EVOLUTIONARY BIOLOGY

Evolution of a chordate-specific mechanism for myoblast fusion

Haifeng Zhang[1], Renjie Shang[1,2], Kwantae Kim[3], Wei Zheng[4,5], Christopher J. Johnson[3], Lei Sun[6], Xiang Niu[7], Liang Liu[8,9], Jingqi Zhou[2], Lingshu Liu[2], Zheng Zhang[1], Theodore A. Uyeno[10], Jimin Pei[11], Skye D. Fissette[12], Stephen A. Green[13], Sukhada P. Samudra[2], Junfei Wen[1], Jianli Zhang[14], Jonathan T. Eggenschwiler[2], Douglas B. Menke[2], Marianne E. Bronner[13], Nick V. Grishin[11,15], Weiming Li[12], Kaixiong Ye[2,9], Yang Zhang[4,5], Alberto Stolfi[3]*, Pengpeng Bi[1,2]*

Vertebrate myoblast fusion allows for multinucleated muscle fibers to compound the size and strength of mononucleated cells, but the evolution of this important process is unknown. We investigated the evolutionary origins and function of membrane-coalescing agents Myomaker and Myomixer in various groups of chordates. Here, we report that *Myomaker* likely arose through gene duplication in the last common ancestor of tunicates and vertebrates, while *Myomixer* appears to have evolved de novo in early vertebrates. Functional tests revealed a complex evolutionary history of myoblast fusion. A prevertebrate phase of muscle multinucleation driven by Myomaker was followed by the later emergence of Myomixer that enables the highly efficient fusion system of vertebrates. Evolutionary comparisons between vertebrate and nonvertebrate Myomaker revealed key structural and mechanistic insights into myoblast fusion. Thus, our findings suggest an evolutionary model of chordate fusogens and illustrate how new genes shape the emergence of novel morphogenetic traits and mechanisms.

INTRODUCTION

A fundamental step in vertebrate muscle development is the fusion of mononucleated myoblasts to form multinucleated myofibers (*1*). Generation of syncytial myofibers allows concerted power outputs to fulfill complex locomotor functions and therefore was likely instrumental for the adaptive radiation of vertebrates. Myomaker (MymK) and Myomixer (MymX) are two recently identified muscle-specific fusogens that drive plasma membrane coalescence during vertebrate myoblast fusion (*2–6*). Deletion of either gene causes perinatal lethality of mice due to fusion defects resulting in muscle malfunction (*2*, *3*). Moreover, forced expression of this duo confers fusogenic activity even onto fibroblasts, which are not normally capable of undergoing cell fusion (*3*).

Here, we report the identification and characterization of MymX and MymK orthologs outside of jawed vertebrates. We demonstrate that the fusogenic activity of MymK likely evolved in the last common ancestor of tunicates and vertebrates (Olfactores) and therefore predates the origin of MymX, which appears to have evolved

de novo specifically in the vertebrate lineage to facilitate the massive multinucleation of skeletal muscles. Coculturing mammalian cells expressing either vertebrate or tunicate MymK revealed that MymK/MymX synergy primarily depends on the presence of either component on a different cell (i.e., in trans). Together, our study provides a crucial insight into the still poorly understood evolutionary and molecular mechanisms underlying vertebrate myogenesis.

RESULTS

Evolutionary origins of MymK

The phylum Chordata is composed of vertebrates together with two nonvertebrate subphyla: Tunicata and Cephalochordata (Fig. 1A). Cephalochordates have mononucleated muscles indicating no myoblast fusion (fig. S1A) (*7*), whereas tunicates exhibit limited multinucleation of certain muscles (*8*) and vertebrates have extensive, obligatory multinucleation (fig. S1B). Therefore, we reasoned that comparative gene function studies of these closely related animal groups might shed insights into the evolutionary history and cellular mechanisms of myoblast fusion.

Originally known as *Tmem8c*, *MymK* belongs to a gene family that in vertebrates also contains two other paralogs: *Tmem8a* and *Tmem8b*. Homology-guided searches revealed that multiple tunicate species have both *MymK* and a *Tmem8a/b*-like gene (herein named *Tmem8*-related) (Fig. 1, A and B). In the cephalochordate *Branchiostoma floridae*, only a single *Tmem8* family gene could be identified. *Tmem8* sequences are found in diverse eukaryotes, including the unicellular filasterean *Capsaspora owczarzaki* (Fig. 1B and fig. S2A). Comparisons of these proteins revealed an epidermal growth factor–like domain that exists in all Tmem8 family proteins except MymK, which appears to have lost this domain (fig. S2B). Therefore, the duplication of an ancestral *Tmem8* gene likely gave rise to *Tmem8a/b* and *MymK* before tunicates and vertebrates diverged (fig. S2C). Although the exact timing of this duplication event cannot be determined (fig. S2C), the lack of multinucleated

[1]Center for Molecular Medicine, University of Georgia, Athens, GA, USA. [2]Department of Genetics, University of Georgia, Athens, GA, USA. [3]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA. [4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [5]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA. [6]The Fifth People's Hospital of Shanghai, and Shanghai Key Laboratory of Medical Epigenetics, Institutes of Biomedical Sciences, Fudan University, Shanghai, China. [7]Tri-Institutional Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, USA. [8]Department of Statistics, University of Georgia, Athens, GA, USA. [9]Institute of Bioinformatics, University of Georgia, Athens, GA, USA. [10]Department of Biology, Valdosta State University, Valdosta, GA, USA. [11]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [12]Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA. [13]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [14]College of Engineering, University of Georgia, Athens, GA, USA. [15]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA.
*Corresponding author. Email: pbi@uga.edu (P.B.); alberto.stolfi@biosci.gatech.edu (A.S.)
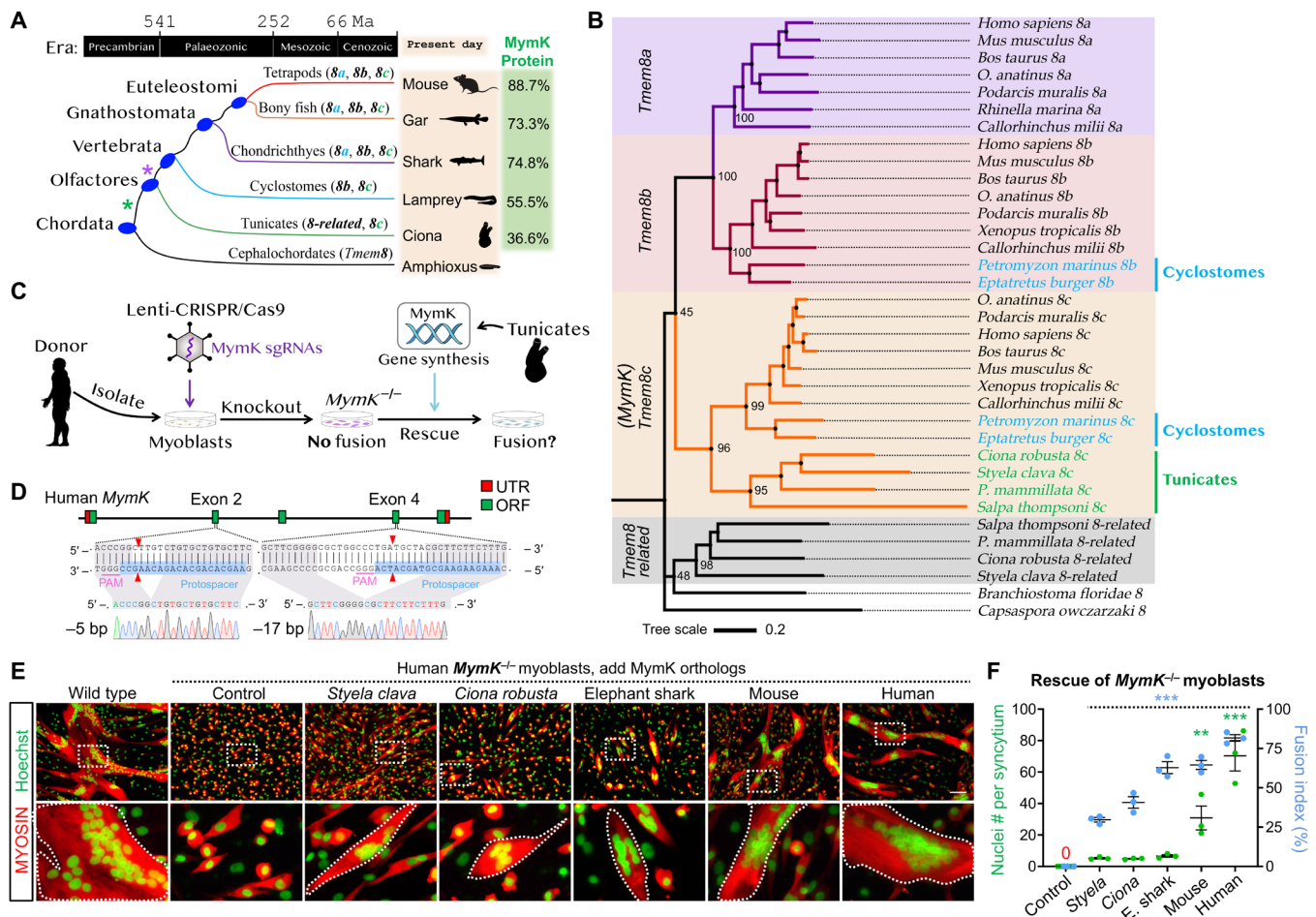
Fig. 1. Tunicate MymK orthologs have weak fusogenic function in vertebrate myoblasts. (**A**) Phylogenetic relationships of various chordate clades used to deduce the evolutionary origins of the *MymK* gene (also known as *Tmem8c*). Asterisks represent two potential duplication events of *Tmem8* genes that give rise to *8-related*, *8a*, *8b*, and *8c* members. Sequence identities of MymK orthologs were compared to human MymK protein. Scale at the top shows approximate date in millions of years (Ma) ago. (**B**) Phylogeny of the Tmem8 gene family inferred by a distance-based method (neighbor joining). The bootstrap percentages were obtained from 1000 replicates. *O. anatinus, Ornithorhynchus anatinus*; *P. mammillata, Phallusia mammillata*. Tmem8 gene members from jawless vertebrates were highlighted in blue and tunicates in green. Extended phylogenetic analysis is seen in fig. S2A. (**C**) Schematic of experimental design to test the fusogenic function of tunicate MymK proteins in human *MymK*⁻/⁻ myoblasts. (**D**) Human *MymK* gene structure, sgRNA positions, and genotyping results that showed biallelic frameshift mutations induced by CRISPR/Cas9. bp, base pair; UTR, untranslated region. (**E**) Myosin immunostaining of human *MymK*⁻/⁻ myoblasts transfected with MymK orthologs. Muscle syncytia (outlined) were observed in nonvertebrate (*Styela* and *Ciona*) MymK expression groups, although smaller than the syncytia induced by vertebrate MymK proteins. Scale bar, 100 μm. (**F**) Measurements of myoblast fusion after 4 days of myogenic differentiation. E. shark, elephant shark. Data are means ± SEM. **$P < 0.01$ and ***$P < 0.001$, compared to control group, one-way analysis of variance (ANOVA).

muscles in cephalochordates and other deuterostomes (fig. S2A) suggested a functional link between muscle multinucleation and the presence of *MymK* in olfactorians (tunicates + vertebrates). Although multinucleation is also a prominent feature of arthropod musculature (*9–11*), the *MymK* gene is absent from this phylum, suggesting convergent evolution of myoblast fusion through different molecular mechanisms.

Functional comparisons between vertebrate and nonvertebrate MymK proteins

We identified *MymK* orthologs in various tunicates, including benthic ascidians and pelagic thaliaceans, ranging from ~26 to 30% amino acid identity in alignments with human MymK (fig. S3, A and B). When expressed in human myoblasts (fig. S3C), tunicate MymK protein can be specifically detected in the membrane fraction

(fig. S3D). To examine their functional conservation as fusogens, we devised a heterologous rescue approach (Fig. 1C). First, we used CRISPR to inactivate *MymK* in myoblasts isolated from different vertebrate species (human, Fig. 1, D to F; mouse, fig. S4, A to D; lizard, fig. S4, E to H), which completely abolished syncytializations (Fig. 1E and fig. S4, C and H). We then expressed tunicate MymK proteins and assayed their ability to rescue the fusion of these MymK-deficient cells. All tunicate MymK orthologs tested can consistently rescue the fusion of *MymK*⁻/⁻ myoblasts, albeit with lower levels of efficiency than vertebrate proteins (Fig. 1F and figs. S4, D and H, and S5). Consistent with the neofunctionalization of *MymK*, tunicate Tmem8-related and cephalochordate Tmem8 proteins did not elicit fusogenic activity (fig. S6).

Although MymK is necessary and sufficient for vertebrate myoblast fusion, a second membrane protein called MymX synergistically

enhances the fusogenic activity of MymK during myogenesis (*6*). Despite extensive searching (Materials and Methods), we were not able to identify MymX homologs in tunicates or any other non-vertebrate species. Thus, we postulated that fusogenic activity of tunicate MymK is independent of MymX. To test this idea, we generated human *MymX/MymK* double knockout (KO) myoblasts by CRISPR. In the absence MymX, tunicate and human MymK induce comparable levels of human myoblast fusion, supporting its conserved function (fig. S7). However, a functional difference between human and tunicate MymK was unmasked by resupplying MymX. Specifically, coexpression of human MymX + human MymK induced massive fusion (fig. S7). Such synergy was not observed when human MymX was paired with tunicate MymK (fig. S7). These results suggest that, although the fusogenic role of MymK predates the emergence of vertebrates and the *MymX* gene, vertebrate-specific changes to MymK were essential for the evolution of functional synergy with MymX.

Temporally and spatially restricted expression of *MymK* drives multinucleation program of *Ciona* muscle

Having established the fusogenic activity of tunicate MymK in vertebrate cells, we next asked whether it plays a role in the development of multinucleated myofibers in the laboratory model tunicate, *Ciona robusta.* The presence of *MymK* in tunicates was intriguing because these nonvertebrate chordates also have multinucleated muscles (*12*, *13*). While the tail muscles from tunicate larvae are mononucleated, the siphon and body wall muscles of postmetamorphic juveniles and adults are composed of multinucleated fibers (*14*). The presence of multinucleated siphon muscles in other tunicate species also correlates with the presence of the *MymK* gene (Fig. 2B). Notably, *MymK* is absent from appendicularians, which have secondarily lost multinucleated siphon and body wall muscles (*15*). In contrast, *MymK* orthologs were found in all other tunicate species with multinucleated siphon or body wall muscles (Fig. 2B).

In *Ciona*, expression of *MymK* was observed exclusively in multinucleated juvenile muscles by in situ hybridization (Fig. 2C) and by a *MymK* promoter green fluorescent protein (GFP) reporter (Fig. 2D and fig. S8, C and D). In contrast, *MymK* expression was not observed in any other cell type including mononucleated larval tail muscle cells (fig. S8A). This was further confirmed by reanalyzing published single-cell transcriptome data (*16*, *17*) collected at different developmental stages, in which we detected *MymK* expression specifically in multinucleated muscle precursor cells (Fig. 2, E and F, and fig. S8B). Together, these results suggest that expression of the *MymK* gene is highly specific to multinucleated muscles in tunicates.

We then performed *MymK* loss-of-function experiments in *Ciona* using tissue-specific CRISPR mutagenesis (*18*) in the cardiopharyngeal mesoderm lineage that gives rise to the multinucleated muscles of the atrial siphon (Fig. 2G and fig. S9A) (*19*). In control juveniles, circular atrial siphon myofibers invariably formed as orderly rings with occasional longitudinal myofibers emanating from the siphon region (Fig. 2G, fig. S9B, and movie S1). In contrast, *MymK* CRISPR resulted in highly disorganized atrial siphon muscles (Fig. 2G; fig. S9, C and D; and movie S2). Moreover, there was a reduction in the frequency of binucleated atrial siphon/longitudinal myofibers in *MymK* CRISPR juveniles (Fig. 2H), suggesting that *MymK* is required for myoblast fusion in *Ciona*. However, overexpression of MymK in mononucleated larval tail muscle cells did not promote obvious multinucleation [fig. S10; 16 hours post-fertilization (hpf)

control, movie S3; 16 hpf *MRF* > *MymK*, movie S4]. This suggests that, as in vertebrates (*6*), the fusogenic activity of MymK in *Ciona* likely requires other factor(s) present in juvenile but not larval tail muscle cells.

A distantly related *MymX* sequence from lamprey genome can replace its mammalian orthologs in enhancing myoblast fusion

Cyclostomes such as lampreys and hagfish diverged from jawed vertebrates (gnathostomes) ∼500 million years ago (*20*). Histological analysis revealed extensive multinucleation of sea lamprey (*Petromyzon marinus*) muscle (Fig. 3A), which can host up to several hundred myonuclei per fiber, a stark contrast to maximally a few dozen in tunicates (*8*). We hypothesized that a protein with MymX function exists in lamprey to robustly induce myoblast fusion in cooperation with MymK.

The search for MymX orthologs is intrinsically challenging due to the small size (<100 amino acids) and high frequency of substitutions (Fig. 3B). Nonetheless, iterative BLAST (basic local alignment search tool) searches identified one hit from a genome shotgun sequence (GenBank: AEFG01021847.1) of the sea lamprey. Alignment of RNA sequencing (RNA-seq) reads revealed a single-exon open reading frame (ORF) that encodes 583 amino acids including the hydrophobic AxLyCxL motif (*21*) that is essential for mammalian MymX function (Fig. 3, B and C, and fig. S11, A to C). A homologous sequence was also found from arctic lamprey (*Lethenteron camtschaticum*, APJL01015224), revealing an ORF of 595 amino acids that shares 93% identity with the sea lamprey sequence (fig. S12). The complete ORF of sea lamprey MymX was codon-optimized, cloned by gene synthesis, and expressed in human myoblasts. Western blot readily detected a 70-kDa band specifically from the membrane fraction (Fig. 3D). Immunofluorescence revealed the presence of lamprey MymX on the cell surface (Fig. 3E and fig. S11D), suggesting a function in this compartment.

Given the low sequence identity between lamprey and gnathostome MymX, we examined its function in a heterologous rescue experiment. While *MymX*$^{-/-}$ myoblasts are weakly fusogenic due to the residual activity of MymK in these cells (*6*), the expression of lamprey MymX robustly enhanced cell fusion (human, Fig. 3, F and G; mouse, fig. S13). Similar to mammalian MymX, this fusion-promoting activity of lamprey MymX strictly requires MymK for function because it failed to induce fusion when *MymK* was deleted from human myoblasts (fig. S14).

Lamprey-specific C terminus from MymX is indispensable for optimal fusogenic activity

For lamprey MymX, only a short region of 52 amino acids at the N terminus (N52) can be aligned to conventional orthologs (Fig. 3B). However, expression of the N52 polypeptide failed to induce myoblast fusion (Fig. 4, A to C). A similarly deleterious effect was observed when the conserved AxLyCxL motif was removed (Fig. 4, A to C). We continued to dissect the function of its unusually long extracellular C-terminal sequence by generating a series of sea lamprey *MymX* mutants (Fig. 4A). As the region of deletions enlarged, MymX function, quantified as nuclei number per syncytium, gradually diminished (Fig. 4C and fig. S15). Therefore, the optimal activity of sea lamprey MymX requires its large, nonconserved C-terminal structure.

We next investigated the expression pattern of *MymX* and *MymK* during lamprey muscle development. Sea lampreys have a complex
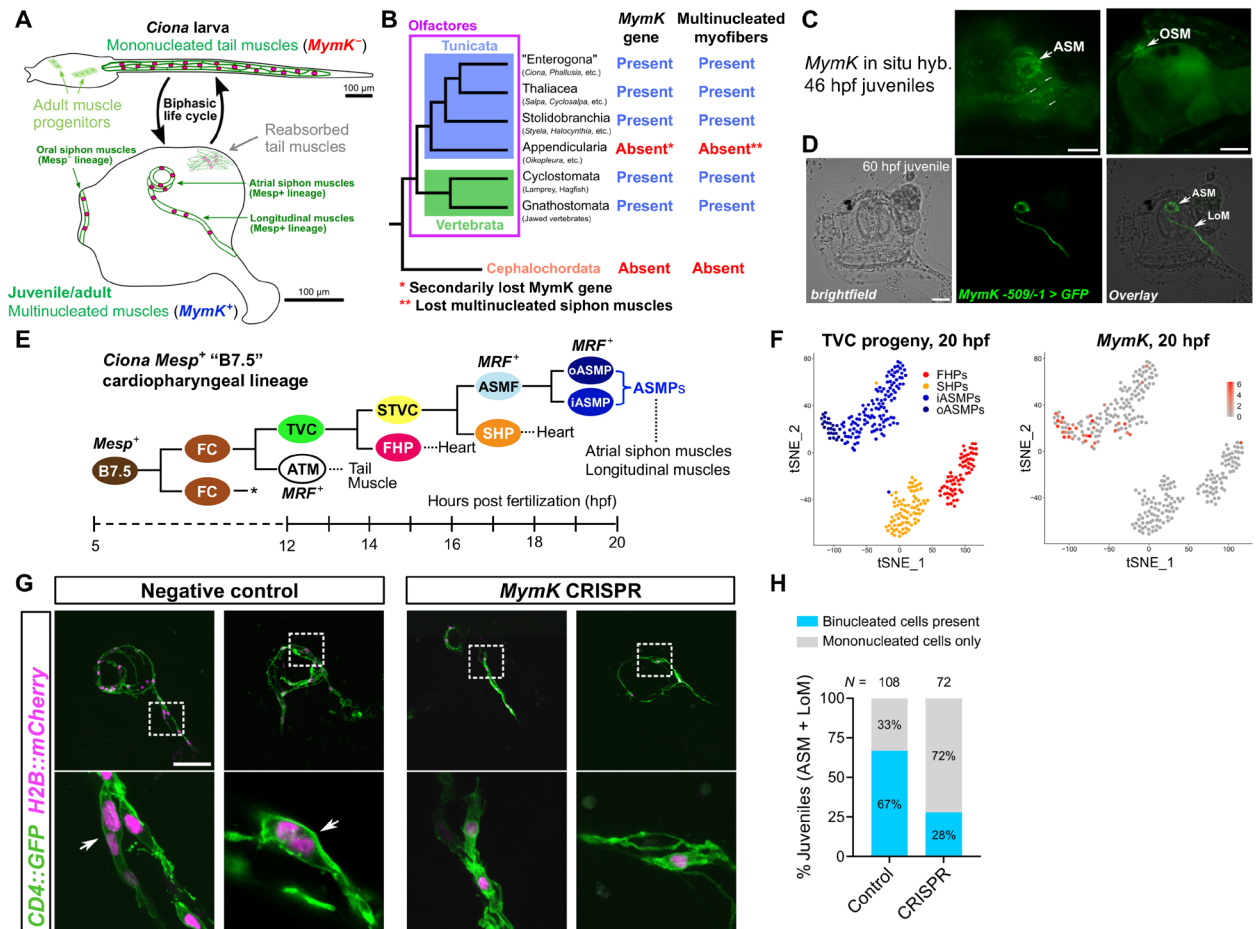
Fig. 2. MymK is required for multinucleation of postmetamorphic muscles in the tunicate *Ciona*. (**A**) Diagram of biphasic life cycle of ascidians (sessile tunicates) like *Ciona*. The motile larvae have strictly mononucleated tail muscles during the dispersal phase. After settlement and metamorphosis, tail muscle cells undergo programmed cell death and are reabsorbed, while dedicated muscle progenitors set aside in the larva differentiate to form the multinucleated siphon and body wall muscles of the juvenile. Muscles surrounding and emanating from the oral and atrial siphons are derived from distinct cell lineages in the larva. Only those from the atrial siphon are derived from the Mesp⁺ B7.5 lineage [in (E)]. (**B**) Cladogram of extant chordates showing correlation between the presence of *MymK* gene and muscle multinucleation in different clades. (**C**) Whole-mount mRNA in situ hybridization showing *MymK* expression in developing atrial siphon muscle (ASM) and oral siphon muscle (OSM) cells in metamorphosing juveniles. Smaller arrows indicate autofluorescent tunic cells. (**D**) *C. robusta* juvenile developed from a zygote transfected with a *MymK* promoter reporter plasmid, labeling ASMs and longitudinal body wall muscles (LoM). (**E**) Diagram of the B7.5 lineage in *C. robusta*, based on conclusions from (*18*). FC, founder cell; TVC, trunk ventral cell; ATM, anterior tail muscle; STVC, secondary TVC; FHP, first heart precursor; SHP, second heart precursor; ASMF, atrial siphon muscle founder cell; ASMP, atrial siphon muscle precursor; oASMP, outer ASMP; iASMP, inner ASMP. Asterisk indicates that both FCs give rise to identical lineages. *MRF*, *myogenic regulatory factor* (*MyoD* ortholog). (**F**) *t*-distributed stochastic neighbor embedding (tSNE) plots based on information from (*16*) showing *MymK* expression mapped onto TVC progeny clusters at 20 hpf. *MymK* is expressed exclusively in ASMPs and especially enriched in outer ASMPs. Abbreviations same in (E). (**G**) Representative Z-projection confocal fluorescence images of 84 hpf negative control (transfected with *Mesp > Cas9* only, no sgRNAs) juveniles alongside same-age juveniles in which *MymK* was targeted for mutagenesis specifically in the B7.5 lineage. *MymK* CRISPR: zygotes transfected with *Mesp > Cas9* and *U6 > MymK*-sgRNA vectors. Muscle plasma membranes and nuclei labeled by *MRF > CD4::GFP* and *MRF > H2B::mCherry*, respectively. Arrows in negative control panels showing development of typical binucleated myofibers that is inhibited upon *MymK* CRISPR. (**H**) Data from scoring of juveniles represented in (G) showing reduced frequency of binucleated atrial siphon/longitudinal myofibers in *MymK* CRISPR juveniles. *N*, numbers of juveniles assayed for each condition. Scale bars, 50 μm.

life cycle that involves a freshwater-based larval period of 2 to 10 years, followed by metamorphosis into a marine-based adult stage. It is unclear when myoblast fusion occurs in lampreys, although it was reported that muscle cells from young larvae remained mononucleated (*22*). Because of their complex life history, we could not examine larvae of defined age or stage to identify the temporal window of myoblast fusion in sea lamprey. Instead, we compared groups of larvae of uncertain age (estimated 2 to 3.5 years of age) but of different sizes (Fig. 4D), assuming that multinucleation might be occurring

primarily during muscle growth (*23*, *24*). Moderate multinucleation was consistently observed in muscles of both groups (Fig. 4D), although a higher number of nuclei per myofiber was associated with larger muscle and body size (Fig. 4, D′ and D″). Last, expression of both *MymX* and *MymK* was detected in larval muscles of both sea and arctic lampreys but not in other tissues or adult muscles (Fig. 4E and fig. S16).

Together, our expression and functional data suggest that these distantly related lamprey sequences are true orthologs of MymX
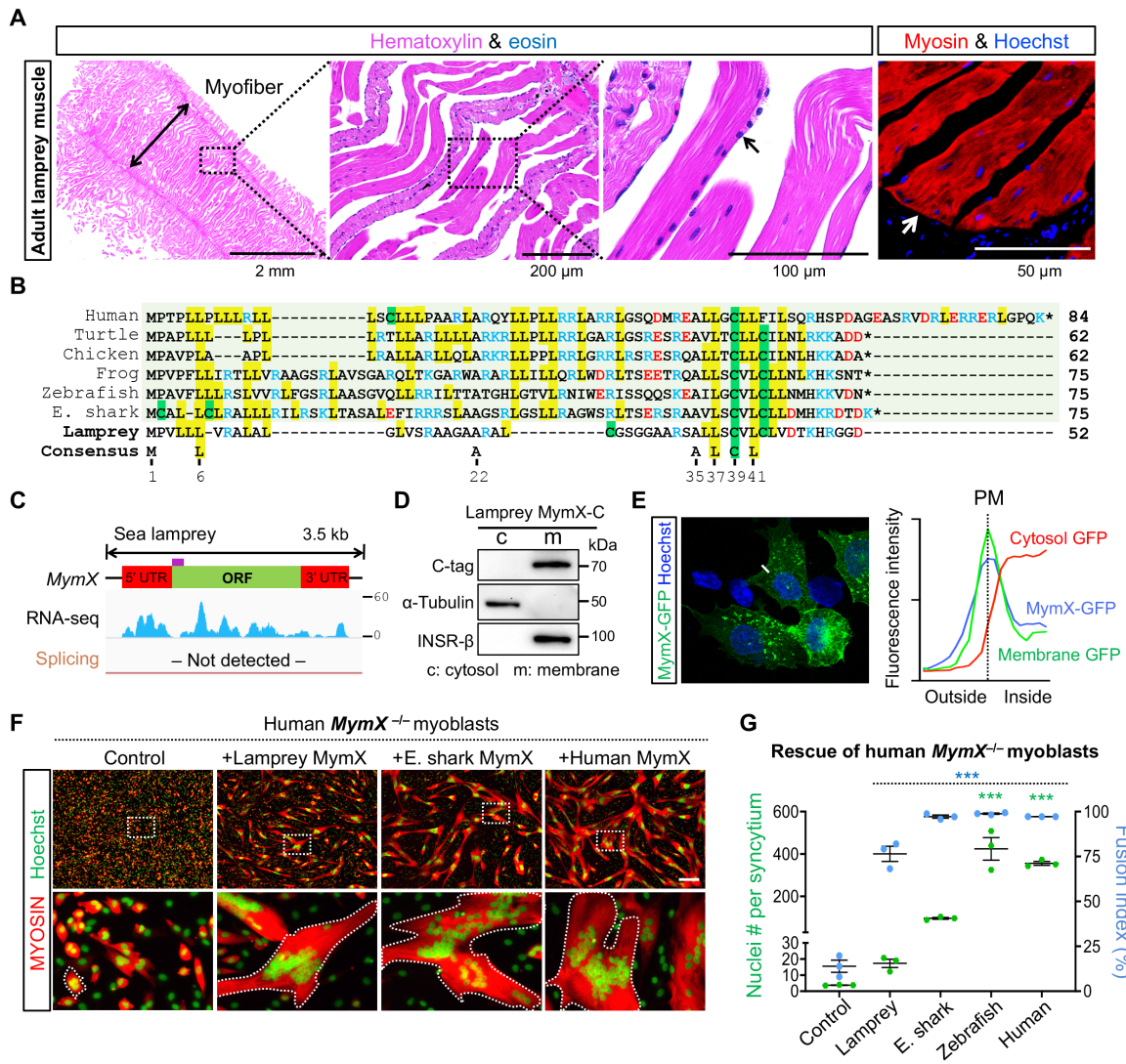
Fig. 3. Discovery of the unusual *MymX* genes from lampreys. (**A**) Histological staining and immunofluorescence of muscle tissues dissected from adult sea lamprey (*P. marinus*). Multinucleated myofibers (arrows) are observed from the longitudinal sections. (**B**) Cross-species homology of lamprey MymX aligned with its orthologs from jawed vertebrates. Only a few residues from the AxLyCxL motif and the N terminus can be aligned. x denotes leucine, valine, or isoleucine, and y denotes serine, threonine, or glycine. The numbers below the consensus sequence refer to the positions in sea lamprey MymX (only the N-terminal 52 amino acids are shown). (**C**) RNA-seq tracks that confirmed transcription of *MymX* gene in sea lamprey (sequence read archive accession: PRJNA497902). No splicing junction was detected in the hypothetical ORF. The purple rectangle highlights the coding region of the N-terminal 52 amino acids of lamprey MymX shown in (B). (**D**) Western blot analyses of cytosolic (c) and membrane (m) fractions of human myoblasts transfected with C-tagged lamprey MymX. C-tag is a small four–amino acid peptide tag E-P-E-A. α-Tubulin blot was used as a positive control of cytosolic proteins. Insulin receptor β (INSR-β) blot was used as a positive control of membrane proteins. (**E**) Human myoblasts transfected with lamprey MymX-GFP fusion protein. Nuclei were counterstained with Hoechst dye. The fluorescence intensity cross the plasma membrane (white bar in the image) was measured. Membrane and cytosol targeting GFPs were added as measurement controls (see images in fig. S11D). (**F**) Myosin immunostaining of human *MymX*$^{-/-}$ myoblasts transfected with MymX orthologs. Note that sea lamprey MymX can rescue fusogenic defects of human *MymX*$^{-/-}$ cells that formed larger muscle syncytia (outlined) than control (empty vector). E. shark, elephant shark. Scale bar, 100 μm. (**G**) Measurement of myoblast fusion in (F) after 4 days of differentiation. Data are means ± SEM. ***$P < 0.001$, compared to control group, one-way ANOVA.

and that the functional cooperativity between MymX and MymK in myoblast fusion is likely to be conserved in cyclostomes. Because MymX does not appear to share homology with any other protein and because our extensive in silico TBLASTN (translated BLAST) search did not identify an AxLyCxL motif containing sequence of interest from genome/transcriptome of multiple tunicate species or any invertebrates, we propose that *MymX* is a vertebrate-specific orphan gene encoding a core molecular component of myoblast

fusion that arose de novo before the split between jawed and jawless vertebrates (fig. S17).

Insights into mechanisms of myoblast fusion obtained from evolutionary comparisons

We next sought to leverage these newly identified muscle fusogens from different chordate groups for insights into the mechanisms of myoblast fusion, which remain poorly understood. Because
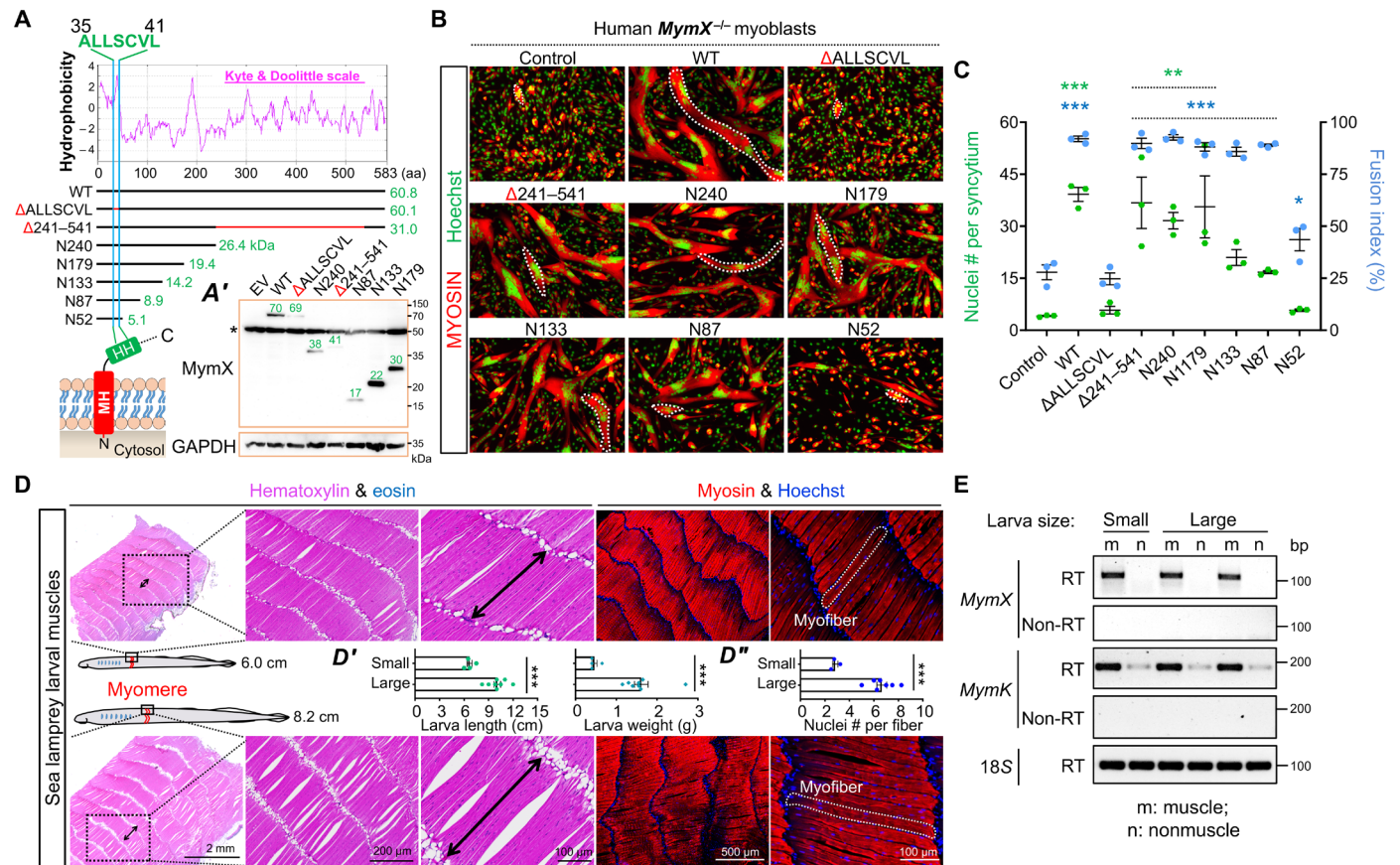
Fig. 4. Mutation and expression analysis of lamprey MymX protein. (**A**) Hydrophobicity map of sea lamprey MymX and a schematic of mutants. Red lines highlight deleted regions. HH, hydrophobic helix; MH, membrane-anchor helix; aa, amino acid; GAPDH, glyceraldehyde-3-phosphate dehydrogenase. (*A'*) Western blot results that confirmed expression of lamprey MymX mutants in human myoblasts. MymX was detected by blotting a diminutive C-tag fused at the C terminus of target. The predicted and detected molecular weights are labeled on the schematics and Western blots, respectively. The four–amino acid epitope tag (E-P-E-A) is 0.4 kDa. Quantifications of blots are seen in fig. S15B. EV, empty vector. (**B**) Myosin immunostaining of human *MymX*$^{-/-}$ myoblasts transfected with full length [wild type (WT)] or truncated lamprey MymX proteins. (**C**) Measurement of myoblast fusion in (B) after 4 days of myogenic differentiation. (**D**) Staining of longitudinal sections of muscles dissected from sea lamprey larvae of two different size groups to identify muscle fusion stage. Measurements of larva body length and weight (*D'*) and nuclei number per myofiber (*D''*). (**E**) Reverse transcription PCR results that validated the muscle-specific expression pattern of *MymX* and *MymK* genes in sea lamprey larval muscle tissues. m, muscle cDNA; n, nonmuscle (intestine and liver) cDNA.

the emergence of MymX and its functional cooperativity with MymK represent a key step in the evolution of vertebrate myogenesis, we focused on dissecting the mechanism of MymX-MymK synergy.

Our previous fusion reconstitution assays revealed that MymK protein is needed in the plasma membrane of both cells undergoing fusion, whereas adding MymX to only one side is sufficient to boost the efficiency of fusion (Fig. 5A) (*6*). This raised the question of whether MymX promotes the function of MymK in trans (between the two cells' membranes) or cis (in the same membrane). The discovery of tunicate MymK proteins that are unable to synergize with mammalian MymX allowed us to design a novel experiment testing the cis/trans basis of the MymX-MymK synergy between myoblasts expressing either tunicate or human MymK (Fig. 5A). Unexpectedly, we found that MymX was only capable of significantly promoting cell fusion when expressed in trans to a cell expressing mammalian MymK protein. In contrast, fusion was not significantly enhanced when mammalian MymK and MymX were expressed in cis (Fig. 5, B and C). This potential in trans synergy is consistent with

the topology of MymX protein where the conserved AxLyCxL motif found in all MymX proteins is located on the extracellular side, where it might be able to interact with MymK and/or other factors on the opposing membrane in trans.

We next sought to better understand the structural basis of both conserved and divergent MymK functions across Chordata. By applying a deep neural network-based structure assembly method (*25–27*), we obtained seven structural models for MymK proteins from representative species of vertebrates and tunicates (fig. S18, A and B). MymK proteins from all taxonomic groups share >80% similarity of the overall structure (fig. S18C) in which seven transmembrane (TM) helices are arranged in an anticlockwise manner when viewed from the extracellular space (Fig. 5D). As part of TM1, the N-terminal residues protrude toward the outside of the cell and form the extracellular face together with three extracellular loops (fig. S18, D and E). The structure of C-terminal residues is disordered and forms the intracellular face together with three intracellular loops (fig. S18, D and E). The TM helices of MymK enclose an internal cavity that goes through the entire structure with a small intracellular
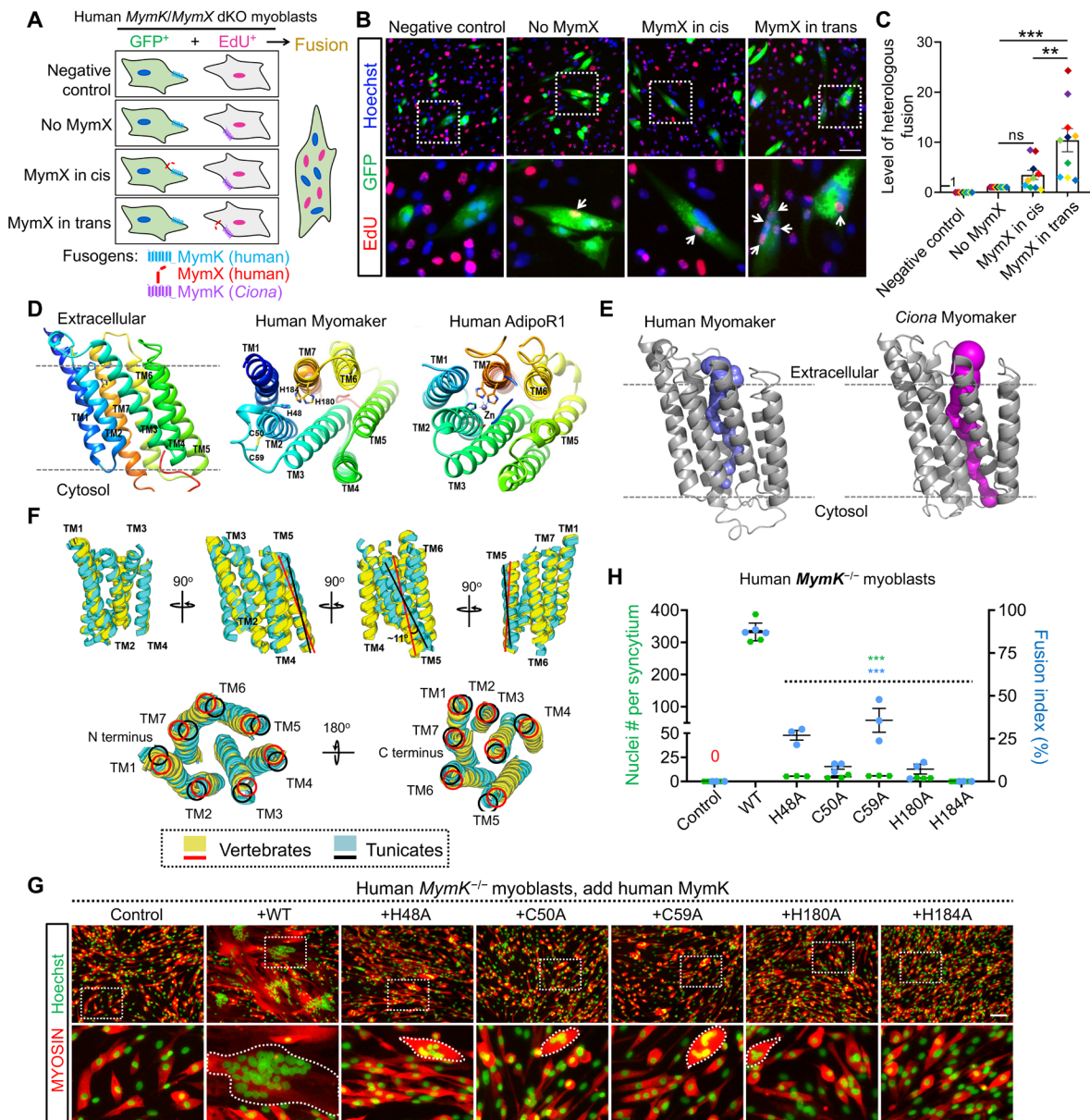
Fig. 5. Evolutionarily distinct Mymk proteins reveal mechanistic insights into structure function and synergy. (**A**) Schematic of experiment design. Note that the basal level of myoblast fusion requires MymK to be present in both cells, while the expression of vertebrate MymX can only boost vertebrate MymK (e.g., human) but not nonvertebrate (e.g., *Ciona*) MymK activity. The natural uncoupling between MymX synergy and fusogenicity observed in tunicate MymK permits the test of vertebrate MymX/MymK synergy using cell mixing cultures. dKO, double knockout. (**B**) Representative fluorescence images of human myoblasts after mixing culture as illustrated in (A). Arrows point to the EdU+ nuclei inside GFP+ cells formed from fusion. Scale bar, 100 μm. (**C**) Measurement of heterologous fusion by counting EdU+ nuclei inside GFP+ syncytia. Data were normalized to the "no MymX" group. Data from the same replicate were highlighted in the same color. $N = 10$. **$P < 0.01$; ns, not significant. (**D**) Ribbon representation of the predicted human MymK structure. The conserved histidine and cysteine residues on human MymK model are highlighted. Zinc-binding motif of adiponectin receptor 1 (AdipoR1; PDB ID: 6KRZ) was shown on the right. TM, transmembrane helix. (**E**) Side views of the predicted cavities inside MymK proteins. (**F**) Superimpositions of the overall structural models for MymK proteins from vertebrates (human, mouse, zebrafish, and elephant shark) and tunicates (*Phallusia*, *Ciona*, and *Styela*). The orientations of TM5 show obvious shifts between the two taxonomic groups. (**G**) Myosin immunostaining of human *MymK*−/− myoblasts that revealed the fusogenic activity of human MymK mutants. Cells were differentiated for 4 days. Scale bar, 100 μm. (**H**) Measurement of myoblast fusion in (G) after myogenic differentiation and compared to WT expression group. Data are means ± SEM. ***$P < 0.001$, one-way ANOVA.

opening and a larger extracellular opening (Fig. 5E). Unsupervised comparison clustered the predicted structures into consistent taxonomic groups (fig. S18C). The major structural differences between tunicate and vertebrate MymK are on the protein surfaces (fig. S18D) and the orientation of TM5 helix, which is tilted by 11° relative to

TM5 in tunicates (Fig. 5F), hinting at a potentially important role of this structural adaption for the synergy with MymX.

Last, our comparative three-dimensional protein modeling predicted a close resemblance between MymK and adiponectin receptor (AdipoR) structures (Fig. 5D and fig. S18F) (*28*, *29*). Stabilization of

AdipoR structure requires a zinc ion coordinated by three histidine (His) residues (Fig. 5D). We identified a similar motif in MymK, near the outer lipid layer of the membrane (His[48], His[180], and His[184]). In addition, two cysteine (Cys) residues from the TM2 (Cys[50]) and extracellular loop 1 (Cys[59]) are predicted to form a disulfide bond. These histidine and cysteine residues are perfectly conserved in all tunicate and vertebrate MymK orthologs, suggesting a crucial contribution to the structure and function of MymK. Mutating these residues in human MymK drastically affected its fusogenic activity (Fig. 5, G and H). In summary, by looking at both conserved and divergent features, we have gained new insights into the mechanisms of MymK and MymX function in myoblast fusion.

DISCUSSION

Our comparative study of MymK and MymX in multiple chordate (vertebrate and nonvertebrate) species sheds light on the evolution of myoblast fusion (Fig. 6). Whereas the *MymK* gene was certainly generated through duplication of an ancestral *Tmem8* gene, *MymX*, as an orphan gene, might have arisen de novo. Our data suggest that tunicate MymK can promote myoblast fusion in both tunicate and vertebrate cells but is unable to synergize with vertebrate MymX proteins to augment fusion levels. In contrast, lamprey MymX

function is conserved enough to synergize with human MymK, despite its highly divergent length and sequence. Together, our data are consistent with a de novo origin of MymX after the tunicate-vertebrate split.

One scenario for the origin of the *MymX* gene could be through a transitory protogene that produced a short polypeptide, given the short length (<100 amino acids) of MymX proteins in most vertebrates. After cyclostomes and gnathsostomes had diverged, MymX may have been secondarily elongated in lampreys (583 amino acids in sea lamprey and 595 amino acids in arctic lamprey). Alternatively, the ancestral MymX protein was closer in size to that of extant lampreys but was secondarily reduced in length in jawed vertebrates. Notably, possibly attributed to the fact that the hagfish (*Eptatretus burgeri*) genome is not complete (30), a *MymX* ortholog has yet to be identified in this species.

This potential stepwise evolution of myoblast fusion in chordates lends support to an updated "new head/new heart" hypothesis (31, 32), which postulates that the active predatory lifestyle of early vertebrates was made possible due to increased sensory capabilities, a chambered heart, and a muscularized pharynx, all derived from mostly cephalic neural crest or cardiopharyngeal progenitor cells. In this context, the evolution of MymK may have been a key innovation of the last common ancestor of vertebrates and tunicates.
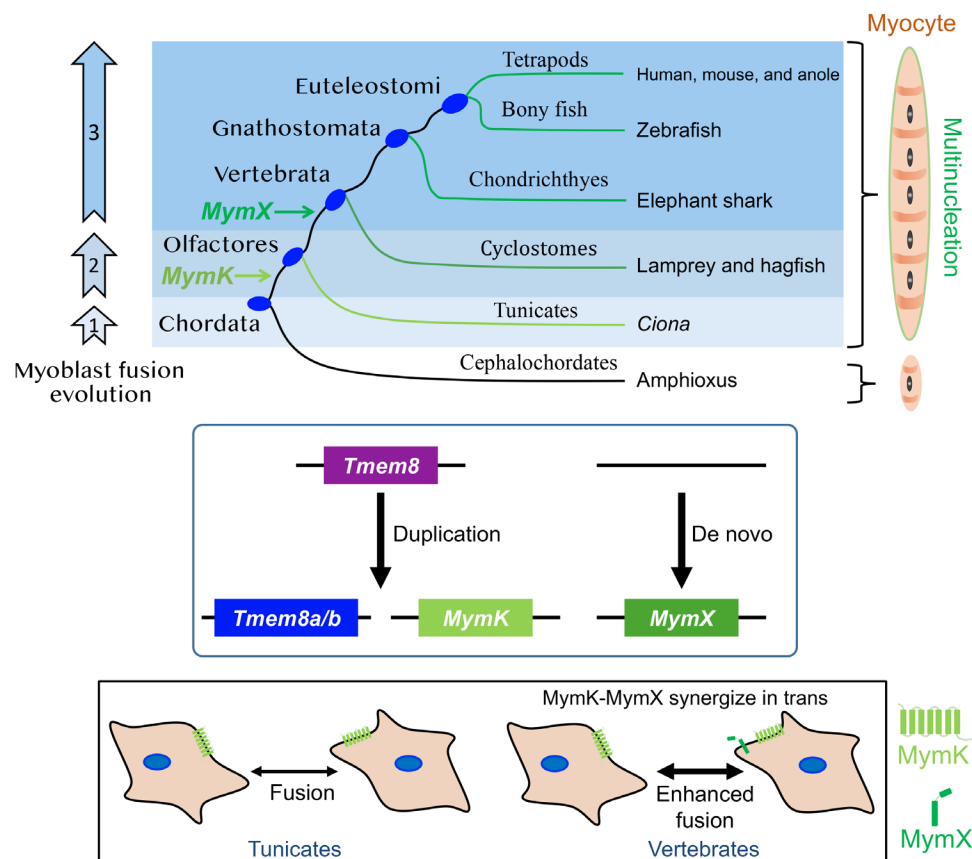
Fig. 6. Evolution and mechanism of chordate-specific control system of myoblast fusion. The emergence of *MymK* gene after duplication of the eukaryotic *Tmem8* gene allowed the multinucleation of muscle cells for common ancestors of tunicates and vertebrates. The emergence of MymX and structural adaptions of MymK proteins drove the extensive and obligatory fusion in vertebrates. Note that Tmem8a/b is called Tmem8-related in tunicates simply due to poor phylogenetic resolution, according to tunicate gene nomenclature rules. In vertebrates, this gene became duplicated again to give rise to Tmem8a and Tmem8b. Evolutionary comparison of vertebrate and tunicate MymK revealed that MymK/MymX synergy primarily depends on the presence of either component on a different cell (i.e., in trans).

Both tunicates and vertebrates have a prominently muscularized pharynx, which may have been facilitated by the emergence of MymK as a driver of myoblast fusion. However, additional enhancement of the myoblast fusion pathway in vertebrates would have been made possible by the de novo evolution of MymX after the split from tunicates. The higher number of nuclei per myofiber seen in vertebrates compared to tunicates suggests that a more "active lifestyle" may have required stronger, longer myofibers formed by more constituent myoblasts. However, considering gene loss is a pervasive source of genetic variation in evolution (33, 34), it is also possible that *MymX* (and by extension, MymK-MymX synergy) arose earlier in chordate evolution but was later lost in tunicates as this group secondarily evolved a biphasic life cycle with a sessile adult phase.

In addition to clarifying the evolutionary history of MymK and MymX, our comparative approach also yielded key insights into the molecular mechanisms of these chordate-specific fusogens. Taking advantage of the natural uncoupling between fusogenicity and cooperativity with MymX observed with tunicate MymK, our cis/trans tests suggest that the functional synergy between MymK and MymX is likely most important in trans. We propose that, while MymK is needed in both cells undergoing fusion, MymX synergizes with MymK in trans and not in cis. Last, our structural modeling and sequence alignment of diverse MymK proteins revealed key residues that are hyperconserved across all chordates and crucial for MymK function. Although it remains unknown whether the functional synergy of MymX/MymK involves a direct physical interaction or indirectly through engaging additional fusion-promoting factors in the same pathway (6), our structure-function analysis of both factors should provide a key basis for the complete understanding of the biophysical mechanism of myoblast fusion.

In summary, our study closes long-existing gaps in the evolutionary history of myoblast fusion, an important developmental mechanism that independently evolved in chordates and ultimately facilitated much of vertebrate evolution. Furthermore, we show that by combining a broad, comparative "evo-devo" approach with genetic interrogation of protein function and in silico protein structure modeling, we can advance our understanding of the molecular and evolutionary mechanisms of key developmental processes.

MATERIALS AND METHODS
Animal husbandry
Standard operating procedures for transporting, maintaining, handling, and euthanizing of sea lamprey and hagfish were approved by the Institutional Committee on Animal Use and Care of Michigan State University and California Institute of Technology, Valdosta State University, and University of Georgia and in compliance with standards defined by the National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Sea lampreys were trapped in tributaries of Lakes Huron and Michigan by the U.S. Fish and Wildlife Service and Fisheries and Oceans Canada. Captured lampreys were transported to the U.S. Geological Survey, Hammond Bay Biological Station (HBBS), Millersburg, Michigan and held in 200- to 1000-liter tanks that were continually fed with ambient temperature, aerated Lake Huron water. Adult lampreys also were transported to the California Institute of Technology where lamprey husbandry was performed as previously described (35) in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, and

protocols were approved by the Institutional Animal Care and Use Committee of the California Institute of Technology (lamprey, protocol no. 1436-17). Adult lamprey muscle was fixed in 4% paraformaldehyde (PFA) and processed using conventional histology.

To produce sexually mature ovulated females and males for embryo collection, sea lampreys were transferred to the Ocqueoc River, Millersburg, Michigan and held in cages (0.5 m^3) constructed of polyvinyl chloride and polyurethane mesh, allowing natural sexual maturation in a riverine environment. Sea lampreys were checked daily for sexual maturation. Sexually mature males were identified by applying abdominal pressure and checking milt expression (36). Sexually mature females were identified by applying abdominal pressure and checking for ovulated oocyte expression along with visual observation of secondary sexual characteristics (37). Sexually mature males and female lampreys were returned to HBBS where they were held until used for collecting and culturing lamprey embryos as previously outlined (35). Embryo viability was determined using techniques established for evaluation of the sterile male release program in the Laurentian Great Lakes (38). Embryos were checked daily for viability, and dead embryos were removed from holding containers. Embryos were pooled together for individual samples according to Piavis stages.

Female Atlantic hagfishes (*Myxine glutinosa*, Linnaeus, 1758) were used in this study (specimen/mass/length; #1/64 g/45 cm; #2/57 g/41 cm; #3/55 g/43 cm). Live specimens were collected at Shoals Marine Laboratory (Appledore Island, ME) and transported to Valdosta State University. Specimens were euthanized using 400 mg of MS222 (Finquel anesthetic, Argent Chemicals, Redmond WA) and 200 mg of NaHCO$_3$ (pH buffer) mixed in 1 liter of filtered artificial seawater. An incision was then made along the ventral midline to collect tissue specimens for the histological analysis. Preserved amphioxus and shark specimens were obtained from VWR (470001-802 and 470001-486). Subsequent paraffin processing, embedding, sectioning, and hematoxylin and eosin staining were performed by standard procedures.

Human and mouse cell cultures
Human myoblasts (hSkMC-AB1190) were isolated and immortalized as previously published (39). These cells were cultured in 15% fetal bovine serum (FBS) (GemCell, 100-500) and 5% growth medium supplement mix (PromoCell, C-39365) in skeletal muscle cell basal medium (PromoCell, C-23260) with GlutaMAX and 1% gentamicin sulfate. Mouse 10T1/2 fibroblasts [American Type Culture Collection (ATCC), CCL-226] and C2C12 myoblasts (ATCC, CRL-1772) were maintained in 10% FBS with 1% penicillin/streptomycin (Gibco, 15140122) in DMEM (Dulbecco's modified Eagle's medium–high glucose, D5796). Myoblast differentiation medium contained 2% horse serum in DMEM with 1% penicillin/streptomycin. Cells have passed mycoplasma test using the Universal Mycoplasma Detection Kit (ATCC, 30-1012 K).

Lizard cell culture and CRISPR experiments in lizard myoblasts
Myogenic single clones (myosin heavy chain+) were isolated from immortalized *Anolis sagrei* embryonic cells ASEC-1 (*A. sagrei* embryonic cell line 1; to be described in detail elsewhere). ASEC-1 and clonally derived myoblasts were cultured in DMEM supplemented with glutamine and 10% FBS (with penicillin/streptomycin and amphotericin B) and cultured at 29°C and 5% CO$_2$.

For CRISPR/Cas9–mediated *MymK* KO experiment, Cas9 and guide RNA were transfected to lizard myoblasts using a Lipofectamine LTX Plus kit (Thermo Fisher Scientific, A12621). pSpCas9(BB) plasmid was a gift from F. Zhang (Addgene, plasmid no. 62988) (*40*). Puromycin (25 μg/ml) selection was performed for 24 hours starting from 48 hours after transfection. Single clone was isolated and allowed to expand. *MymK* genotypes for each clone were analyzed by polymerase chain reaction (PCR), followed by Sanger sequencing. Sequences for single guide RNAs (sgRNAs) and genotyping PCR primers are provided in table S1.

Lentivirus preparation and CRISPR experiments in human and mouse myoblasts

sgRNAs that target the coding regions of human and mouse *MymX* and *MymK* genes were individually cloned into the Lenti-CRISPR v2 vector and validated by Sanger sequencing. Lenti-CRISPR v2 vector was a gift from F. Zhang (Addgene, plasmid no. 52961) (*41*). sgRNA sequences are provided in table S1.

For lentivirus production, Lenti-X 293 T cells (Clontech, 632180) were cultured in DMEM (containing 1% penicillin/streptomycin and 10% FBS). Transfection was performed using FuGENE 6 (Promega, E2692) with psPAX2 and pMD2.G plasmids. At 48 hours after transfection, the lentivirus supernatant was collected, filtered, and concentrated by the Lenti-X Concentrator (Clontech, PT4421-2) following the manufacturer's protocol. psPAX2 vector was a gift from D. Trono (Addgene, plasmid no. 12260). pMD2.G vector was a gift from D. Trono (Addgene, plasmid no. 12259). Human and mouse myoblasts were infected by lentivirus in growth medium. Human *MymX*/*MymK* double KO myoblast line was generated from a *MymX*^KO clone by infecting lenti-CRISPR MymK sgRNAs. Single clone was isolated, expanded, and genotyped by PCR and Sanger sequencing. Sequences for genotyping PCR primers are provided in table S1. Human *MyoD*^−/− myoblasts were generated and authenticated in a previous study.

Retroviral vector preparations and gene expression

Retroviral expression vector pMXs-Puro (Cell Biolabs, RTV-012) was used for cloning and expressing MymX and MymK orthologs. ORF inserts were codon-optimized and synthesized by Integrated DNA Technologies. The DNA sequences were verified by Sanger sequencing. For rescue experiments, the sgRNA-insensitive DNA cassettes were used. pLOVE-GFP plasmid was a gift from M. Ramalho-Santos (Addgene, plasmid no. 15949) (*42*). pMXs-Cherry plasmid was generated and described previously. Membrane targeting GFP was cloned from Addgene (plasmid no. 17787).

To produce retrovirus, retroviral plasmid was transfected to human embryonic kidney 293 cells using FuGENE 6 (Promega, E2692). Two days after transfection, viral medium was collected, filtered, and used to infect cells assisted by polybrene (Sigma-Aldrich, TR-1003-G). One day after viral infection, cells were switched to growth medium. To induce myogenic differentiation, cells were switched to myoblast differentiation medium (2% horse serum in DMEM with 1% penicillin/streptomycin). Human myoblasts can be fully differentiated 3 days after switching to differentiation medium. Mouse and lizard myoblasts were differentiated by switching to differentiation medium for at least 7 and 9 days, respectively.

MymK-MymX synergy tests

A total of 5×10^4 human MymX/MymK double KO myoblasts were seeded into a 24-well plate. Cells were infected with retrovirus-expressing proteins indicated in Fig. 5A. Two days after infection, one group is cultured in medium containing 0.2 μM 5-ethynyl-2'-deoxyuridine (EdU) for 18 hours. GFP-labeled cells were then detached by trypsin and mixed at a 1:10 ratio with the EdU-labeled cells for coculture in growth medium. The mixing culture was induced for myogenic differentiation by switching to horse-serum medium for 4 days. Cells were then fixed by 4% PFA. Nuclei were counterstained with Hoechst. EdU staining was performed by following the previous protocol (*43*). EdU/GFP/Hoechst fluorescence images from the same regions were merged using ImageJ. The level of heterologous fusion was quantified by enumerating EdU^+ nuclei inside the GFP^+ syncytia. For each replicate, the average quantification results for three randomly chosen imaging areas were shown. A total of 10 independent replicates were performed.

Differentiation index and fusion index measurements

Differentiation index was measured as the percentage of nuclei in MF20^+ cells in relative to the total number of nuclei. Fusion index was measured as the percentage of nuclei number in myotubes (≥3 nuclei) in relative to total number of muscle nuclei. Differentiation and fusion indexes were calculated on the basis of the result of manual counting, while treatment information was blinded.

RNA extraction, cDNA synthesis, and real-time PCR

Total RNA was extracted from cells, tissues, or embryos using a TRIzol reagent (Thermo Fisher Scientific, 15-596-018) according to the manufacturer's instructions. The RNA quality and concentration were assessed by a spectrophotometer (NanoDrop, Thermo Fisher Scientific) for absorbance at 260 and 280 nm. cDNA was synthesized from 2 μg of total RNA by reverse transcription using random primers with Moloney murine leukemia virus reverse transcriptase (Thermo Fisher Scientific, 28025013). Real-time PCR was performed on the QuantStudio 3 Real-Time PCR System (Thermo Fisher Scientific) using SYBR Green Master Mix (Roche) and gene-specific primers. The $2^{\Delta\Delta Ct}$ method was used to compare gene expression levels after normalization to 18*S* ribosomal RNA. Primer sequences are listed in table S1.

Membrane fractionation

Membrane fractionations were performed using the Mem-PERTM Plus Membrane Protein Extraction Kit (Thermo Fisher Scientific, 89842). Briefly, human myoblasts were scrapped off the culture dish into ice-cold phosphate-buffered saline (PBS) with a cell scraper. After centrifugation, cell pellets were washed twice in PBS and permeabilized in cytosol fraction buffer with constant mixing for 10 min at 4°C. After centrifugation at 16,000*g* for 15 min, the cytosol protein fraction was collected as the supernatant. The pellet was resuspended in membrane protein solubilization buffer and incubated at 4°C for 30 min with constant mixing. The membrane protein fraction was collected as the supernatant after 16,000*g* centrifugation for 15 min at 4°C.

Western blotting analyses

Cells were lysed in radioimmunoprecipitation assay buffer (Sigma-Aldrich, R0278) supplemented with complete protease inhibitor (Sigma-Aldrich, 04693159001) and incubated on ice for 15 min. Lysates were then centrifuged at 16,000*g* for 15 min at 4°C. The protein supernatant was collected and mixed with 4× Laemmli sample buffer (Bio-Rad, 161-0747). A total 20 to 40 μg of protein was loaded

and separated by SDS–polyacrylamide gel electrophoresis gel electrophoresis. The proteins were transferred to a polyvinylidene fluoride (PVDF) membrane (Sigma-Aldrich, ISEQ00010) and blocked in 5% fat-free milk for 1 hour at room temperature and then incubated with the following primary antibodies diluted in 5% milk overnight at 4°C: glyceraldehyde-3-phosphate dehydrogenase (Santa Cruz Biotechnology, sc-32233), α-tubulin (Santa Cruz Biotechnology, sc-8035), biotin anti–C-tag conjugate (Thermo Fisher Scientific, 7103252100), insulin receptor β (Cell Signaling Technology, 3020S), MymX (Thermo Fisher Scientific, PA5-47639), and MymR (mouse monoclonal antibody). After washes in tris-buffered saline with 0.1% Tween 20 (TBST), PVDF membrane was incubated with the following secondary antibody in blocking buffer for 1 hour at room temperature: horseradish peroxidase (HRP) streptavidin (Vector Laboratories, SA-5004), donkey anti-sheep immunoglobulin G (IgG)–HRP conjugate (Santa Cruz Biotechnology, sc-2473), goat anti-mouse IgG–HRP conjugate (Invitrogen, A28177), and goat anti-rabbit IgG–HRP conjugate (Invitrogen, A27036). Immunodetection was performed using the Western Blotting Luminol Reagent (Thermo Fisher Scientific, 34075).

Immunostaining and microscopy of vertebrate cells

Cells were fixed in 4% PFA/PBS for 10 min at room temperature, permeabilized with 0.2% Triton X-100 in PBS, and blocked with 3% bovine serum albumin/PBS for 1 hour at room temperature. Cells were incubated with the primary antibody overnight at 4°C, followed by incubation with Alexa Fluor–conjugated secondary antibodies: myosin (Developmental Studies Hybridoma Bank, MF20) and MyoD (Santa Cruz Biotechnology, sc-304); goat anti-mouse IgG (H + L), Superclonal recombinant secondary antibody, and Alexa Fluor 555 (Invitrogen, A28180); goat anti-mouse IgG (H + L), Superclonal recombinant secondary antibody, and Alexa Fluor 488 (Invitrogen, A28175); goat anti-rabbit IgG (H + L), Superclonal recombinant secondary antibody, and Alexa Fluor 555 (Invitrogen, A27039); and goat anti-rabbit IgG (H + L), Superclonal recombinant secondary antibody, and Alexa Fluor 488 (Invitrogen, A27034). The nucleus was counterstained with Hoechst 33342. The staining was visualized on a BioTek Lionheart FX automated microscope. Fluorescence images were collected by a camera on the BioTek Microscope System or Olympus FLUOVIEW FV1200 confocal laser scanning microscope.

The search of MymX orthologs in tunicates

Cross-species alignment of MymX sequences (21) revealed conservation of an N-terminal hydrophobic domain, which is predicted to be the membrane anchoring region, and a C-terminal hydrophobic AxLyCxL motif, in which x denotes L, V, or I, and y denotes S, T, or G. After searching MymX from vertebrate species where genome or transcriptome data are available, rare versions of the AxLyCxL motif were found. We then iteratively TBLASTN-searched (E value: 1000) all the possible combinations of the hydrophobic motif in the genome and transcriptome databases of tunicates that were available from ANISEED and the National Center for Biotechnology Information (NCBI). Hit sequences were examined individually and excluded if it contains a stop codon or a hydrophilic residue (R, K, D, and E) or does not contain a hydrophobic motif [predicted by TOPCONS (44) at the N terminus of the predicted coding regions by GENESCAN (45), with 1-kb DNA sequences from either end of the AxLyCxL motif being used]. By this criterion, MymX

ortholog was not found from the tunicate species. Consistently, a recent study reported the absence of both MymX and MymK genes in amphioxus (46).

Molecular phylogenetic analysis

Protein sequences were retrieved from the GenBank, Refseq, Ensembl, and ANISEED databases or by BLAST search of the genome and transcriptome databases. All sequences were provided in table S2. To construct the phylogenetic tree, protein sequences were first aligned using MUSCLE (47) with the default setting. Alignment files were provided as file S1. The maximum number of iterations was set to 8. Neighbor joining (NJ) trees were reconstructed from the alignments by the software Geneious Prime (www.geneious.com/prime/). The maximum likelihood (ML) trees were built from the alignments using RAxML (version 8.2.11) (48) with either the JTT + GAMMA or LG + GAMMA model. Bootstrap analysis was carried out with 1000 replicates for both NJ and ML trees. The RAxML command line for the bootstrap analysis is raxmlHPC -N100 -m PROT-GAMMALGF -fa -s tmem8.aln.fasta -n tmem8 -p470940 -x680848. Bootstrap support values for internal nodes on the ML phylogenetic trees were calculated by a Python program sumtrees.py (49) with the command line sumtrees.py -f0 -p -t RAxML_bestTree.tre --replace -F newick -o RAxML_bootstrap.con.tre --no-annotations RAxML_bootstrap.tre.

Molecular modeling of MymK protein structures

The tertiary structure prediction of the MymK orthologs is based on the D-I-TASSER pipeline (https://zhanggroup.org/D-I-TASSER/) (27, 50), which is an extension of I-TASSER and C-I-TASSER and integrates the deep learning–based distance and hydrogen bonding network models with iterative threading assembly simulations. The D-I-TASSER algorithm (named as "Zhang-Server") has participated in the most recent 14th critical assessment of protein structure prediction experiment (CASP14), which is a blind test to assess the protein folding ability of different participated algorithms, and was ranked as the best automatic protein structure prediction server (www.predictioncenter.org/casp14/zscores_final.cgi?gr_type=server_only).

The D-I-TASSER pipeline consists of four consecutive steps: (i) multiple sequence alignment (MSA) generation by DeepMSA2 (51), (ii) Protein Data Bank (PDB) template detection by LOMETS3 (52) and deep learning–based residue-residue distance map/hydrogen bonding prediction by DeepPotential, (iii) structure conformation (decoy) sampling by replica exchange Monte Carlo (REMC) simulation, and (iv) full-length model construction and atomic-level model refinement.

First, starting from the input sequences, DeepMSA2 is used to create a set of MSAs by iteratively searching the query sequence through whole-genome [Uniref90 (53)] and metagenome sequence databases [Metaclust (54), BFD (55), Mgnify (56), and IMG/M (57)]. The MSA with the highest accumulative probability obtained by the TripletRes-predicted (58) top 10 L (L is the protein length) contacts is selected. In the second step, the selected MSA is used as the input for template detection by LOMETS3 and distance map and hydrogen bonding prediction by DeepPotential. LOMETS3, a newly developed meta-server program combining both profile- and contact-based threading programs, is used to identify structural templates from a nonredundant PDB structural library, while DeepPotential is a newly developed deep residual neural network-based predictor to create multiple spatial restraints, including Cα-Cα and Cβ-Cβ distances

and hydrogen bonding networks. In the DeepPotential pipeline, a set of coevolutionary features are extracted from the MSA obtained by DeepMSA2. These coevolutionary features, which are inherently two-dimensional, include the raw coupling parameters from the pseudo-likelihood maximized (PLM) 22-state Potts model and the raw mutual information (MI) matrix. The 22 states of the Potts model represent the 20 standard amino acids, the nonstandard amino acid type, and the gap state. The corresponding parameters for each residue pair in the PLM and MI matrices are also extracted as additional features that measure query-specific coevolutionary information in an MSA. The field parameters and the self-MI are considered as the one-dimensional features, incorporated with hidden Markov model features. The one-hot representation of the MSA and other descriptors, such as the number of sequences in the MSA, are also considered. These one-dimensional features and two-dimensional features are fed into deep convolutional neural networks separately, where each of them goes through a set of one-dimensional and two-dimensional residual blocks, respectively, and are then tiled together. The feature representations are considered as the inputs of another fully residual neural network, which outputs several inter-residue terms. The Cα-Cα distances, Cβ-Cβ distance, and Cα-Cα–based hydrogen bond network geometry descriptors between residues are considered as prediction terms. The distance and hydrogen bond geometry values are discretized into binary descriptors; using these binary values, the neural networks were trained using cross-entropy loss. In the third step, the continuous fragments excised from the LOMETS3 templates are used as the initial conformations for full-length structure assembly using REMC simulations under the guidance of a composite force field, including (i) optimized knowledge-based energy term, (ii) spatial restraints collected from LOMETS3 templates, and (iii) deep learning distance and hydrogen bonding restraints obtained from DeepPotential. Last, at least 10,000 decoys generated by the low-temperature replicas are submitted to SPICKER (59) for structure clustering and model selection based on the energy and structure similarity. The largest SPICKER cluster is further refined by the atomic-level fragment-guided molecular dynamic (60) simulations, with the side-chain rotamer structure repacked by FASPR (61). All MymK structural models are provided in file S2.

Comparison of TM5 helix orientation of MymK proteins

We examined the alignment of threading template (PDB ID: 3wxvA) used for computing MymK structures and did not notice a template shift. Therefore, the difference of TM5 angle between vertebrate and tunicate MymK models is originated from deep learning. As illustrated in the distance map obtained from deep learning (fig. S18B), the distance between TM5 and TM6 varies between human (A) and *Ciona* (B) MymK. Specifically, TM5 and TM6 from human but not Ciona MymK form an obvious constant "contact" (~5 Å, white boxes shown in fig. S18B) throughout the two helixes, thus allowing the antiparallel conformation of TM5 and TM6.

Estimation of structural model quality and similarity

The global quality of structural model is usually appraised by the TM score (62) between model and the experimental determined structure

$$\text{TM score} = \frac{1}{L}\sum_{i=1}^{L_{\text{ali}}} \frac{1}{1 + (d_i/d_0)^2} \tag{1}$$

where L is the number of residues, d_i is the distance between the ith aligned residue pair between the model and the experimental structure, and $d_0 = 1.24 \cdot \sqrt[3]{L-15} - 1.8$ is a scaling factor. A TM score ranges between 0 and 1, and a TM score greater than 0.5 indicates a structure model of correct global topology (63).

Because the experimental structure is absent in the present study, instead of actual TM score, an estimated TM score (eTM score) was calculated using LOMETS3 threading template quality, contact map satisfaction rate, mean absolute error of the distance, and simulation convergence in D-I-TASSER

$$\text{eTM score}(m) = w_1 ln\left(\frac{M(m)}{M_{\text{total}}} \cdot \frac{1}{\langle \text{RMSD} \rangle_m}\right) +$$

$$w_2 ln\left(\frac{1}{K}\sum_{i=1}^{K}\frac{Z(i)}{Z_0(i)}\right) + w_3 w_{\text{neff}} ln\left(\frac{O\left(\text{CM}^m, \text{CM}^{\text{pred}}\right)}{N\left(\text{CM}^{\text{pred}}\right)}\right) +$$

$$w_4 w_{\text{neff}} ln\frac{1}{5L}\sum_{(i,j)}^{5L}|d_{i,j}^{\text{pred}} - d_{i,j}^m| + w_5 \tag{2}$$

$$w_{\text{neff}} = \min\left(\max\left(0.66, \frac{\log_2(\text{neff}) - 3}{10}\right), 1\right) \tag{3}$$

$$\text{Neff} = \frac{1}{\sqrt{L}}\sum_{n=1}^{N}\frac{1}{1 + \sum_{m=1, m \neq n}^{N} I\left[S_{m,n} \geq 0.8\right]} \tag{4}$$

where $w_1 = 0.047$, $w_2 = 0.062$, $w_3 = 0.077$, $w_4 = -0.185$, and $w_5 = 0.740$ are fitting parameters retrieved by regression on the large-scale benchmark test data (64, 65), C is the overall confidence score of structural assembly; here, we only care about the first model, thus m is equal to 1 by default. M_{total} is the total number of decoy conformations submitted to SPICKER clustering, M is the number of decoys in the mth cluster, <RMSD> is the average RMSD of decoys in the largest cluster, $Z(i)$ is the significance score of the ith template K equals to 11, which is the number of the component threading methods in LOMETS3, $N(\text{CM}^{\text{pred}})$ is the number of DeepPotential-predicted contacts (predicted distance < 8 Å) used to guide the REMC simulation, $O(\text{CM}^{\text{model}}, \text{CM}^{\text{pred}})$ is the number of overlapping contacts between final mth models and predicted contacts, $d_{i,j}^m$ is the Cα-Cα distance between residue i and j extracted from the mth D-I-TASSER structural model, $d_{i,j}^{\text{pred}}$ is the predicted Cα-Cα distance between residue i and j from DeepPotential, L is the length of protein, and Neff is used for measuring the quality of MSA. The eTM score highly correlates with the actual TM score relative to the experimental structures, with a Pearson correlation coefficient of 0.84 based on a 797 training proteins dataset (27, 64, 65). The similarity among the MymK structural models was calculated as TM score by superimposing the structural models using TM align (66), a sequence order–independent protein structure alignment tool.

RNA-seq data analysis

For bulk RNA-seq data, the FASTQ files generated from previous studies (67, 68) were downloaded from the NCBI GenBank database with the accession number provided in the figure legends. Sequence reads were aligned to genomes by alignment method STAR (version 2.7.2a) using default setting (69). Integrative Genomics Viewer (70) was used to view the sequencing reads and identify splicing sites and new transcript isoforms.

Single-cell RNA-seq data generated from published studies (*16*, *17*) were reanalyzed for MymK gene expression. Reference genome and annotation files for *C. robusta* were obtained from ANISEED database (*71*). MymK locus was manually added to the annotation file. Whole larva (18 hpf) single-cell RNA-seq data (GSM3764784, GSM3764785, and GSM3764786) were used to examine MymK expression in tail muscle cells. Gene-barcode matrices for each sample were generated by 10x Genomics Cell Ranger 3.1.0 using count pipeline under default settings (*72*). Downstream analyses were performed by R package Seurat (version 4.0.) (*73*). Cells with fewer than 1000 expressed genes and genes expressed in fewer than three cells were removed. A total of 15,043 genes across 13,067 cells were kept in total. Three individual Seurat objects were merged, and read counts were normalized and log-transformed for subsequent analysis. The top 1000 genes with the highest SDs were selected to exhibit high cell-to-cell variation in the dataset using the FindVariableFeatures function on variance stabilizing transformation method. Principal components analysis was performed on the scaled data, and statistically significant principal components were determined by heatmap pairwise comparison. FindClusters function was used to iteratively group cells by adjusting resolution parameter to 1.4. Expression patterns of genes were visualized by VlnPlot function. Larval tail muscle cell cluster was identified by checking marker genes' expression.

For reanalysis of the single-cell RNA-seq data of the fluorescence-activated cell sorting–purified cardiopharyngeal lineage cells (GSE99844), sequence reads of each cell were individually mapped to reference genome using TopHat 2.1.2 with parameter–no-coverage-search (*74*). MymK FPKM (fragments per kilobase of transcript per million fragments mapped) values were calculated by Cufflinks 2.2.1. Clustering results and developmental pseudotime were obtained from the original study (*16*). Gene expression patterns were visualized using Seurat R package (*73*).

Software for image and protein sequence analyses

The topology of membrane protein was predicted by TOPCONS (*44*). The secondary structure of protein was predicted by PSIPRED (*75*). Protein hydrophobicity and similarity were calculated by Expasy (*76*) (https://web.expasy.org/cgi-bin/sim/sim.pl?prot). Cell and nucleus enumerations for measuring fusion and differentiation indexes were performed using ImageJ (1.52q) (*77*).

Ciona embryo handling, electroporation, and immunostaining

Adult *C. robusta* (intestinalis type A) were collected by M-REP (San Marcos, USA). Gametes were isolated for in vitro fertilization and dechorionation and subsequent electroporation following the standard protocols (*78*). All plasmid sequences and mixes are described in file S3. Embryos were raised at 20°C and fixed at the desired stage as calculated by hpf. To obtain juveniles, larvae were allowed to metamorphose on (but not attach to) agarose-coated petri dishes in filtered/buffered artificial sea water supplemented with 1× penicillin/streptomycin (Omega Scientific, catalog number PS-20), followed by daily changes of penicillin/streptomycin sea water. For direct visualization of fluorescent proteins, embryos were fixed in MEM-FA [3.7% formaldehyde, 0.1 M Mops (pH7.4), 0.5 M NaCl, 1 mM EGTA, 2 mM MgSO4, and 0.05% Triton X-100] for 15 min, rinsed in 1× PBS/0.4% Triton X-100/50 mM NH$_4$Cl and 1× PBS/0.05% Triton X-100. For immunostaining of CD4::GFP, embryos were fixed and rinsed as above and incubated in mouse anti-GFP (clones 7.1 and

13.1, Roche) at 1:500 dilution for 1 hour and Alexa Fluor 488 goat anti-mouse IgG secondary (Thermo Fisher Scientific, catalog number A11001) at 1:500 dilution for 1 hour. Both incubations were done in 1× PBS/0.05% Triton X-100/2% normalized goat serum and rinsed in 1× PBS/0.05% Triton X-100. All samples were mounted in 1× PBS/50% glycerol/2% DABCO (1,4-diazabicyclo[2.2.2]octane).

In situ hybridization in *Ciona* juveniles

Ciona juveniles were raised as described above, fixed in MEM-PFA [4% PFA, 0.1 M Mops (pH7.4), 0.5 M NaCl, 1 mM EGTA, 2 mM MgSO$_4$, and 0.05% Tween-20] for 2 hours at room temperature or overnight at 4°C, gradually dehydrated in serial dilutions of ethanol, and lastly stored in 75% ethanol at −20°C. Whole-mount mRNA in situ hybridization was carried out as previously described (*79*), using the TSA Plus fluorescein detection kit (Akoya Biosciences, catalog number NEL741001KT). Fluorescein-labeled *MymK* riboprobes were prepared by in vitro transcription with T7 RNA polymerase from unpurified PCR amplicons of custom-synthesized *MymK* cDNA (Twist Bioscience, see sequence in file S3).

Imaging of *Ciona*

Images were acquired on a Leica DMi8 and DM IL light-emitting diode epifluorescence or an Olympus FLUOVIEW FV1200 confocal laser scanning microscope. The single focal plane images for the representative *Ciona* were used to produce focal plane videos (movies S1 to S4).

Peakshift assay for sgRNA validation

Embryos were subjected to CRISPR sgRNA validation following the "peakshift" method (*80*). Briefly, embryos were electroporated with 25 μg of Eef1a>Cas9 (*81*) and 75 μg of a given U6>sgRNA(F+E) (*81*) expression plasmid per 700 μl of electroporation volume. As a negative control for Sanger sequencing chromatogram analysis (see below), U6>Gsx.4(F+E) vector was used instead to drive expression of a sgRNA designed against the unrelated *Gsx* gene instead (see sgRNA sequences below). Embryos were allowed to grow to hatching, and then genomic DNA was extracted from each sample of pooled embryos using a QIAamp DNA mini kit (Qiagen) following the manufacturers' recommendations. Purified genomic DNA was then used as template for PCR using Accuprime Pfx (Thermo Fisher Scientific) following the manufacturer's recommendations and using a touch-down genomic PCR program as previously described (*82*).

Amplicons were PCR-amplified using MymK Peakshift forward (CGCGATCACAAATGACGAAAC) and MymK Peakshift reverse (CCCGCAATTACAACATGCTAG) primers. PCR reactions were verified on an agarose gel stained with ethidium bromide and then purified using a QIAquick PCR purification kit (Qiagen). Amplicons were sequenced by Sanger sequencing using Exon2seqRev (CCCG-CAATTACAACATGCTAG) and Exon4seqFwd (GCATAAGGT-GCTGTATGAAACAG) to detect indels in exons 2 and 4, respectively. Sanger sequencing chromatograms were compared between embryos electroporated with MymK sgRNAs and Gsx.4 (negative control) sgRNA using the web application TIDE (*83*). Additional sgRNAs targeting exon 3 were tested by sequencing with Exon3seqRev primer (ATTTTGCGTGTCTGAACCTC) but failed to generate any detectable indels.

Quantification and statistical analysis

Quantification results for each experiment were based on at least three independent experiments. For image analysis, randomly chosen

views were analyzed. Statistical analyses were carried out with GraphPad Prism 8.3.0. Data are presented as means ± SEM. For experiments involving multiple groups, one-way analysis of variance (ANOVA) with Tukey's multiple comparisons test was performed. For experiments involving only two treatment groups, Student's t test with a two-tail distribution was performed. $P < 0.05$ was considered statistically significant.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.add2696

REFERENCES AND NOTES

1. J. H. Kim, P. Jin, R. Duan, E. H. Chen, Mechanisms of myoblast fusion during muscle development. *Curr. Opin. Genet. Dev.* **32**, 162–170 (2015).
2. D. P. Millay, J. R. O'Rourke, L. B. Sutherland, S. Bezprozvannaya, J. M. Shelton, R. Bassel-Duby, E. N. Olson, Myomaker is a membrane activator of myoblast fusion and muscle formation. *Nature* **499**, 301–305 (2013).
3. P. Bi, A. Ramirez-Martinez, H. Li, J. Cannavino, J. R. McAnally, J. M. Shelton, E. Sánchez-Ortiz, R. Bassel-Duby, E. N. Olson, Control of muscle formation by the fusogenic micropeptide Myomixer. *Science* **356**, 323–327 (2017).
4. Q. Zhang, A. A. Vashisht, J. O'Rourke, S. Y. Corbel, R. Moran, A. Romero, L. Miraglia, J. Zhang, E. Durrant, C. Schmedt, S. C. Sampath, S. C. Sampath, The microprotein minion controls cell fusion and muscle formation. *Nat. Commun.* **8**, (2017).
5. M. E. Quinn, Q. Goh, M. Kurosaka, D. G. Gamage, M. J. Petrany, V. Prasad, D. P. Millay, Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. *Nat. Commun.* **8**, (2017).
6. H. Zhang, J. Wen, A. Bigot, J. Chen, R. Shang, V. Mouly, P. Bi, Human myotube formation is determined by MyoD-Myomixer/Myomaker axis. *Sci. Adv.* **6**, (2020).
7. L. Z. Holland, Muscle development in amphioxus: Morphology, biochemistry, and molecular biology. *Is. J. Zool.* **42**, S235–S246 (1996).
8. F. Razy-Krajka, A. Stolfi, Regulation and evolution of muscle development in tunicates. *Evodevo* **10**, 13 (2019).
9. S. Kreissl, A. Uber, S. Harzsch, Muscle precursor cells in the developing limbs of two isopods (Crustacea, Peracarida): An immunohistochemical study using a novel monoclonal antibody against myosin heavy chain. *Dev. Genes Evol.* **218**, 253–265 (2008).
10. D. M. Lee, E. H. Chen, *Drosophila* myoblast fusion: Invasion and resistance for the ultimate union. *Annu. Rev. Genet.* **53**, 67–91 (2019).
11. R. Paniagua, M. Royuela, R. M. Garcia-Anchuelo, B. Fraile, Ultrastructure of invertebrate muscle cell types. *Histol. Histopathol.* **11**, 181–201 (1996).
12. P. A. Toselli, G. R. Harbison, The fine structure of developing locomotor muscles of the pelagic tunicate, *Cyclosalpa affinis* (Thaliacea: Salpidae). *Tissue Cell* **9**, 137–156 (1977).
13. Y. Shinohara, K. Konishi, Ultrastructure of the body-wall Muscle of the Ascidian *Halocynthia roretzi*: Smooth Muscle Cell With Multiple Nuclei. *J. Exp. Zool.* **221**, 137–142 (1982).
14. K. Terakado, T. Obinata, Structure of multinucleated smooth muscle cells of the ascidian *Halocynthia roretzi. Cell Tissue Res.* **247**, 85–94 (1987).
15. A. Ferrandez-Roldan, M. Fabregà-Torrus, G. Sánchez-Serna, E. Duran-Bello, M. Joaquín-Lluís, P. Bujosa, M. Plana-Carmona, J. Garcia-Fernàndez, R. Albalat, C. Cañestro, Cardiopharyngeal deconstruction and ancestral tunicate sessility. *Nature* **599**, 431–435 (2021).
16. W. Wang, X. Niu, T. Stuart, E. Jullian, W. M. Mauck III, R. G. Kelly, R. Satija, L. Christiaen, A single-cell transcriptional roadmap for cardiopharyngeal fate diversification. *Nat. Cell Biol.* **21**, 674–686 (2019).
17. C. Cao, L. A. Lemaire, W. Wang, P. H. Yoon, Y. A. Choi, L. R. Parsons, J. C. Matese, W. Wang, M. Levine, K. Chen, Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* **571**, 349–354 (2019).
18. F. Razy-Krajka, K. Lam, W. Wang, A. Stolfi, M. Joly, R. Bonneau, L. Christiaen, Collier/OLF/EBF-dependent transcriptional dynamics control pharyngeal muscle specification from primed cardiopharyngeal progenitors. *Dev. Cell* **29**, 263–276 (2014).
19. A. Stolfi, T. B. Gainous, J. J. Young, A. Mori, M. Levine, L. Christiaen, Early chordate origins of the vertebrate second heart field. *Science* **329**, 565–568 (2010).
20. J. J. Smith, S. Kuraku, C. Holt, T. Sauka-Spengler, N. Jiang, M. S. Campbell, M. D. Yandell, T. Manousaki, A. Meyer, O. E. Bloom, J. R. Morgan, J. D. Buxbaum, R. Sachidanandam, C. Sims, A. S. Garruss, M. Cook, R. Krumlauf, L. M. Wiedemann, S. A. Sower, W. A. Decatur, J. A. Hall, C. T. Amemiya, N. R. Saha, K. M. Buckley, J. P. Rast, S. Das, M. Hirano, N. M. Curley, P. Guo, N. Rohner, C. J. Tabin, P. Piccinelli, G. Elgar, M. Ruffier, B. L. Aken, S. M. J. Searle, M. Muffato, M. Pignatelli, J. Herrero, M. Jones, C. T. Brown, Y.-W. Chung-Davidson, K. G. Nanlohy, S. V. Libants, C.-Y. Yeh, D. W. McCauley, J. A. Langeland, Z. Pancer, B. Fritzsch, P. J. de Jong, B. Zhu, L. L. Fulton, B. Theising, P. Flicek, M. E. Bronner, W. C. Warren, S. W. Clifton, R. K. Wilson, W. Li, Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat. Genet.* **45**, 415–421 (2013).
21. J. Shi, P. Bi, J. Pei, H. Li, N. V. Grishin, R. Bassel-Duby, E. H. Chen, E. N. Olson, Requirement of the fusogenic micropeptide myomixer for muscle formation in zebrafish. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 11950–11955 (2017).
22. T. Nakao, Electron microscopic studies on the myotomes of larval lamprey, *Lampetra japonica. Anat. Rec.* **187**, 383–403 (1977).
23. C. M. Fan, L. Li, M. E. Rozo, C. Lepper, Making skeletal muscle from progenitor and stem cells: Development versus regeneration. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 315–327 (2012).
24. H. Yin, F. Price, M. A. Rudnicki, Satellite cells and the muscle stem cell niche. *Physiol. Rev.* **93**, 23–67 (2013).
25. A. Roy, A. Kucukural, Y. Zhang, I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
26. Y. Li, W. Zheng, C. Zhang, E. Bell, X. Huang, R. Pearce, X. Zhou, Y. Zhang, Protein 3D structure prediction by D-I-TASSER in CASP14. CASP14 Conference abstract, 339–341 (2020).
27. W. Zheng, C. Zhang, Y. Li, R. Pearce, E. W. Bell, Y. Zhang, Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Rep. Methods* **1**, 100014 (2021).
28. H. Tanabe, Y. Fujii, M. Okada-Iwabu, M. Iwabu, Y. Nakamura, T. Hosaka, K. Motoyama, M. Ikeda, M. Wakiyama, T. Terada, N. Ohsawa, M. Hato, S. Ogasawara, T. Hino, T. Murata, S. Iwata, K. Hirata, Y. Kawano, M. Yamamoto, T. Kimura-Someya, M. Shirouzu, T. Yamauchi, T. Kadowaki, S. Yokoyama, Crystal structures of the human adiponectin receptors. *Nature* **520**, 312–316 (2015).
29. D. P. Millay, D. G. Gamage, M. E. Quinn, Y. L. Min, Y. Mitani, R. Bassel-Duby, E. N. Olson, Structure-function analysis of myomaker domains required for myoblast fusion. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2116–2121 (2016).
30. K. Yamaguchi, Y. Hara, K. Tatsumi, O. Nishimura, J. J. Smith, M. Kadota, S. Kuraku, Inference of a genome-wide protein-coding gene set of the inshore hagfish *Eptatretus burgeri. bioRxiv* , 2020.07.24.218818 (2020).
31. C. Gans, R. G. Northcutt, Neural crest and the origin of vertebrates: A new head. *Science* **220**, 268–273 (1983).
32. R. Diogo, R. G. Kelly, L. Christiaen, M. Levine, J. M. Ziermann, J. L. Molnar, D. M. Noden, E. Tzahor, A new heart for a new head in vertebrate cardiopharyngeal evolution. *Nature* **520**, 466–473 (2015).
33. R. Albalat, C. Canestro, Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
34. F. H. Ruddle, K. L. Bentley, M. T. Murtha, N. Risch, Gene loss and gain in the evolution of the vertebrates. *Dev. Suppl.*, 155–161 (1994).
35. N. Nikitina, M. Bronner-Fraser, T. Sauka-Spengler, Culturing lamprey embryos. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5122 (2009).
36. M. J. Siefkes, A. P. Scott, B. Zielinski, S. S. Yun, W. Li, Male sea lampreys, *Petromyzon marinus* L., excrete a sex pheromone from gill epithelia. *Biol. Reprod.* **69**, 125–132 (2003).
37. V. C. Applegate, Natural history of the sea lamprey, *Petromyzon marinus*, in Michigan, *Doctoral dissertation*, (University of Michigan, 1950).
38. B. Campbell, *Nest Survey Procedures and Examples for the Sterile Male Release Program* (Department of Fisheries & Oceans, Sea Lamprey Control Centre, 2003).
39. K. Mamchaoui, C. Trollet, A. Bigot, E. Negroni, S. Chaouch, A. Wolff, P. K. Kandalla, S. Marie, J. di Santo, J. L. St Guily, F. Muntoni, J. Kim, S. Philippi, S. Spuler, N. Levy, S. C. Blumen, T. Voit, W. E. Wright, A. Aamiri, G. Butler-Browne, V. Mouly, Immortalized pathological human myoblasts: Towards a universal tool for the study of neuromuscular disorders. *Skelet. Muscle* **1**, (2011).
40. F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, F. Zhang, Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
41. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
42. R. Blelloch, M. Venere, J. Yen, M. Ramalho-Santos, Generation of induced pluripotent stem cells in the absence of drug selection. *Cell Stem Cell* **1**, 245–247 (2007).
43. A. Salic, T. J. Mitchison, A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2415–2420 (2008).
44. K. D. Tsirigos, C. Peters, N. Shu, L. Kall, A. Elofsson, The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **43**, W401–W407 (2015).
45. C. B. Burge, S. Karlin, Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346–354 (1998).
46. M. E. Aase-Remedios, C. Coll-Llado, D. E. K. Ferrier, Amphioxus muscle transcriptomes reveal vertebrate-like myoblast fusion genes and a highly conserved role of insulin signalling in the metabolism of muscle. *BMC Genomics* **23**, 93 (2022).
47. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

48. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

49. J. Sukumaran, M. T. Holder, DendroPy: A Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).

50. W. Zheng, Y. Li, C. Zhang, X. Zhou, R. Pearce, E. W. Bell, X. Huang, Y. Zhang, Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins* **89**, 1734–1751 (2021).

51. C. Zhang, W. Zheng, S. M. Mortuza, Y. Li, Y. Zhang, DeepMSA: Constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105–2112 (2020).

52. Wei Zheng, Y. L., Xiaogen Zhou, Chengxin Zhang, Robin Pearce, Yang Zhang, Template-based protein folding guided by residue-residue distance and hydrogen-bond network prediction from deep-learning. CASP14 abstract, 342–344 (2020).

53. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu; UniProt Consortium, UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

54. M. Steinegger, J. Söding, Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).

55. M. Steinegger, M. Mirdita, J. Söding, Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).

56. A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, E. Sakharova, M. Scheremetjew, A. Korobeynikov, A. Shlemov, O. Kunyavskaya, A. Lapidus, R. D. Finn, MGnify: The microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).

57. I. M. A. Chen, K. Chu, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, M. Huntemann, N. Varghese, J. R. White, R. Seshadri, T. Smirnova, E. Kirton, S. P. Jungbluth, T. Woyke, E. A. Eloe-Fadrosh, N. N. Ivanova, N. C. Kyrpides, IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).

58. Y. Li, C. Zhang, E. W. Bell, D.-J. Yu, Y. Zhang, Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins* **87**, 1082–1091 (2019).

59. Y. Zhang, J. Skolnick, SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.* **25**, 865–871 (2004).

60. J. Zhang, Y. Liang, Y. Zhang, Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **19**, 1784–1795 (2011).

61. X. Huang, R. Pearce, Y. Zhang, FASPR: An open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **36**, 3758–3765 (2020).

62. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).

63. J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).

64. J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).

65. W. Zheng, Y. Li, C. Zhang, R. Pearce, S. M. Mortuza, Y. Zhang, Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149–1164 (2019).

66. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

67. M. L. Martik, S. Gandhi, B. R. Uy, J. A. Gillis, S. A. Green, M. Simoes-Costa, M. E. Bronner, Evolution of the new head by gradual acquisition of neural crest regulatory circuits. *Nature* **574**, 675–678 (2019).

68. J. Pascual-Anaya, I. Sato, F. Sugahara, S. Higuchi, J. Paps, Y. Ren, W. Takagi, A. Ruiz-Villalba, K. G. Ota, W. Wang, S. Kuratani, Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates. *Nat. Ecol. Evol.* **2**, 859–866 (2018).

69. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

70. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

71. O. Tassy, D. Dauga, F. Daian, D. Sobral, F. Robin, P. Khoueiry, D. Salgado, V. Fox, D. Caillol, R. Schiappa, B. Laporte, A. Rios, G. Luxardi, T. Kusakabe, J. S. Joly, S. Darras, L. Christiaen, M. Contensin, H. Auger, C. Lamy, C. Hudson, U. Rothbächer, M. J. Gilchrist, K. W. Makabe, K. Hotta, S. Fujiwara, N. Satoh, Y. Satou, P. Lemaire, The ANISEED database: Digital representation, formalization, and elucidation of a chordate developmental program. *Genome Res.* **20**, 1459–1468 (2010).

72. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. Mc Dermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. M. Farland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

73. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2020).

74. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

75. D. W. A. Buchan, D. T. Jones, The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).

76. E. Gasteiger, C. Hoogland, A. Gattiker, S'everine Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, *The Proteomics Protocols Handbook* (Springer, 2005), pp. 571–607.

77. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

78. L. Christiaen, E. Wagner, W. Shi, M. Levine, Isolation of sea squirt (*Ciona*) gametes, fertilization, dechorionation, and development. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5344 (2009).

79. L. Christiaen, E. Wagner, W. Shi, M. Levine, Whole-mount in situ hybridization on sea squirt (*Ciona intestinalis*) embryos. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5348 (2009).

80. S. Gandhi, M. Haeussler, F. Razy-Krajka, L. Christiaen, A. Stolfi, Evaluation and rational design of guide RNAs for efficient CRISPR/Cas9-mediated mutagenesis in *Ciona*. *Dev. Biol.* **425**, 8–20 (2017).

81. A. Stolfi, S. Gandhi, F. Salek, L. Christiaen, Tissue-specific genome editing in *Ciona* embryos by CRISPR/Cas9. *Development* **141**, 4115–4120 (2014).

82. S. Gandhi, F. Razy-Krajka, L. Christiaen, A. Stolfi, in *Transgenic Ascidians* (Springer, 2018), pp. 141–152.

83. E. K. Brinkman, T. Chen, M. Amendola, B. van Steensel, Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168–e168 (2014).