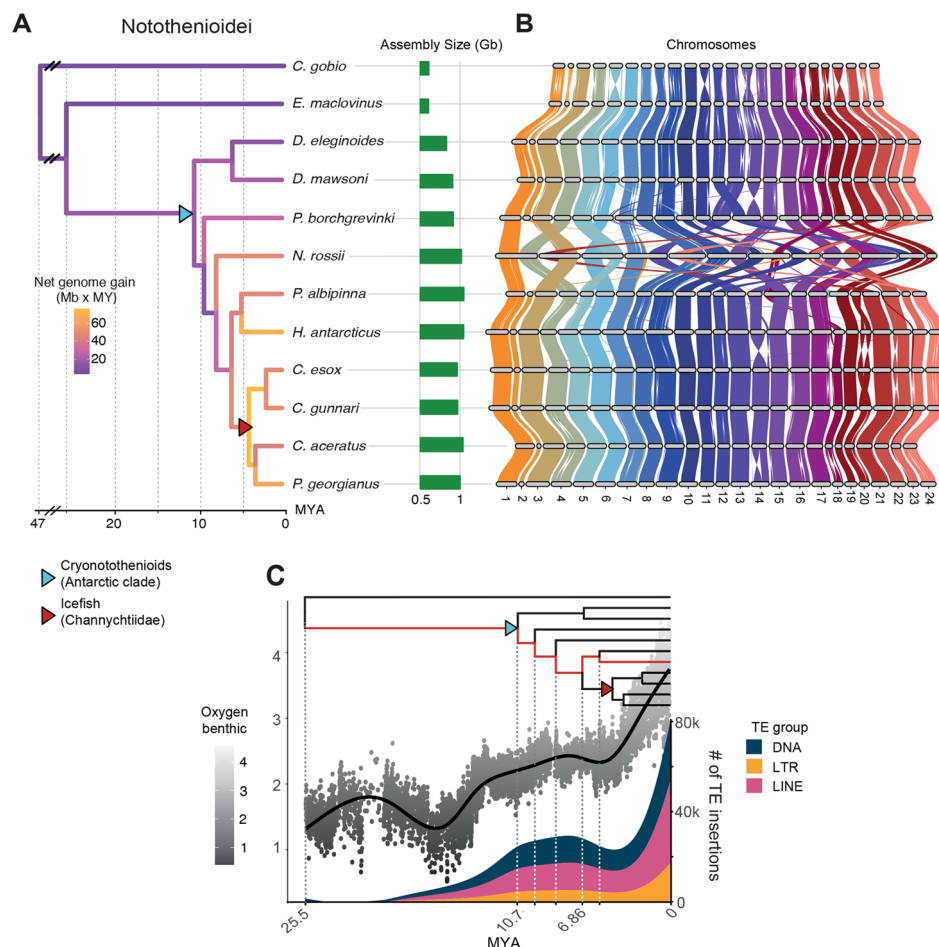


Fig. 4 Genome evolutionary dynamics during notothenioid diversification. **A** Dated phylogeny of notothenioid species included in all comparative genomic analyses. Branch colours show the rate of net genomic gain per million years as megabases per million years (Mb x MY). **B** Syntenic relationships among notothenioid chromosomes. Chromosomes were named based on synteny with *C. gobio* assembly. **C** Number of TE-related insertions along *H. antarcticus* ancestral and terminal branches (highlighted in red in the phylogenetic tree above), shown as a smoothed curve. Scatterplot and black curve show temperature variation over time based on deep-sea $\delta^{18}\text{O}$ records (data from [29])

(Arteidraconidae) ($n=23$). In *N. rossi* 10 of the chromosomes appear to have originated from the Robertsonian fusion of 10 ancestral acrocentric pairs, one from the fusion of three chromosomes, and one remained unfused leading to a karyotype of $n=12$. In *P. borchgrevinki* and *P. albipinna*, only two different chromosomes have been fused, resulting in $n=23$ karyotype in both species. Across all notothenioid genomes, detailed molecular mapping detected only a few large inversions and intrachromosomal translocations were detected, mostly at chromosome ends (Fig. 4B; Additional file 1: Fig. S4).

We used a whole genome alignment of 12 chromosomal assemblies to understand patterns of DNA gain and loss during notothenioid diversification and quantified the amount of bp involved in insertion and deletion events across all internal and terminal branches. Overall, genomic gain greatly outweighs genomic loss which accelerates towards recent times, reaching a maximum net rate of 76 Mb per MY along the ancestor of the Channichthyidae family (Fig. 4A; Additional file 2: Table S9). Along the cryonotothenioid stem branch we observed a net genome expansion of 256 Mb. Although this represents the fourth largest absolute gain across all branches, surpassed only by the terminal expansions in *N. rossii*, *H. antarcticus*, and *P. borchgrevinki*, the long branching time resulted in a relatively modest rate of genome gain of 16.46 Mb per MY (Fig. 4A; Additional file 2: Table S9).

Between 3,107 and 51,757 of the insertions inferred along *H. antarcticus* ancestors, map to an annotated TE element of its genome, requiring a reciprocal overlap of 75% between the two annotations and thus likely correspond to genuine TE insertions (Additional file 2: Table S10). The number of TE insertions per MY greatly increased during early evolution of cryonotothenioids, between 10.7–6.86 MYA, slightly decreased along the lineage leading to the *Harpagiferidae* and *Arteidraconidae* sister species, and reached a main spike of activity along *H. antarcticus* terminal branch (10,456 insertions per MY; Fig. 4C). DNA transposons were the most active along most branches, but LINEs were predominant along the cryonotothenioid stem branch (41% DNA vs 42% LINEs) and, particularly, *H. antarcticus* terminal branches (33% DNA vs 46% LINEs) (Additional file 2: Table S10). Patterns of TE activity, especially for LINE elements, mirror variations in temperature data, based on $\delta^{18}\text{O}$ variability [29] (Fig. 4C), with higher values related to lower temperatures.

Conserved elements and selection analyses on protein-coding genes

We identified genes and non-coding regulatory elements that may have facilitated the successful colonization and diversification of cryonotothenioids in the cold waters of the Southern Ocean. First, we analysed intergenic conserved elements (CEs) which are arising along their stem branch and are conserved among all analysed species based on the notothenioid whole genome alignments, and second, we investigated signs of diversifying selection on protein coding genes along the same ancestral branch using a gene-wise branch-site model. We expect that genomic innovations underlying cold adaptation arose along the notothenioid ancestral lineage. Although some cryonotothenioids such as *D. eleginoides* and *C. esox*, inhabit more temperate regions, by colonizing Patagonia and several sub-Antarctic islands [30, 31], these are nevertheless considered secondary adaptations [32]. Prediction of CEs yielded a total of 675,818 conserved genomic regions accounting for a total of 78,685,553 bp (Additional file 1: Fig. S7A). Most CEs (78%) are

shared by all notothenioids, whereas only 10,182 emerged in the cryonotothenioid ancestor and are shared by all members of the Antarctic clade (cryonotothenioid-specific CEs) (Additional file 1: Fig. S7). Among the cryonotothenioid-specific CEs, most correspond to intergenic (42.9%) and exonic (29.3%) genomic regions. Intergenic, non-coding CEs (CNEs) might play a regulatory role in gene expression. We found that the 1,261 genes associated with at least one cryonotothenioid-specific CNE were enriched for a total of 236 Gene Ontology (GO) terms, most of which are related to signalling, cell differentiation and developmental processes (Additional file 2: Table S11). However, we also identified 120 genes involved in biological functions known to be important for cold adaptation, in processes such as antioxidant activity, lipid metabolism, protein folding, and cytoskeleton organization (Additional file 2: Table S12). Some genes of particular interest including *prdx5* associated with antioxidant activity (Additional file 1: Fig. S7B), *ube2e1* and *ube2h*, involved in protein ubiquitination; *ahsa1* which has chaperone-binding activity, and *tubb4a*, with Antarctic fish tubulins previously having been identified as cold adapted at the amino acid level [33].

For selection analyses we used an extended dataset which includes an additional seven temperate or tropical fishes tagging on the cryonotothenioid stem branch (Additional file 1: Fig. S8, Additional file 2: Table S7). To avoid conflating the signal with lineage-specific evolutionary dynamics, we did not test for diversifying selection in any cryonotothenioid child branch. Furthermore, to establish a stronger link between the identified genes and cold adaptation, we also excluded genes that were found to be under diversifying selection in any other non-cryonotothenioid temperate or tropical fish, based on a control analysis (Additional file 1: Fig. S7). While we acknowledge that this latest approach might increase the false negative rates, excluding genes that might be related to both cold-adaptation and other phenotypic traits due to for example substitution on different codons [34] we prefer to automatically exclude genes that might be subjected to pervasive diversifying selection along multiple branches of the phylogeny. We identified a total of 334 genes under selection on the cryonotothenioid ancestor but not in temperate-adapted species, of which 256 were functionally annotated (Additional file 2: Table S13). A total of 54 GO terms were enriched among these genes (Additional file 2: Table S14). Of these, several are potentially linked to cold adaptation, particularly to oxidative stress (GO:0070301, GO:0010039), lipid metabolism (GO:0006084, GO:1,903,727, GO:0006690) and proteostasis (GO:0051260, GO:0051603, GO:0006509, GO:1,901,984). Example gene include the sirtuin 1 (*sirt1*), peroxiredoxin 3 (*prdx3*), mitochondrial superoxide dismutase 2 (*sod2*), and dihydrolipoamide dehydrogenase (*dld*) genes.

Characterization and evolution of the antifreeze gene locus

The antifreeze gene locus was located on chromosome 16, spanning approximately 699 kb on each haplotype, in a region flanked by the *hsl* and *tomm40* genes (Additional file 2: Table S15). Using a combination of automated and manual approaches, we identified 17 and 16 *afgp* gene copies on Hap1 and Hap2, respectively, as well as two chimeric *afgp/tlp* genes on each haplotype (Fig. 5A). The gene annotation completely matched the *afgp*-specific k-mer signature identified by Klumpy suggesting a complete representation of the locus (Additional file 1: Fig. S9). Repeat coverage of the *afgp* locus was 58.7%

(Hap1) and 59.1% (Hap2), approximately 1.43 times higher than the genome-wide estimation (Fig. 5A).

The typical structure of a functional *afgp* gene encodes for an exon-1 signal peptide and an exon-2 AFGP polyprotein with XLF spacer and AARG terminal motif [4] (Fig. 5B). On Hap1, 14 of the *afgp* genes identified contained the correct structure and were thus considered to be functional. All complete *afgp* copies except one were automatically predicted by GeMoMa, highlighting its strong ability to detect intact *afgp* genes (Additional file 1: Fig. S9). Three more identified copies were putative pseudogenes (Additional file 2: Table S15), two of which miss the exon-1 signal peptide (*afgp1* and *afgp17*) (Fig. 5A). The third copy (*afgp4*) encodes a partial AFGP polyprotein with an unrelated amino acid tail sequence that lacks the canonical AARG terminal motif. The two chimeric *afgp/tlp* genes contain a signal peptide, but while *afgp/tlp1* encodes the XLF spacer and a short TLP segment, *afgp/tlp2* gene lacks the *tlp*-derived exons, encoding an unrelated amino acid tail sequence and lacks the XLF spacer, making it a putative pseudogene (Additional file 1: Fig. S10).

The structure of the *afgp* locus of fHarAnt1.2 is almost identical in both haplotypes, without evidence of major rearrangements. One *afgp* gene copy is missing in Hap2 (*afgp6*) compared to Hap1, but this may be likely due to a mis-assembly resulting in a gap (Fig. 5A). Only insertions and deletions were identified between the two haplotypes, with 48 short variants (<50 bp), involving a total of 428 bp and other 16 with a length from 51 bp to 7,477 bp affecting a total of 32,109 bp. The previously mentioned pseudogenization of *afgp4* on Hap1 resulted from a single-base deletion that shifts the reading frame relative to Hap2, which retains a functional copy. Here, 22 insertions/deletions occur on exon2 of *afgp* or *afgp/tlp* genes affecting a total of 3,27 kb. These variants do not disrupt the reading frame but generate length allelic variants of the AFGP polyprotein. Specifically, we identified five large indels with a length between 51 bp and 1,224 bp affecting a large portion of exon-2 of *afgp7*, *afgp9* and *afgp11* genes, respectively (Fig. 5B). Finally, the chimeric *afgp/tlp2* gene on Hap1 is affected by the largest identified SV within the locus: a large deletion of 7,334 bp that involves all *tlp*-derived exons encoding the TLP mature protein (Fig. 5B).

The locus is characterized by a clear and repetitive organization with 14 *afgp* gene copies found in an area spanning 287,044 bp. The genes and TEs in this region are arranged in a characteristic motif on both haplotypes consisting of two pairs of *afgp* genes with alternating orientation, and a specific set of TE insertions (Fig. 5Aii). The motif is consistently repeated along the region suggesting that it mainly consists of segmental duplications (Fig. 5A). Mapping of ONT, and 10X reads (minimum mapping quality >0, and excluding secondary alignments) from [7] which come from the same individual, revealed sharply reduced coverage within the previously mentioned central region of the locus (Additional file 1: Fig. S11A, B, and C). On average, coverage decreased by 16.8% for ONT reads and by 55.5% for 10X reads. These values reach the 85.4% and 32.8% when only reads with a minimum mapping quality of 20 are considered. Under this threshold 32.9% of the locus show no coverage of 10X reads.

To better understand the evolutionary dynamics and duplication history of *afgp* and *afgp/tlp* genes we inferred a maximum likelihood tree based on the conserved exon-1 and the starting region of intron-1 which is shared between *afgp* and chimeric genes

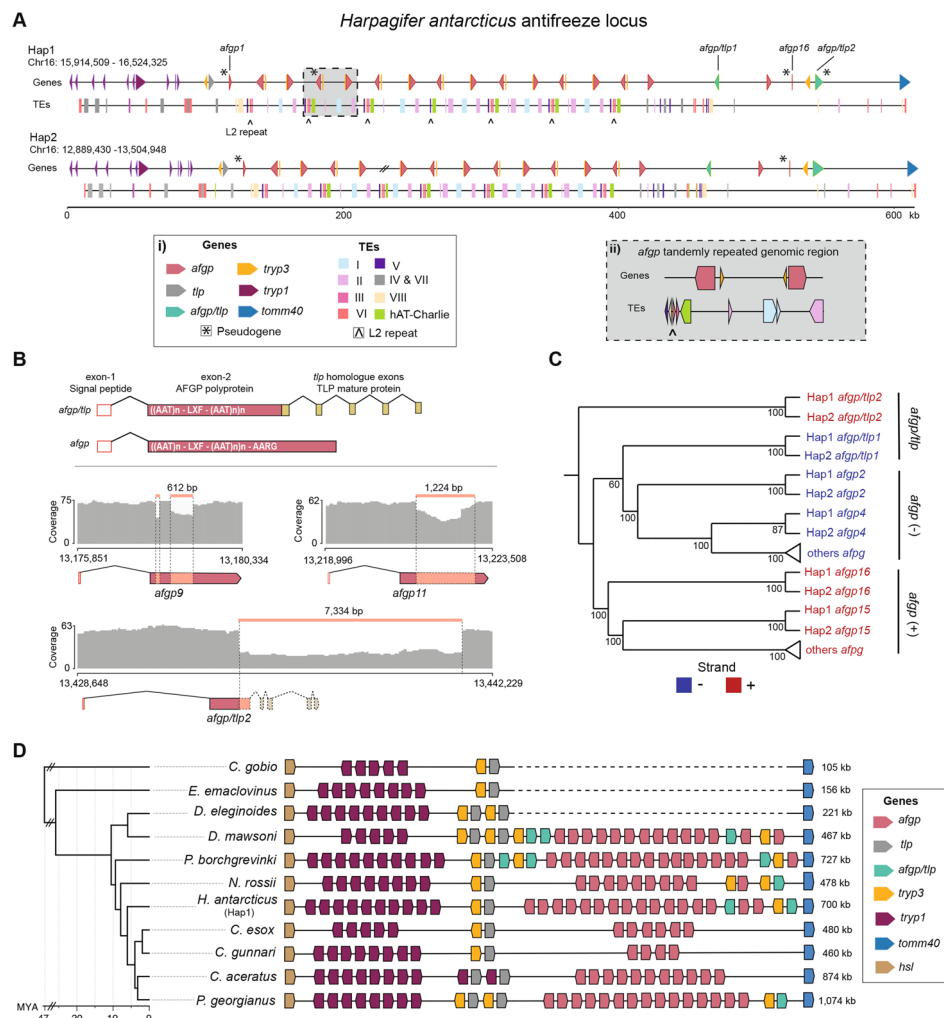


Fig. 5 Structure and evolution of the antifreeze locus. **A** Structure of the antifreeze gene locus on both *H. antarcticus* haplotypes. Triangles represent different genes, and rectangles represent transposable elements. *Afgp* genes are named from 1 to 16 based on their position on the locus. Asterisks highlight pseudogenes and double-slash lines indicate assembly gaps. The grey panel (ii) on the right corner shows a zoomed in version of the tandemly repeated genomic unit marked by the characteristic TE motif. **B** Top: simplified representation of *afgp* and chimeric *afgp/tlp* genes (exon and intron lengths are not proportional to their actual size). Bottom: representative *afgp* and *afgp/tlp* chimeric genes subjected to insertion/deletion events between the two haplotypes. Coverage refers to ONT reads mapped back to assembly. **C** Phylogenetic tree of *afgp* and chimeric *afgp/tlp* gene copies. Different colours correspond to strand ($\pm$). **D** Structure of the *afgp* locus across notothenioids. All genes, including pseudogenised copies are reported. Locus for *P. borchgrevinkii*, *C. esox*, *C. gunnari*, *C. aceratus*, *P. georgianus* and *D. mawsoni* is based on published work [7, 9, 10, 35]. *P. albinpinna* could not be included in this analysis due to the locus being highly fragmented in the corresponding assembly (Additional file 1: Fig. S9)

(Additional file 1: Fig. S12; Additional file 1: Fig. S12). Due to the difficulties in resolving relationships among highly similar *afgp* gene copies, based on few alignment positions (391, of which 56 are parsimony informative) (Additional file 1: Fig. S13A), we additionally performed a ML tree inference excluding chimeric genes (Additional file 1: Fig. S13B). This allowed us to include the full intron-1 and analyse a total of 1,964 positions with 3.2 times more parsimony-informative sites. Because both trees recovered the same two main *afgp* clades, we combined the two topologies by placing the chimeric genes

onto the *afgp*-only tree, preserving their phylogenetic positions (Fig. 5C). The chimeric *afgp/tlp2* genes from both haplotypes form a clade separated by a long branch from all other genes (Additional file 1: Fig. S13A). Considering that *afgps* likely originated from a chimeric gene [4], *afgp/tlp2* probably represents the most ancestral gene and was therefore used to root the tree. All other analysed genes form two clades that perfectly separate genes on the plus strand from those on the minus strand, including the chimeric *afgp/tlp1*, which is also the first to diverge within its clade.

To have a broader overview of the locus across cryonotothenioids we also annotated the locus in additional species. We used an automated re-annotation process of the *afgp* locus by employing GeMoMa and Klumpy, followed by visual inspection of candidate genes. To benchmark this approach we used the *afgp* locus of species *P. georgianus* which had previously been manually annotated, which we then re-annotated through the automated process. The automated approach identified the same number of chimeric *afgp/tlp* (1) and *afgp* (15) genes in *P. georgianus*, organized in the same structure as described by [7]. We then auto-annotated two additional species: *P. albipinna* and *N. rossii* (Fig. 5D; Additional file 1: Fig. S9). In *P. albipinna* the locus was distributed across different unlocalized scaffolds despite the use of long-read sequencing technologies and Hi-C data (Additional file 1: Fig. S9). In *N. rossii* we detected two chimeric genes and 10 *afgp* genes. We performed comparison of the structure of the locus along a number of species including our new and published annotations for other included species (Fig. 5D). Overall we detected some notable similarities between species *C. esox*, *C. gunnari*, *H. antarcticus*, and *N. rossii*. Specifically, in all these species the majority of *afgp* genes are arranged in arrays with alternating orientation. However, each pair of *afgp* genes lacks the previously described TE motif that was identified in the *H. antarcticus afgp* locus.

Discussion

Haplotype-resolved assembly of the Antarctic spiny plunderfish genome reveals low haplotypic structural variation

Here, we present the first haplotype-resolved genome assembly of a notothenioid species and one of the most contiguous notothenioid genome assemblies to date. The assembly was generated with ONT reads which supported assembly more successfully compared to sequencing attempts with HiFi data. The size of the two assembled haplotypes was 1,069–1,222 Mb in Hap1 and Hap2 respectively, with this size variation corresponding to higher number of unplaced contigs in Hap2 (230 Mb) compared to Hap1 (67 Mb). The size variation between assembled haplotypes could be due to assembly difficulties in regions rich in tandem repeats (e.g., satellite repeats HarAnt-Sat1), which could be asymmetrically split during Hifiasm assembly [36], though biological variation cannot be excluded. The two haplotypes exhibit low levels of divergence and a relatively small number of structural variations, which is consistent with the low *k*-mer based and SNP based heterozygosity (0.37% and 0.15%, respectively) of the *H. antarcticus* genome (Additional file 2: Table S8; Additional file 1: Fig. S4). Few exons and genes were found to be absent from one of the two haplotypes, in contrast to findings in other metazoans, where highly divergent and hemizygous genomic regions can represent more than 50% of a diploid genome [37, 38].

Complex SVs, such as translocations, duplications, and inversions were found to colocalize towards the repeat-rich and gene-poor chromosome ends (Fig. 2A, Additional file 1: Fig. S4). This was even more so in terms of tandem repeat content, consistent with a model of elevated mutation rates in sub-telomeric genomic regions [39]. Accumulation of repeats towards chromosome ends was also observed in another cryonotothenioid species, *T. borchgrevinki* [35], so we suggest that this might be a plesiomorphic feature of the clade. Furthermore, recent TE insertions, most notably LINE-L2 elements, were found to significantly contribute to the emergence of insertions and deletions between haplotypes across intergenic, intronic, and gene-flanking regions. Moreover, young heterozygous LINE-L2 insertions are markedly depleted in introns and enriched in intergenic regions compared to those shared between haplotypes. This pattern mirrors observations in zebrafish, where gene bodies accumulate mainly old LINE insertions [40]. In few cases, insertions within the exons of protein-coding genes were detected, which have led to the truncation and putative pseudogenization of the affected genes on one of the haplotypes. Overall, our results suggest that despite the low levels of overall haplotypic structural variability, recently active TEs likely represent important contributors in driving population-level differentiation in *H. antarcticus*, promoting gene-loss and reshaping *cis*- and *trans*- regulatory elements.

ZP genes were suggested to be subjected to population-level variability in the icefish *Chionodraco hamatus* [41]. This gene family underwent neofunctionalization and lineage-specific duplications in cryonotothenioids as well as in other cold-adapted teleost fish lineages, promoting egg survival during early developmental stages in freezing conditions [5]. Our phased assembly revealed a high number of ZP genes in *H. antarcticus*, similarly to other cold-adapted notothenioids (Additional file 1: Fig. S6) with multiple putatively functional copies in seven out of the 11 types previously described, including ZPAX1 and ZPC5, which possess the highest ice melt-promoting activity [42, 43]. However, both haplotypes encode a comparable number of genes across all ZP types (Additional file 1: Fig. S6) suggesting that population-level copy number variability might be limited in this species.

Recent accumulation of ancient LINE-L2 elements in *H. antarcticus*

Different TE clades can have varying evolutionary trajectories across species, reflecting host–element co-evolution, lineage-specific regulatory constraints, and variation in genomic defence mechanisms [44, 45]. LINE-L2 transposons were previously shown to be particularly abundant in *H. antarcticus* compared to other notothenioids, covering about 8% of the genome [7]. Analysis of the new genome assembly similarly showed high LINE-L2 levels deriving mainly from the activity of eight different LINE-L2 families. Both the high number of heterozygous LINE-L2 insertions and the repeat-divergence landscape analyses suggest a recent, lineage-specific accumulation of these elements (Fig. 2). Eight evolutionary distinct families of LINE-L2 elements were identified, with concurrent activity within the last four million years (MY) (Fig. 2D). Moreover, most of these LINE-L2 families, including the most abundant ones, are also shared with sister species of the cryonotothenioid clade, suggesting their ancient emergence. Despite the recent activity of LINE-L2 elements, we observe a negative effect of their genomic content around breakpoint regions and occurrence of inversions, duplications, and

translocations between haplotypes suggesting that their accumulation has not resulted in the emergence of other more complex structural variants.

In another LINE transposon superfamily, the LINE-L1 retrotransposons, it has been shown that either a single lineage was usually active at high replication rate at any given time in mammals, or multiple, ancient LINE-L1 lineages were concurrently active in fishes, reptiles, and bivalves [45, 46]. The evolutionary path of LINE-L2s in notothenioids appear therefore to follow the generally observed evolution of LINE-L1 of fishes.

Rampant genomic gain and low levels of genomic deletion during cryonotothenioid diversification

TE-mediated genome expansion has been shown to predate the diversification of extant cryonotothenioids and that enhanced TE activity continued in more derived clades such as icefishes [7, 35]. To further test this hypothesis, we have calibrated into absolute time TE landscape profiles and quantified genomic gains through TEs along the notothenioid ancestral branches using a comparative genomic approach (Fig. 4A). Despite the strong assumptions and inherent uncertainties associated with dating TE landscapes, such as relying on neutral substitution rates assumed to be constant throughout the genome, and assuming that consensus sequences represent ancestral repeats, both methods suggest that most TE activity occurred recently, concurrently with the recent cryonotothenioid diversification. A single massive burst of TE activity was also detected in repeat profiles of other cryonotothenioids [7, 16, 35]. Low levels of ancient TE activity followed by recent peaks could reflect either genuinely increased transposition rates or extensive genomic deletions that removed older insertions, thereby erasing evidence of ancient bursts. However, net genome-gain estimates reveal low deletion rates during notothenioid evolution, supporting the former scenario: a slow accumulation of TEs before the emergence of extant cryonotothenioids, followed by a more recent increased rate of TE accumulation. Our reconstruction also revealed an ancient burst of genome expansion in the ancestor of icefishes explaining their exceptionally large genomes, which may be explained by sea temperature history in the region [7].

Paleoclimatic based temperature calculations (Fig. 4C) indicate an initial temperature drop at approximately 16 MYA which predates the emergence time of the extant cryonotothenioid clade, and a second, stronger reduction around 3 MYA which could have contributed to their recent diversification [7], similarly to other marine clades [47]. Here we show that the timing of these temperature variations appears to coincide with our reconstructions of TE levels of activity (Fig. 4C). LINE elements were found to be particularly expanded not only in *H. antarcticus* but also in other cryonotothenioids (Fig. 2E). The presence of environment induced stressors, such as fluctuations of temperature, have been linked to increased TE activity due to likely disrupting mechanisms of epigenetic regulation [48]. In the cryonotothenioid species *D. mawsoni*, it has been shown that the levels of LINE element transposition are increased through cold induced cell transfection [49]. We provide evidence that compared to other TEs, LINES have been particularly active along the cryonotothenioid ancestor and starting from 4 MYA, concurrently with the two main temperature drops documented in the region (Fig. 4C). Overall, our comparative genomics analysis suggests that the

increased genome size of cryonotothenioids may have occurred through a two-step process: one ancient, slow TE activity in the absence of major genomic deletions that predate the emergence of extant species and one more recent, lineage-specific massive bursts of transposition due to loss of epigenetic control.

Macrosyntenic relationships and chromosomal fusions in notothenioid genomes

Our synteny analyses of 12 chromosome-scale notothenioid genomes demonstrated conserved macrosyntenic relationships between notothenioid chromosomes in accordance with the presence of a stable karyotype with low structural divergence during notothenioid radiation [28]. In the Nototheniidae, the most common chromosome number is $n=24$, mainly comprising acrocentric chromosomes. Indeed *H. antarcticus* has a karyotype of $n=24$ comprising 21 acrocentric, 1 metacentric, and 2 submetacentric chromosomes [28]. Lineage-specific Robertsonian chromosomal fusions have occurred in some notothenioids [50], with the most notable case being *N. rossii* where only 12 chromosomes ($n=12$) are observed (Fig. 4B). The few other identified smaller scale rearrangements tended to colocalize towards repeat-rich chromosome ends, similar to haplotypic structural variants identified in the *H. antarcticus* genome. These genomic regions are therefore thought to be subjected to higher rates of structural variation compared to the rest of the chromosomes, with repetitive elements that may drive this variability. Furthermore, the existence of LTR retrotransposon–insertion hotspots near chromosome ends and at chromosome fusion points, suggests that such repetitive sequences may have facilitated Robertsonian fusions, as has occurred in other more distant species [17, 51]. In cryonotothenioids, reduction in chromosome number from the ancestral karyotype of $n=24$, caused by multiple independent chromosomal fusions, appears to be common within specific subfamilies, with such cases observed in the genera *Trematomus*, *Notothenia*, and *Pogonophryne* [25].

Tandem segmental duplications and haplotype length variants shape the structure of the antifreeze locus

The antifreeze locus is a very challenging genomic region to assemble both because of its repetitive structure due to the presence of tandemly duplicated gene families such as the *afgps*, and trypsinogens, as well as due to its high TE content [7]. Using ONT reads we were able to assemble and annotate the complete *afgp* locus for *H. antarcticus* for both distinct haplotypes. The locus comprises 17 *afgp* copies (Hap1) and a higher repeat content compared to the genome average (59% vs 43%). Two copies of the chimeric *afgp/tlp* gene were identified on each haplotype (Fig. 5A). Of these, one copy on Hap1 and both copies on Hap2 exhibit a canonical structure in accordance with previously described genes [4, 8, 9] and are considered to be functional copies. These chimeric genes were previously considered to be putative evolutionary intermediates of the *afgp* genes [8]. Annotation of the locus on members of the icefish family showed either lack (*C. gunnari* and *C. aceratus*) or pseudogenisation (*P. georgianus*) of chimeric gene copies, suggesting that chimeric *afgp/tlp* genes might have been lost in low latitudes of the Southern Ocean by relaxed selection due to redundancy function with *afgp* genes [7, 10]. However, chimeric genes have been retained in *H. antarcticus* which has a partially

overlapping distribution range with these species, leading us to question the drivers of retaining these genes for this species.

BAC library-based analyses of the *afgp* locus of *D. mawsoni* suggested the presence of massive variation between haplotypes in both the overall size of the locus, and in the number of *afgp* gene copies [9]. Tandem arrays are generally expected to be frequently subjected to copy number variation due to recombination between homologue chromosomal regions which could also promote high levels of haplotypic variation [52]. Through this haplotype resolved assembly for *H. antarcticus* we were able to directly investigate the potential presence of haplotypic divergence in the locus. Nevertheless, we found that in *H. antarcticus*, both haplotypes exhibit almost identical structure and gene copy number. The main structural variations identified were pseudogenizations of one *afgp* gene due to a single base-pair deletion and of one *afgp/tlp* gene due to large deletion of all *tlp* derived exons on Hap1.

Despite low levels of variation at gene copy and functionality levels, we also identified allelic variation of the *afgp* gene copies, with regards to the actual length of the AFGP polyprotein coding exons. Potential variation on the length of the AFGP polyprotein could have important biological implications. The notothenioid AFGP proteins function through an “adsorption–inhibition” mechanism [53], adsorbing to the surface of ice and preventing the growth of the crystals as a function of glycoprotein concentration, size, and shape [54, 55]. It has been observed that smaller protein molecules are less effective than large ones to interfere with ice crystal growth [54]. A potential scenario that could be further examined in future analysis would be the effect of these size variations in the AFGP polyproteins to the adaptive potential of the species in *H. antarcticus* populations and how this could contribute to local adaptation in more extreme conditions. Length allelic variants have also been observed in other highly repetitive gene families, such as silk genes in caddisflies, butterflies, and spiders [56], where sequence length was linked to the properties of the silk fibres [57]. Such types of allelic variants may represent a convergence source of variation in proteins composed of repeated motifs across all metazoans. In these cases we could consider that because the gene itself provides a substrate for recombination, and the repetitive structure can weaken purifying selection against insertions and deletions, these could create a standing pool of genetic variation on which positive selection can eventually act.

In the case of the *afgp* locus, the characteristic TE plus *afgp* gene motif that was detected in *H. antarcticus* likely supports a specific mode of expansion of the locus. The *afgp* copies are arranged in opposite orientation separated into two strand-specific clusters (Fig. 5A, C), suggesting that the expansion of the genes occurred through tandem segmental duplications, which duplicated as two pairs of genes at a time. Furthermore, the high sequence homology identity across duplicated regions and the short branch lengths separating gene copies in phylogenetic analyses (Fig. 5, Additional file 1: Fig. S13) suggest that these duplications are either evolutionarily recent or have been strongly homogenized by concerted evolution.

After their emergence in the cryonotothenioid ancestor, reduction and loss of *afgp* genes has occurred multiple independent times during their radiation [58]. However, the exact time at which *afgp* genes begun duplicating and expanding remains an open question. Comparison of the *afgp* locus structure across multiple notothenioid genomes

(Fig. 5D) showed a similar genomic configuration of some genomic features in *P. borchgrevinki*, *H. antarcticus*, *C. esox*, *C. gunnari*, and *N. rossii*. Nevertheless, the distinct TE motif appears to be specific to the *H. antarcticus* genome (Fig. 5Aii). We hypothesize that the tandemly duplicated genomic region marked by the characteristic TE motif of *H. antarcticus* represents a derived, apomorphic feature of this lineage, with different rates of duplication and loss driving variation of *afgp* gene copy number. We therefore suggest that from a hypothetical cryonotothenioid ancestral locus with few *afgp* genes, independent tandem duplications increased gene-copy number in different lineages.

Diversifying selection and origin of novel conserved elements in cold-adaptive pathways during cryonotothenioid emergence

To detect additional genomic features underlying cold adaptation in notothenioids we searched for both protein-coding genes that experienced diversifying selection and CNEs that first emerged along their ancestral branch. Among the genes showing evidence of diversifying selection in the cryonotothenioid ancestor, many were found to be associated with biological processes known to be important for survival in extreme cold conditions (Additional file 2: Table S14). We found that most CNEs were shared across all notothenioid species, suggesting that a large portion of the genome expansion characterizing the cryonotothenioid ancestral lineage may have evolved neutrally. Additionally, most cryonotothenioid-specific CNEs were found to be associated with genes involved in signalling, cell differentiation, and developmental processes (Additional file 2: Table S11). CNEs are known to have non-random distribution within the genome, because they have the tendency to cluster near genes with regulatory functions in multicellular development and differentiation. A subset of the genes which we have found to contain newly acquired CNEs are related to two biological processes which are overrepresented among genes under diversifying selection. These biological processes are antioxidant activity and proteostasis and they have been linked to evolutionary cold adaptation [59, 60]. We suggest that because these genomic innovations arose in the cryonotothenioid ancestor, they could be amongst the key components of these pathways that enabled this clade to survive and become established in the Southern Ocean. Additional analysis, and experimental work could be done in the future to further verify the implications of the genes to cold adaptation in notothenioids.

Antioxidant activity and proteostasis

Life in freezing sea water poses particular biological problems for the organisms living there. The amount of oxygen increases considerably with reduced sea water temperatures, producing naturally high levels of cellular oxidative stress for Antarctic marine species [59, 61]. In addition, cold denaturation of proteins impacts proteostasis [60]. Therefore, in order to survive in this harsh environment, organisms need to modify their biochemical processes.

Our analysis revealed four key genes related to oxidative stress that underwent diversifying selection in the cryonotothenioid ancestral lineage, including the sirtuin 1 (*sirt1*), peroxiredoxin 3 (*prdx3*), mitochondrial superoxide dismutase 2 (*sod2*), and dihydrolipoamide dehydrogenase (*dld*) genes. Moreover, we found that the peroxiredoxin

5 (*prdx5*) gene acquired novel CNEs concurrently with extant cryonotothenioid emergence. Gene *sirt1* is a class III histone deacetylase that is recruited in the promoter region of multiple antioxidant genes, including *sod2*, *prdx3*, and *prdx5*, promoting their expression and their protein concentration [62]. These three genes encode mitochondrial enzymes that act in concert to control reactive oxygen species (ROS) levels. *Sod2* converts superoxide radicals into hydrogen peroxide, which is then detoxified to water by *prdx3* and *prdx5* [63]. The interplay between protein modifications and the emergence of novel, potentially regulatory CNEs within this pathway may have strengthened the antioxidant capacity of the cryonotothenioid ancestor, enabling it to cope with elevated oxidative stress in the mitochondria.

Signs of diversifying selection have been found [64] also in other cold-adapted notothenioids acting on genes involved in antioxidant activity, such as in *N. coriiceps* [65] and *D. mawsoni* [16], and across species. Moreover, genes such *sod2* and *prdx5*, which are involved in ROS scavenging, have been found to be overexpressed in cold-adapted notothenioids compared to their temperate-adapted relatives [66]. Superoxide dismutase appears to be particularly important in combatting oxidative stress in Antarctic species. Gene duplications used to enhance the cellular activity of this enzyme have been identified in Antarctic invertebrates [67]. Furthermore, evolution to function efficiently in the cold have resulted in at least some members of this gene family being particularly thermally sensitive in some Antarctic marine invertebrates [68]. This would suggest that specific protein modifications might help to mitigate cellular oxidative stress in Antarctic species.

In terms of proteostasis, there is evidence that Antarctic species, including fish, face challenges in producing, folding and maintaining proteins at such low temperatures [60]. Whole animal RNA to protein ratios, ubiquitylation rates, and levels of chaperone proteins are much higher in Antarctic compared to temperate species [69–71]. Some of this activity is due to a mutation in the promotor region of the classical inducible form of *hsp70* resulting in constitutive expression and cold-enhanced activity of the proteasome in Antarctic notothenioids [72, 73]. In our analyses on genes that were found to be under diversifying selection, we identified genes involved in protein folding and degradation, such as members of heat shock protein family member (*hspa4*), ubiquilin/ubiquitin proteins (*ubqln4* and *ube2z*) and the chaperon *psmc1*, similar to genes with acquired CNEs, such as the ubiquitin *ube2h*, *ube2e1*, *itch* and the co-chaperon *ahsa1*. Furthermore, there was evidence of selection pressures on genes involved in transcription, splicing, and the proteasome (e.g. *eif3b*, *sf3a2* and *psmb10*) indicating wider scale cold adaptation of the proteostasis pathway.

It should be noted that these examples are likely minimal identifications due to the stringency of orthologue assignment, which may not accurately identify duplicated genes, the number of which are considerable in the notothenioids [66]. Such analyses will also fail to identify those genes which are poorly conserved between species, such as genes involved in the immune function, intrinsically disordered proteins and those containing intrinsically disordered regions.

Conclusions

This haplotype-resolved genome assembly of *H. antarcticus* thus provides a valuable resource for future functional investigations into cold adaptation and resilience to climate change. Overall, haplotype diversity was low and contrary to expectations haplotypic diversity in the *afgp* locus was also low despite a high number of transposon insertions. However, we did identify sequence length variation in the *afgp* genes between the two haplotypes, which could be biologically relevant. Furthermore, we show that genomic gain was linked to transposon activity which correlates with drops in temperature in the Southern Ocean, and this activity resulted in the generation of novel genomic elements. Finally, we show that there has been significant selection pressure on notothenioid genomes, both in genes and control regions, to enable these fish to thrive for much of their life below 0 °C.

Methods

Sequencing data generation

The specimen of *Harpagifer antarcticus* that was used for genome assembly was collected at Ryder Bay in the Antarctic peninsula (coordinates: −67,59,053 −68,28,863), and was preserved by flash freezing. This specimen was the same that was used for the generation of the fHarAnt1.1 assembly (accession: GCA_902827135.1; BioSample: SAMEA104132831) [7].

High Molecular Weight (HMW) DNA was extracted using the Sanger Tree of Life Manual MagAttract v1 protocol [74] in the Tree of Life Core Laboratory at the Wellcome Sanger Institute. To assess the quantity of the extracted DNA the Qubit™ 1X dsDNA High Sensitivity assay (ThermoFisher, Cat No. Q33231) on the Qubit Flex Fluorometer (ThermoFisher, Cat No. Q33327) was used. The quality of the DNA was then determined using the Femto Pulse (Agilent, PN: M5330AA) instrument with the gDNA 165 kb Analysis kit 275 samples (Agilent, Cat No. FP-1002–0275). Approximately 3 µg of the extracted DNA underwent library preparation for ONT sequencing using the Ligation Sequencing Kit V14 (SQK-LSK114). Sequencing was performed on one flow cell on the PromethION 24 in Sequencing Operations at the Wellcome Sanger Institute.

Three *H. antarcticus* specimens were Schedule 1 killed and the gills rapidly dissected, added to tissue culture media (L15 without phenol red, 10% foetal bovine serum and antibiotics) and placed on ice. The gills were chopped up in culture media and the resultant cell suspensions of each sample added to individual wells in 6-well Nunclon™ tissue culture plates. For each fish, one of each of the gill samples was incubated at 0 °C with the paired sample incubated at 6 °C for 6 days. Cells were harvested and RNA extracted using TRI Reagent according to manufacturer's instructions. 1 µg of rRNA depleted RNA was used for RNAseq library preps with Kapa Hyperprep UDI kit (Cat#KK8544). Paired end sequencing was carried out by Novogene on the Novaseq PE150 platform.

Genome assembly

The ONT reads were sequenced at an estimated 23.4X coverage per haplotype and based called with dorado 7.2.13 (<https://github.com/nanoporetech/dorado>) in super accurate mode. The reads were assembled using Hifiasm version 0.24.0-r702 [75] with --ont option in Hi-C phasing mode. This produced a phased assembly with contig N50

of 19.3 Mb and 14.8 Mb in each haplotype. The sizes of both haplotypes were 1,072 Gb and 1,222 Gb which were close to the kmer-based estimate from the GenomeScope at 1,152 Gb.

Furthermore Hi-C reads were mapped to each haplotype independently using bwa-mem2 [76]. Each haplotype was then scaffolded with Hi-C data in YaHS [77], using the --break option for handling potential mis-assemblies in contigs. The produced assembly was evaluated for kmer completeness using MerquyFK (<https://github.com/thejenemyers/MERQUERY.FK>) using kmers from the ONT reads. This indicated 99.39% kmer completeness and QV value 42.6.

The individual haplotype assemblies were screened for contamination using the ASCC pipeline (<https://github.com/sanger-tol/ascc>) removing 5.8 Mb Hap1 and 1.8 Mb Hap2 non-target and mitochondrial sequences. The assemblies were then combined and prepared for curation using the treeval pipeline (<https://github.com/sanger-tol/treeval>) generating a Hi-C contact map with supplementary analysis tracks (telomeres, gaps and long read coverage). Assembly errors were corrected by manipulating the contact map in PretextView (<https://github.com/sanger-tol/PretextView>) an AGP of the resolved chromosomal haplotypes were exported and a curated fasta for each haplotype generated using pretext-to-asm (<https://github.com/sanger-tol/agp-tpf-utils>) with each super scaffold named according to descending size. Contact maps for each resolved haplotype were generated by curationpretext (<https://github.com/sanger-tol/curationpretext>) and images generated with PretextViewSnapshot (<https://github.com/sanger-tol/PretextViewSnapshot>).

Repeat annotation

Repetitive elements were *de-novo* mined with RepeatModeler v.2.0.5 [78] adding the *-LTRStruct* extension to improve LTR detection. Potential multicopy host genes included in the repetitive library were removed with ProtExcluder v.1.2 after blasting (Blastx, e-value 1e-10) the repetitive library against a protein database consisting of all vertebrate proteins included in the Swiss-Prot database (<https://www.uniprot.org/>) and the predicted proteomes of five published notothenioid genomes available on NCBI RefSeq: *C. gobio*, *P.georgianus*, *Gymnodraco acuticeps*, *Trematomus bernacchii*, *Eleginops maclovinus*. Furthermore, a TE library was made with MCHelper v.1.7.0 [79] with default parameters. Both haplotypes were annotated with RepeatMasker in sensitive mode (-s). RepeatMasker derived annotation was post-processed with: (a) parseRM.pl script (<https://github.com/4ureliek/Parsing-RepeatMasker-Outputs/tree/master>) to summarize the TE content and obtain TE landscapes describing transposon activity through relative time, and (b) the RM2Bed.py script from RepeatMasker to resolve overlaps in the TE annotation (--overlap resolution 'higher score') and produce bed files describing the coordinates of TE insertion % of CpG corrected Kimura. Divergence of TE landscapes were translated in absolute time using a neutral substitution rate of 3.2E-3 as obtained from divergence time analyses (see [Methods: Species tree reconstruction and divergence time estimation](#)).

Tandem repeats were identified with the Centrominer function of quarTeT v.1.2.5 [80]. rDNA genes were annotated with barrnap v.0.9 [81] <https://github.com/tseemann/barrnap>

ap and telomeric repeats with tiddk v0.2.31 [82] screening the assembled chromosomes for the TTAGGG telomeric motif.

Gene annotation

Gene annotation of Hap1 was performed with BRAKER3 [83]. Repetitive elements were soft masked and both protein and transcriptomic data were used as external evidence to train AUGUSTUS [84] and GeneMark-ETP [85] *ab-initio* gene predictors. Illumina RNAseq reads from muscle, skin, liver, kidney, heart, and brain [6] (Additional file 2: Table S16) and from gills (data generated herein) were mapped with HISAT2 [86]. The sorted bam files were supplied to BRAKER3 together with the protein database previously used for cleaning the repetitive libraries from host genes. The quality of the annotation was assessed with BUSCO v.5 [87]. To obtain a gene annotation for Hap2, we mapped BRAKER-derived genes with Liftoff [88] (-polish -copies) to obtain their lifted coordinates. Functional annotation was performed on the Hap1 predicted proteome with eggNOG-mapper v2 [89] transferring only annotations with experimental evidence and auto-adjusting the taxonomic scope based on the query sequence.

Zona Pellucida (ZP) genes were extracted from the predicted proteomes of both *H. antarcticus* haplotypes by identifying all proteins containing a ZP domain (IPR001507) following annotation with InterProScan v5.47–82.0 [90]. We further confirmed missing genes by aligning *D. mawsoni* ZP proteins to the *H. antarcticus* haplotypes using mini-prot v0.18-r281 [91].

Annotation of the *afgp* locus

To annotate the *afgp* locus on each haplotype, a combination of automated and manual approaches was used. The annotation of the *D. mawsoni afgp* locus (accession HQ447059.1, haplotype 1) [9] was used as a reference for homology-based gene prediction with GeMoMa v.1.7.1 [92] under default parameters. This initial annotation was supplemented by manually blasting *D. mawsoni* gene sequences with BLASTN v.2.12.0 [93] and reconstructing the gene structure based on alignment coordinates. Each identified *afgp* copy was then manually inspected for frameshifts and gaps to identify complete and pseudogenised gene copies. Finally, we used Klumpy [94] with the *find_klumps* function to verify the completeness of the annotation, using the exon-1 and exon-2 of one randomly chosen functional *afgp* gene copy previously annotated as query.

To infer evolutionary relationships of *afgp* genes, we used the short exon-1 (signal peptide) and intron-1 nucleotide sequences, excluding the highly repetitive exon-2 encoding the AFGP polyprotein, following the Nicodemus-Johnson et al. [9] approach. Analysis was performed both with and without the chimeric *afgp/tlp* genes. Sequences were aligned with MAFFT v.7.525 [95] using the E-INS-i. Because *afgp* and *afgp/tlp* genes share only exon-1 and an initial part of intron-1, when these latest genes were included in the analyses, we removed the non-homologue region based on visual inspection of the alignment. Phylogenetic inference was performed with IQ-TREE 2 [96] with 1,000 ultra-fast bootstrap replicates [97], to assess nodal support, selecting the best-fit evolutionary model with ModelFinder [98].

Characterization of LINE-L2 elements

LINE-L2 insertions annotated on Hap1 and longer than 1,000 bp were extracted with bedtools *getfasta* and subjected to Blastx (-evalue 1e-05 -max_target_seqs 1 -max_hsps 1) against all reverse transcriptase (RT) amino acid sequences that compose the seed alignment of the Reverse transcriptase PFAM [99] profile PF00078. Based on the BLASTX v.2.12.0 [93] alignment coordinates, the amino acid sequences of the RT segments were extracted, discarding those shorter than 100 amino acids. To decrease the size of the dataset for the subsequent phylogenetic analyses all RT segments were clustered at 80% sequence identity with CD-HIT [100] keeping the longest element as representative of the cluster. All representative RT segments together with reference sequences of the LINE-L2, L2A, L2b, Crack, Daphne and CR1 clades described in [26] were aligned with MAFFT in G-INS-i mode and trimmed with TrimAl v.1.4.1 [101] in *gappyout* mode. Phylogenetic inference was performed with IQ-TREE v2.4.0 [96] with 1,000 UltraFast bootstrap replicates to assess nodal support and selecting the best-fit evolutionary model with ModelFinder.

CD-HIT [100] clustering identified eight clusters, each comprising between 600 and 3,669 RTs, together accounting for 73% of all extracted RTs longer than 100 amino acids. The complete insertion corresponding to the representative element of each one of these clusters was used as query in a “Blast-Extend-Extract” analyses [102–104] to reconstruct a consensus sequence representative of the family. The curated consensi were used in an additional RepeatMasker analyses, and to reconstruct TE landscapes calibrated in absolute time as previously described for all repeats. We finally looked for the identified and reconstructed LINE-L2 family across genomes (Additional file 2: Table S7) through BLASTN searches requiring an alignment length of at least 500 nucleotides with at least 80% of identity to the query consensus sequence.

Haplotype comparison

A whole genome alignment of the two haplotypes was performed with Minimap2 v.2.26 [105] and SNPs, SVs (insertions, deletions, and duplications), and rearranged genomic regions (inversions and translocations) were inferred between the two haplotypes using Syri v.1.7.1 [106]. In each case Hap1 was used as the reference. We identified genes and exons subject to structural variation between the two haplotypes by intersecting the gene annotation with Syri-derived SV and rearrangement coordinates using bedtools v2.30 [107] *intersect*. We tested whether the genomic distribution of SVs, considering only translocations, duplications and inversions, was associated with specific genomic features using a generalized linear mixed model (GLMM) with a binomial error structure. SVs were compared to an equal number of randomized genomic intervals generated with bedtools *shuffle*. For each interval we quantified the following: distance to the nearest chromosome end, distance to the closest assembly gap, length of the SV, and the proportion of bases within 50 kb flanking windows annotated as TEs, tandem repeats, or LINE elements. All predictors were standardized prior to analysis. The model included chromosome as a random intercept to account for non-independence of intervals on the same chromosome. GLMMs were fitted in R using the lme4 package [108]. Odds ratios and 95% confidence intervals were obtained by exponentiating model coefficients.

To identify putative heterozygous TE insertions, we overlapped the RepeatMasker annotations of Hap1 and Hap2 with deletions and insertion events, respectively, requiring that both insertions/deletions and the TE annotation overlap by at least 75% of their length.

To assess whether the SVs identified between haplotypes might result from assembly errors, we additionally mapped ONT reads back to the assembly with Minimap2 and called heterozygous SVs with Sniffles2 [109]. We then quantified the overlap between the genome-based and read-based heterozygous SVs larger than 50 bp using SURVIVOR [110]. Overlapping SVs genotyped as homozygous for the alternate allele were considered as putative mis-assemblies.

Species tree reconstruction and divergence time estimation

Divergence times were estimated across a set of genomes of notothenioids, including our newly generated *H. antarcticus* assembly (Hap1), and 11 publicly available chromosomal assemblies: *Cottoperca gobio* (*Cottoperca trigloides*) [24], *Notothenia rossii* [111], *Pseudochaenichthys georgianus* [7], *Eleginops maclovinus* [112], *Dissostichus mawsoni* [113], *Dissostichus eleginoides*, *Chaenocephalus aceratus* [114], *Pagothenia borchgrevinki* (*Trematomus borchgrevinki*) [115], *Champsocephalus gunnari*, *Champsocephalus esox* [10], *Pogonophryne albipinna* [116] (Additional file 2: Table S7). Additionally another seven marine temperate and tropical fishes' representative of four, closely related orders (Additional file 2: Table S7).

First, we manually constructed a topology of the included species following the phylogenetic relationships inferred in [7, 10]. Branch lengths were then estimated using the nucleotide sequences of orthologous single-copy BUSCO genes extracted with BUSCO v.5.7.1 [87] and the actinopterygii_odb10 reference dataset. Briefly, single-copy genes present in all analysed species, were separately aligned using MAFFT [95] with the E-INS-i algorithm. Each alignment was cleaned of ambiguous positions using TrimAl in automated mode, and sequences sharing only a reduced region with the others in the alignment were removed using the parameters -resoverlap 0.8 and -seqoverlap 75. Finally, only alignments with at least 500 positions and for which no sequences were excluded by TrimAl were retained. To reduce computational time, 1,000 alignments were randomly selected, concatenated, and subjected to constrained phylogenetic inference with IQ-TREE v2.4.0, selecting the best-fit evolutionary model with ModelFinder. The resulting tree was dated with LSD2 [117] under a fixed substitution rate using [7] estimates as priors of the following nodes: 1) split between *Takifugu rubripes* and *Nelussetta ayraudi* from all other species: 94 MYA, 2) notothenioid emergence: 47 MYA, and 3) cryonotothenioid diversification: 10.7 MYA. A neutral substitution rate of notothenioids was obtained by re-running the same analyses on the notothenioid-only subtree and using fourfold degenerate sites extracted from the BUSCO concatenated alignment.

Selection analyses

Publicly available predicted proteomes of fish species included in divergence time estimation were clustered into orthogroups (OG) with OrthoFinder v2.5.4 [118] in *ultra-sensitive* mode. We then isolated OGs in which at least one gene for all species was present. Multi-copy OGs (i.e. more than one gene was present for at least one species) were decomposed into single-copy OGs using DISCO [119]. Decomposed and original

single-copy OGs were then aligned with Prank v250331 [120] under a codon model, cleaned of ambiguous positions with Gblocks taking into consideration the codon structure [121] and from sequences with high percentage of gappy positions with Trimal (-resoverlap 0.6 -seqoverlap 55) [101]. After filtering, only OGs for which all species were present were kept (i.e. not affected by Trimal filtering) and with an alignment longer than 300 codon positions. IQ-TREE2 v2.4.0 was used to infer gene trees under a GTR+I+G substitution model. Because our aim was to detect genes that might be strictly related to cold adaptation in extant cryonotothenioids, we used aBSREL from the HyPhy package v2.5.8 [122] testing for diversifying selection acting on their stem branch. Prior to this, to ensure that we were testing the same evolutionary event we discarded any gene trees in which cryonotothenioids were not monophyletic and not in a sister relationship with their true temperate notothenioid sister species *E. maclovinus* using a custom R script. Finally, because aBSREL analyses does not test that selection has not occurred outside tested branches we re-ran it for all genes previously identified as under diversifying selection along the cryonotothenioid stem tagging all temperate-adapted species (including non-cryonotothenioid notothenioids) and all their ancestral branches. OGs that were found to be under positive selection in at least one temperate-adapted branch were then removed from aBSREL significant results on the cryonotothenioid ancestor. GO term enrichment was performed using all genes analysed by aBSREL as background based on EGG-NOG annotation with the TopGO R package [123] with a Fisher exact test and an *elim* algorithm.

Syntenic detection and multi-genome alignment of notothenioids

The same set of notothenioid genomes included in divergence time estimation was used to infer syntenic relationships based on gene order conservation using GENESPACE v1.2.3 [124] with default parameters and Progressive Cactus v2.9.7 [125] to perform a multi whole-genome alignment. Prior to running Progressive Cactus, the genome assemblies were soft masked with RepeatMasker, using a repeat library generated for each genome using RepeatModeler2 with default parameters. To guide the whole-genome alignment we used the notothenioid-only subtree obtained with the constrained tree search and the 1,000 randomly selected BUSCO genes.

We inferred insertion and deletion events along all branches of the species tree with the *halBranchMutations* function of the HAL package [126]. Net genome gain rates per MY were calculated by subtracting the total number of base pairs involved in deletion events from those involved in insertion events and dividing the result by the branch length in time. Insertions inferred along the *H. antarcticus* ancestors were then lifted off to the *H. antarcticus* Hap1 assembly using *halLiftover* (--noDuples). Due to the fragmentation of the lifted insertions, we merged intervals closer than 100 bp when they derive from the same ancestral insertion events. To detect putative TE-derived insertion events, lifted annotations were intersected with the MCHelper-derived TE annotation requiring a reciprocal overlap of the TE annotation and lifted insertion of 75%. PhyloFit from the PHAST package v1.5 [127] was used under a REV substitution model to train an initial non-conserved model of evolution on the same set of 4-fold degenerate sites extracted from BUSCO genes that was used to estimate the neutral substitution rate. PhastCons was then applied to a MAF representation of the whole-genome multiple alignment,

generated with cactus-hal2maf using *H. antarcticus* as the reference genome (--noAncestors --refGenome HarAnt1.hap1 --dupeMode single), to refine the non-conserved model and train a conserved model for each chromosome. Parameters from the per-chromosome conserved and non-conserved models were then averaged using PhyloBoot, and PhastCons was re-run to predict discrete Conserved Elements (CEs). CEs closer than 5 bp were then merged into single elements. To identify CEs shared by all analysed notothenioids, and those that emerged along the cryonotothenioid stem branch, we used halAlignmentDepth to calculate per-base alignment depth for the *H. antarcticus* genome. Elements covered over at least 75% of their length by all other notothenioids were considered ancestrally present, whereas elements covered over at least 75% of their length only by cryonotothenioids were considered to have been acquired along their stem branch. We assigned each CE to different genomic features (exons, introns, promoters, transcription terminating site, and intergenic regions) with the HOMER annotatePeaks.pl script [128]. GO term enrichment analyses was performed similarly to what was performed for selection analyses but using all *H. antarcticus* genes as background.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-026-04293-7>.

Additional file 1: Figures S1—S13.

Additional file 2: Tables S1—S16.

Acknowledgements

We thank the Wellcome Sanger Institute DNA pipelines for support with sequencing data generation of the *H. antarcticus* genome. We would like to thank the Rothera Marine Team for sample collection.

Peer review information

Alison Monroe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

I.B. conceptualised the study and supervised the work. I.B., J.M. designed analysis. A.D., M.S.C., N.F., I.B., generated data. I.B., R.D., D.L.B., N.F. contributed resources, K.K. performed genome assembly, T.M., J.M.D.W. performed genome curation, J.M. performed data analysis, with J.M.D.W. input. J.M. and I.B. wrote the original manuscript with contributions from M.S.C. All co-authors read, provided feedback, and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. We thank the following for funding support. IB and JM were supported by the Centre for Translational Biodiversity Genomics (LOEWE-TBG) funded through the program LOEWE—Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz of Hesse's Ministry of Higher Education, Research, and the Arts. NF and DB were supported by NIH grant R35 GM144336.

Data availability

The phased genome assembly of *Harpagifer antarcticus* (fHarAnt1.2) is available under accession GCA_985557965.1 (fHarAnt1.hap1.2: https://www.ebi.ac.uk/ena/browser/view/GCA_985557965.1 GCA_985557965.1, fHarAnt1.hap2.2: https://www.ebi.ac.uk/ena/browser/view/GCA_986562475.1 GCA_986562475.1) [129], Bioproject PRJEB113216 [130], Biosample: <https://www.ncbi.nlm.nih.gov/biosample/19870983> SAMEA104132831. The gill RNAseq data are available under accession E-MTAB-17028 in the EBI Biostudies archive <https://doi.org/10.6019/E-MTAB-17028> [131]. The repeat libraries and gene annotation are available on Zenodo <https://zenodo.org/records/21629332> [132], and custom code on GitHub (https://github.com/jacopoM28/Check_monophyly, MIT license) [133]. All externally available data used are listed in supplementary tables (Additional File 2) and are referenced in the text (Assemblies: Table S7, RNAseq data: Table S16).

Declarations

Ethics approval and consent to participate

All research was conducted under Antarctic Treaty permits issued by the UK Foreign, Commonwealth and Development Office to British Antarctic Survey, United Kingdom of Great Britain and Northern Ireland, Antarctic Act 1994 Antarctic Act 2013, Antarctic Regulations 1995/490 (as amended). Permit number BAS-S7-2018/3 granted under Section 7 of the Antarctic Act 1994. All experimental work was approved by the British Antarctic Survey Ethical Review Committee, who confirmed that no animal ethics approval was necessary.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 9 March 2026 Accepted: 22 September 2026

**References**

1. Lecointre G, Améziane N, Boisselier M-C, Bonillo C, Busson F, Causse R, et al. Is the Species Flock Concept Operational? The Antarctic Shelf Case. *PLOS ONE*. 2013;8:e68787. <https://doi.org/10.1371/journal.pone.0068787>. Public Library of Science. Cited 2026 Jan 9.
2. Eastman JT, Eakin RR. Checklist of the species of notothenioid fishes. *Antarct Sci*. 2021;33:273–80. <https://doi.org/10.1017/S0954102020000632>.
3. Di Prisco G, Pisano E, Clarke A. Fishes of Antarctica: a biological overview Edited by G. di Prisco, E. Pisano & A. Clarke Springer-Verlag, Milano (1998). 363 pages. Price DM 198. ISBN 88 470 0028 9. *Antarctic Sci*. 1999;11:379–80. <https://doi.org/10.1017/S0954102099210462>. Cited 2026 Jan 15.
4. Chen L, DeVries AL, Cheng C-HC. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences. Proc Natl Acad Sci*. 1997;94:3811–6. <https://doi.org/10.1073/pnas.94.8.3811>. Cited 2026 Jan 14.
5. Cao L, Huang Q, Wu Z, Cao D, Ma Z, Xu Q, et al. Neofunctionalization of zona pellucida proteins enhances freeze-prevention in the eggs of Antarctic notothenioids. *Nat Commun*. 2016;7:12987. <https://doi.org/10.1038/ncomms12987>. Nature Publishing Group. Cited 2026 July 27.
6. Berthelot C, Clarke J, Desvignes T, William Detrich H III, Flicek P, Peck LS, et al. Adaptation of Proteins to the Cold in Antarctic Fish: A Role for Methionine? *Genome Biol Evol*. 2019;11:220–31. <https://doi.org/10.1093/gbe/evy262>. Cited 2025 Nov 19.
7. Bista I, Wood JM, Desvignes T, McCarthy SA, Matschiner M, Ning Z, et al. Genomics of cold adaptations in the Antarctic notothenioid fish radiation. *Nature communications*. 2023;14:3412. <https://www.nature.com/articles/s41467-023-38567-6>. Nature Publishing Group UK London. Cited 2025 Aug 25.
8. Cheng C-HC, Chen L. Evolution of an antifreeze glycoprotein. *Nature*. 1999;401:443–4. <https://doi.org/10.1038/46721>.
9. Nicodemus-Johnson J, Silic S, Ghigliotti L, Pisano E, Cheng C-HC. Assembly of the antifreeze glycoprotein/trypsinogen-like protease genomic locus in the Antarctic toothfish *Dissostichus mawsoni* (Norman). *Genomics*. 2011;98:194–201. <https://doi.org/10.1016/j.ygeno.2011.06.002>. Cited 2025 Nov 30.
10. Rivera-Colón AG, Rayamajhi N, Minhas BF, Madrigal G, Bilyk KT, Yoon V, et al. Genomics of secondarily temperate adaptation in the only non-Antarctic icefish. *Mol Biol Evol*. 2023;40:msad029. <https://doi.org/10.1093/molbev/msad029>.
11. Ryder D, Stone D, Minardi D, Riley A, Avant J, Cross L, et al. De novo assembly and annotation of the Patagonian toothfish (*Dissostichus eleginoides*) genome. *BMC Genomics*. 2024;25:233. <https://doi.org/10.1186/s12864-024-10141-4>. Cited 2026 Jan 15.
12. Fraser KPP, Peck LS, Clark MS, Clarke A, Hill SL. Life in the freezer: protein metabolism in Antarctic fish. *R Soc Open Sci*. 2022;9:211272. <https://doi.org/10.1098/rsos.211272>. Cited 2026 Jan 15.
13. Fraser KPP, Peck LS, Clark MS, Clarke A. A comparative study of tissue protein synthesis rates in an Antarctic, *Harpagifer antarcticus* and a temperate, *Lipophrys pholis* teleost. *Compar Biochem Physiol Part A Mol Integ Physiol*. 2024;295:111650. <https://doi.org/10.1016/j.cbpa.2024.111650>. Cited 2026 Jan 9.
14. Marty A-PM, Ward EN, Lamb JR, van Tartwijk FW, Peck LS, Clark MS, et al. A High-Resolution Microscopy System for Biological Studies of Cold-Adapted Species Under Physiological Conditions. *Small Methods*. 2025;9:2401682. <https://doi.org/10.1002/smt.202401682>. Cited 2026 Mar 2.
15. Detrich HW, Stuart A, Schoenborn M, Parker SK, Methé BA, Amemiya CT. Genome enablement of the notothenioid: genome size estimates from 11 species and BAC libraries from 2 Representative Taxa. *J Exp Zoolology Part B Mol Dev Evol*. 2010;314B:369–81. <https://doi.org/10.1002/jez.b.21341>. Cited 2025 Nov 12.
16. Chen L, Lu Y, Li W, Ren Y, Yu M, Jiang S, et al. The genomic basis for colonizing the freezing Southern Ocean revealed by Antarctic toothfish and Patagonian robalo genomes. *Gigascience*. 2019;8:giz016. <https://doi.org/10.1093/gigascience/giz016>. Cited 2026 Jan 14.
17. Auvinet J, Graça P, Ghigliotti L, Pisano E, Dettai A, Ozouf-Costaz C, et al. Insertion Hot Spots of DIRS1 Retrotransposon and Chromosomal Diversifications among the Antarctic Teleosts Nototheniidae. *Int J Mol Sci*. 2019;20:701. <https://doi.org/10.3390/ijms20030701>. Multidisciplinary Digital Publishing Institute. Cited 2026 Jan 14.
18. Coronado-Zamora M, González J. Transposable elements and adaptation. In: *Transposable elements and genome evolution*. John Wiley & Sons, Ltd; 2024. p. 115–33. <https://doi.org/10.1002/9781394312467.ch5>.
19. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. *Genome Biol*. 2018;19:199. <https://doi.org/10.1186/s13059-018-1577-z>. Cited 2026 Feb 5.
20. Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. Adaptation to Global Change: A Transposable Element–Epigenetics Perspective. *Trends Ecol Evol*. 2016;31:514–26. <https://doi.org/10.1016/j.tree.2016.03.013>. Cited 2026 Jan 14.
21. Xu H, Fang C, Xu W, Wang C, Song Y, Zhu C, et al. Evolution and genetic adaptation of fishes to the deep sea. *Cell*. 2025;188:1393–1408.e13. <https://doi.org/10.1016/j.cell.2025.01.002>. Elsevier. Cited 2026 Jan 14.
22. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*. 2014;346:1320–31. <https://doi.org/10.1126/science.1253451>.

23. Cicconardi F, Milanetti E, Pinheiro de Castro EC, Mazo-Vargas A, Van Belleghem SM, Ruggieri AA, et al. Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies. *Nat Commun.* 2023;14:5620. <https://doi.org/10.1038/s41467-023-41412-5>. Nature Publishing Group. Cited 2026 Jan 14.
24. Bista I, McCarthy SA, Wood J, Ning Z, Detrich Iii HW, Desvignes T, et al. The genome sequence of the channel bull blenny, *Cottoperca gobio* (Günther, 1861). Wellcome open research. 2020;5:148. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7649722/>. Cited 2025 Aug 25.
25. Pisano E, Ozouf-Costaz C. Chromosome change and the evolution in the Antarctic fish suborder Notothenioidei. *Antarct Sci.* 2000;12:334–42. <https://doi.org/10.1017/S0954102000000390>.
26. Kapitonov VV, Tempel S, Jurka J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene.* 2009;448:207–13. <https://doi.org/10.1016/j.gene.2009.07.019>. Cited 2026 Jan 14.
27. Pisano E, Ghigliotti L. Ribosomal genes in notothenioid fishes: Focus on the chromosomal organisation. *Mar Genomics.* 2009;2:75–80. <https://doi.org/10.1016/j.margen.2009.03.006>.
28. Pisano E, Ozouf-Costaz C, Prirodina V. Chromosome diversification in Antarctic fish (Notothenioidei). In: Di Prisco G, Pisano E, Clarke A, editors. *Fishes of Antarctica: a biological overview*. Milano: Springer Milan; 1998. p. 275–85. https://doi.org/10.1007/978-88-470-2157-0_24.
29. Miller KG, Browning JV, Schmelz WJ, Kopp RE, Mountain GS, Wright JD. Cenozoic sea-level and cryospheric evolution from deep-sea geochemical and continental margin records. *Sci Adv.* 2020;6:eaa31346. <https://doi.org/10.1126/sciadv.aaz1346>. American Association for the Advancement of Science. Cited 2026 Jan 14.
30. Stankovic A, Spalik K, Kamler E, Borsuk P, Weglenski P. Recent origin of sub-Antarctic notothenioids. *Polar Biol.* 2002;25:203–5. <https://doi.org/10.1007/s00300-001-0327-x>.
31. Collins M, Brickle P, Brown J, Belchier M. The Patagonian toothfish: biology, ecology and fishery. In: *Advances in marine biology*. Academic Press; 2010. p. 227–300. <https://doi.org/10.1016/B978-0-12-381015-1.00004-6>.
32. Near TJ, Dornburg A, Harrington RC, Oliveira C, Pietsch TW, Thacker CE, et al. Identification of the notothenioid sister lineage illuminates the biogeographic history of an Antarctic adaptive radiation. *BMC Evol Biol.* 2015;15:109. <https://doi.org/10.1186/s12862-015-0362-9>. Cited 2026 Jan 14.
33. Detrich HW, Parker SK, Williams RC, Nogales E, Downing KH. Cold Adaptation of Microtubule Assembly and Dynamics: structural interpretation of primary sequence changes present in the α - and β -tubulins of antarctic fishes *. *J Biol Chem.* 2000;275:37038–47. <https://doi.org/10.1074/jbc.M005699200>. Elsevier. Cited 2026 Jan 14.
34. Kowalczyk A, Chikina M, Clark NL. A cautionary tale on proper use of branch-site models to detect convergent positive selection. *bioRxiv*; 2021. p. 2021.10.26.465984. <https://doi.org/10.1101/2021.10.26.465984>. Cited 2026 Feb 5.
35. Rayamajhi N, Rivera-Colón AG, Minhas BF, Cheng CHC, Catchen JM. The genome of the cryopelagic Antarctic baid notothen, *Trematomus borchgrevinki*. *G3 Genes[Genomes]Genetics.* 2025;15:jkae267. <https://doi.org/10.1093/g3journal/jkae267>. Cited 2026 Jan 14.
36. Li H, Durbin R. Genome assembly in the telomere-to-telomere era. *Nat Rev Genet.* 2024;25:658–70. <https://doi.org/10.1038/s41576-024-00718-w>. Nature Publishing Group. Cited 2026 July 30.
37. Liu S, Shi C, Chen C, Tan Y, Tian Y, Macqueen DJ, et al. Haplotype-resolved genomes provide insights into the origins and functional significance of genome diversity in bivalves. *Cell Rep.* 2025;44:115697. <https://doi.org/10.1016/j.celrep.2025.115697>. Cited 2026 Jan 14.
38. Rosani U, Gerdol M, Krupovic M. The highly dynamic pangenome of basal chordates is enriched in defence and immunity genes and is inherited following the Mendelian law. *PLOS Genetics.* 2025;21:e1011833. <https://doi.org/10.1371/journal.pgen.1011833>. Public Library of Science. Cited 2026 Feb 5.
39. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell.* 2019;176:663–675.e19. <https://doi.org/10.1016/j.cell.2018.12.019>. Elsevier. Cited 2026 Jan 14.
40. Chang N-C, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res.* 2022;32:1408–23. <https://doi.org/10.1101/gr.275655.121>. Cold Spring Harbor Lab. Cited 2026 Jan 14.
41. Lu Y, Li W, Li Y, Zhai W, Zhou X, Wu Z, et al. Population genomics of an icefish reveals mechanisms of glacier-driven adaptive radiation in Antarctic notothenioids. *BMC Biol.* 2022;20:231. <https://doi.org/10.1186/s12915-022-01432-x>. Cited 2026 July 30.
42. Cao L, Huang Q, Wu Z, Cao D, Ma Z, Xu Q, et al. Neofunctionalization of zona pellucida proteins enhances freeze-prevention in the eggs of Antarctic notothenioids. *Nat Commun.* 2016;7:12987. <https://doi.org/10.1038/ncomm12987>. Nature Publishing Group. Cited 2026 July 30.
43. Nishiguchi T, Ishikawa A. Convergent Gene Duplication in Arctic and Antarctic Teleost Fishes. *jzoo.* 2025;42:116–26. <https://doi.org/10.2108/zs240098>. Zoological Society of Japan. Cited 2026 July 30.
44. Han S, Dias GB, Basting PJ, Viswanatha R, Perrimon N, Bergman CM. Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. *Nucleic Acids Res.* 2022;50:e124. <https://doi.org/10.1093/nar/gkac794>. Cited 2026 Jan 14.
45. Boissinot S, Sookdeo A. The evolution of LINE-1 in vertebrates. *Genome Biol Evol.* 2016;8:3485–507. <https://doi.org/10.1093/gbe/evw247>.
46. Martellosi J, Nicolini F, Subacchi S, Pasquale D, Ghiselli F, Luchetti A. Multiple and diversified transposon lineages contribute to early and recent bivalve genome evolution. *BMC Biol.* 2023;21:145. <https://doi.org/10.1186/s12915-023-01632-z>. Cited 2026 Jan 14.
47. Davis KE, Hill J, Astrop TI, Wills MA. Global cooling as a driver of diversification in a major marine clade. *Nat Commun.* 2016;7:13003. <https://doi.org/10.1038/ncomms13003>. Nature Publishing Group. Cited 2026 Jan 14.
48. Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene.* 2012;509:7–15. <https://doi.org/10.1016/j.gene.2012.07.042>. Cited 2026 Jan 14.

49. Chen S, Yu M, Chu X, Li W, Yin X, Chen L. Cold-induced retrotransposition of fish LINEs. *J Genet Genom*. 2017;44:385–94. <https://doi.org/10.1016/j.jgg.2017.07.002>. Cited 2026 Jan 14.
50. Amores A, Wilson CA, Allard CAH, Detrich HW III, Postlethwait JH. Cold Fusion: Massive Karyotype Evolution in the Antarctic Bullhead Notothen *Notothenia coriiceps*. *G3 Genes[Genomes]Genetics*. 2017;7:2195–207. <https://doi.org/10.1534/g3.117.040063>. Cited 2026 Mar 9.
51. Gerton JL. A working model for the formation of Robertsonian chromosomes. *J Cell Sci*. 2024;137:jcs.261912. <https://doi.org/10.1242/jcs.261912>. Cited 2026 Jan 11.
52. Read LR, Raynard SJ, Rukšć A, Baker MD. Gene repeat expansion and contraction by spontaneous intrachromosomal homologous recombination in mammalian cells. *Nucleic Acids Res*. 2004;32:1184–96. <https://doi.org/10.1093/nar/gkh280>.
53. Raymond JA, DeVries AL. Adsorption inhibition as a mechanism of freezing resistance in polar fishes. *Proc Natl Acad Sci*. 1977;74:2589–93. <https://doi.org/10.1073/pnas.74.6.2589>. Cited 2026 Jan 21.
54. DeVries AL. The role of antifreeze glycopeptides and peptides in the freezing avoidance of Antarctic fishes. *Compar Biochem Physiol Part B Compar Biochem*. 1988;90:611–21. [https://doi.org/10.1016/0305-0491\(88\)90302-1](https://doi.org/10.1016/0305-0491(88)90302-1). Cited 2026 Jan 21.
55. Schrag JD, DeVries AL. The effect of freezing rate on the cooperativity of antifreeze glycopeptides. *Compar Biochem Physiol Part A Physiol*. 1983;74:381–5. [https://doi.org/10.1016/0300-9629\(83\)90619-9](https://doi.org/10.1016/0300-9629(83)90619-9). Cited 2026 Jan 21.
56. Frandsen PB, Hotaling S, Powell A, Heckenhauer J, Kawahara AY, Baker RH, et al. Allelic resolution of insect and spider silk genes reveals hidden genetic diversity. *Proc Natl Acad Sci*. 2023;120:e2221528120. <https://doi.org/10.1073/pnas.2221528120>. Cited 2026 Jan 14.
57. Arakawa K, Kono N, Malay AD, Tateishi A, Ifuku N, Masunaga H, et al. 1000 spider silkomes: linking sequences to silk physical properties. *Sci Adv*. 2022;8:eabo6043. <https://doi.org/10.1126/sciadv.abo6043>.
58. Miya T, Gon O, Mwale M, Cheng C-HC. Multiple independent reduction or loss of antifreeze trait in low Antarctic and sub-Antarctic notothenioid fishes. *Antarct Sci*. 2016;28:17–28. <https://doi.org/10.1017/S0954102015000413>.
59. Abele D, Puntarulo S. Formation of reactive species and induction of antioxidant defence systems in polar and temperate marine invertebrates and fish. *Comp Biochem Physiol A Mol Integr Physiol*. 2004;138:405–15. <https://doi.org/10.1016/j.cbpb.2004.05.013>.
60. Peck LS. A Cold Limit to Adaptation in the Sea. *Trends Ecol Evol*. 2016;31:13–26. <https://doi.org/10.1016/j.tree.2015.09.014>. Elsevier. Cited 2026 Jan 11.
61. Carpenter JH. New Measurements of Oxygen Solubility in Pure and Natural Water. *Limnol Oceanography*. 1966;11:264–77. <https://doi.org/10.4319/lo.1966.11.2.0264>. Cited 2026 Jan 11.
62. Olmos Y, Sánchez-Gómez FJ, Wild B, García-Quintans N, Cabezudo S, Lamas S, et al. SirT1 Regulation of Antioxidant Genes Is Dependent on the Formation of a FoxO3a/PGC-1 α Complex. *Antioxidants Redox Signal*. 2013;19:1507–21. <https://doi.org/10.1089/ars.2012.4713>. Mary Ann Liebert, Inc., publishers. Cited 2026 Jan 14.
63. Ferko M, Alanova P, Janko D, Opletalova B, Andelova N. Mitochondrial Peroxiredoxins and Monoamine Oxidase-A: Dynamic Regulators of ROS Signaling in Cardioprotection. *Physiol Res*. 2024;73:887–900. <https://doi.org/10.33549/physiolres.935513>. Cited 2026 Jan 14.
64. Bilyk KT, Zhuang X, Papetti C. Positive and Relaxed Selective Pressures Have Both Strongly Influenced the Evolution of Cryonotothenioid Fishes during Their Radiation in the Freezing Southern Ocean. *Genome Biol Evol*. 2023;15:evad049. <https://doi.org/10.1093/gbe/evad049>. Cited 2026 Jan 23.
65. Shin SC, Ahn DH, Kim SJ, Pyo CW, Lee H, Kim M-K, et al. The genome sequence of the Antarctic bullhead notothen reveals evolutionary adaptations to a cold environment. *Genome Biol*. 2014;15:468. <https://doi.org/10.1186/s13059-014-0468-1>. Cited 2026 Jan 14.
66. Chen Z, Cheng C-HC, Zhang J, Cao L, Chen L, Zhou L, et al. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proc Natl Acad Sci*. 2008;105:12944–9. <https://doi.org/10.1073/pnas.0802432105>. Cited 2025 Nov 19.
67. Clark MS, Thorne MAS, Toullec J-Y, Meng Y, Guan LL, Peck LS, et al. Antarctic Krill 454 Pyrosequencing Reveals Chaperone and Stress Transcriptome. *PLOS ONE*. 2011;6:e15919. <https://doi.org/10.1371/journal.pone.0015919>. Public Library of Science. Cited 2026 Jan 11.
68. Abele D, Tesch C, Wencke P, Pörtner HO. How does oxidative stress relate to thermal tolerance in the Antarctic bivalve *Yoldia eightsi*? *Antarct Sci*. 2001;13:111–8. <https://doi.org/10.1017/S0954102001000189>.
69. Fraser KPP, Rogers AD. Protein Metabolism in Marine Animals: The Underlying Mechanism of Growth. *Adv Marine Biol*. 2007. p. 267–362. [https://doi.org/10.1016/S0065-2881\(06\)52003-6](https://doi.org/10.1016/S0065-2881(06)52003-6). Academic Press. Cited 2026 Jan 11.
70. Todgham AE, Hoaglund EA, Hofmann GE. Is cold the new hot? Elevated ubiquitin-conjugated protein levels in tissues of Antarctic fish as evidence for cold-denaturation of proteins in vivo. *J Comp Physiol B*. 2007;177:857–66. <https://doi.org/10.1007/s00360-007-0183-2>. Cited 2026 Jan 11.
71. Hofmann GE, Lund SG, Place SP, Whitmer AC. Some like it hot, some like it cold: the heat shock response is found in New Zealand but not Antarctic notothenioid fishes. *J Exp Marine Biol Ecol*. 2005;316:79–89. <https://doi.org/10.1016/j.jembe.2004.10.007>. Cited 2026 Jan 11.
72. Place SP, Zippay ML, Hofmann GE. Constitutive roles for inducible genes: evidence for the alteration in expression of the inducible *hsp70* gene in Antarctic notothenioid fishes. *Am J Physiol Regul Integr Comp Physiol*. 2004. <https://doi.org/10.1152/ajpregu.00223.2004>.
73. Todgham AE, Crombie TA, Hofmann GE. The effect of temperature adaptation on the ubiquitin–proteasome pathway in notothenioid fishes. *J Exp Biol*. 2017;220:369–78. <https://doi.org/10.1242/jeb.145946>.
74. Strickland M, Moll R, Cornwell C, Smith M, Howard C. Sanger Tree of Life HMW DNA Extraction: Manual MagAttract. 2023. <https://www.protocols.io/view/sanger-tree-of-life-hmw-dna-extraction-manual-maga-cy3ixyke>. Cited 2026 Feb 16.
75. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–5. <https://doi.org/10.1038/s41592-020-01056-5>. Nature Publishing Group. Cited 2026 Jan 15.

76. Vasimuddin Md, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2019. p. 314–24. <https://doi.org/10.1109/IPDPS.2019.00041>. Cited 2026 Jan 15.
77. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*. 2023;39:btac808. <https://doi.org/10.1093/bioinformatics/btac808>. Cited 2026 Jan 14.
78. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*. 2020;117:9451–7. <https://doi.org/10.1073/pnas.1921046117>. Cited 2025 Dec 9.
79. Orozco-Arias S, Sierra P, Durbin R, González J. MCHelper automatically curates transposable element libraries across eukaryotic species. *Genome Res*. 2024;34:2256–68. <https://doi.org/10.1101/gr.278821.123>. Cold Spring Harbor Laboratory Press. Cited 2026 July 30.
80. Lin Y, Ye C, Li X, Chen Q, Wu Y, Zhang F, et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res*. 2023;10:uhad127. <https://doi.org/10.1093/hr/uhad127>. Cited 2026 Feb 5.
81. Seemann T. tseemann/barrnap. 2026. <https://github.com/tseemann/barrnap>. Cited 2026 Mar 5.
82. Brown MR, Manuel Gonzalez de La Rosa P, Blaxter M. tldk: a toolkit to rapidly identify telomeric repeats from genomic datasets. 2025;41:btaf049. <https://doi.org/10.1093/bioinformatics/btaf049>. Cited 2026 Jan 14.
83. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res*. 2024;34:769–77. <https://doi.org/10.1101/gr.278090.123>. Cold Spring Harbor Lab. Cited 2026 Jan 14.
84. Hoff KJ, Stanke M. Predicting Genes in Single Genomes with AUGUSTUS. *Current Protocols in Bioinformatics*. 2019;65:e57. <https://doi.org/10.1002/cpbi.57>. Cited 2026 Jan 14.
85. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform*. 2020;2:lqaa026. <https://doi.org/10.1093/nargab/lqaa026>. Cited 2026 Jan 14.
86. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37:907–15. <https://doi.org/10.1038/s41587-019-0201-4>.
87. Manni M, Berkeley MR, Seppely M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol*. 2021;38:4647–54. <https://doi.org/10.1093/molbev/msab199>. Cited 2026 Jan 14.
88. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. 2021;37:1639–43. <https://doi.org/10.1093/bioinformatics/btaa1016>. Cited 2026 Jan 14.
89. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol*. 2021;38:5825–9. <https://doi.org/10.1093/molbev/msab293>. Cited 2026 Jan 14.
90. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30:1236–40. <https://doi.org/10.1093/bioinformatics/btu031>. Cited 2026 July 30.
91. Li H. Protein-to-genome alignment with miniprot. *Bioinformatics*. 2023;39:btad014. <https://doi.org/10.1093/bioinformatics/btad014>. Cited 2026 July 30.
92. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. In: Kollmar M, editor. *Gene prediction: methods and protocols*. New York, NY: Springer; 2019. p. 161–77. https://doi.org/10.1007/978-1-4939-9173-0_9.
93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
94. Madrigal G, Minhas BF, Catchen J. Klumpy: A tool to evaluate the integrity of long-read genome assemblies and illusive sequence motifs. *Mol Ecol Res*. 2025;25:e13982. <https://doi.org/10.1111/1755-0998.13982>. Cited 2026 Jan 14.
95. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772–80. <https://doi.org/10.1093/molbev/mst010>.
96. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37:1530–4. <https://doi.org/10.1093/molbev/msaa015>.
97. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol*. 2018;35:518–22. <https://doi.org/10.1093/molbev/msx281>. Cited 2026 Jan 14.
98. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14:587–9. <https://doi.org/10.1038/nmeth.4285>. Nature Publishing Group. Cited 2026 Jan 14.
99. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49:D412–9. <https://doi.org/10.1093/nar/gkaa913>. Cited 2026 Jan 15.
100. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:3150–2. <https://doi.org/10.1093/bioinformatics/bts565>. Cited 2026 Jan 15.
101. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25:1972–3. <https://doi.org/10.1093/bioinformatics/btp348>. Cited 2026 Jan 14.
102. Platt RN II, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biol Evol*. 2016;8:403–10. <https://doi.org/10.1093/gbe/evw009>.
103. Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. A beginner's guide to manual curation of transposable elements. *Mob DNA*. 2022;13:7. <https://doi.org/10.1186/s13100-021-00259-7>.

104. Peona V, Martelossi J, Almojil D, Bocharkina J, Brännström I, Brown M, et al. Teaching transposon classification as a means to crowd source the curation of repeat annotation – a tardigrade perspective. *Mobile DNA*. 2024;15:10. <https://doi.org/10.1186/s13100-024-00319-8>. Cited 2026 Jan 15.
105. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100. <https://doi.org/10.1093/bioinformatics/bty191>. Cited 2025 Oct 27.
106. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRl: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*. 2019;20:277. <https://doi.org/10.1186/s13059-019-1911-0>. Cited 2026 Jan 14.
107. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2. <https://doi.org/10.1093/bioinformatics/btq033>. Cited 2026 Jan 14.
108. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Software*. 2015;67:1–48. <https://doi.org/10.18637/jss.v067.i01>. Cited 2026 Jan 14.
109. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol*. 2024;42:1571–80. <https://doi.org/10.1038/s41587-023-02024-y>. Nature Publishing Group. Cited 2026 Jan 14.
110. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun*. 2017;8:14061. <https://doi.org/10.1038/ncomms14061>. Nature Publishing Group. Cited 2026 Jan 14.
111. Bista I, Collins M. The genome sequence of the marbled rockcod, *Notothenia rossii* Richardson, 1844. *Wellcome Open Res*. 2024;9:227. <https://doi.org/10.12688/wellcomeopenres.21270.1>. Cited 2025 Nov 15.
112. Cheng C-HC, Rivera-Colón AG, Minhas BF, Wilson L, Rayamajhi N, Vargas-Chacoff L, et al. Chromosome-Level Genome Assembly and Circadian Gene Repertoire of the Patagonia Blennie *Eleginops maclovinus*—The Closest Ancestral Proxy of Antarctic *Cryonotothenioids*. *Genes*. 2023;14:1196. <https://doi.org/10.3390/genes14061196>. Multidisciplinary Digital Publishing Institute. Cited 2026 Jan 15.
113. Jae Lee S, Kim J-H, Jo E, Choi E, Kim J, Choi S-G, et al. Chromosomal assembly of the Antarctic toothfish (*Dissostichus mawsoni*) genome using third-generation DNA sequencing and Hi-C technology. *Zool Res*. 2021;42:124–9. <https://doi.org/10.24272/j.zissn.2095-8137.2020.264>. Cited 2026 Jan 15.
114. Lee SJ, Kim J, Choi EK, Jo E, Cho M, Kim J-H, et al. A chromosome-level reference genome of the Antarctic blackfin icefish *Chaenocephalus aceratus*. *Sci Data*. 2023;10:657. <https://doi.org/10.1038/s41597-023-02561-w>. Nature Publishing Group. Cited 2026 Jan 15.
115. Rayamajhi N, Rivera-Colón AG, Minhas BF, Cheng CHC, Catchen JM. The genome of the cryopelagic Antarctic bald notothen, *Trematomus borchgrevinkii*. G3 (Bethesda). 2025;15:jkae267. <https://doi.org/10.1093/g3journal/jkae267>.
116. Jo E, Choi S, Lee SJ, Kim J, Choi EK, Cho M, et al. Chromosome-level genome assembly and annotation of the Antarctica whitefin plunderfish *Pogonophryne albigipinna*. *Sci Data*. 2023;10:891. <https://doi.org/10.1038/s41597-023-02811-x>. Nature Publishing Group. Cited 2026 Jan 15.
117. To T-H, Jung M, Lycett S, Gascuel O. Fast dating using least-squares criteria and algorithms. *Syst Biol*. 2016;65:82–97. <https://doi.org/10.1093/sysbio/syv068>.
118. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20(1):238. <https://doi.org/10.1186/s13059-019-1832-y>.
119. Willson J, Roddur MS, Liu B, Zaharias P, Warnow T. DISCO: species tree inference using multicopy gene family tree decomposition. *Syst Biol*. 2022;71:610–29. <https://doi.org/10.1093/sysbio/syab070>.
120. Löytynoja A. Phylogeny-aware alignment with PRANK. In: Russell DJ, editor. Multiple sequence alignment methods. Totowa, NJ: Humana Press; 2014. p. 155–70. https://doi.org/10.1007/978-1-62703-646-7_10.
121. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*. 2007;56:564–77. <https://doi.org/10.1080/10635150701472164>.
122. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Mol Biol Evol*. 2015;32:1342–53. <https://doi.org/10.1093/molbev/msv022>. Cited 2026 Jan 15.
123. Alexa A, Rahnenführer J, Marini F. topGO. Bioconductor. <http://bioconductor.org/packages/topGO/>. Cited 2026 Jan 15.
124. Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. Weigel D, editor. *eLife*. 2022;11:e78526. <https://doi.org/10.7554/eLife.78526>. eLife Sciences Publications, Ltd. Cited 2026 Jan 15.
125. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*. 2020;587:246–51. <https://doi.org/10.1038/s41586-020-2871-y>. Nature Publishing Group. Cited 2026 Jan 15.
126. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*. 2013;29:1341–2. <https://doi.org/10.1093/bioinformatics/btt128>.
127. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform*. 2011;12:41–51. <https://doi.org/10.1093/bib/bbq072>.
128. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38:576–89. <https://doi.org/10.1016/j.molcel.2010.05.004>.
129. Martelossi J, Krashennikova K, Denton A, Wood JMD, Mathers T, Durbin R, Fong N, Bentley DL, Clark MS, Bista I. A new phased assembly of the Antarctic spiny plunderfish provides novel insights into the evolution of the notothenioid radiation. *Genome Assembly fHarAnt1*, European Nucleotide Archive (ENA). 2026. <https://www.ebi.ac.uk/ena/browser/text-search?query=fHarAnt1>.
130. Martelossi J, Krashennikova K, Denton A, Wood JMD, Mathers T, Durbin R, Fong N, Bentley DL, Clark MS, Bista I. A new phased assembly of the Antarctic spiny plunderfish provides novel insights into the evolution of the notothenioid radiation. *Nanopore reads for genome Assembly fHarAnt1.2*, European Nucleotide Archive (ENA). 2026. <https://www.ncbi.nlm.nih.gov/sra/ERX16667749> [acn].

131. Martelossi J, Krashennikova K, Denton A, Wood JMD, Mathers T, Durbin R, Fong N, Bentley DL, Clark MS, Bista I. A new phased assembly of the Antarctic spiny plunderfish provides novel insights into the evolution of the notothenioid radiation. RNAseq data, BioStudies database. 2026. <https://doi.org/10.6019/E-MTAB-17028>.
132. Martelossi J, Krashennikova K, Denton A, Wood JMD, Mathers T, Durbin R, Fong N, Bentley DL, Clark MS, Bista I. Annotations for: A new phased assembly of the Antarctic spiny plunderfish provides novel insights into the evolution of the notothenioid radiation. Zenodo. 2026. <https://doi.org/10.5281/zenodo.21629331>.
133. Martelossi J. Check monophyly script. GitHub. 2026. https://github.com/jacopoM28/Check_monophyly.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.