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

Xuan Zhuang[a,1,2], Chun Yang[a], Katherine R. Murphy[a], and C.-H. Christina Cheng[a,1]

[a]School 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 noncoding 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 noncoding 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 readthrough open reading frame (ORF). It became functionalized when a putative translocation event conferred the essential *cis* promoter for transcriptional initiation. We experimentally proved that all genic components of the extant gadid *AFGP* originated from entirely nongenic 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 phylostratigraphy 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 inoculative 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 sialic acid synthase followed by sequence tinkering by natural selection produced 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 genic 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.

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

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

[2]Present 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.1817138116/-/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 size 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 trypsinogenlike 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 straddled 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 genic 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 *AFGP*s 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 [*sensu* (18)] includes both AFGP-bearing and non-AFGP-bearing species; the former occurs in two subclades within the subfamily Gadinae (Fig. 1).

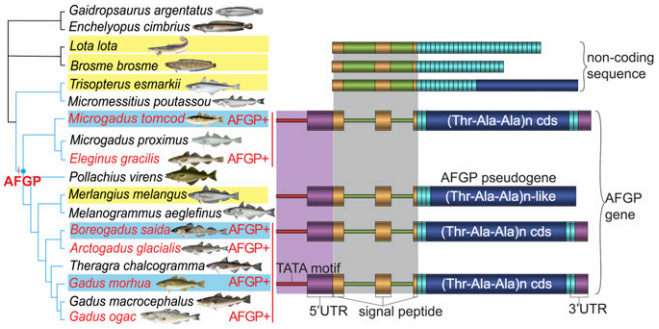

Fig. 1. Gadid phylogeny and AFGP gene/homolog structures. The phylogenetic tree of Gadidae is a congruent cladogram derived from Bayesian and maximum likelihood trees using complete ND2 gene sequences (*SI Appendix*, Fig. S1). Light blue branches indicate lineages of the Gadinae subfamily. The two gadine subclades containing AFGP-bearing species (red vertical bars), their most recent common ancestor (blue dot), and the emergence of the AFGP trait are as indicated. The three AFGP-bearing species (AFGP+) and four AFGP-lacking species analyzed in this paper are shaded in blue and yellow, respectively. The structure of their AFGP gene or nongenic homolog is shown to the right. Gray and purple shaded areas indicate homologous regions. Cyan segments are sequence repeats. The dark blue segment is a repetitive AFGP cds or AFGP-like sequence.

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 $(Thr\text{-}Ala\text{-}Ala)_n$ 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 *AFGP*s 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 propeptide and the long AFGP polyprotein. 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 *AFGP*s

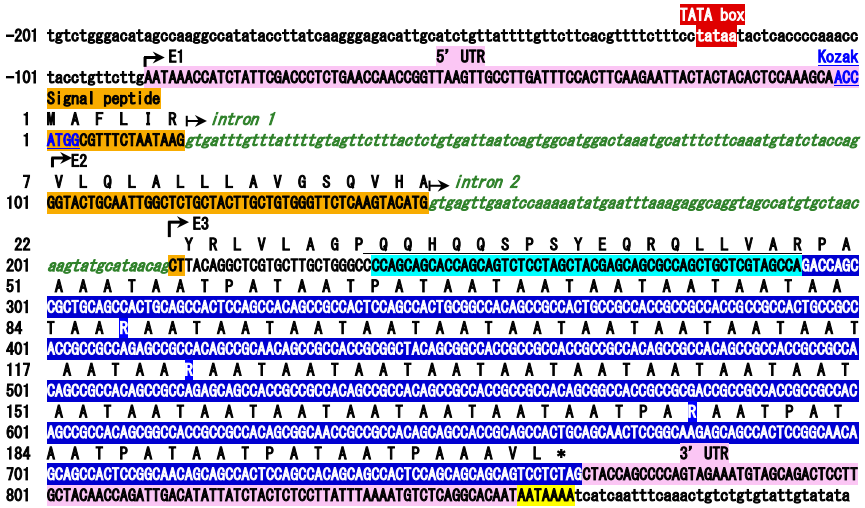

Fig. 2. An annotated functional AFGP polyprotein gene sequence from the polar cod *B. saida* (Bs). The color scheme for sequence features follows Fig. 1. An Arg (blue highlighted) (or Lys in other *AFGPs*) occasionally replaces Thr and may serve as the cleavage site of the polyprotein precursor. The 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.

from the three gadids (*SI Appendix*, Fig. S6), supporting the accuracy of the new structural delineation. The encoded AFGP polyprotein 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 *AFPGs* 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 cds rich in Gln(Q) codons (Fig. 2 and *SI Appendix*, Fig. S6) connects the SP and the AFGP tripeptide cds. It is absent in mature AFGP molecules and is presumably removed post-translationally.

To characterize *AFGP* genomic homologs from the AFGP-lacking gadids, we sequenced positive recombinant clones isolated from the smaller-insert genomic DNA phage libraries for *M. merlangus*, *T. esmarkii*, and *L. lota* and from a BAC library for *B. brosme*. The clones were isolated by screening the libraries with a probe specific to the 5′ region of a polar cod *AFGP* (nt-141–299 in Fig. 2) excluding $(Thr\text{-}Ala\text{-}Ala)_n$ cds (hereon called the *AFGP* 5′ probe). One representative sequence of the *AFGP* homolog from each species is shown in *SI Appendix*, Fig. S7. *AFGP* homologs from these four species share a 5′ region (~240 nt) of high nucleotide sequence identities (75–95%), which corresponds to the SP cds and the two introns of functional *AFGPs* (gray shading, Fig. 1) followed by a repetitive sequence of one or two varieties (cyan and dark blue segments, Fig. 1). The *M. merlangus* homolog most closely resembles a functional AFGP gene because it shares additional upstream sequence identities (64–74%) with the *AFGP* 5′ UTR and promoter region (purple shading, Fig. 1). However, its repetitive $(Thr\text{-}Ala\text{-}Ala)_n$-like cds (E3) contains various mutations that disrupt the tripeptide repeats (*SI Appendix*, Fig. S7*A*). These structural similarities and differences typify pseudogenization. The *AFGP* homologs of the other three: *T. esmarkii*, *L. lota*, and *B. brosme* lack the counterpart of the *AFGP* 5′UTR and promoter region (Fig. 1), indicating they are nontranscribed sequences. The repeat regions in the basal gadine *T. esmarkii* comprise a 5′ CAG (Q)-rich segment (cyan, Fig. 1 and *SI Appendix*, Fig. S7*B*) and a 3′ segment (dark blue, Fig. 1) that in silico translates as $(Thr\text{-}Pro\text{-}Ala_{(2\text{–}7)})_n$ repeats (*SI Appendix*, Fig. S7*B*). The homologs of the basal lotines *L. lota* and *B. brosme* contain tandem CAG(Q)-rich duplicates only (cyan, Fig. 1) and no $(Thr\text{-}Ala\text{-}Ala)_n$-like cds (*SI Appendix*, Fig. S7 *C* and *D*, respectively). The increasing deviation in sequence structure from the functional *AFGP* with increasing species evolutionary age allowed us to deduce the evolutionary origin and history of the gadid *AFGP*.

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 cds 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 $(Thr\text{-}Ala\text{-}Ala)_n$ cds in functional *AFGP*s (Fig. 1, cyan segments; Fig. 3 *A* and *B*, and *SI Appendix*, Fig. S8, cyan blocks). These four 27-nt duplicates share high sequence similarities with each other (Fig. 3*C*) 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-Ala (Fig. 3*D*). A chance 1-nt substitution (C → A or G → A) in the first position could give rise to the incipient three codons for the Thr-Ala-Ala tripeptide unit of AFGP (Fig. 3*D*). We hypothesize that, upon the onset of selective pressure from cold polar marine conditions, duplications of a 9-nt ancestral element in the midst of the four GCA-rich duplicates occurred. As the number of AFGP tripeptide-coding repeats increased, the antifreeze function would become augmented. With the expansion of the tripeptide cds, the 5′ and 3′ 27-nt duplicate pairs would be spread apart to their respective current flanking positions. The 5′ pair of duplicates became the cds of the Q(CAG)-rich (1-nt shift in the reading frame relative to the GCA repeats) propeptide (Fig. 3*A*). The 3′ end of the developing AFGP tripeptide cds was appropriately delimited by an existing in-frame termination codon (TAG) in the first 3′ duplicate (Fig. 3 *B* and *C*).

The homolog of the 27-nt GCA-rich duplicates exists in the AFGP-lacking gadids. It occurs upstream of the $(Thr\text{-}Ala\text{-}Ala)_n$-like repeats in the *M. merlangus* AFGP pseudogene (Mm_AFGPΨ) and the *T. esmarkii* (Te_AFGP-like) sequence (Fig. 3*A* 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 minisatellitelike 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. 3*A*) 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

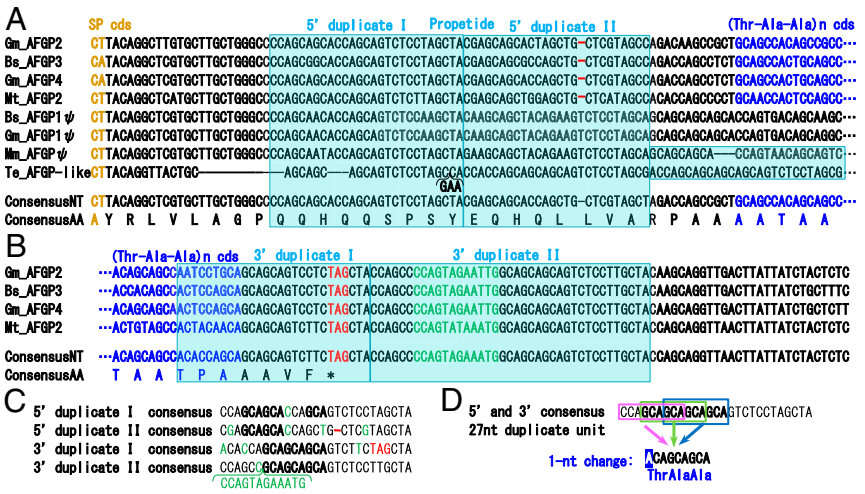

Fig. 3. Sequence components in the origination of the AFGP-coding region. Species names: Bs, *B. saida*; Gm, *G. morhua*; Mm, *M. merlangus*; Mt, *M. tomcod*; and Te, *T. esmarkii*. (*A* and *B*) Alignment of the 5′ and 3′ 27-nt GCA-rich duplicates region in representative *AFGP* or homologs from five species (full alignment in *SI Appendix*, Fig. S8). The consensus nt (ConsensusNT) and amino acid (ConsensusAA) sequences below each alignment are based on the full alignment. (*A*) Alignment of the 5′ duplicate pair (cyan shaded blocks). The putative ancestral T nucleotide persists in the 5′ duplicate II of the *T. esmarkii* noncoding homolog and AFGP pseudogenes (Ψ). T deletion (red dashes) produced a 1-nt frameshift linking the SP, propeptide, and the AFGP tripeptide cds in a single read-through ORF for an AFGP pre-proprotein. (*B*) Alignment of a 3′ duplicate pair (cyan shaded blocks). The *M. merlangus* pseudogene and *T. esmarkii* noncoding homolog have no 3′ duplicates. The 3′ duplicate II provided the termination codon (TAG, in red). (*C*) Alignment of consensus sequences of 5′ and 3′ duplicates. Nt variations among duplicates denoted in green. (*D*) The overall consensus 27-nt duplicate unit. Color rectangles frame the possible 9-nt elements that could become the incipient (Thr-Ala-Ala) codons via a single nt change in the first position.

and frameshift mutations such that a read-through ORF of the emerging downstream (Thr-Ala-Ala)$_n$ cds could be maintained. This is, indeed, observed in all functional *AFGP*s, 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(Ala)-rich element for it to become the founder codons for one Thr-Ala-Ala tripeptide. Following this, microsatellitelike 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(Ala) 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 *AFGP*s 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 *AFGP*s lack a "T" nucleotide that is present in the consensus sequence and persists in *M. merlangus* AFGP pseudogenes (ψ) and the *T. esmarkii* noncoding *AFGP* homolog (Fig. 3 *A* 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)$_n$ 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 *AFGP*s have a TATA box, the core promoter for transcriptional activation, appropriately placed at 25–30 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 *AFGP*s 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)$_n$-like cds exists in *T. esmarkii* (*SI Appendix*, Fig. S7*B*), 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 $(Thr\text{-}Ala\text{-}Ala)_n$ 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 *AFGP*s (*SI Appendix*, Fig. S3), and, thus, they are evolutionarily unrelated to *AFGP*. The absence of a potential protein gene ancestor nearby is consistent with gadid *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. brosme*, 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 $(Thr\text{-}Ala\text{-}Ala)_n$ cds. Thus, it represents a subsequent nonfunctionalization 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 forgoing deductions on the noncoding origins of the essential *AFGP* 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. 4*A*) existed in the gadid ancestor as a short noncoding genomic sequence comprising a segment (∼240 nt) with latent-coding 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. 4*B*). One of the 9-nt AFGP tripeptide-coding elements in the midst of

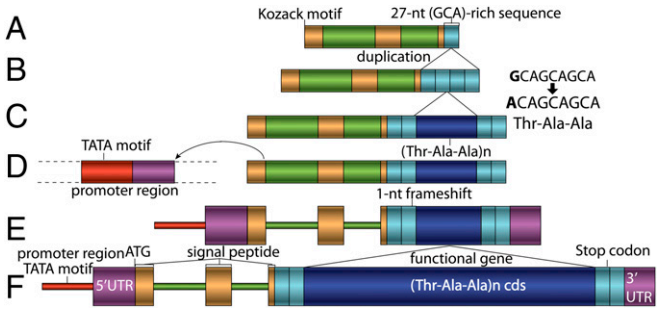

Fig. 4. Evolutionary mechanism of the gadid AFGP gene from noncoding DNA. The color codes of the sequence components follow Fig. 1. (*A*) The ancestral noncoding DNA contained latent signal peptide-coding exons with a 5′ Kozak motif, adjacent to a duplication-prone 27-nt GCA-rich sequence. (*B*) The 27-nt GCA(Ala)-rich sequence duplicated forming four tandem copies. (*C*) A 9-nt in the midst of the four 27-nt duplicates became the three codons for one AFGP Thr-Ala-Ala unit and underwent microsatellitelike duplication forming a proto-ORF. (*D*) A proximal upstream regulatory region acquired through a putative translocation event. (*E*) A 1-nt frameshift led to a contiguous SP, a propeptide, and a Thr-Ala-Ala-like cds in a read-through ORF. (*F*) Intragenic $(Thr\text{-}Ala\text{-}Ala)_n$ cds amplification, fulfilling the antifreeze function under natural selection.

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. 4*C*). A putative translocation 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. 4*D*). Concurrently or subsequently, a 1-nt frameshift deletion in the second 5′ 27-nt duplicate likely occurred and served to link the latent cds for the SP and the downstream AFGP $(Thr\text{-}Ala\text{-}Ala)_n$ repeats in a single read-through ORF. Expression and secretion of the nascent antifreeze protein became possible (Fig. 4*E*). The smallest (and often the most abundant) functional AFGP isoform (AFGP8) comprises only four tripeptide repeats (10), which could be achieved through only two tandem duplications. The fledgling 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 $(Thr\text{-}Ala\text{-}Ala)_n$ cds expansion forming large AFGP polyprotein genes (Fig. 4*F*) 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 SP 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 genelike 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 genic components of the northern cod AFGP gene could have developed from noncoding DNA and the union of these emerging coding parts into a new 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 lasting 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. brosme* 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)$_n$ cds or the 5′ sequence of the AFGP gene to isolate clones containing *AFGP* or *AFGP* homologs, respectively. For the AFGP-bearing *G. morhua*, we deduced the AFGP-positive 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 rescreened 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* homologs using species-specific probes to verify the hypothesis that they are untranscribed 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 Jørgen Christiansen for the opportunity to participate in the TUNU cruises on the R/V 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. Special thanks go to Melody Clark, Lloyd Peck, 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.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, Guerzoni D (2015) New genes from non-coding sequence: The role of *de novo* protein-coding genes in eukaryotic evolutionary innovation. *Phil Trans R Soc B* 370: 20140332.
4. Domazet-Lošo T, Brajković J, Tautz D (2007) A phylostratigraphy 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 *de novo* genes: What, how and why. *Nat Rev Genet* 17:567–578.
7. Guerzoni D, McLysaght A (2016) *De novo* genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biol Evol* 8:1222–1232.
8. Moyers BA, Zhang J (2015) Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol* 32:258–267.
9. Cheng C-HC (1998a) Evolution of the diverse antifreeze proteins. *Curr Opin Genet Dev* 8:715–720.
10. DeVries AL, Cheng C-HC (2005) Antifreeze proteins and organismal freezing avoidance in polar fishes. *The Physiology of Polar Fishes*, eds Farrell AP, Steffensen JF (Elsevier Academic Press, San Diego), Vol 22, pp 155–201.
11. Liu Y, et al. (2007) Structure and evolutionary origin of Ca$^{(2+)}$-dependent 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* 107:21593–21598.
13. Chen L, DeVries 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. DeVries AL (1971) Glycoproteins as biological antifreeze agents in antarctic fishes. *Science* 172:1152–1155.
15. Chen L, DeVries AL, Cheng CHC (1997b) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc Natl Acad Sci USA* 94:3811–3816.
16. Cheng CHC, 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.
18. Teletchea F, Laudet V, Hänni C (2006) Phylogeny of the Gadidae (sensu Svetovidov, 1948) based on their morphology and two mitochondrial genes. *Mol Phylogenet Evol* 38:189–199.
19. Osuga DT, Feeney RE (1978) Antifreeze glycoproteins from Arctic fish. *J Biol Chem* 253:5338–5343.
20. Hew CL, Slaughter D, Fletcher GL, Joshi SB (1981) Antifreeze glycoproteins in the plasma of Newfoundland Atlantic cod (*Gadus morhua*). *Can J Zool* 59:2186–2192.
21. Fletcher GL, Hew CL, Joshi SB (1982) Isolation and characterization of antifreeze glycoproteins from the frostfish, *Microgadus tomcod*. *Can J Zool* 60:348–355.
22. Zhuang X, Yang C, Fevolden S-E, Cheng CH (2012) Protein genes in repetitive sequence-antifreeze glycoproteins in Atlantic cod genome. *BMC Genomics* 13:293.
23. O'Grady SM, Schrag JD, Raymond JA, Devries AL (1982) Comparison of antifreeze glycopeptides from Arctic and Antarctic fishes. *J Exp Zool* 224:177–185.
24. Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* 22:253–259.
25. Clark MB, et al. (2011) The reality of pervasive transcription. *PLoS Biol* 9:e1000625.
26. Kozak M (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* 15:8125–8148.
27. Nicodemus-Johnson J, Silic S, Ghigliotti L, Pisano E, Cheng CHC (2011) Assembly of the antifreeze glycoprotein/trypsinogen-like protease genomic locus in the Antarctic toothfish Dissostichus mawsoni (Norman). *Genomics* 98:194–201.
28. Cheng CC, Cziko PA, Evans CW (2006) Nonhepatic origin of notothenioid antifreeze reveals pancreatic synthesis as common mechanism in polar fish freezing avoidance. *Proc Natl Acad Sci USA* 103:10491–10496.
29. Andersson DI, Jerlström-Hultqvist J, Näsvall J (2015) Evolution of new functions *de novo* and from preexisting genes. *Cold Spring Harb Perspect Biol* 7:a017996.
30. Reinhardt JA, et al. (2013) *De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet* 9:e1003860.
31. Tautz D (2014) The discovery of *de novo* gene evolution. *Perspect Biol Med* 57:149–161.
32. Carvunis A-R, et al. (2012) Proto-genes and de novo gene birth. *Nature* 487:370–374.
33. Zhao L, Saelao P, Jones CD, Begun DJ (2014) Origin and spread of *de novo* genes in *Drosophila melanogaster* populations. *Science* 343:769–772.
34. Palmieri N, Kosiol C, Schlötterer C (2014) The life cycle of *Drosophila* orphan genes. *eLife* 3:e01311.