A Complete Assembly and Annotation of the American Shad Genome Yields Insights into the Origins of Diadromy

Jonathan P. Velotta [1,*,†], Azwad R. Iqbal [2,†], Emma S. Glenn [1], Ryan P. Franckowiak [2], Giulio Formenti [3], Jacquelyn Mountcastle [3], Jennifer Balacco [3], Alan Tracey [4], Ying Sims [4], Kerstin Howe [4], Olivier Fedrigo [3], Erich D. Jarvis [3], Nina O. Therkildsen [2]

[1]Department of Biological Sciences, University of Denver, Denver, CO 80210, USA

[2]Department of Natural Resources and the Environment, Cornell University, Ithaca, NY 14853, USA

[3]The Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10021, USA

[4]Tree of Life, Wellcome Sanger Institute, Cambridge, UK

*Corresponding author: E-mail: jonathan.velotta@du.edu.

[†]Both authors contributed equally.

Accepted: December 16, 2024

Abstract

Transitions across ecological boundaries, such as those separating freshwater from the sea, are major drivers of phenotypic innovation and biodiversity. Despite their importance to evolutionary history, we know little about the mechanisms by which such transitions are accomplished. To help shed light on these mechanisms, we generated the first high-quality, near-complete assembly and annotation of the genome of the American shad (*Alosa sapidissima*), an ancestrally diadromous (migratory between salinities) fish in the order Clupeiformes of major cultural and historical significance. Among the Clupeiformes, there is a large amount of variation in salinity habitat and many independent instances of salinity boundary crossing, making this taxon well-suited for studies of mechanisms underlying ecological transitions. Our initial analysis of the American shad genome reveals several unique insights for future study including: (i) that genomic repeat content is among the highest of any fish studied to date; (ii) that genome-wide heterozygosity is low and may be associated with range-wide population collapses since the 19th century; and (iii) that natural selection has acted on the branch leading to the diadromous genus *Alosa*. Our analysis suggests that functional targets of natural selection may include diet, particularly lipid metabolism, as well as cytoskeletal remodeling and sensing of salinity changes. Natural selection on these functions is expected in the transition from a marine to diadromous life history, particularly in the tolerance of nutrient- and ion-devoid freshwater. We anticipate that our assembly of the American shad genome will be used to test future hypotheses on adaptation to novel environments, the origins of diadromy, and adaptive variation in life history strategies, among others.

Key words: whole-genome sequencing, comparative genomics, fish, anadromy.

Significance

Despite their importance to evolutionary history, we know little about the genetic mechanisms by which transitions across salinity boundaries have been accomplished. The order Clupeiformes is a well-suited model, yet few genomic resources are available. We generated the first high-quality, near-complete assembly and annotation of the genome of the American shad (*Alosa sapidissima*), a migratory Clupeiform. Our analysis of the American shad genome reveals unique characteristics including high genomic repeat content and low genome-wide diversity. Using comparative genomic approaches, we identified natural selection in putative functional targets in the branch leading to the migratory genus *Alosa*.

Introduction

Transitions across ecological boundaries are fundamental to the generation of phenotypic novelty and biodiversity as they subject organisms to a changed landscape of selection. G.G. Simpson wrote in 1944 that transitions between "adaptive zones"—now thought of as open ecological niches—lead to rapid adaptive radiation (Simpson 1944). As prime examples, the transitions from water to land (Shubin, Daeschler, and Jenkins 2006) and from saltwater to freshwater (Lee and Bell 1999) have led to major diversification events that have generated biodiversity and shaped species distributions around the globe. The repeated evolution of novel behavioral, morphological, and/or physiological traits often follows such shifting selection regimes (Betancur-R 2010; Velotta et al. 2014, 2022; Velotta, McCormick, and Schultz 2015; Sabatino et al. 2022). Understanding the genomic basis of such adaptive evolutionary changes is a major goal of modern evolutionary biology.

Among the fishes, habitat transitions across salinity boundaries have facilitated diversification and helped pave the way for evolution of land-dwelling tetrapods (Lee and Bell 1999; Schultz and McCormick 2013). Because of the extreme differences in salt concentration between freshwater (<5 mOsm/kg) and the ocean (~1,000 mOsm/kg), the boundary between the two is a formidable one that few taxa have crossed (Lee and Bell 1999). One potential evolutionary step in the transition across these salinity zones for fishes is the evolution of diadromy, characterized by migration between freshwater or saltwater during particular life stages. Rare, but distributed broadly across taxa (McDowall 1988), diadromy has contributed to biodiversity, either acting as an evolutionary link between freshwater and saltwater, or as itself a unique ecological niche leading to adaptive radiation (Corush 2019). Adaptations that permit this remarkable life history strategy include changes to osmoregulatory flexibilities that allow for tolerance of both saltwater and freshwater (known as euryhalinity; McCormick 2013), as well changes to reproductive investment (Crespi and Teo 2002) and body size (Burns and Bloom 2020).

Despite its overwhelming importance to biodiversity, little research into the genetic changes that underlie the evolution of diadromy has been conducted. Studies in fishes in the order Salmoniformes, a group with widespread diadromy, suggest that the evolution of osmoregulatory flexibility among its diadromous members may be the result of a whole-genome duplication and subsequent neofunctionalization in gene families involved in ion exchange (Norman et al. 2011, 2012). The origin of diadromy in this group, however, precedes the whole-genome duplication coincident with their diversification by 50 to 55 million years, suggesting that whole-genome duplication may not be the sole explanation for the evolution of complex behavioral and physiological traits that permitted diadromy in this group (Alexandrou et al. 2013).

American shad (*Alosa sapidissima*) is a diadromous clupeid (family *Clupeidae*) that spawns in rivers along the east coast of North America, from Florida to Quebec (Walburg and Nichols 1967). Shad are anadromous, which is a type of diadromy where individuals are born in freshwater but spend their adult lives at sea. Anadromous fishes such as shad serve as a key ecological bridge between freshwater and saltwater, providing a large source of marine-derived nutrients to inland rivers and streams (Garman and Macko 1998). Depending on latitude, adult American shad make the return migration to freshwater spawning grounds either exactly once (semelparity; only among rivers south of Cape Fear, North Carolina, USA), or on multiple occasions (iteroparity), the frequency of which increases with latitude (Leggett and Carscadden 1978). Such population-level variation in spawning frequency is rare in migratory fishes, and, if genetically based, may be unique among clupeids. Across their broad range, shad breeding habitat is exceptionally heterogeneous, spanning three biogeographical provinces (Engle and Summers 1999), a fact that makes local adaptation likely, although underexplored. In closely related Eurasian shad (*Alosa alosa* and *Alosa fallax*), for example, repeated local adaptation to varying salinity environments has been found (Sabatino et al. 2022). Moreover, shad have recently and successfully colonized the west coast of North America after being introduced to California in the late 1800s and even developed novel freshwater-living landlocked populations in certain reservoirs (Hasselman et al. 2012). Recent colonization history and rapid adaptation to novel environments, as well as the aforementioned extreme native-range heterogeneity and phylogenetic position within the fish tree of life, make the American shad a unique model for studying the comparative and population genomic mechanisms of adaptation and life history transitions.

Here, we generated the first high-quality, chromosome-level, reference assembly of the American shad genome. This is among the first assemblies in the genus *Alosa*, a group well-known for their anadromous life history, and one of broad ecological importance as a source of marine-derived nutrients in freshwater (Garman and Macko 1998). It is now one of five publicly available (https://www.ncbi.nlm.nih.gov/datasets/genome) chromosome-level assemblies in the *Clupeidae*, a taxon in which salinity habitat transitions are widespread (Bloom and Lovejoy 2014; Bloom and Egan 2018). With the American shad genome, we provide an improved opportunity to understand the genomic basis of variation in life history strategy, migratory behavior, physiological flexibility, and adaptation to novel environments. As an initial investigation, we compared shad genome sequence variation and gene family evolution to that

of other species in the order Clupeiformes and related clades (Betancur-R et al. 2013; Bloom and Egan 2018) in order to identify loci, gene families, and biological functions that may have contributed to the evolution of diadromy and other unique features of this species.

Results and Discussion

Genome Assembly and Synteny

The final genome assembly for the American shad is composed of 24 chromosomes and an additional 45 scaffolds with a scaffold N50 of 38 megabases (Mb) and spanning 903.6 Mb in total length (Fig. 1a; supplementary table S1, Supplementary Material online). Genome completeness assessed with the BUSCO (Simão et al. 2015) Actinopterygii gene set indicates the assembly is highly complete with only 2.7% of genes missing and measuring 95.6% complete (supplementary table S1, Supplementary Material online). Contiguity was also high, with N's constituting 0.6% of the final assembly (Fig. 1a). Overall, this assembly represents a near-complete, contiguous, chromosome-level genome for the American shad.

We assessed the degree of synteny between the American shad and Allis shad (*Alosa alosa*) genome assemblies to identify the degree of conserved genomic structure between the two congeners and possible chromosomal rearrangements. We found that individual chromosomes between the two species exhibited a high degree of synteny based on sequence alignments (Fig. 1b), as 84% alignments represented relationships between single chromosome pairs. The identity (chromosome number) of chromosomes exhibiting high synteny varied due to differences in chromosome naming conventions used for each assembly. Rearrangements beyond pairwise chromosomal relationships were also identified (representing ~16% of alignments), though these were typically short in length and broadly dispersed among chromosomes, constituting only 5.28% of the total length of alignments.

Genome Annotation

Annotations generated by the NCBI Eukaryotic Genome Annotation Pipeline detected 47,628 genes and pseudogenes, of which 25,730 were protein coding and 21,044 were noncoding, along with 55,994 mRNAs and 56,159 coding sequences (CDSs) (NCBI Annotation release ID: 100) (Table 1). Genic content varied substantially across each annotated chromosome, with certain 1 Mb windows harboring up to 2% of the total annotated genic content (supplementary fig. S3, Supplementary Material online).

Repetitive Region Annotation

Repetitive element annotation using a custom database of known teleost repeats and species-specific repeats generated de novo indicates that repetitive elements make up 45.79% of the total American shad genome (Fig. 2), one of the highest proportions of repeats among surveyed fish genomes of similar size (Yuan et al. 2018). The vast majority of repetitive elements were classified as interspersed repeats (39.89% of the total genome), with DNA transposons composing the bulk of these interspersed repeats (18.68% of the total genome) (Table 2). The distribution of repetitive elements by Kimura distance—a measure of evolutionary distance between alignments (Kimura 1980)—indicates that the bulk of elements arose recently, with activity centered around 0% to 5% divergence relative to their consensus sequences (Fig. 2b). A comparison of repetitive element annotations between American shad and other members of the family *Clupeidae* (Allis shad—*Alosa alosa* and Atlantic herring—*Clupea harengus*) shows an expansion of repetitive element content in the genus *Alosa*, with the congener *Alosa alosa* exhibiting a similar composition and overall percentage of repetitive content to the American shad (Fig. 2c). The increase in repetitive content in the *Alosa* species relative to the Atlantic herring appears to be driven by an expansion of repetitive DNA elements—a broad class encompassing various types of DNA transposons—which is the largest general class of repetitive element by percentage of sequence annotated in both *Alosa* genomes (Fig. 2c).

Genome-wide Diversity

We directly assessed genome-wide heterozygosity by calculating the proportion of variant (heterozygous) sites per kilobase of genome length, accounting for genomic windows with excessively low coverage. Mean genome-wide heterozygosity was calculated to be 1.93 variants per kilobase or 0.193%, comparable to the degree of heterozygosity found in Atlantic cod (*Gadus morhua*) (Star et al. 2011) and Big-eye Mandarin fish (Lu et al. 2020) and particularly low among fish genomes with available data (Fig. 3c) (Tigano et al. 2021). Overall standing variation can thus be inferred to be low and may be associated with range-wide population collapses of American shad since the late 19th century (Limburg and Waldman 2009). The distribution of heterozygous sites varied from chromosome to chromosome, with some chromosomes concentrating regions of high heterozygosity around chromosome ends, while other chromosomes had more widely distributed heterozygosity "hotspots" (Fig. 3a). Elevated heterozygosity at some chromosome ends may be associated with repeat-rich telomeres, as we find chromosome ends enriched for SINE, LINE, DNA, and LTR elements (supplementary fig. S5, Supplementary Material online). Despite low genome-wide diversity and demographic declines in recent history, runs of homozygosity (ROH) in this individual were uncommon and short (<1 Mb in

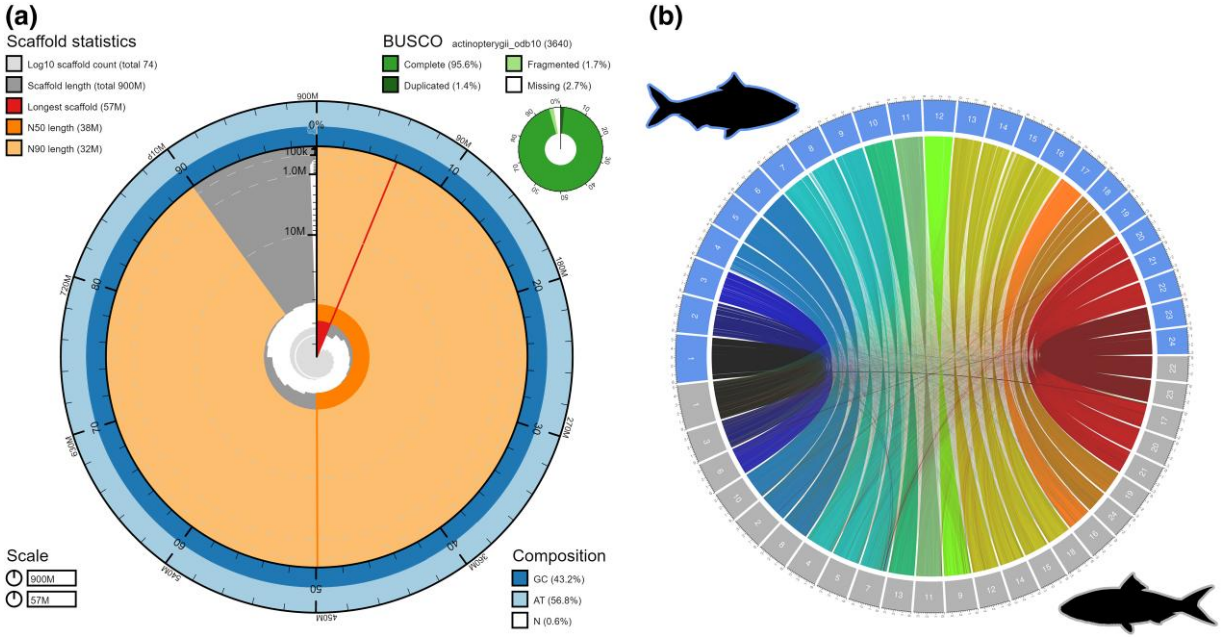

Fig. 1. Genome assembly metrics and synteny plots indicate that the American shad genome assembly is highly complete and exhibits a high degree of synteny with a congener, the Allis shad (*Alosa alosa*). a) Snail plot summary of assembly statistics for assembly fAloSap1. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 903,581,644-bp assembly. The distribution of chromosome lengths is shown in dark gray with the plot radius scaled to the longest chromosome present in the assembly (56,504,578 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (38,440,066 and 31,742,902 bp), respectively. The pale gray spiral shows the cumulative chromosome count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT, and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the actinopterygii_odb10 set is shown in the top right. b) Circos plot showing synteny between American shad (blue) and Allis shad (gray). Alignments shorter than 200bp were excluded. Ribbons are color coded to the chromosomes of sequences aligned to the American shad genome. Allis shad chromosomes have been reordered based on syntenic relationships with American shad for visual clarity. Fish images were download from the public domain *via* phylopic.org.

length), suggesting that the degree of inbreeding remains low (supplementary fig. S6, Supplementary Material online). American shad are broadcast spawners, meaning that both sexes release their gametes directly into the water column to produce fertilized zygotes (Limburg 2003). This behavior may contribute to reduced inbreeding despite population declines as parentage is often distributed among multiple males per spawning female.

Demographic History

To better understand possible historic drivers of contemporary genetic diversity and to compare demographic histories between diadromous and marine clupeids, we employed the pairwise sequentially Markovian coalescent model (PSMC) (Li and Durbin 2011) to infer the historical effective population sizes (Ne) of the American shad, its diadromous congener the Allis shad, and the marine Atlantic herring (Fig. 3d). American shad N_e peaked in deep time ~15 MYA, before gradually declining until ~5 MYA and stabilizing briefly before declining again until ~400 KYA. A period of expansion followed, peaking at ~71 KYA, before a rapid decline shortly after. This rapid decline in N_e after 100 KYA coincides with the onset of the Last Glacial Period, which may have restricted the availability of suitable coastal habitats due to the recurrence of large North American ice sheets (Andersen et al. 2004), while the preceding expansion of N_e appears to coincide with an interglacial period with relatively high global temperatures (Petit et al. 1999). Such declines in N_e at the onset of the Last Glacial Period are fairly common among surveyed fish species with available genomic data (Li et al. 2021). In comparison, the Allis shad, a diadromous congener native to Europe, had N_e consistently lower than the other two species across the time period inferred by PSMC, beginning with a peak N_e ~5 MYA and gradually declining afterward with the exception of a small expansion of N_e at ~685 KYA. The Atlantic herring exhibited a distinct historical pattern of N_e, beginning with a period of relative stability until a contraction around ~1.5 MYA, a subsequent large expansion peaking at ~500 KYA, before a subsequent decline. Both the Atlantic herring and the Allis shad experienced expansions in the same broad historical era, which coincided with a period of decline in the American shad, possibly

Table 1 Summary table of repetitive elements by type

	Number of elements	Length occupied (bp)	Percentage of sequence
Retro elements	343,346	97,456,414	10.79
SINEs:	55,311	8,031,937	0.89
Penelope	8414	1,986,747	0.22
LINEs:	151,449	47,753,253	5.28
L2/CR1/Rex	66,628	23,386,097	2.59
R1/LOA/Jockey	361	270,729	0.03
R2/R4/NeSL	2277	545,157	0.06
RTE/BOV-B	11,432	2,305,727	0.26
L1/CIN4	5937	2,551,977	0.28
LTR elements:	136,586	41,671,224	4.61
BEL/PAO	4707	1,593,170	0.18
Ty1/Copia	3586	492,919	0.05
Gypsy/DIRS1	89,616	30,336,009	3.36
Retroviral	14,339	3,178,166	0.35
DNA transposons	950,803	168,781,716	18.68
hobo-Activator	313,254	55,810,962	6.18
Tc1-IS630-Pogo	21,467	5,220,080	0.58
MuDR-IS905	1228	202,968	0.02
Piggybac	2132	444,772	0.05
Tourist/Harbinger	44,745	12,309,012	1.36
Other	4925	2,053,684	0.23
Rolling-circles	490	422,622	0.05
Unclassified:	408,642	94,237,080	10.43
Total interspersed repeats:		360,475,210	39.89
Small RNA:	47,604	6,772,234	0.75
Satellites:	4173	920,778	0.10
Simple repeats:	777,200	46,246,753	5.12
Low complexity:	64,872	5,114,659	0.57

Table summary of repetitive elements by type detected and masked by RepeatMasker during repetitive element annotation.

relating to differences in available habitat between Europe and North America during that period.

It is worth noting that while Allis and American shads share similar anadromous life histories, Atlantic herring is a marine species inhabiting coastal habitats in large, highly fecund, and migratory schools (Whitehead 1985). These ecological and life history differences may underlie apparent differences in demographic history between herring and shads, as herring are not reliant on coastal rivers for spawning habitat, possibly reducing the effect of glacial cycles on N_e patterns. Previous work inferring the demographic history of Atlantic herring using genetic or genomic data has found large effective Ne, with microsatellite/allozyme data estimating N_e overlapping with infinity (Larsson et al. 2010), and whole-genome data estimating N_e in the millions (Martinez Barrio et al. 2016). In contrast, the N_e estimates presented here are substantially more modest, with N_e values ranging in the tens of thousands. Another study (Li et al. 2021) estimated the demographic history of Atlantic herring using the same genome assembly and inference method (PSMC) as our study, finding similar demographic patterns but with different absolute values of N_e. This is likely due to that study opting to use different model parameters and to directly estimate species mutation rates based on divergence time models, affecting the scaling of PSMC outputs. Martinez Barrio et al. (2016) used a multiple-genome approach to estimate Atlantic herring N_e with diCal (Sheehan et al. 2013) a method that can have improved demographic inference in recent time compared with PSMC. PSMC is limited in its ability to recover and estimate recent demographic changes (<10 KYA) due to its reliance on single-genome data and can be sensitive to assembly quality, model parameterization, and population structure (Li et al. 2021; Hilgers et al. 2024).

Comparative Phylogenetic Analysis: Identifying Signatures of Natural Selection

We first conducted an exploratory aBSREL selection analysis across each branch in our phylogenetic tree (Fig. 4a). This analysis revealed few genes overall bearing the signature of natural selection in the branches leading to American shad (3), Atlantic herring (1), and the nodes representing the ancestor of *Alosa* (1), *Clupeidae* (0), and Clupeiformes (4; Fig. 4a). The branch leading to the Allis shad (*Alosa alosa*) exhibited more loci bearing the signature of selection than

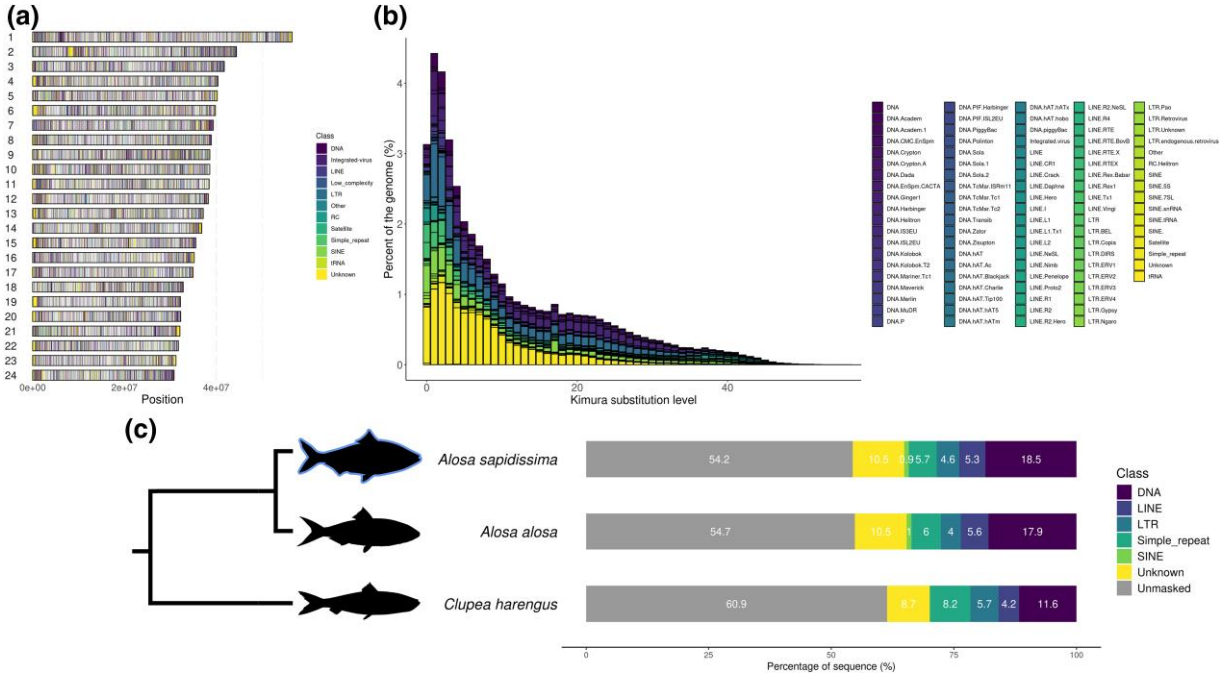

Fig. 2. Repetitive element annotation indicates that an especially high percentage of the American shad genome consists of repetitive content. a) Distribution of repetitive elements across chromosomes, colored by general type for American shad. b) Composition of repetitive elements. Repetitive element types are color coded according to their specific type and arranged according to their Kimura substitution level, which can be used to infer the relative age of the repetitive element. c) Comparison of total repeat content as a percentage of genome sequence between the American shad (*A. sapidissima*), Allis shad (*Alosa alosa*), and the Atlantic herring (*C. harengus*) colored by general type. Fish images were download from the public domain *via* phylopic.org.

any other branch or node in the tree (22; Fig. 4a). Note that we cannot rule out the possibility that detection of selection is inflated in Allis shad owing to increased error rates resulting from Oxford nanopore sequencing (Sahlin and Medvedev 2021), which was the primary method of sequencing used for the Allis shad assembly (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_017589495.1/). We detected no selection on genes in any other branch of the tree (Fig. 4a). No genes were shared across any branches or nodes bearing at least one gene under selection (supplementary table S2, Supplementary Material online). Of the four genes under selection in the branch leading to fishes in the order Clupeiformes (Node 3; Fig. 4a) *Aquaporin-1* (*Aqp1*) may be the most relevant to the evolution of flexible salinity tolerances in this group since it is a crucial component of the osmoregulatory process (see below). We detected no functional enrichment on any branch in the exploratory selection analysis.

Selection on *Apq1* in the branch representing the ancestor of the Clupeiformes may help to explain the extreme flexibility in salinity tolerance in this group. The order Clupeiformes consists of roughly 400 species, many of which are diadromous or otherwise occupy varied salinity habitats at some portion of their life history (Schultz and McCormick 2013). This is rare among fishes in which most clades, especially in more derived taxa, specialize in saltwater or freshwater but generally not both (Schultz and McCormick 2013). Although Clupeiformes evolved in the sea (Bloom and Egan 2018), numerous taxa within the group have transitioned to diadromy, at least 10 times independently (Bloom and Lovejoy 2014; Bloom and Egan 2018; DeHaan et al. 2023). Aquaporins are specialized water transport proteins that facilitate passive water movement across epithelial. Aqp1 in particular is highly expressed in intestines (Aoki et al. 2003; Giffard-Mena et al. 2007; Martinez et al. 2005) and most other tissues (Finn and Cerdà 2011) and is critical to the process of hypo-osmoregulation in seawater (Grosell et al. 2011); in this process, water is absorbed across the gut through Aqp1 channels to prevent dehydration. Other aquaporins have been shown to be repeatedly under natural selection in populations of species that vary in salinity tolerance (Velotta et al. 2022). Thus, selection on aquaporin may help explain success across salinity environments. Whether and how precisely genetic variation in *Aqp1* may underlie variation in hypo-osmoregulatory ability in Clupeiformes should be the subject of future research.

A targeted aBSREL analysis was conducted to detect selection on the branches leading to American shad, Allis shad, and the putative ancestor of the two (Node 5). This analysis identified 13 loci bearing the signature of natural selection in American shad, 42 in Allis shad, and 99 in

Table 2 Summary table of NCBI genome annotation results

Feature	Counts
Genes and pseudogenes	**47,628**
Protein coding	25,730
Noncoding	21,044
Nontranscribed pseudogenes	702
Genes with variants	12,758
Immunoglobulin/T-cell receptor gene segments	152
mRNAs	**55,994**
Fully supported	54,224
With > 5% ab initio	728
Partial	107
Model RefSeq (XM_)	55,994
Noncoding RNAs	**24,119**
Fully supported	8,416
Model RefSeq (XR_)	14,393
CDSs	**56,159**
Fully supported	54,224
With > 5% ab initio	838
Partial	107
With major correction(s)	781
Model RefSeq (XP_)	56,007

"Genes and pseudogenes" refers to gene and transcribed and nontranscribed pseudogene features. Features labeled ">5% *ab initio*" refers to features supported only partially by experimental evidence, for which more than 5% of their length was predicted by Gnomon using a hidden Markov model. "Noncoding RNAs" refers to misc_RNA, tRNA, rRNA, and ncRNA of all classes and does not include pseudogenes. "Model RefSeq (XR_)" refers to noncoding transcripts (misc_RNA, lncRNA, rRNA, snRNA, snoRNA, and guide_RNA) predicted by Gnomon or RFAM and cmsearch and assigned XR_* accessions. "Model RefSeq (XP_)" refers to proteins predicted by Gnomon and assigned XP_* accessions. Features labeled "with major correction(s)" refers to CDSs with correction for premature stop codons, frameshifts, or internal gaps.

Node 5 (Fig. 4b). The majority of genes under selection in this analysis were unique to individual branches. A single gene was shared by American shad and Node 5 (*Dcaf13;* supplementary table S2, Supplementary Material online) while two were shared between Allis shad and Node 5 (*Trim69* and a B-like apolipoprotein; supplementary table S2, Supplementary Material online). No genes bearing the signature of selection were shared between American and Allis shad (Fig. 4b). This result is not surprising given that these two species share an anadromous life history that arose only once in the common ancestor of all *Alosa sp.* We detected significant functional enrichment only in the branch leading to the ancestor of *Alosa* (Node 5), including two GO Biological Process terms: "response to lipid, and smooth muscle cell differentiation." We speculate that responses to lipids in particular may reflect selection for more efficient fat metabolism in freshwater. In threespine stickleback, for instance, duplication and copy number variation in the fatty acid desaturase gene (*Fads2*) leads to more efficient synthesis of long-chain fatty acids in lipid-poor freshwater environments (Ishikawa et al. 2019). Evolution of an anadromous life history requires that larva feed and grow in freshwater, before returning to the ocean, which may have selected for increased metabolism of fats during development. One gene contributing to this term is a fatty acid translocase known as *CD36*, which is known to be a high-efficiency receptor for long-chain fatty acid cellular uptake (Pepino et al. 2014).

Comparative Phylogenetic Analysis: Rapid Gene Family Expansion and Contraction

We used CAFE 5 to identify gene families experiencing expansion or contraction at a "rapid" evolutionary rate relative to a background rate of change (Mendes et al. 2021). Rapid gene family expansion and contraction was not distributed evenly across the phylogeny (Fig. 5a): most gene family evolution occurred in either extant *Alosa* branch. First, we detected minimal overlap in the gene families exhibiting rapid expansions or contractions between species of *Alosa* and the branch representing their common ancestor (Fig. 5b and c). Second, we found that Allis shad exhibited an overwhelming majority of gene family contractions (Fig. 5c), while American shad exhibited the majority of gene family expansions (Fig. 5b). This result suggests that gene family-level adaptation may be different in each species of *Alosa*, resulting in extensive family loss in Allis shad but simultaneous gain in different gene families in American shad. This is unexpected given that Allis and American shads share a recent common ancestor (~3 MYA *via* timetreeorg: Wilson et al. 2008; Rabosky et al. 2013, 2018) and that all *Alosids* share a single ancestor in which anadromy evolved once (Bloom and Egan 2018). However, we cannot rule out the possibility that this result is instead an artifact of differences in gene annotation between species (supplementary table S1, Supplementary Material online). While BUSCO scores are nearly identical (~95%), the American shad genome annotation contains about 2,000 additional features compared with Allis shad (supplementary table S1, Supplementary Material online). Moreover, ~50% of American shad expanded gene families were detected as gene family contractions in Allis shad. Of these expanded orthogroup genes in American shad, 95% of them blast to a unique location in the Allis shad transcriptome (Pasquier et al. 2016; data not shown). Thus, we urge caution when interpreting our results, and any results, where gene annotations may be imbalanced among species comparisons.

We performed functional enrichment analysis on all rapidly evolving gene families in American shad, Allis shad, and the node representing their common ancestor (Node 5). Because of the above-mentioned bias in gene annotation in Allis shad, we omitted gene families that were both contracted in Allis shad and expanded in American shad, thus restricting the analysis to just those gene families that were uniquely expanded/contracted in each branch. For Allis shad gene family expansions and contractions, as well as American shad gene family contractions, we detected enrichment of Gene Ontology Molecular Function

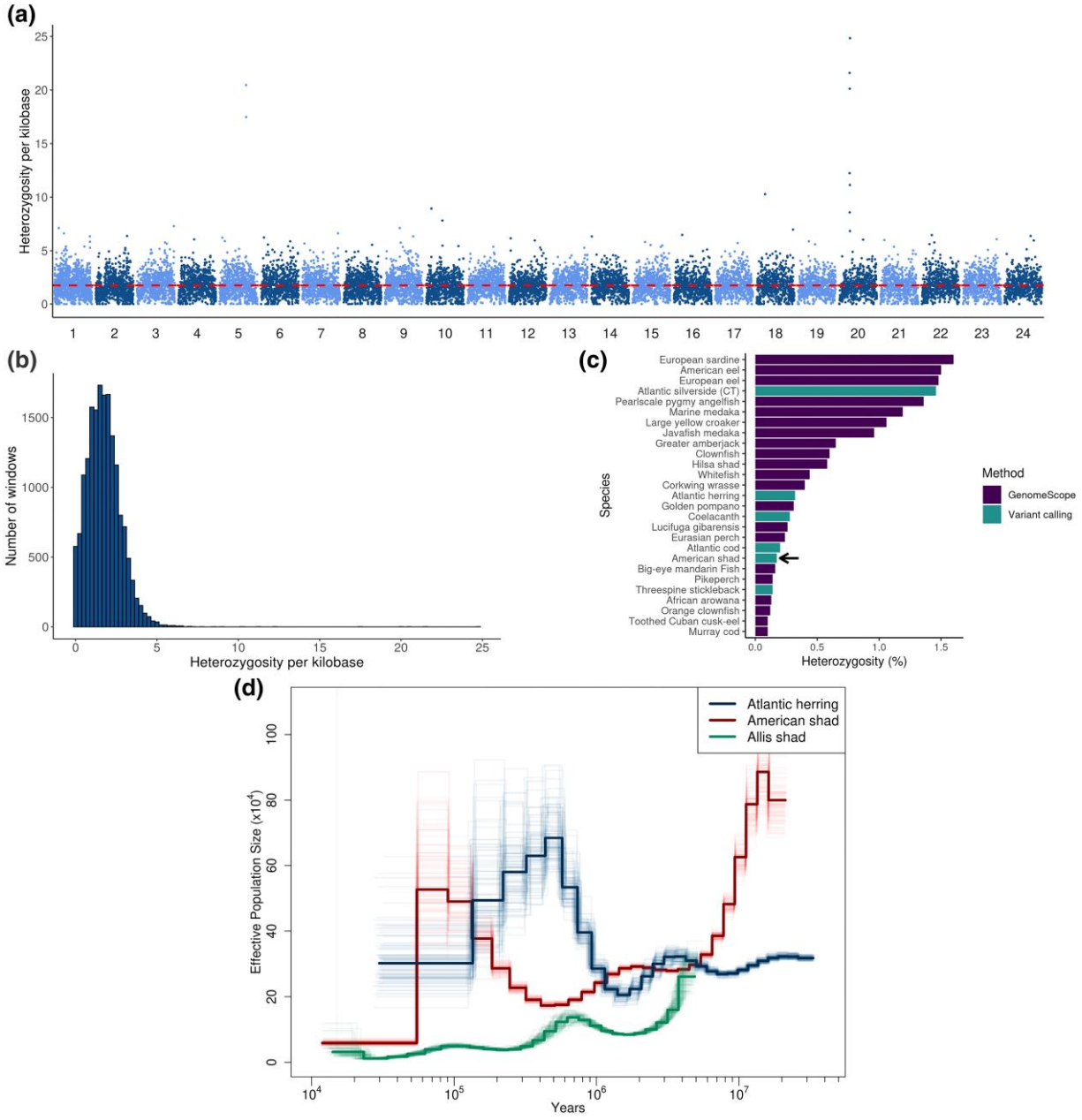

Fig. 3. Measuring genome-wide diversity reveals low heterozygosity and historical population contractions in American shad. a) Heterozygosity plotted for each chromosome in nonoverlapping 50-kb windows for the American shad. Mean genome-wide heterozygosity is marked with a dashed line. b) A histogram view of heterozygosity with counts of nonoverlapping 50-kb windows in American shad. c) Comparison of mean genome-wide heterozygosity among fish species with available data, adapted from Tigano et al. (2021). American shad is denoted with an arrow. d) Demographic histories (effective population size -*Ne*) of American shad, Allis shad, and Atlantic herring inferred by PSMC. Bootstrap replicates are represented by semi-transparent lines for each species.

term "GTPase signal transduction, and GTPase regulator activity," among others (supplementary table S5, Supplementary Material online). Enriched genes include several subfamilies of dedicator of cytokinesis proteins, which play a role in the regulation of cytoskeleton organization (Benson and Southgate 2021). This is potentially relevant since cytoskeletal strain and subsequent remodeling are used as an osmosensory signal of changing salinities (Kültz 2012). Thus, it is not surprising that gene families involved in cytoskeleton remodeling would have rapidly evolved upon transition from marine to diadromous life history, as appears to be the case in *Alosa*. We detected no enriched functions in unique gene families that exhibited rapid expansion in American shad.

In the branch leading to the *Alosa* ancestral node, we detected significant enrichment of the Gene Ontology

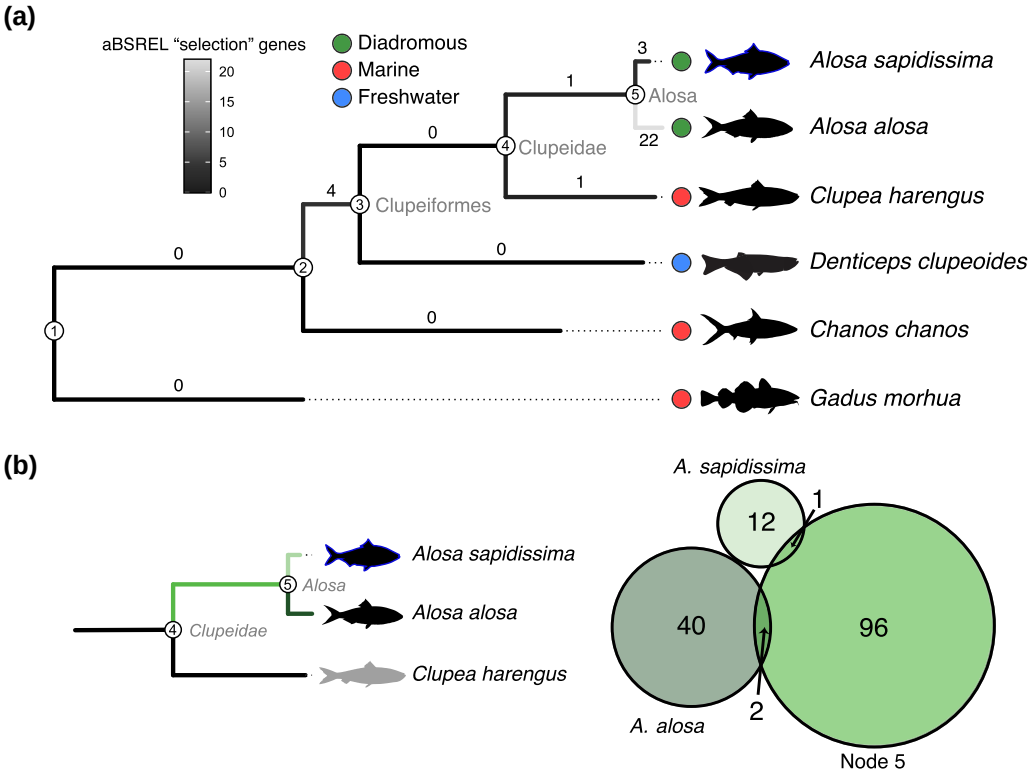

Fig. 4. Results of *dN/dS* selection tests reveal genes bearing the signature of natural selection across the Clupeiformes and in the branches leading the *Alosa*. a) Results of exploratory analysis reveal few selection genes at each branch. Branches are shaded to represent the number of selection genes following figure legend. b) Results of targeted selection tests on branches leading to American shad (*A. sapidissima*), Allis shad (*A. alosa*), and the branch leading to the putative ancestor of the two (Node 5). Colors of Venn diagram represent specific branches presented in phylogeny. Venn diagram shows overlap in the number of selection genes for each branch. Th phylogeny was created using OrthoFinder and selection tests were performed in aBSREL (see Materials and Methods). Fish images were download from the public domain *via* phylopic.org except for *Denticeps clupeoides* by Emma S. Glenn, and *Chanos chanos* by Erika Schumacher (https://creativecommons.org/licenses/by-nc/3.0/; no changes were made).

Molecular Function term "Calcium ion transmembrane transport activity" (supplementary table S5, Supplementary Material online) for those gene families exhibiting significant contractions but not expansions. Gene families involved in functional enrichment include those coding for Ca^{2+} transporting proteins that help regulate intracellular Ca^{2+} homeostasis (e.g. *ATP2B4*), voltage-gated Ca^{2+}, which regulate membrane potentials (e.g. *CACNA1AB*) and those involved in Ca^{2+} release at neuromuscular junctions (e.g. *RYR3*). Genes in the family of calcium ATPases, in particular, have been shown to be under strong, repeated natural selection across several species in which salinity tolerance varies across populations (Velotta et al. 2022). Calcium ATPases are important for uptake in freshwater environments where environmental calcium is inherently low. Taken together, these results suggest that, at the branch leading to anadromous Alosids, there has been selection on gene families involved in voluntary skeletal muscle contraction—putatively for swimming and migration—as well as osmoregulation in low-ion environments.

Conclusions

As a broad assessment of the American shad genome, we analyzed repetitive element frequency, genome-wide diversity, and demographic history. We also conducted a comparative genomic analysis using available Clupeiformes genomes to detect signatures of natural selection. We found that repeat content is among the highest of fishes of similar genome size. In addition, shad genome-wide heterozygosity is low compared with other species which we suspect may be associated with range-wide population collapses since the 19th century. Despite this, ROH were uncommon and short, suggesting that the level of inbreeding remains low. We inferred a precipitous decline in effective N_e around 100 KYA, which has not rebounded. This decline coincides with the onset of the Last Glacial Period, which likely restricted access to breeding habitat (Andersen et al. 2004) and is fairly common among fishes of the Northern Hemisphere (Li et al. 2021).

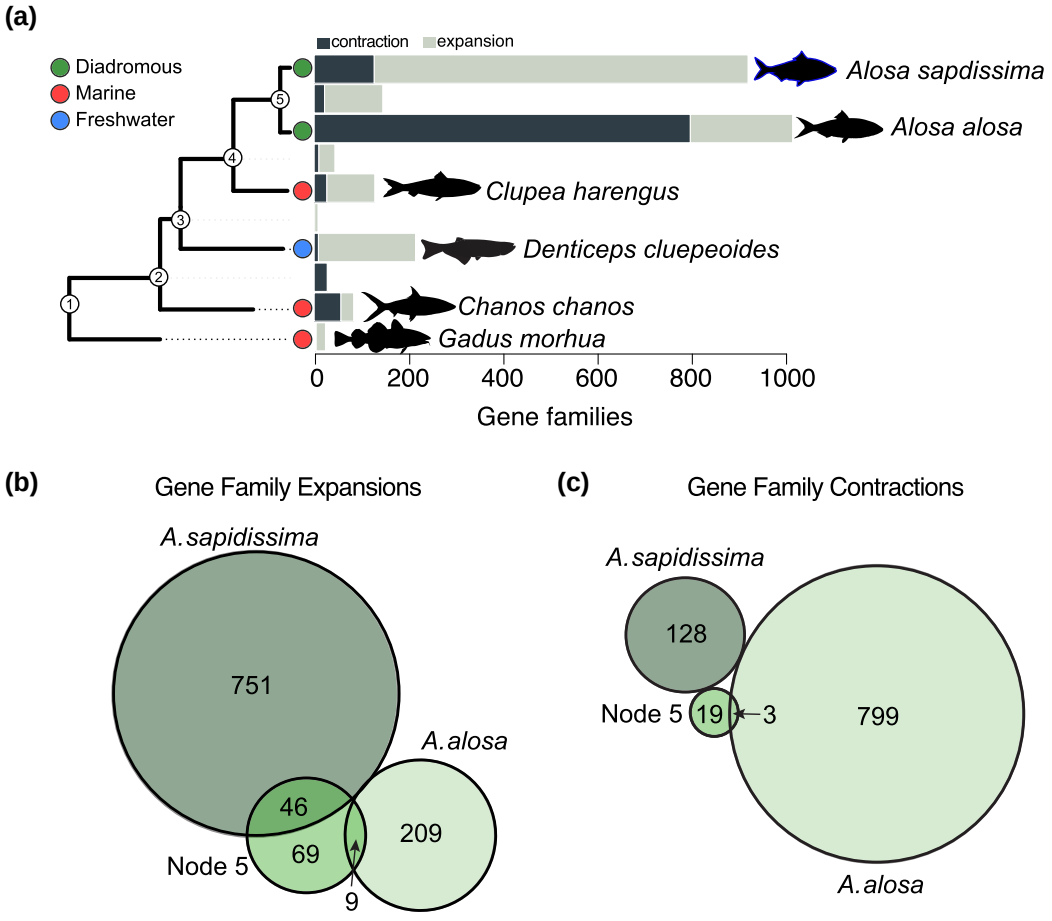

Fig. 5. Results of gene family expansion and contraction across the phylogeny (a) and in branches leading to *Alosa* (b, c). Venn diagrams represent overlap in numbers of gene families exhibiting significant expansion (b) or contraction (c). Fish images were download from the public domain *via* phylopic.org except for *Denticeps clupeoides* by Emma S. Glenn, and *Chanos chanos* by Erika Schumacher (https://creativecommons.org/licenses/by-nc/3.0/; no changes were made).

Our results suggest that natural selection has acted on the branch leading to the genus *Alosa*, in terms of both protein-coding evolution and rapid expansion of functional gene families. Some of the key targets of selection are likely involved in pathways regulating diet and metabolic fuel use (lipid metabolism) as well as cytoskeletal remodeling involved in osmoregulation. We hypothesize that natural selection on fatty acid metabolism, in particular, may improve fitness for developing juveniles in nutrient-poor, natal freshwater environments. Moreover, rapid gene family expansion of pathways involved in cytoskeletal remodeling suggests evolution of osmosensation, the mechanisms by which cells sense and respond to changes in osmotic pressure.

As is likely the case with many studies of nonmodel organisms, we are underpowered to detect functional consequences of selection tests, since few genes have Gene Ontology (GO) annotations. Moreover, across our phylogeny, we detected merely 1,800 single-copy orthologs with confidence, reducing our ability to detect selection to just over 7% of the ~25,000 protein-coding genes. This is likely due to the expansive evolutionary time periods separating the species in our tree. It is clear from this work that much additional sequencing across the order Clupeiformes is needed. Nevertheless, we expect the assembly and analysis of the American shad genome serves to support future research in testing the hypotheses generated above and in broader understandings of the genomic basis of adaptation and life history evolution using this well-suited but understudied system.

Materials and Methods

Genome Sequencing, Assembly, and Annotation

Sample Collection

An adult female American shad (*Alosa sapidissima*) was collected from the St. Johns River, Florida, USA (28.438892 N, 81.894728 W) for use in generating a high-quality reference genome assembly. Female muscle tissue was used for DNA extraction and subsequent library construction

and sequencing. To assess tissue-specific transcript expression and generate genome annotations, additional tissues were collected from an adult male individual from the same location to supplement tissues available from the collected female. In total, the tissues collected included dissections from the brain, muscle, gonads of both sexes, and liver, all of which were flash frozen in liquid nitrogen prior to RNA extraction and sequencing.

DNA Extraction

We used 200 mg of skeletal muscle for high molecular weight DNA isolation following the Circulomics Nanobind method using the Tissue Big DNA Kit (Circulomics NB-900-701-01). Tissue was kept on dry ice until disrupted with the Qiagen TissueRuptor II (cat. no. 9002755). Following purification, the DNA was kept at room temperature for 1 week prior to quantification to allow for homogenization of viscous DNA. We quantified DNA with the Qubit 3 fluorometer (Invitrogen Qubit dsDNA Broad Range Assay cat no. Q32850), and the size was measured with a pulsed-field gel electrophoresis (Pippin Pulse, SAGE Science, Beverly, MA).

RNA Extraction

Total RNA was extracted and purified from multiple tissues including brain, skeletal muscle, testes, ovaries, and liver, using the QIAGEN RNAeasy Protect kit (cat. no. 74124). A total of 20 to 30 mg of each tissue was homogenized with the Qiagen TissueRuptor II (cat. no. 9002755) as input. The quality of all RNA was assessed with a Fragment Analyzer (Agilent Technologies, Santa Clara, CA), and quantity was measured with a Qubit 3 Fluorometer (Qubit RNA BR Assay Kit—catalog no. Q33216).

Genome Sequencing and Assembly

As part of the Vertebrate Genomes Project (VGP), a high-quality reference genome assembly of *A. sapidissima* was generated at the Vertebrate Genomes Lab of the Rockefeller University using the VGP Pipeline 2.0 (Larivière et al. 2024) using a combination of long-read Pacific Biosciences (PacBio) HiFi long-read sequencing, Hi-C short read, and optical genome mapping (OGM) technologies (supplementary fig. S1, Supplementary Material online).

PacBio Library Preparation and Sequencing

To generate HiFi reads, large insert libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (PN 100-938-900) following the manufacturer's instructions (PN 101-853-100 version 03) and sizes selected between 15 and 20 kb with Sage BluePippin (Sage Science, USA). Sequence reads were generated from these libraries using the PacBio Sequel II HiFi system with a 30-h movie time.

Bionano Library Generation and Optical map Imaging

We used OGM to allow for high-quality scaffolding using Bionano imaging instruments. Libraries for optical mapping were generated using unfragmented ultra-HMW DNA that was labeled using DLE-1 following Bionano Prep Labeling NLRS (document number 30024) and DLS protocols (document number 30206). Labeled DNA samples were then imaged using the Bionano Saphyr instrument.

Hi-C Library Preparation and Sequencing

To generate high-quality scaffolds, we used chromatin interaction (Hi-C) libraries prepared by Arima Genomics (https://arimagenomics.com/) using the two-enzyme Arima-HiC kit (P/N: A510008). The proximally ligated DNA was then sheared and size selected between 200 and 600 bp using SPRI beads and enriched for biotin-labeled DNA using streptavidin beads. Illumina-compatible libraries were generated from the resulting fragments using the KAPA Hyper Prep kit (P/N: KK8504), which were subsequently PCR amplified and purified using SPRI beads. The libraries were then sequenced on Illumina HiSeq X at ~60× coverage following manufacturer protocols.

HiFi reads were used to generate haplotype-separated contigs using Hifiasm v0.14 (Cheng et al. 2021) with default parameters. Haplotypic false duplications were identified and purged from primary contigs with purge_dups v1.0.1 (Guan et al. 2020) using default parameters. Primary contigs were then scaffolded with Bionano optical maps using Bionano Solve v3.6.1 with default parameters. A further round of scaffolding was performed with long-range Hi-C data using the Salsa v2.2 program (Ghurye et al. 2019) with default parameters. The genome was manually curated with gEVAL v2021-04-21 (Chow et al. 2016) and Hi-C maps against the scaffolds to correct potential assembly structural errors, to manually join and align unplaced scaffolds, and to name chromosomes (Howe et al. 2021). The manual curation efforts of the primary haplotype resulted in decontamination, a 5.2% reduction of assembly length due to the removal of remnant artificial duplication, a 5.4% increase in scaffold N_{50} (68.6% increase in N_{90}), and assignment of 99.4% of the assembly sequence to the 24 identified chromosomes. The primary haplotype was named fAloSap1.pri (GCA_018492685.1) and the alternate fAloSap1.alt (GCA_018492705.1) using the VGP naming scheme (Rhie et al. 2021): f for fish, AloSap for species name, 1 for first VGP quality assembly, and pri and alt for primary and alternate, respectively.

Genome annotation was completed by the National Center for Biotechnology Information (NCBI) using the NCBI Eukaryotic Genome Annotation Pipeline (Thibaud-Nissen et al. 2013) run on the fAloSap1.pri

assembly (GCF_018492685.1) generated by this project. To assess annotation completeness, BUSCO v4.1.4 (Simão et al. 2015) was run in "protein" mode on the annotated gene set picking one longest protein per gene and run using the actinopterygii_odb10 lineage dataset.

Genome Assembly Analyses

The final primary assembly was analyzed for completeness using BUSCO (Simão et al. 2015) run on the actinopterygii_odb10 lineage dataset, along with the vertebrata_odb10, metazoa_odb10, and eukaryota_odb10 datasets, respectively. The genome assembly was run through NCBI BLAST using the NCBI nucleotide database using blastn to generate taxonomic annotations for each sequence in the assembly for output using BlobTools2 (Challis et al. 2020), with -max_target_seqs set to 10, -max_hsps set to 1, and -evalue set to 1e−25 as per BlobToolkit guidelines. To examine the sequencing depth across scaffolds/chromosomes, we mapped the PacBio HiFi reads onto the genome assembly with minimap2 2.24 (Li 2018) using the map-hifi setting. The resulting alignments were sorted using SAMtools sort (Li et al. 2009) and output in BAM format. The BUSCO outputs, BLAST hits, and BAM alignments were added into a BlobTools2 dataset to generate visualizations of genome assembly statistics (Fig. 1a), scaffold/chromosome base coverage (supplementary fig. S2, Supplementary Material online), and BUSCO completeness scores (Fig. 1a).

Investigating Heterozygosity

PacBio HiFi reads used to generate the reference sequence for fAloSap1.pri were mapped back onto the reference genome with minimap2 2.24 (Li 2018) with the map-hifi setting and sorted using SAMtools sort. To investigate the degree of heterozygosity present in the genome, we used GATK 4.5.0 (McKenna et al. 2010) for genotype calling using HaplotypeCaller followed by GenotypeGVCFs. All sites including both variants and invariant sites were called using the -ERC BP_RESOLUTION and flag in HaplotypeCaller and the -allSites flag in GenotypeGVCFs. SNPs, indels, and invariant sites were filtered separately according to GATK best practices, and we further filtered out sites with excessively low or high read depth (1/3× and 2× the genome-wide average, respectively) and SNPs within 10 bp of an indel. To estimate heterozygosity, we calculated SNP density over 50 kb fixed windows, filtering out windows with a genotyping rate < 0.5 indicating an excess of sites with no coverage. We examined the effect of larger and smaller window sizes (10 kb to 1 mb) on heterozygosity estimates and found that smaller window sizes captured localized regions with especially high heterozygosity (supplementary fig. S7a, Supplementary Material online). We observed limited differences (~2.6%) in mean estimated heterozygosity using windows larger than 10 kb (supplementary fig. S7b, Supplementary Material online). We ultimately chose 50-kb windows due to their ability to represent fine-scale patterns without presenting excessive bias toward highly localized, high-diversity regions. ROH were identified using the PLINK 1.9 –homozyg function (Chang et al. 2015) using default parameters except for homozyg-kb set to 100 to allow for detection of ROH > 100 kb in length.

The PacBio HiFi-mapped reads were also analyzed using GenomeScope2 (Ranallo-Benavidez, Jaron, and Schatz 2020) to generate estimates of heterozygosity and genome size. However, the resulting *k*-mer spectra plots exhibited poor model fit, and the resulting heterozygosity and genome size estimates were substantially higher than estimates generated using more traditional methods of nuclei size measurements and short-read data. This is likely due to repeat content high in GA repeats in the genome (see Results) that is challenging to sequence for PacBio HiFi, biasing *k*-mer distributions away from GenomeScope model assumptions.

Annotating Repetitive Regions

To annotate the repetitive regions in the genome, RepeatModeler v2.0.1 (Flynn et al. 2020) was used to build a repeat library for the American shad genome employing the Repbase database (Bao, Kojima, and Kohany 2015) using the latest version 28.04 (2023 April 27). The repeat consensus database was then filtered to exclude known protein sequences with the uniref90 database (known transposases provided with RepeatMasker were preremoved from the uniref90 database). The remaining repeat database was then classified with Repbase and used to generate a repeat annotation GFF file using RepeatMasker (Materials and Methods adapted from Stanhope et al. 2023). The repeat sequences in the genome assembly were soft masked with RepeatMasker (version 4.1.0) using a repeat library created by merging the Repbase database with de novo TEs identified by RepeatModeler. This process was repeated for the genome assemblies for the Atlantic herring (*C. harengus,* GCF_900700415.1) and Allis shad (*Alosa alosa*, GCF_017589495.1) to compare repeat content between the American shad and other members of the family *Clupeidae*.

Demographic History Inference

To better contextualize the observed patterns of genome-wide diversity, we inferred historical shifts in effective N_e using the diploid genome of the American shad (generated via variant calling to investigate heterozygosity) with the PSMC (Li and Durbin 2011). We also inferred the demographic history of two related species—*C. harengus* (Atlantic herring), a marine clupeid native to the North

Atlantic (GCF_000966335.1, a short-read draft quality genome), and *Alosa alosa* (Allis shad), an anadromous congener native to Europe (GCF_017589495.1, an unphased ONT long-read genome).

PacBio HiFi reads used to generate the American shad genome were aligned to it using minimap2 (Li 2018) with the map-hifi preset, while the Oxford Nanopore reads used to generate the Allis shad genome were aligned with minimap2 using the map-ont preset. As the recommended PSMC settings for generating a diploid consensus sequence were not designed for long-read sequencing, a diploid consensus genome was generated by instead calling variants using bcftools (Danecek et al. 2021) mpileup 1.18 with the pacbio-ccs preset for the American shad alignments and the ont preset for the Allis shad alignments. Variants were called from the generated pileup files with bcftools call using the -c -Ov parameters. The resulting VCF was used to generate a diploid consensus FASTQ file with the vcfutils vcf2fq function with a minimum and maximum depth of ⅓ and 2 times the mean depth across the sample, respectively.

For the Atlantic herring assembly, we used the Illumina short-read dataset by first trimming reads using TRIMMOMATIC 0.39 and mapping trimmed reads with BWA-MEM (Li 2013). The resulting alignments were then deduplicated using Picard 2.9.0 (https://broadinstitute.github.io/picard/) and merged into a single alignment file using SAMtools merge (Danecek et al. 2021) and filtered to only retain alignments from scaffolds longer than 50 kb. A diploid consensus genome was generated by first generating a pileup file bcftools 1.18 mpileup and downgrading mapping quality for reads with excessive mismatches with the -C50 flag. We then called variants on the pileup file using bcftools call consensus caller function (-c) and converted the resulting VCF file to a diploid consensus FASTQ file with the vcfutils vcf2fq function with a minimum and maximum depth of ⅓ and 2 times the mean depth across the sample, respectively.

PSMC was run on a maximum of 25 iterations with parameters: -N25 -t15 -r5 -p "4 + 25 × 2 + 4 + 6" with 100 bootstrapped replicates to account for variation in Ne estimates for all three species. Time estimates on the PSMC output were scaled using the published Atlantic herring mutation rate of 2.0×10^{-9} (Feng et al. 2017) for all species, as there are currently no published estimates of American shad or Allis shad mutation rate. Generation times were set to 4 years for American shad, 4 years for Allis shad, and 6 years for Atlantic herring based on published life history estimates (Feng et al. 2017; IUCN 2024).

Assessing Synteny with the Allis Shad Genome Assembly

To identify the degree of synteny between the American shad genome assembly and that of a congener, the Allis shad (*Alosa alosa*), we aligned the American shad genome (GCF_018492685.1) to the Allis shad genome (Accession GCF_017589495.1) using LAST, a *lastal* program (Kiełbasa et al. 2011; Frith and Kawaguchi 2015) with the *-m100* parameter to increase alignment sensitivity and *last-split* to identify the unique best alignments for each part of each queried sequence. Alignments shorter than 200 bp were filtered, and syntenic blocks were visualized with circos plots generated with the R package *circlize* (Gu et al. 2014).

Comparative Phylogenetic and Selection Analyses

We conducted a comparative phylogenetic analysis to identify genetic signatures of natural selection in the branch leading to American shad as well as across species in the order Clupeiformes and the genus *Alosa*. To do so, we analyzed genome-wide ratios of nonsynonymous to synonymous substitution rates (*dN/dS*). We calculated *dN/dS* from alignments of orthologous protein CDSs from six species including four from the Clupeiformes: *Alosa sapidissima* (American shad), *Alosa alosa* (Allis shad), *Clupea harengus* (Atlantic herring), and *Denticeps clupeoides* (denticle herring), and two more distantly related outgroups: *Chanos chanos* (milkfish) and *Gadus morhua* (Atlantic cod; supplementary fig. S4, Supplementary Material online and supplementary table S1, Supplementary Material online). We included all genomes from Clupeiformes that had protein sequences available and were >75% complete (BUSCO score against actinopterygii_odb10 lineage dataset). *Alosa alosa* was previously the only other species in the genus *Alosa* for which a whole-genome, chromosome-level assembly existed. Incomplete genomes for other Clupeiformes species were omitted from the analysis. Single-copy orthologs were identified from protein sequences downloaded from NCBI in the program *OrthoFinder* (Emms and Kelly 2015). Only the longest transcript variant for each gene was used in the orthology analysis (Emms and Kelly 2019). A species tree was inferred from orthogroup results in *OrthoFinder* following the method by Emms and Kelly (2019). Orthologs were aligned using the Python program Guidance (Penn et al. 2010). Sequences that contained a percentage of gaps over 25% were excluded, as were alignments without all six species. We were left with 1,800 high-confidence single-copy orthologs for selection testing.

For each orthologous gene, we tested for signatures of natural selection using the model aBSREL (adaptive branch-site random effects likelihood model) implemented in HyPhy (Pond, Frost, and Muse 2005; Smith et al. 2015). In the aBSREL model, observed *dN/dS* ratios are tested against a null expectation where loci are constrained to not evolve by natural selection. *P*-values are obtained from a likelihood ratio test comparing the observed and null models. We

conducted two analyses for genes under selection using the aBSREL model: (i) An exploratory analysis in which a series of tests for selection at each branch and node in the tree was conducted. We applied a false-discovery rate (FDR) correction for multiple testing and the *P*-values generated from our exploratory analysis. (ii) A targeted series of selection tests in the branch leading to American shad. This was conducted by specifying the branch leading to American shad as the "foreground" branch in the model and is equivalent to the exploratory test results without correcting for multiple testing.

We used CAFE 5 (Mendes et al. 2021) to identify rates of expansion and contraction of gene families across our phylogeny. Gene families were identified in *OrthoFinder* as "orthogroups," which represent the set of genes in a group of species descended from a single gene (Emms and Kelly 2019). A table of gene counts for each orthogroup for each species was input into CAFE 5. CAFE 5 was run using the Gamma model with a single lambda estimated by CAFE 5 and parameter $k = 6$, indicating the use of six gamma rate categories, yielding a maximum likelihood value of −12.40, $\alpha = 1.18$, and $\lambda = 2.47$. Parameter $k = 6$ was selected as it produced the highest maximum likelihood when CAFE 5 was run with parameters $k = 1$ through $k = 10$. *P*-values were calculated on the rate of evolution of gene family expansion and contraction by comparing empirical rates to a stochastic birth/death model (Hahn et al. 2005). Raw *P*-values were corrected for multiple testing using FDR correction in *P.adjust* in R (R Core Team 2023). We considered "rapidly evolving" gene families to be any gene family with a significant FDR-corrected *P*-value < 0.05.

To perform functional enrichment analysis of rapidly evolving gene families, accession numbers from *C. harengus* were taken from significant orthogroups and used in conjunction with the NCBI Datasets command-line tool (www.ncbi.nlm.nih.gov/datasets) to isolate the corresponding gene symbol in *C. harengus*. To avoid bias in functional enrichment, we took the first gene accession from each orthogroup in cases where there was more than one for *C. harengus* (as in Prost et al. 2019). Functional enrichment tests were performed using these gene symbols. Rapidly evolving orthogroups without associated *C. harengus* accession numbers were omitted from this analysis. Functional enrichment tests were performed in *gProfiler* (Kolberg et al. 2023) using the g:GOSt function. Queries were tested against the *C. harengus* genome, and *P*-values were corrected using *gProfiler's* native g:SCS algorithm, which is designed to account for hierarchical, nonindependent associations implicit in GO terms.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We are grateful to Reid Hyle from the Florida Fish and Wildlife Conservation Commission for collecting the tissue samples used to generate the genome assembly. Our deepest thanks to Gregg Thomas of the Harvard Faculty of Arts and Science Informatics Group for providing invaluable guidance on comparative phylogenomic analyses. We thank Stephen Sabatino and Devin Bloom for their helpful suggestions during the review process.

Author Contributions

N.O.T., J.P.V., and A.R.I. conceived of the work. E.D.J. oversaw genome sequencing, assembly, and annotation. G.F., J.M., J.B., A.T., Y.S., K.H., and O.F. conducted the genome sequencing, assembly, annotation, and curation. J.P.V., A.R.I., E.S.G., and R.P.F. analyzed the data. J.P.V. and A.R.I. wrote the manuscript with input from all authors. N.O.T. funded the work.

Funding

Funding for this work was provided through (i) a State Wildlife Grant awarded to the New York State Department of Environmental Conservation by the US Fish and Wildlife Service and from the New York State Environmental Protection Fund (grant to N.O.T.) and (ii) HHMI, which helped support parts of the genome sequencing effort at the Rockefeller VGL.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability

The reference genome for American shad is published on the NCBI genomes database under the accession GCF_018492685.1.

Literature Cited

Alexandrou MA, Swartz BA, Matzke NJ, Oakley TH. Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. Mol Phylogenet Evol. 2013:69(3):514–523. https://doi.org/10.1016/j.ympev.2013.07.026.

Andersen KK, Azuma N, Barnola JM, Bigler M, Biscaye P, Caillon N, Chappellaz J, Clausen HB, Dahl-Jensen D, Fischer H, et al. High-resolution record of northern hemisphere climate extending into the last interglacial period. Nature. 2004:431(7005):147–151. https://doi.org/10.1038/nature02805.

Aoki M, Kaneko T, Katoh F, Hasegawa S, Tsutsui N, Aida K. Intestinal water absorption through aquaporin 1 expressed in the apical membrane of mucosal epithelial cells in seawater-adapted Japanese eel. J Exp Biol. 2003:206(19):3495–3505. https://doi.org/10.1242/jeb.00579.

Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015:6(1):11. https://doi.org/10.1186/s13100-015-0041-9.

Benson CE, Southgate L. The DOCK protein family in vascular development and disease. Angiogenesis. 2021:24(3):417–433. https://doi.org/10.1007/s10456-021-09768-8.

Betancur-R R. Molecular phylogenetics supports multiple evolutionary transitions from marine to freshwater habitats in ariid catfishes. Mol Phylogenet Evol. 2010:55(1):249–258. https://doi.org/10.1016/j.ympev.2009.12.018.

Betancur-R R, Broughton RE, Wiley EO, Carpenter K, López JA, Li C, Holcroft NI, Arcila D, Sanciangco M, Cureton Ii JC, et al. The tree of life and a new classification of bony fishes. PLOS Curr. 2013:5: ecurrents.tol.53ba26640df0ccaee75bb165c8c26288. https://doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.

Bloom DD, Egan JP. Systematics of clupeiformes and testing for ecological limits on Species richness in a trans-marine/freshwater clade. Neotr Ichthyol. 2018:16:e180095. https://doi.org/10.1590/1982-0224-20180095.

Bloom DD, Lovejoy NR. The evolutionary origins of diadromy inferred from a time-calibrated phylogeny for Clupeiformes (herring and allies). Proc R Soc Lond B: Biol Sci. 2014:281(1778):20132081. https://doi.org/10.1098/rspb.2013.2081.

Burns MD, Bloom DD. Migratory lineages rapidly evolve larger body sizes than non-migratory relatives in ray-finned fishes. Proc R Soc Lond B Biol Sci. 2020:287(1918):20192615. https://doi.org/10.1098/rspb.2019.2615.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. 2020:10(4):1361–1374. https://doi.org/10.1534/g3.119.400908.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015:4(1):s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de Novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021:18(2):170–175. https://doi.org/10.1038/s41592-020-01056-5.

Chow W, Brugger K, Caccamo M, Sealy I, Torrance J, Howe K. gEVAL—a web-based browser for evaluating genome assemblies. Bioinformatics. 2016:32(16):2508–2510. https://doi.org/10.1093/bioinformatics/btw159.

Corush JB. Evolutionary patterns of diadromy in fishes: more than a transitional state between marine and freshwater. BMC Evol Biol. 2019:19(1):168. https://doi.org/10.1186/s12862-019-1492-2.

Crespi BJ, Teo R. Comparative phylogenetic analysis of the evolution of semelparity and life history in salmonid fishes. Evolution. 2002:56(5):1008–1020. https://doi.org/10.1111/j.0014-3820.2002.tb01412.x.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021:10(2): giab008. https://doi.org/10.1093/gigascience/giab008.

DeHaan LM, Burns MD, Egan JP, Bloom DD. Diadromy drives elevated rates of trait evolution and ecomorphological convergence in Clupeiformes (Herring, Shad, and Anchovies). Am Nat. 2023:202(6):830–850. https://doi.org/10.1086/726894.

Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015:16(1):157. https://doi.org/10.1186/s13059-015-0721-2.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019:20(1):238. https://doi.org/10.1186/s13059-019-1832-y.

Engle VD, Summers JK. Latitudinal gradients in benthic community composition in Western Atlantic estuaries. J Biogeogr. 1999:26(5):1007–1023. https://doi.org/10.1046/j.1365-2699.1999.00341.x.

Feng C, Pettersson M, Lamichhaney S, Rubin C-J, Rafati N, Casini M, Folkvord A, Andersson L. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. eLife. 2017:6:e23907. https://doi.org/10.7554/eLife.23907.

Finn RN, Cerda J. Aquaporin evolution in fishes. Front physiol. 2011:2. https://doi.org/10.3389/fphys.2011.00044.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020:117(17):9451–9457. https://doi.org/10.1073/pnas.1921046117.

Frith MC, Kawaguchi R. Split-alignment of genomes finds orthologies more accurately. Genome Biol. 2015:16(1):106. https://doi.org/10.1186/s13059-015-0670-9.

Garman GC, Macko SA. Contribution of marine-derived organic matter to an Atlantic coast, freshwater, tidal stream by anadromous clupeid fishes. J North Am Benthol Soc. 1998:17(3):277–285. https://doi.org/10.2307/1468331.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, PhillippyAM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019:15(8): e1007273. https://doi.org/10.1371/journal.pcbi.1007273.

Giffard-Mena I, Boulo V, Aujoulat F, Fowden H, Castille R, Charmantier G, Cramb G. Aquaporin molecular characterization in the sea-bass (*Dicentrarchus labrax*): The effect of salinity on AQP1 and AQP3 expression. Comp Biochem Physiol A Mol Integr Physiol. 2007:148(2):430–444. https://doi.org/10.1016/j.cbpa.2007.06.002.

Grosell M, Farrell AP, Brauner CJ. Fish Physiology: The Multifunctional Gut of Fish. Vol. 30. Academic Press; 2011.

Gu Z, Gu L, Eils R, Schlesner M, Brors B. *circlize* implements and enhances circular visualization in R. Bioinformatics. 2014:30(19): 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020:36(9):2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 2005:15(8):1153–1160. https://doi.org/10.1101/gr.3567505

Hasselman DJ, Hinrichsen RA, Shields BA, Ebbesmeyer CC. American shad of the Pacific coast: a harmful invasive species or benign introduction? Fisheries (Bethesda). 2012:37(3):115–122. https://doi.org/10.1080/03632415.2012.659941.

Hilgers L, Liu S, Jensen A, Brown T, Cousins T, Schweiger R, Guschanski K, Hiller M. Avoidable false PSMC population size peaks occur across numerous studies. bioRxiv. 2024. https://www.biorxiv.org/content/10.1101/2024.06.17.599025v2.

Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, Torrance J, Tracey A, Wood J. Significantly improving the quality of genome assemblies through curation. GigaScience. 2021:10(1):giaa153. https://doi.org/10.1093/gigascience/giaa153.

Ishikawa A, Kabeya N, Ikeya K, Kakioka R, Cech JN, Osada N, Leal MC, Inoue J, Kume M, Toyoda A, et al. A key metabolic gene for recurrent freshwater colonization and radiation in fishes. Science. 2019:364(6443):886–889. https://doi.org/10.1126/science.aau5656.

IUCN. The IUCN Red List of Threatened Species. Version 2024-2. 2024. https://www.iucnredlist.org.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011:21(3): 487–493. https://doi.org/10.1101/gr.113985.110.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980:16(2):111–120. https://doi.org/10.1007/BF01731581.

Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. G:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res.2023:51(W1):W207–W212. https://doi.org/10.1093/nar/gkad347.

Kültz D. The combinatorial nature of osmosensing in fishes. Physiology. 2012:27(4):259–275. https://doi.org/10.1152/physiol.00014.2012.

Larivière D, Abueg L, Brajuka N, Gallardo-Alba C, Grüning B, Ko BJ, Ostrovsky A, Palmada-Flores M, Pickett BD, Rabbani K, et al. Scalable, accessible and reproducible reference genome assembly and evaluation in galaxy. Nat Biotechnol. 2024:42:367–370. https://doi.org/10.1038/s41587-023-02100-3.

Larsson LC, Laikre L, André C, Dahlgren TG, Ryman N. Temporally stable genetic structure of heavily exploited Atlantic herring (Clupea harengus) in Swedish waters. Heredity. 2010:104(1): 40–51. https://doi.org/10.1038/hdy.2009.98.

Lee CE, Bell MA. Causes and consequences of recent freshwater invasions by saltwater animals. Trends Ecol Evol. 1999:14(7):284–288. https://doi.org/10.1016/S0169-5347(99)01596-7.

Leggett WC, Carscadden JE. Latitudinal variation in reproductive characteristics of American shad (*Alosa sapidissima*): evidence for population specific life history strategies in fish. J Fish Res Board Can. 1978:35(11):1469–1478. https://doi.org/10.1139/f78-230.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. https://doi.org/10.48550/arXiv.1303.3997.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li H, Durbin R. Inference of human population history from whole genome sequence of a single individual. Nature. 2011:475(7357): 493–496. https://doi.org/10.1038/nature10231.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 genome project data processing subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009:25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li J, Bian C, Yi Y, Yu H, You X, Shi Q. Temporal dynamics of teleost populations during the Pleistocene: a report from publicly available genome data. BMC Genomics. 2021:22(1):490. https://doi.org/10.1186/s12864-021-07816-7.

Limburg K, Hattala KA, Kahnle A. American shad in its native range. Am Fish Soc Symp. 2003:35:125–140.

Limburg KE, Waldman JR. Dramatic declines in north Atlantic diadromous fishes. BioScience. 2009:59(11):955–965. https://doi.org/10.1525/bio.2009.59.11.7.

Lu B, Jin H, Fu J. Molecular convergent and parallel evolution among four high-elevation anuran species from the Tibetan region. BMC Genomics. 2020:21(1). https://doi.org/10.1186/s12864-020-07269-4.

Martinez AS, Cutler CP, Wilson GD, Phillips C, Hazon N, Cramb G. Regulation of expression of two aquaporin homologs in the intestine of the European eel: effects of seawater acclimation and cortisol treatment. Am J Physiol Regul Integr Comp Physiol. 2005:288(6):R1733–R1743. https://doi.org/10.1152/ajpregu.00747.2004.

Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang HE, Dainat J, Ekman D, Höppner M, Jern P, et al. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife. 2016:5. https://doi.org/10.7554/eLife.12081.

McCormick SD. Smolt physiology and endocrinology. In: McCormick SD, Farrell AP, Brauner CJ, editors. Euryhaline fishes: vol. 32, fish physiology. New York: Elsevier; 2013. p. 199–251.

McDowall RM. Diadromy in fishes: migrations between freshwater and marine environments. London: Croom Helm; 1988.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010:20(9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2021:36(22–23):5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.

Norman JD, Danzmann RG, Glebe B, Ferguson MM. The genetic basis of salinity tolerance traits in Arctic charr (*Salvelinus alpinus*). BMC Genet. 2011:12(1):81. https://doi.org/10.1186/1471-2156-12-81.

Norman JD, Robinson M, Glebe B, Ferguson MM, Danzmann RG. Genomic arrangement of salinity tolerance QTLs in salmonids: a comparative analysis of Atlantic salmon (*Salmo salar*) with Arctic charr (*Salvelinus alpinus*) and rainbow trout (Oncorhynchus mykiss). BMC Genomics. 2012:13(1):420. https://doi.org/10.1186/1471-2164-13-420.

Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P, Klopp C, Postlethwait JH, et al. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC Genomics. 2016:17(1):368. https://doi.org/10.1186/s12864-016-2709-z.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res. 2010:38(suppl_2):W23–W28. https://doi.org/10.1093/nar/gkq443.

Pepino MY, Kuda O, Samovski D, Abumrad NA. Structure-function of CD36 and importance of fatty acid signal transduction in fat metabolism. Annu Rev Nutr. 2014:34(1):281–303. https://doi.org/10.1146/annurev-nutr-071812-161220.

Petit JR, Jouzel J, Raynaud D, Barkov NI, Barnola JM, Basile I, Bender M, Chappellaz J, Davis M, Delaygue G, et al. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. Nature. 1999:399(6735):429–436. https://doi.org/10.1038/20859.

Pond SLK, Frost SDW, Muse SV. Hyphy: hypothesis testing using phylogenies. Bioinformatics. 2005:21(5):676–679. https://doi.org/10.1093/bioinformatics/bti079.

Prost S, Armstrong EE, Nylander J, Thomas GWC, Suh A, Petersen B, Dalen L, Benz BW, Blom MPK, Palkopoulou E, et al. Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise. GigaScience. 2019:8(5):giz003. https://doi.org/10.1093/gigascience/giz003.

Rabosky DL, Chang J, Cowman PF, Sallan L, Friedman M, Kaschner K, Garilao C, Near TJ, Coll M, Alfaro ME. An inverse latitudinal gradient in speciation rate for marine fishes. Nature. 2018:559(7714): 392–395. https://doi.org/10.1038/s41586-018-0273-1.

Rabosky DL, Santini F, Eastman J, Smith SA, Sidlauskas B, Chang J, Alfaro ME. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat Commun. 2013:4(1). https://doi.org/10.1038/ncomms2958.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.

Nat Commun. 2020:11(1):1432. https://doi.org/10.1038/s41467-020-14998-3.
R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. https://www.R-project.org.
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021:592(7856):737–746. https://doi.org/10.1038/s41586-021-03451-0.
Sabatino SJ, Pereira P, Carneiro M, Dilytė J, Archer JP, Munoz A, Nonnis-Marzano F, Murias A. The genetics of adaptation in freshwater Eurasian shad (*Alosa*). Ecol Evol. 2022:12(5):e8908. https://doi.org/10.1002/ece3.8908.
Sahlin K, Medvedev P. Error correction enables use of Oxford nanopore technology for reference-free transcriptome analysis. Nat Commun. 2021:12(1):2. https://doi.org/10.1038/s41467-020-20340-8.
Schultz ET, McCormick SD. Euryhalinity in an evolutionary context. In: McCormick SD, Farrell AP, Brauner CJ, editors. Euryhaline fishes: vol. 32, fish Physiology. New York: Elsevier; 2013. p. 477–533.
Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. Genetics. 2013:194(3):647–662. https://doi.org/10.1534/genetics.112.149096.
Shubin NH, Daeschler EB, Jenkins FA. The pectoral fin of Tiktaalik roseae and the origin of the tetrapod limb. Nature. 2006:440(7085):764–771. https://doi.org/10.1038/nature04637.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015:31(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
Simpson GG. Tempo and mode in evolution. New York: Columbia University Press; 1944.
Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 2015:32(5):1342–1353. https://doi.org/10.1093/molbev/msv022.
Stanhope MJ, Ceres KM, Sun Q, Wang M, Zehr JD, Marra NJ, Wilder AP, Zou C, Bernard AM, Pavinski-Bitar P, et al. Genomes of endangered great hammerhead and shortfin Mako sharks reveal historic population declines and high levels of inbreeding in great hammerhead. iScience. 2023:26(1):105815. https://doi.org/10.1016/j.isci.2022.105815.
Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, et al. The genome sequence of Atlantic cod reveals a unique immune system. Nature. 2011:477(7363):207–210. https://doi.org/10.1038/nature10342.
Thibaud-Nissen F, Souvorov A, Murphy T, DiCuccio M, Kitts P. Eukaryotic genome annotation pipeline. The NCBI handbook. 2. Bethesda (MD): National Center for Biotechnology Information (US); 2013.
Tigano A, Jacobs A, Wilder AP, Nand A, Zhan Y, Dekker J, Therkildsen NO. Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation. Genome Biol Evol. 2021:13(6):evab098. https://doi.org/10.1093/gbe/evab098.
Velotta JP, McCormick SD, O'Neill RJ, Schultz ET. Relaxed selection causes microevolution of seawater osmoregulation and gene expression in landlocked alewives. Oecologia. 2014:175(4):1081–1092. https://doi.org/10.1007/s00442-014-2961-3.
Velotta JP, McCormick SD, Schultz ET. Trade-offs in osmoregulation and parallel shifts in molecular function follow ecological transitions to freshwater in the alewife. Evolution. 2015:69(10):2676–2688. https://doi.org/10.1111/evo.12774.
Velotta JP, McCormick SD, Whitehead A, Durso CS, Schultz ET. Repeated genetic targets of natural selection underlying adaptation of fishes to changing salinity. Integr Comp Biol. 2022:62(2):357–375. https://doi.org/10.1093/icb/icac072.
Walburg CH, Nichols PR. Biology and management of the American shad and status of the fisheries, Atlantic coast of the United States. U.S. Department of the Interior, Bureau of Commercial Fisheries; 1960.
Whitehead PJP. FAO species catalogue, vol. 7. Clupeoid fishes of the world (Suborder Clupeoidei). An annotated and illustrated catalogue of the herrings, sardines, pilchards, anchovies, and wolf-herrings. Part 1. Chirocentridae, Clupeidae, and Pristigasteridae. FAO Fisheries Synopsis. 1985:7(125):1–303.
Wilson AB, Teugels GG, Meyer A. Marine incursion: the freshwater herring of Lake Tanganyika are the product of a marine invasion into West Africa. PLoS ONE. 2008:3(4):e1979. https://doi.org/10.1371/journal.pone.0001979.
Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z. Comparative genome analysis of 52 fish Species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics. 2018:19(1):1–10. https://doi.org/10.1186/s12864-018-4516-1.

Associate editor: Bonnie Fraser