Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids

Xuan Zhuanga,1,2, Chun Yanga, Katherine R. Murphys, and C.-H. Christina Chenga,1

*aSchool of Integrative Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved January 17, 2019 (received for review October 9, 2018)

A fundamental question in evolutionary biology is how genetic novelty arises. De novo gene birth is a recently recognized mechanism, but the evolutionary process and function of putative de novo genes remain largely obscure. With a clear life-saving function, the diverse antifreeze proteins of polar fishes are exemplary adaptive innovations and models for investigating new gene evolution. Here, we report clear evidence and a detailed molecular mechanism for the de novo formation of the northern gadid (codfish) antifreeze glycoprotein (AFGP) gene from a minimal non-coding sequence. We constructed genomic DNA libraries for AFGP-bearing and AFGP-lacking species across the gadid phylogeny and performed fine-scale comparative analyses of the AFGP genomic loci and homologs. We identified the noncoding founder region and a nine-nucleotide (9-nt) element therein that supplied the codons for one Thr-Ala-Ala unit from which the extant repetitive AFGP-coding sequence (cds) arose through tandem duplications. The latent signal peptide (SP)-coding exons were fortuitous non-coding DNA sequence immediately upstream of the 9-nt element, which, when spliced, supplied a typical secretory signal. Through a 1-nt frameshift mutation, these two parts formed a single read-through open reading frame (ORF). It became functionalized when a putative translational event conferred the essential Cys promoter for transcriptional initiation. We experimentally proved that all genetic components of the extant gadid AFGP originated from entirely non-genic DNA. The gadid AFGP evolutionary process also represents a rare example of the proto-ORF model of de novo gene birth where a fully formed ORF existed before the regulatory element to activate transcription was acquired.

de novo gene | proto-ORF | adaptive evolution | noncoding origin | codfish AFGP

Evolutionary innovation of new genetic elements is recognized as a key contributor to organismal adaptation. For decades, the treatise of Ohno (1) shaped the paradigm of new gene creation in that it relies on the duplication of a preexisting protein gene. When subjected to selection, adaptive sequence changes in one copy may occur from which a gene with a novel function may emerge (1, 2). Creating new protein-coding genes de novo from noncoding DNA sequences was considered extremely rare. In recent years, however, examples of de novo genes have been reported in diverse animals and plants (see review ref. 3 and studies referenced therein). De novo gene births were generally deduced using a combination of phylogenetics and comparative genomic/transcriptomic analyses or the phylodiversity approach (4), which revealed evidence for lineage- or species-specific gene transcripts, whereas the orthologous sequences in sister species were nongenic. These revelations have spurred considerable interest and hypotheses of how de novo genes arise and evolve as well as questions regarding their functional importance (5, 6). Validating new genes identified from sequence-based comparisons is complicated by uncertainties around how comprehensive the genome assemblies and gene expression data are (7, 8). More challenging yet is identifying the selective pressures and molecular mechanisms that created these putative new genes, and the adaptive functions and species fitness they may confer.

In contrast, antifreeze protein genes of polar teleost fishes are unequivocal new genes that confer a clear life-saving function and fitness benefit. The selective pressure that compelled their evolution is also clear. They evolved in direct response to polar marine glaciations, preventing death of fish from in situ freezing by environmental ice crystals in subzero waters (9, 10). Such a strong life-or-death selective pressure has driven the independent evolution of multiple structurally distinct types of antifreezes: antifreeze peptide (AFP) types I, II, and III and antifreeze glycoprotein (AFGP) in diverse fish lineages where they perform the same ice-growth inhibition function (10). The structural differences lie in their distinct genetic ancestry. Thus, fish antifreezes as a group can richly inform on the diversity of molecular origins and evolutionary mechanisms that produced a vital function.

The well-known mechanism of evolution by gene duplication from a preexisting ancestor as diverse as C-type lectin and stalic acid synthase followed by sequence tinkering by natural selection produced AFP II (11) and AFP III (12), respectively. AFGPs product AFP II (11) and AFP III (12), respectively. AFGPs

Significance

The diverse antifreeze proteins enabling the survival of different polar fishes in freezing seas offer unparalleled vistas into the breadth of genetic sources and mechanisms that produce crucial new functions. Although most new genes evolved from preexisting genetic ancestors, some are deemed to have arisen from noncoding DNA. However, the pertinent mechanisms, functions, and selective forces remain uncertain. Our paper presents clear evidence that the antifreeze glycoprotein gene of the northern codfish originated from a noncoding region. We further describe the detailed mechanism of its evolutionary transformation into a full-fledged crucial life-saving gene. This paper is a concrete dissection of the process of a de novo gene birth that has conferred a vital adaptive function directly linked to natural selection.

Author contributions: X.Z. and C.-H.C.C. designed research; X.Z., C.Y., K.R.M., and C.-H.C.C. performed research; X.Z. and C.Y. analyzed data; and X.Z. and C.-H.C.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. MK011258–MK011272, MH992395–MH992397, and MK011291–MK011308).

1To whom correspondence may be addressed. Email: zhuangxuan@gmail.com or c-cheng@illinois.edu.

2Present address: Department of Ecology and Evolution, The University of Chicago, Chicago, IL 60637.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1817381116/-/DCSupplemental.

Published online February 14, 2019.
have evolved independently in two unrelated fish lineages at opposite poles: the Antarctic notothenioid fishes (Notothenioidei) and the Arctic/northern codfishes (Gadidae), providing a striking example of protein sequence convergence (9, 13). In both lineages, AFGPs occur as a family of isoforms composed of varying numbers of repeats of a basic tripeptide unit (Thr-Ala-Ala) with each Thr glycosylated with a disaccharide (10, 14). They are encoded by a family of polyprotein genes, each of which produces a large polyprotein precursor consisting of many tandemly linked AFGP molecules that are then post-translationally cleaved to yield mature AFGPs (13, 15). The Antarctic notothenioid AFGP evolved through a more innovative process than gene duplication and sequence divergence. It exemplifies partial de novo gene evolution. A functionally unrelated ancestral trypsin-nogenlike protease (TLP) gene provided the secretory signal and a 3′ untranslated sequence of the incipient AFGP. The large repetitive AFGP polyprotein-coding region was generated de novo from duplications of a partly non-sense 9-nt sequence that stranded an intron–exon junction in the TLP, which happened to comprise the three codons for one Thr-Ala-Ala unit (15, 16).

Where from and how the northern gadid AFGP evolved have remained lasting enigmas. Despite voluminous collections of genes and genome sequences available in databases, there are no meaningful homologs to any part of the gadid AFGP to hint at ancestry. This peculiar absence of related genes suggests that the gadid AFGP gene may have originated from nonprotein-coding DNA. Gadid AFGP presumably evolved very recently, in response to the cyclic northern hemisphere glaciation that commenced in the late Pliocene about 3 Mya. We reason it is unlikely that mutational processes could completely obscure even noncoding sequences within such a short evolutionary time such that the extant form of the AFGP nongenic ancestor should remain identifiable. We, therefore, decided to track the AFGP genotype and its homologs within the gadid phylogeny to pinpoint the ancestral DNA site of origin and reconstruct the gadid AFGP evolutionary path. Here, we report the identification of the noncoding founder sequence and the mechanism by which it gave rise to a new functional gadid AFGP gene. Our results also show that the gadid AFGP evolutionary process likely represents a rare example of the proto-ORF model of de novo gene birth (6, 17) where the noncoding founder ORF existed well before the novel gene arose.

Results and Discussion

At the minimum, de novo formation of a functional protein gene requires the acquisition of an ORF encoding the new protein and the basic cis-regulatory elements to activate its transcription and translation. AFGPs are secreted plasma proteins, thus, a signal peptide (SP) will also be needed to instruct cellular export of AFGP molecules into the blood circulation. Thus, to reconstruct the formation of the gadid AFGP gene requires elucidating how these essential genetic components were generated and became properly linked into a functional whole gene. We began with precise delineation of these components that make up the structure of functional AFGPs in AFGP-bearing gadids. We then juxtaposed them against the structures of the AFGP homologs in more basal non-AFGP-bearing species representing progressively more ancestral states. This enabled us to decipher the essential molecular steps and timing in the de novo formation of the AFGP gene in the gadid lineage.

Phylogenetic Context for the Selected Gadid Species. The phylogenetic tree in Fig. 1 (detailed in SI Appendix, Fig. S1) depicts the relationships of the northern cod species used in this paper. We characterized the AFGP genes or noncoding homologs of seven species (gene structures to the right of the tree, Fig. 1), which were chosen for the strategic positions they occupy in the gadid tree. The monophyletic Gadidae family [senso (18)] includes both AFGP-bearing and non-AFGP-bearing species; the former occurs in two subclades within the subfamily Gadinae (Fig. 1).

The selected AFGP-bearing Boreogadus saida (polar cod) (19) and Gadus morhua (Atlantic cod) (20) represent one gadine subclade, and Microgadus tomcod (Atlantic tomcod) (21) represents the other. The four AFGP-lacking gadids were chosen for their evolutionary distances from the AFGP bearers. Two of them are gadines; Merlangius merlangus (whiting) nests within the AFGP-bearing subclade containing B. saida and G. morhua and thus shares the last common ancestor with all AFGP-bearing species (Fig. 1, blue dot), whereas Trisopterus esmarkii (Norway pout) is basal to the two AFGP-bearing subclades. The other two, Brosme brosme (cusk) and the freshwater Lota lota (burbot) belong to the subfamily Lotinae and are basal species that serve as ancestral proxies before the AFGP trait emerged (Fig. 1).

Gadid AFGP Genomic Regions and AFGP Gene Structure. We isolated AFGP-positive large-insert genomic DNA clones from the respective Bacterial Artificial Chromosome (BAC) library of B. saida and M. tomcod by screening with a probe specific to the AFGP (Thr-Ala-Ala)n cds and sequenced the minimal tiling path clones spanning the AFGP genomic region. For G. morhua, a draft genome was available, but the AFGP genomic region was incomplete (22). We bioinformatically deduced the pertinent BAC clones (SI Appendix, Fig. S2) and obtained them (available from the vendor) for sequencing. The reconstructed genomic regions contained 12 functional and four pseudo AFGP genes in B. saida, five and two in G. morhua, and three and one in M. tomcod, spanning ~510, 190, and 80 kbp, respectively, in the three species (SI Appendix, Fig. S3).

To determine the gene structure of functional AFGPs we used supporting transcript sequences obtained by 5′ rapid amplification of cDNA ends (RACE) (SI Appendix, Fig. S4). A functional AFGP consists of three exons and two introns (Fig. 2). The first two small exons (E1 and E2) and the first two nts of the large third exon (E3) encode the SP, and the rest of E3 encodes a short pro-peptide and the long AFGP polypeptide. These demarcations differ from the only known full-length gadid AFGP gene sequence to date (13). In that sequence, the predicted SP contained many atypical hydrophilic residues, indicating inaccurate assignment of splice junctions and reading frames, hence, the need for reassessment. In this paper, the predicted SP contained the requisite stretch of hydrophobic residues followed by a putative cleavage site with high prediction scores (SI Appendix, Fig. S5), which is characteristic of a secretory signal. It is conserved in all functional AFGPs.
from the three gadids (SI Appendix, Fig. S6), supporting the accuracy of the new structural delineation. The encoded AFGP polypeptide in E3 contains -22-550 (Thr-Ala-Pro/Ala) tripeptide repeats in different AFGP genes with occasional Thr/Arg substitutions (Fig. 2) or Thr/Lys substitutions in other AFGPs in this paper. Some of the Arg or Lys residues must serve as cleavage sites of the polyprotein precursor as intermittent Arg (or Lys) remain in the mature protein (23). A putative propeptide cd (hereon called the propeptide encoded by the 5′→3′ upstream sequence identities (64-75%) with the 27-nt GCA-rich duplicates that now flank each end of the repetitive (Thr-Ala-Ala), cds in functional AFGPs (Fig. 1, cyan segments; Fig. 3A and B, and SI Appendix, Fig. S8, cyan blocks). These four 27-nt duplicates share high sequence similarities with each other (Fig. 3C) indicating they resulted from the duplication of an initial copy. Within these four duplicates, we found multiple 9-nt sequence elements with an in silico translation of Pro/Ala-Ala-GCA repeats in the polyprotein precursor. The putative core promoter TATA box (red highlight), the Kozak consensus sequence ACCATGG (underlined blue letters), and the polyadenylation signal sequence AATAAAA (yellow highlight) are as indicated.

Fig. 2. An annotated functional AFGP polypeptide gene sequence from the polar cod B. saida (Bs). The color scheme for sequence features follows Fig. 1. An Arg (blue highlighted) or Lys in other AFGPs occasionally replaces Thr and may serve as the cleavage site of the polyprotein precursor. The putative core promoter TATA box (red highlight), the Kozak consensus sequence ACCATGG (underlined blue letters), and the polyadenylation signal sequence AATAAAA (yellow highlight) are as indicated.

Origination of AFGP-Coding Sequence from Noncoding DNA. A priori, the tandem Thr-Ala-Ala tripeptide-coding repeats in AFGP genes strongly suggest that the AFGP cd evolved from repeated duplications of an ancestral 9-nt Thr-Ala-Ala-coding element. We scrutinized the sequences of all of the AFGP genes and homologs from the seven gadids and discovered that the ancestral 9-nt element likely originated within a pair of conserved 27-nt GCA-rich duplicates that now flank each end of the repetitive (Thr-Ala-Ala), cds in functional AFGPs (Fig. 1, cyan segments; Fig. 3A and B, and SI Appendix, Fig. S8, cyan blocks). These four 27-nt duplicates share high sequence similarities with each other (Fig. 3C) indicating they resulted from the duplication of an initial copy. Within these four duplicates, we found multiple 9-nt sequence elements with an in silico translation of Pro/Ala-Ala-GCA repeats in the polyprotein precursor. The putative core promoter TATA box (red highlight), the Kozak consensus sequence ACCATGG (underlined blue letters), and the polyadenylation signal sequence AATAAAA (yellow highlight) are as indicated.

The homolog of the 27-nt GCA-rich duplicates exists in the AFGP-lacking gadids. It occurs upstream of the (Thr-Ala-Ala), like repeats in the M. merlangus AFGP pseudogene (Mm_AFGPP) and the T. esmarkii (Te_AFGP-like) sequence (Fig. 3A and SI Appendix, Fig. S7 A and B), whereas in L. lota and B. brosme, it proliferated as ~30-nt duplicates (Fig. 1, cyan segments and SI Appendix, Fig. S7 C and D). These tandem copies are characteristic of nonprotein-coding minisatelitellite DNA. Thus, the GCA-rich region was most likely a noncoding region that was prone to duplicative expansion in gadids. Its presence in the basal lotines L. lota and B. brosme indicates that it existed in the gadid ancestor before the emergence of the AFGP. The substantial variations in nucleotide sequences (Fig. 3A) and translated amino acids (SI Appendix, Fig. S6) in this region suggest a lack of functional constraint, which is consistent with it being a noncoding site in the gadid ancestor and in the extant species without AFGP. In the AFGP-bearing species, the propeptide encoded by the 5′-27 nt GCA-rich duplicate pair is not relevant for antifreeze activity, thus, its nucleotide sequence could also drift. However, it must be constrained against non-sense
and frameshift mutations such that a read-through ORF of the emerging downstream (Thr-Ala-Ala) cds could be maintained. This is, indeed, observed in all functional AFGPs, whereas AFGP pseudogenes have suffered frameshift mutations (SI Appendix, Figs. S6 and S8).

In short, the emergence of a functional AFGP polyprotein cds required only a single nucleotide change in the ancestral 9 nt GCA(A-la)-rich element for it to become the founder codons for one Thr-Ala-Ala tripeptide. Following this, microsatellite tandem duplications of the tripeptide cds unit would be required, but both processes could occur with relative ease. Sequence complexity could increase over time via additional 1-nt substitutions, such as GCA(A-la) to CCA(Pro), ACA(Thr) to AGA(Arg) (Fig. 2), or to AAA(Lys). Such substitutions occur throughout the AFGP genes sequenced in this paper. They create imperfect tandem repeats, which importantly serve to reduce sequence similarity between duplicons, preventing the expansion and contraction of the repeats by homologous recombination (24).

Formation of an In-Frame SP-Coding Sequence. As a secreted protein that functions in extracellular fluids to arrest the growth of invading ice crystals, all functional AFGPs have a proper SP cds (Fig. 2 and SI Appendix, Figs. S5 and S6). We examined the 5′ sequence region to determine how a SP cds was acquired. In sequence alignment, we discovered that the 5′ GCA-rich duplicate II of functional AFGPs lack a "T" nucleotide that is present in the consensus sequence and persists in M. merlangus AFGP pseudogenes (w) and the T. esmarkii noncoding AFGP homolog (Fig. 3A and C and SI Appendix, Fig. S8). Thus, this indel very likely resulted from a deletion event. The impact of this 1-nt deletion was that it produced a 1-nt reading frameshift in the presumptive propeptide (encoded by the 5′-27 nt duplicate pair), resulting in the upstream sequence that could supply a SP being linked with the downstream (Thr-Ala-Ala) cds in a single read-through ORF. The emerging AFGP gene was thus endowed with the necessary secretory signal.

Functionalization of the Emergent AFGP Gene. Forming proper coding regions of the AFGP gene alone would not lead to a gene product unless a minimal promoter was acquired to activate transcription thereby functionalizing the gene. All extant functional AFGPs have a TATA box, the core promoter for transcriptional activation, appropriately placed at 25–35 nt upstream from the presumptive transcription start site, but it is missing in the AFGP-like sequences of the basal species (Fig. 1 and SI Appendix, Fig. S9).

To uncover the origin of the promoter region, we examined the sequences of this upstream region in all sequenced species. As already noted above, all AFGP homologs share an ~240 nt 5′ region of high sequence similarities with functional AFGPs from the Met start codon through the SP cds (gray highlight, Fig. 1 and SI Appendix, Fig. S9), but only M. merlangus shares sequence similarities further upstream with the AFGP 5′ UTR and upstream regulatory region inclusive of the TATA box (SI Appendix, Fig. S9). The counterparts of this further upstream sequence in the basal gadine T. esmarkii and the lotines B. brosme and L. lota are drastically divergent and lack promoter elements (SI Appendix, Fig. S9, nt 1–120/130), but, interestingly, they share high similarities with each other. These results suggest the following. First, the divergent upstream sequences of the AFGP homologs shared by the basal T. esmarkii, B. brosme, and L. lota indicate they occupy a homologous genomic location (the putative ancestral site) that is distinct from the location of the extant AFGP genotype in the more derived M. merlangus and species of the AFGP-bearing clade. This is supported by a complete lack of microsynteny in our comparison of the sequence contigs of the AFGP-homolog loci of B. brosme with those of B. saida AFGP loci. We found none of the predicted neighboring genes are shared between the two species (SI Appendix, Table S1). Second, the acquisition of the proximal 5′ promoter region and functionalization of the emerging AFGP gene occurred after the divergence of the Trisopterus lineage. Since a (Thr-Ala-Ala)₆-like cds exists in T. esmarkii (SI Appendix, Fig. S7B), expansion of the Ala(GCA)-rich codons that could lead to an AFGP ORF likely began at the ancestral site before the recruitment of a cis-regulatory region. Without it, an emerging AFGP-like cds could not be transcribed and would remain as non-sense repetitive DNA in the Trisopterus and more basal lineages. We propose the possibility that the cis-promoter region was acquired in the most recent common ancestor of the AFGP-bearing clade through a stochastic translocation of the ancestral AFGP founder region to a new genomic site that happened to contain a TATA motif thereby conferring transcriptional capability. Although the specific mechanism of the translocation is currently unclear, cryptic transcriptional initiation sites and regulatory signals are deemed prevalent throughout genomes as increasing evidence suggests large portions of genomes become transcribed at some time (25). Regarding translational activation, all examined AFGP and homologs contain the Kozak consensus sequence ACCATGG for eukaryotic translation initiation (26) (Fig. 2 and SI Appendix, Fig. S9). Therefore, this motif likely...
existed in the founder genomic site and became functionalized when the promoter region was acquired.

**Nongenic Origin of SP and Promoter Region.** We experimentally verified that the AFGP SP cds and promoter sequence did not originate from any existing protein-coding genes in the gadid genome. We hybridized the genomic BAC library macroarrays of *B. saida* and *M. tomcod* with the AFGP 5′ probe that is specific to this region. The hybridized clones were exactly the same clones that hybridized to the (Thr-Ala-Ala)_n cds probe (SI Appendix, Fig. S10). This strongly supports that no homologs of the SP, 5′ UTR, and promoter regions of AFGP exist outside of the AFGP genomic loci. Thus, the promoter and SP cds of functional AFGP also originated de novo, unassociated with any preexisting protein gene.

**Further Verifications of Nongenic Origin of AFGP.** Recently evolved genes and the extant homologs of their genetic ancestor often remain as near neighbors in the genome. For example, the AFGP gene family of the Antarctic notothenioids closely clusters with its ancestral homologs: the trypsinogenlike protease genes along with the broader trypsin gene family within an ~400 kbp region (27). In contrast, we found none of the neighboring genes (e.g., MAK16 and RAB14) in the AFGP genomic regions of the three AFGP-bearing gadids *B. saida*, *G. morhua*, and *M. tomcod* share any sequence similarity with AFGPs (SI Appendix, Fig. S3), and, thus, they are evolutionarily unrelated to AFGP. The absence of a potential protein gene ancestor nearby is consistent with gadd AFGP having evolved de novo.

We further reasoned that an absence of transcription of the AFGP homologs in the AFGP-lacking gadids would provide compelling support that they are nonfunctional or nongenic DNA. Thus, we performed Northern blot hybridizations of RNA from pancreatic tissue [the site of AFGP synthesis (28)] of the four AFGP-lacking species using their respective species-specific AFGP-homolog sequence as probes and included *B. saida* for comparison. No transcripts of AFGP homologs were detectable in any of the three AFGP-lacking gadids. Only *B. saida* pancreatic RNA showed hybridization with strong intensity to its own AFGP cds probe and in varying intensity to the AFGP homolog probes from the other species due to various degrees of nt sequence identity (SI Appendix, Fig. S11). Since *L. lota*, *B. breviceps*, and *T. esmarkii* are basal to the AFGP-bearing clade, their AFGP homologs must represent the ancestral transcriptionally inactive noncoding form. The AFGP-lacking *M. merlangus* is nested within the AFGP-bearing clade, and its AFGP homolog most closely resembles a functional AFGP except for inactivating mutations in the (Thr-Ala-Ala)_n cds. Thus, it represents a subsequent non-functionalization into a nontranscribed pseudogene after the emergence of AFGP in the common ancestor of the AFGP-bearing clade. The loss of function relates to the nonfreezing water (Tromsø fjord in this study) *M. merlangus* inhabits today where antifreeze protection is not needed.

**Gadid AFGP Evolved from Entirely Nongenic DNA.** Fig. 4 summarizes the foregoing deductions on the noncoding origins of the essential AFGP's sequence components and the possible molecular steps in the evolutionary transformation of these components into a complete new functional AFGP. The AFGP founder structure (Fig. 4A) existed in the gadid ancestor as a short noncoding genomic sequence comprising a segment (~240 nt) with laterincoding exons (bronze segments) that have the potential to form a peptide sequence with properties for a secretory signal. The adjoining 27-nt GCA(Ala)-rich sequence (cyan segment) contained multiple nested 9-nt elements, any of which could become the three codons for the AFGP tripeptide (Thr-Ala-Ala) building block through a 1-nt substitution. Chance duplications of this ancestral 27-nt GCA-rich sequence produced four tandem copies (Fig. 4B). One of the 9-nt AFGP tripeptide-coding elements in the midst of the four copies likely underwent microsatellitelike duplications producing a budding ORF for the repetitive AFGP tripeptide cds, which began spreading the two pairs of 27-nt GCA-rich duplicates apart to the flanking positions (Fig. 4C). A putative translaction event in the last common ancestor of AFGP-bearing gadids moved the hitherto unexpressed AFGP precursor to a new genomic location that fortuitously contained a TATA motif thereby enabled transcription (Fig. 4D). Concurrently or subsequently, a 1-nt frameshift deletion in the second 5′-terminal 27-nt duplicate likely occurred and served to link the latent cds for the SP and the downstream AFGP (Thr-Ala-Ala)_n repeats in a single read-through ORF. Expression and secretion of this nascent antifreeze protein became possible (Fig. 4E). The smallest (and often the most abundant) functional AFGP isoform (AFGPs) comprises only four tripeptide repeats (10), which could be achieved through only two tandem duplications. The fledging antifreezing protection could, therefore, augment fitness in the individual at the onset of northern hemisphere marine glaciation. Subsequent intensification of environmental selection pressures likely drove the intragenic (Thr-Ala-Ala)_n cds expansion forming large AFGP polyprotein genes (Fig. 4F) as well as additional whole gene duplications. The result manifests in the multigene family of AFGP polyproteins (SI Appendix, Fig. S3) and the robust antifreeze activities the AFGP-bearing gadids possess today (10, 13).

**Proto-ORF Model of de Novo Evolution of Gadid AFGP.** The deduced evolutionary process of the gadid AFGP gene from non-sense DNA adds valuable insights into how adaptive functional genes could arise “from scratch.” The birth of de novo genes involves two fundamental events: the formation of an ORF and the acquisition of regulatory signals for transcription. In principle, these events could occur in either order. This prompted two major competing models: the protogene versus the proto-ORF model (3, 17, 29–31). The occurrence of a protogene is generally easier to detect as the de novo gene has a noncoding ortholog with demonstrable transcripts in the out-group species. Thus, the model has found ample support in studies that showed transcription preceded the emergence of an ORF (6, 17, 30, 32). The proto-ORF model states that an ORF was present before regulatory signals for expression were acquired. The existence of proto-ORFs is challenging to prove as they likely accumulate mutations that would interrupt the ORF before they could become transcribed for selection to act upon (3, 29, 30, 33). The history and mechanism of gadid AFGP evolution deduced in this paper (Fig. 4)
fits the proto-ORF model. This is because the (Thr-Ala-Ala)n-like repeats and SC cDs were formed in the basal lineage-lacking AFGP (represented by T. esmarkii) before the regulatory signal for transcription appeared in the more derived gadids in the AFGP-bearing clade (Fig. 1). Thus, we suggest that the recently evolved gadid AFGP serves as a clear and rare supporting example of the proto-ORF model of de novo gene birth.

Although the emergence of de novo genes has been well documented, the selective pressure and functional necessity that compelled their birth remain largely unknown (32). Most de novo genes are deemed unlikely to gain or retain function before their gene-like properties decay (5, 34) unless a timely major shift in the fitness landscape allows them to be sufficiently useful for selection to take hold. The de novo gadid AFGP is a rare example where the affecting fitness shift is clear. Its emergence correlated with strong risk-of-death selection pressures from environmental changes in the form of plunging ocean temperatures and formation of ice in the water column during the Pliocene/Pleistocene northern hemisphere glaciation. The gene multiplied in gadid species that remained in frigid habitats (e.g., B. saida, M. tomcod, and G. morhua) but gradually decayed in species that no longer experience the threat of freezing (e.g., M. merlangus).

Conclusion

We have characterized the evolutionary process and the details of the underlying molecular mechanisms through which all of the essential genetic components of the northern cod AFGP gene could have developed from noncoding DNA and the union of these emerging coding parts into a functional whole gene. We provide evidence that latent-coding components existed before the acquisition of the necessary cis-regulatory region for transcriptional activation. Thus, the gadid AFGP evolutionary history is a rare example supporting the proto-ORF hypothesis of de novo gene birth. With this paper, we fully resolved the last remaining question of how two unrelated groups of fish at opposite poles, the Antarctic notothenioid fishes and the northern codfishes, invented a near-identical AFGP. The notothenioid AFGP evolved within the structural framework of a preexisting gene ancestor but constructed a new AFGP cDs from de novo expansion of a rudimentary partly non-sense tripeptide-coding element. Northern gadid displayed even greater evolutionary ingenuity, constructing all parts of a functional AFGP gene entirely from noncoding DNA.

Materials and Methods

Detailed materials and methods are given in the SI Appendix, Materials and Methods. Briefly, we constructed large genomic DNA-insert BAC libraries for two AFGP-bearing gadids B. saida and M. tomcod and the AFGP-lacking basal B. brasme and smaller-insert phage libraries for three other AFGP-lacking species M. merlangus, T. esmarkii, and L. lota. The libraries were screened with radiolabeled probes derived from (Thr-Ala-Ala), cDs or the 5’ sequence of the AFGP gene to isolate clones containing AFGP or AFGP homologues, respectively. For the AFGP-bearing G. morhua, we deduced the Meloby/Clarck/BAC clones from published genome data and obtained them from a commercial vendor. The relevant positive clones or clone fragments were sequenced using various sequencing strategies, and the assembled sequences were analyzed. To correctly determine the gene structure of the functional AFGP, we obtained 5’ RACE map intron-exon junctions. To verify that the SP and promoter region evolved de novo, we screened the B. saida and M. tomcod BAC libraries with probes specific to this region to detect whether they hybridized elsewhere outside of the AFGP loci. We conducted Northern blot hybridizations to test for mRNA expression of AFGP homologues using species-specific probes to verify that the hypoxia were not from transcribed noncoding DNA. Fish collection and sampling followed University of Illinois at Urbana-Champaign institutional approved protocol as described in SI Appendix. All sequence data have been deposited in National Center for Biotechnology Information.

ACKNOWLEDGMENTS

We sincerely thank our colleagues Kim Praebel, Svein-Erik Fevolden, Arthur DeVries, Howard Reisman, Kevin Bilyk, Shannon Zellerhoff, as well as the Cornell University Biological Field Station for their kind assistance in collecting the gadid species in this study. We thank Jorgen Christiansen for the opportunity to participate in the TUNU cruises on the RV Helmer Hanssen to the Svalbard and East Greenland coasts to collect high Arctic species. We also thank Dr. Chris Amemiya and his previous lab member Andrew Stuart for their insightful advice on the BAC library construction and for making the lab facility available for our use. We also thank Konrad Meister, and Arthur DeVries for their help in editing the paper. This work was supported by the US National Science Foundation Grant DEB 0919496 (to C.H.C.).

1. Ohno S (1970) Evolution by Gene Duplication (George Allen & Unwin Ltd., London; Springer, New York).

2. Jacob F (1977) Evolution and tinkering. Science 196:1161–1166.

3. McLysaght A, Hurst LD (2016) Open questions in the study of genomic history of major adaptations in metazoan lineages. Mol Biol Evol 33:258–327.

4. Domazet-Lošo T, Brajković J, Tautz D (2007) A phylogenostatic approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 23:533–539.

5. Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat Rev Genet 12:692–702.

6. McLysaght A, Hurst LD (2016) Open questions in the study of genomic history of major adaptations in metazoan lineages. Mol Biol Evol 33:258–327.

7. Moyers BA, Zhang J (2015) Phylogenostatic bias creates spurious patterns of genome evolution. Mol Biol Evol 32:258–267.

8. Cheng C-HC (1998a) Evolution of the diverse antifreeze proteins. Curr Opin Genet Dev 8:715–720.

9. Devises AL, Cheng C-HC (2005) Antifreeze proteins and organismal freezing avoidance in polar fishes. The Physiology of Polar Fishes. Phil Trans R Soc B 370:269–282.

10. Devises AL, Cheng C-HC (2005) Antifreeze proteins and organismal freezing avoidance in polar fishes. The Physiology of Polar Fishes. Phil Trans R Soc B 370:269–282.

11. Liu Y, et al. (2007) Structure and evolutionary origin of Ca(2+)-dependent freezing herring type II antifreeze protein. PLoS One 2:e548.

12. Deng C, Cheng C-HC, Ye H, He X, Chen L (2010) Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc Natl Acad Sci USA 107:1595–21598.

13. Chen L, Devises AL, Cheng C-HC (1997a) Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc Natl Acad Sci USA 94:3817–3822.

14. Devises AL (1971) Glycoproteins as biological antifreeze agents in antarctic fishes. Proc Natl Acad Sci USA 107:1595–21598.

15. Chen L, Devises AL, Cheng C-HC (1997b) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic nototheniid fish. Proc Natl Acad Sci USA 94:3811–3816.

16. Cheng C-HC, Chen L (1999) Evolution of an antifreeze glycoprotein. Nature 401:443–444.

17. Schlötterer C (2015) Genes from scratch–The evolutionary fate of de novo genes. Trends Genet 31:215–219.