Does genome size increase with water depth in marine fishes?

Aline P. M. Medeiros,1,2,3,∗ Bráulio A. Santos,4 and Ricardo Betancur-R.2,5,∗

1Programa de Pós-Graduação em Ciências Biológicas, Universidade Federal da Paraíba, João Pessoa 58051-900, Brazil
2Department of Biology, The University of Oklahoma, Norman, Oklahoma 73019
3E-mail: alinepaivamm@gmail.com
4Departamento de Sistemática e Ecologia, Centro de Ciências Exatas e da Natureza, Universidade Federal da Paraíba, João Pessoa 58051-900, Brazil
5E-mail: ricardo.betancur@ou.edu

Received June 10, 2021
Accepted March 28, 2022

A growing body of research suggests that genome size in animals can be affected by ecological factors. Half a century ago, Ebeling et al. proposed that genome size increases with depth in some teleost fish groups and discussed a number of biological mechanisms that may explain this pattern (e.g., passive accumulation, adaptive acclimation). Using phylogenetic comparative approaches, we revisit this hypothesis based on genome size and ecological data from up to 708 marine fish species in combination with a set of large-scale phylogenies, including a newly inferred tree. We also conduct modeling approaches of trait evolution and implement a variety of regression analyses to assess the relationship between genome size and depth. Our reanalysis of Ebeling et al.’s dataset shows a weak association between these variables, but the overall pattern in their data is driven by a single clade. Although new analyses based on our “all-species” dataset resulted in positive correlations, providing some evidence that genome size evolves as a function of depth, only one subclade consistently yielded statistically significant correlations. By contrast, negative correlations are rare and nonsignificant. All in all, we find modest evidence for an increase in genome size along the depth axis in marine fishes. We discuss some mechanistic explanations for the observed trends.

KEY WORDS: Actinopterygii, adaptive macroevolution, depth, genome size, marine fishes.

Genome size, or haploid C-value, has been the focus of several studies for over 50 years, with research showing that it can be affected by intrinsic (e.g., nucleus size, cell size and division rate, metabolic rate) and extrinsic (e.g., ecological constraints) factors (Grime and Mowforth 1982; Minelli et al. 1996; Griffith 2003; Roddy et al. 2020; Seymour and Gaut 2020). In eukaryotes, much of the variation in genome size stems from whole genome duplications (or deletions) and the proliferation of transposable elements in plant and animal genomes (Volf 2005; Phillips et al. 2016; Takuno et al. 2016; Johnson et al. 2018; Musilova et al. 2019). The evolution of larger genomes can occur due to passive accumulation resulting from genetic drift or to adaptive modulation via selection, with the origin of new genes being an important source of evolutionary novelty. Accordingly, larger genomes might be selected, and ultimately favored, because they can enhance adaptability. These include, for example, opsin genes duplication events that have enhanced vision in deep-water fish lineages (Musilova et al. 2019), and gene expansions in metabolic genes associated with the detoxication process of eucalyptus foliage consumption in koalas (Johnson et al. 2018).

A link between genome size and cellular traits is thought to explain adaptability mechanisms in plant species subjected to geographical, physiological, and functional constraints, including fluctuations in microclimate zones (Price 1988; Mowforth and Grime 1989; Minelli et al. 1996; Roddy et al. 2020). Several studies have also identified correlations between genome size and...
metabolic rate in plants (e.g., Roddy et al. 2020) and vertebrates (Hardie and Hebert 2004; Gardner et al. 2020), although a direct functional connection between these traits has not yet been established (Hardie and Hebert 2004; Gardner et al. 2020). In fishes, the influence of genome size on cytological parameters is known to affect ecological and life-history traits, providing support for an adaptive interpretation. For example, a positive association between egg diameter and genome size suggests a linkage to the evolution of parental care (Hardie and Hebert 2004). Other studies in fishes have assessed the relationship between genome size and biological or environmental characteristics such as habitat and physiology (Hinegardner and Rosen 1972; Vitturi et al. 1998; Hardie and Hebert 2003, 2004; Smith and Gregory 2009), finding, for example, that water temperature is negatively correlated with genome size (Hardie and Hebert 2003). In a paper published half a century ago, Ebeling et al. (1971; EEA71 hereafter) investigated the relationship between genome size and depth in teleost fishes. The authors obtained data from 13 species representing four order-level clades (Argentiniformes, Aulopiformes, Myctophiformes, and Lophiiformes) spanning three depth categories (shallow, mesopelagic, and bathypelagic). They identified an increase in genome size along depth categories for argentiniforms but not in the other clades examined. A more recent study by Smith and Gregory (2009) found that bathypelagic species that inhabit open waters at depths between 1000 and 4000 m have larger genomes than their shallower counterparts, although this pattern was nonsignificant. The trend toward larger genomes in cold-water species (Hardie and Hebert 2003) also appears to be consistent with EEA71’s findings (at least for argentiniforms), as water temperature decreases with depth. EEA71 offered a number of biological explanations for the relationship between genome size and depth, including acclimation to different bathymetric clines, and specialization to habitat-stable environments (Somero and Hochachka 1969). Other mechanistic processes underlying the possible increase of genome size over depth discussed by the authors include tandem duplications, saltatory replications, or even neutral accumulation over time (Sun and Mueller 2014; Liedtke et al. 2018). Regardless of the cellular mechanisms that resulted in an increase in genome size, the physical size of the genome has consequences for organismal fitness and may thus be subject of selection (Liedtke et al. 2018), which would explain ecological correlations such as acclimation, specialization, and broader niche breadth (Johnson et al. 2018; Musilova et al. 2019; dos Santos et al. 2021). Accordingly, deep zones are temporarily stable environments, allowing a natural tendency of DNA replication, and passive accumulation over time (Nei 1969). Although stable through time, these environments can also operate as strong ecological filters due to their hostile environmental conditions, including dim (or absence of) light, elevated hydrostatic pressure, low temperatures, and scarce food resources. Depth variation is thus an excellent factor to test for modes of evolution of genome size in a macroevolutionary framework.

Although EEA71 was important in highlighting the effect of depth on fish genome size, in light of current best practices in comparative biology the study had several limitations including small sample sizes, lack of statistical analyses, and omission of phylogenetic covariation. Even more recent studies looking at associations between genome size and ecological or life-history traits have also largely failed to account for phylogenetic non-independence (e.g., Vitturi et al. 1998; Gregory and Hebert 1999; Neafsey and Palumbi 2003; Smith and Gregory 2009; Roddy et al. 2020; but see Brainerd et al. 2001; Gardner et al. 2020). Many tools to access genome size data have been developed in recent years, contributing to the ease of data compilation for hundreds of species (Gregory et al. 2007). Additionally, the advent of large-scale phylogenies for many groups (Hinchliff et al. 2015; Betancur-R et al. 2017; Rabosky et al. 2018) coupled with the increasing availability of phylogenetic comparative approaches to account for evolutionary non-independence (Felsenstein 1985; Lynch 1991) provides a framework to test for correlations between biological (e.g., genome size) and ecological (e.g., depth) variables at large, macroevolutionary scales.

Here, we revisit the hypothesis that genome size increases with depth in marine fishes. We use a suite of phylogenetic comparative methods (PCMs) applied to data from hundreds of marine fish species for which we could obtain genome size, depth, and phylogenetic placement based on alternative trees. If genome size evolves mostly by genetic drift, a passive accumulation through time may occur in some clades but not in others regardless of depth. Conversely, genome size could be positively (or even negatively) correlated with depth, and this correlation can either have or lack an adaptive basis. To test these ideas, we first reanalyzed the EEA71 dataset using both phylogenetic and nonphylogenetic ANOVAs. We then obtained data on genome size, maximum depth, and phylogenetic placement based on alternative trees. If genome size evolves mostly by genetic drift, a passive accumulation through time may occur in some clades but not in others regardless of depth. Conversely, genome size could be positively (or even negatively) correlated with depth, and this correlation can either have or lack an adaptive basis. To test these ideas, we first reanalyzed the EEA71 dataset using both phylogenetic and nonphylogenetic ANOVAs. We then obtained data on genome size, maximum depth, and phylogenetic placement based on alternative trees. Using these expanded datasets, we conducted model-fitting approaches for trait evolution in conjunction with a series of regression analyses to examine the relationship between genome size and depth.

Materials and Methods

Genome Size and Depth Data

We first retrieved genome size data from NCBI (221 species) and the Genome Size (1308 species) databases (Gregory et al. 2007; Gregory 2008), which combined yielded information for 1447 unique ray-finned fish species, after averaging values...
for 82 species in common and discarding data from 59 fish individuals with uncertain species identifications. We then standardized genome size data, given in either picograms or megabases (Mb) in the corresponding databases, by multiplying the value in picograms times 978 (Dolezel et al. 2003; see also Fig. S1).

From this species list, we retrieved data on habitat and depth ranges, including maximum and minimum depths, from a total of 1444 fish species using the R package rfishbase (Boettiger et al. 2012; Froese and Pauly 2019). We then winnowed the resulting datasets to include marine species only, for a final genome/habitat database comprising 708 marine fish species in 60 families. We also gauged sensitivity of phylogenetic comparative analyses to ecological data error by using depth information obtained from other reliable sources (see Supporting Information).

**PHYLOGENETIC TREES AND TREE UNCERTAINTY**

Our downstream comparative analyses are based on two sets of phylogenetic trees. First, we augmented the taxonomic sampling of a previous phylogenetic analysis of bony fishes (Betancur-R et al. 2017) to include 648 of the 708 species in our database (B17+GB tree hereafter; see Supporting Information for details). To this end, we obtained mitochondrial cytochrome oxidase subunit I (COI) and cytochrome b (Cytb) sequences from NCBI for 648 of the 708 species with genome/habitat data, and aligned the sequences using MUSCLE version 3.8.425 (Edgar 2004) as implemented in Geneious Prime 2019.1.1 (https://www.geneious.com). To summarize our knowledge of relationships and divergence times for fish species in the database, we estimated a maximum likelihood (ML) tree using backbone constraint analyses. Of the 648 species with COI and/or Cytb data, 242 were previously placed in the backbone tree (Table S1). The goal was to obtain phylogenetic placement for the remaining previously unexamined 406 species. We conducted constraint ML searches in RAXML version 8.1.20 using by-codon partitions and 10 independent iterations (Stamatakis 2006; Stamatakis et al. 2008), and time-calibrated the resulting ML tree in TreePL version 1.0 (Smith and O’Meara 2012). The TreePL analysis used secondary calibrations extracted from the reference backbone tree via “congruification” (Eastman et al. 2013), a function (“congruity”) implemented in the R package geiger (Harmon et al. 2008). Using the R package phytools (Revell 2012), we then pruned the resulting tree, which included non-target species inherited from the backbone tree, to retain the 648 target species. See Table S1 for a list of accession numbers of sequences obtained from NCBI.

Second, we selected a set of trees from the R18Gen tree (Rabosky et al. 2018) that estimated a time-calibrated phylogeny for 11,638 fish species based on genetic data (R18Gen tree hereafter) and a set of 100 taxonomically imputed trees with simulated placement based on taxonomic constraints for 19,888 additional fish species (31,526 total species; R18Sim trees hereafter). Based on the list of 648 species included in the B17+GB tree, we searched for shared species in the R18Gen tree, finding 594 species in common that we earmarked for downstream analyses. To obtain phylogenetic placement for all 708 species in the genome size/habitat database, we used the collection of 100 R18Sim trees that include imputed phylogenetic placement for 114 species not included in the R18Gen tree.

We also assessed sensitivity of analyses to terminal nodes with shallow divergences (<0.1%), which may be indicative of miss-identifications or taxonomic over-splitting. When two sister species separated by short terminal lengths have different phenotypic values, a model is “forced” to fit extremely fast rates of evolution to accommodate this difference that emerged instantaneously relative to the time scale of the tree. We identified terminal nodes with <0.1% divergences in the B17+GB (n = 2), R18Gen (n = 2), and R18Sim (n = 3) trees. We thus repeated all comparative analyses after randomly pruning out one of the species with shallow divergences (in all cases, tip pairs had similar depths and genome sizes). Detailed information on this step is available in the Supporting Information.

**REANALYSIS OF THE EEA71 DATASET**

We reanalyzed the EEA71 dataset by performing both regular and phylogenetically informed ANOVAs using discrete depth categories with different depth cutoffs, as follows: (i) the cutoff defined by EEA71 for shallow (S; <200 m), mesopelagic (M; 600 m > M > 200 m), and bathypelagic (B; >600 m) species; (ii) a 200 m cutoff for shallow (<200 m) and deep-sea (>200 m) species; and (iii) a 600 m cutoff for shallow (<600 m) and deep-sea (>600 m) species. We assigned species categorized as mesopelagic to either the deep-sea category in the 200 m cutoff or to the shallow category in the 600 m cutoff. Note that table 1 of Ebeling et al. (1971) listed data for 15 specimens in 14 species and three fish orders, but one of their species (Argentina sialis) lacks genome size data, and their three fish orders (Salmoniformes, Myctophiformes, and Lophiiformes) correspond to four orders according to current fish classifications (Argentiniformes, Aulopiformes, Myctophiformes sensu stricto, and Lophiiformes). These analyses used the B17+GB and R18Sim trees. To test whether the pattern of genome size increasing with depth is driven by a single clade (Argentiniformes) in the EEA71 dataset (see Results), we conducted sampling with replacement analyses using data from four other clades (Zeioigadaria, Pelagiaria, Perciformes, and Carangaria) with similar depth profiles to the argentiniform species in that dataset, including six different species combinations from each clade (24 total replacements). We did not simply exclude argentiniforms from these analyses because that would have resulted in a substantial decrease of statistical power due to small sample size (n = 8). We also built 24
randomized datasets by reassigning depth categories for species in Argentiniformes while keeping the reported depth ranges for species in other clades. The main goal of both sets of analyses (clade replacement and randomized depth reassignment) was to assess whether the significant trend of genome size increasing with depth identified by EEA71 for argentiniforms had occurred by chance. Phylogenetic and nonphylogenetic ANOVAs were conducted in the R packages phytools (function phylANOVA) and stats (function aov), respectively.

**EXPANDED DATASET ANALYSES: MODEL-FITTING COMPARISONS OF GENOME SIZE EVOLUTION**

Model fitting and regression (see below) analyses used the datasets including 648 species in the B17+GB tree, 594 fish species in the R18Gen tree, and the complete dataset including all 708 species in the R18Sim trees. Using the functions mvBM and mvOU, implemented in the R package mvMORPH (Clavel et al. 2015), we first fitted two single-regime and two multi-regime models of genome size evolution: (i) Brownian motion (BM) depth-independent model (BMDI), (ii) Ornstein-Uhlenbeck (OU) depth-independent model (OUDI), (iii) BM depth-dependent model (BMDD), and (iv) OU depth-dependent model (OUDD). The multi-regime models fitted (BMDD and OUDD) used the three discrete depth cutoffs previously defined. We compared the fit to the four models using Akaike information criterion weights (AICw).

**EXPANDED DATASET ANALYSES: REGRESSIONS BETWEEN GENOME SIZE AND DEPTH**

We also performed a number of regressions between genome size and depth (modeled as a continuous trait), including OLS regressions that ignore phylogeny and a set of phylogenetic regressions. The phylogenetic regressions (PGLS) were modeled under BM (PGLS-BM) and OU (PGLS-OU) processes. Other sets of regressions conducted, including the adaptation regression of Hansen et al. (2008) that allows a lag in adaptation when modeling evolutions conducted, including the adaptation regression of Hansen et al. (2008) that allows a lag in adaptation when modeling evolutionary relationships between predictor (e.g., depth) and response (e.g., genome size) variables, often resulted in model inadequacy issues and are thus reported in the Supporting Information. The OLS and PGLS regressions were conducted in the R package phylnm (Tung Ho and Ané 2014), using the function phylnm. Because phylnm requires an input tree, OLS regressions were run using star phylogenies as input after transforming species covariances by a Pagel’s λ value of 0 using the lambdaTree function in phytools. We compared the fit to the different regression models using AICw scores and assessed statistical significance of the resulting correlations using P-values.

In addition to global analyses based on all species, we performed the same set of analyses on 11 subclades containing at least 20 representatives (species) from different depths. Although none of the orders examined by EEA71 met this condition, the 11 subclades identified provide phylogenetic replication to further test their hypothesis. For this type of analyses only, one subclade (Lutjaniformes) in the B17+GB tree was replaced with the corresponding clade grafted from two recent studies (Tavera et al. 2018; Rincon-Sandoval et al. 2020) that included more complete species samplings. To further test for differences in slopes and intercepts across subclades, we conducted phylogenetic ANCOVA (pANCOVA) analyses (modeled as BM) using the R package evomap (function gls.ancova; Smaers and Rohlf 2016).

All continuous data (depth in meters and genome size in Mb) were log-transformed prior to the analyses. Datasets used are available from the Dryad data repository https://datadryad.org/stash/share/dhecD71xFxVyBk03-OJ-9gYtGFN5QyGHHkw4xKjBYhP4; GenBank accession numbers are listed in Table 1.

**Results**

**REANALYSIS OF THE EEA71 DATASET**

Our reanalysis of the EEA71 dataset using both phylogenetic and nonphylogenetic ANOVAs show marginal differences in genome size among the three depth categories (Fig. 1a; no tree, P = 0.041; B17+GB tree, P = 0.045; R18Sim trees, P = 0.044; see also Table S2). Similar results are also obtained using different depth cutoffs (Fig. 1b; 200 m cutoff: no tree, P = 0.01; B17+GB tree, P = 0.03; R18Sim trees, P = 0.04; 600 m cutoff: B17+GB tree, P = 0.01; R18Sim trees, P = 0.04). These differences appear to be driven, however, mostly by argentiniform species in their dataset (Fig. 1a). Analyses using sampling with replacement data from four other clades and different datasets from each clade with similar depth properties resulted in nonsignificant differences in all 24 comparisons (P = 0.0793–0.99; Table S2). Moreover, analyses based on randomized depth reassignments for argentiniforms resulted in only eight (out of 144) significant phylogenetic ANOVAs from two randomized datasets (P = 0.02–0.04; Table S2). Taken together, these analyses corroborate a strong influence of argentiniforms in the results reported by Ebeling et al. (1971).

**EXPANDED DATASET ANALYSES: MODEL-FITTING COMPARISONS OF GENOME SIZE EVOLUTION**

The mvMORPH analyses based on the expanded dataset including genome size and depth data from the complete set of fish species produced the worst fit among the competing models for BMDI and BMDD (AICw = 0; Tables 1 and S3), rejecting the idea that genome size evolves following a random walk. In fact, these results offer some evidence that genome size evolves adaptively, largely producing split support among OUDI (AICw = 0.63–0.72) and OUDD (AICw = 0.27–0.33). Note that although the depth-dependent model is overall less
Figure 1. (a) Ebeling et al.’s (1971; EEA71) original dataset with average genome size (Mb) for either all species or order-level comparisons. Depth categories as originally defined include shallow (S; <200 m), mesopelagic (M; 600 m > M > 200 m), and bathypelagic (B; >600 m) depths; sample size (n) is indicated for each group/clade. (b) ANOVA (no tree) and phylogenetic ANOVAs, based on two sets of trees, for the EEA71 dataset using three different depth cutoffs: left, SMB (as originally defined); center, 200 m cutoff, where S < 200 m and D (deep-sea) > 200 m; right, 600 m cutoff, where S < 600 m and D > 600 m. The set of trees used are B17+GB (this study) and 100 R18Sim trees (Rabosky et al. 2018) including taxa with simulated phylogenetic placement. Significant P-values are provided for regular and phylogenetic ANOVAs. Percentage (%) values represent the number of trees from the 100 R18Sim trees that produced significant phylogenetic ANOVAs. P-values from R18Sim phylogenetic ANOVAs were averaged from trees with significant results. Note that statistically significant comparisons become nonsignificant when data from Argentiniformes are replaced with data from other clades with similar depth profiles, and randomized datasets (see text) produced only a few (8%) significant outcomes (Table S2), suggesting that statistical significance is mostly driven by argentiniform species in the EEA71 dataset.

Expanded Dataset Analyses: Regressions Between Genome Size and Depth

We conducted a series of nonphylogenetic (OLS) and phylogenetic (PGLS-BM and PGLS-OU) regression analyses using the new dataset encompassing 648 and 594 species with phylogenetic placement in the B17+GB (Fig. 2) and R18Gen trees, respectively, for either all species or 11 target subclades (Figs. 3 and 4; Tables S4 and S5). Although PGLS-OU regressions based on all species produced the best fit among competing models (AICw = 1.0), indicating positive correlations between genome size and depth (Figs. 3 and 4), only those based on the R18Gen tree were significant (P = 0.013 vs. P = 0.252 with the B17+GB tree). Among individual subclade analyses, the PGLS-OU model is favored in most cases (between four and eight clades, depending on the tree used; see details below), followed by PGLS-BM (two to four clades) and OLS (two to three clades; Table S6; Figs. 3 and 4). These analyses reveal some degree of positive relationships between genome size and depth in eight of the 11 subclades (Figs. 3 and 4; Table S4), one of which is statistically significant (Labriformes; P = 0.013−0.037) when using the corresponding best-fit models (PGLS-BM with the R18Gen tree [AICw = 0.82] and PGLS-OU with the B17+GB tree [AICw = 0.84]). Note that although two other clades also produced significant correlations (Carangaria, P = 0.002−0.002; Ovalentaria, P = 0.0002), these were obtained with PGLS-BM, which had a negligible fit in both cases (AICw = 0; Table S6; Figs. 3 and 4). All in all, most PGLS correlations performed using individual subclades had the best fit relative to OLS regressions, yielding in most cases positive (although largely nonsignificant) correlations. None of the few negative correlations obtained were significant (e.g., Pelagiaria and Spariformes).

Although phylogenetic regressions tend to be robust to tree choice (B17+GB, R18Gen, and R18Sim; Figs. 3, 4, and S3; Tables S4–S6), including terminal nodes with or without shallow divergences (Tables S4–S6), and also to depth information obtained from different sources (Tables S4–S6), a few individual subclade analyses produced mixed correlation results depending on whether phylogenetic history is accounted for or the type of tree used with phylogenetic regression analyses (see above). For example, although genome sizes for Ovalentaria are negatively correlated with depth using both OLS and R18Gen-based PGLS regressions, the B17+GB tree produced positive correlations for this clade (Figs. 3 and 4; Table S4). Tree choice had also an important impact in model-fitting comparisons, with PGLS-OU favored than its depth-independent counterpart, their resulting AICc scores are not significantly different from one another (delta AICc < 2; Tables 1 and S3). These results are robust to both phylogenetic (Tables 1 and S3) and also ecological data (Table S3) uncertainty.
Table 1. Model fitting comparisons based on two single-regime and two multi-regime models of genome size evolution: (i) Brownian motion (BM) depth-independent model (BMDI), (ii) Ornstein-Uhlenbeck (OU) depth-independent model (OUDI); (iii) BM depth-dependent model (BMDD), and (iv) OU depth-dependent model (OUDD). The multi-regime models fitted (BMDD and OUDD) used the three depth cutoffs previously defined (shallow: <200 m; mesopelagic: 200 m and <600 m; bathypelagic: >600 m). AICc = corrected Akaike Information Criterion; AICw = AIC weights. See Table S3 for additional results.

| Dataset            | Depth cutoff | AICc  | BMDI | BMDD | OUDI | OUDD | AICc  | BMDI | BMDD | OUDI | OUDD |
|--------------------|--------------|-------|------|------|------|------|-------|------|------|------|------|------|
| B17+GB tree        | 1-S, 2-M, 3-B| 379.32| 277.29| 266.5| 268.34| 0    | 0     | 0.72 | 0.27 |
| B17+GB tree        | 200 m        | 379.32| 307.64| 266.5| 268.48| 0    | 0     | 0.71 | 0.29 |
| B17+GB tree        | 600 m        | 379.32| 275.32| 266.5| 267.16| 0    | 0     | 0.67 | 0.33 |
| R18Gen tree        | 1-S, 2-M, 3-B| 1370.71| 1288.77| 636.85| 638.3 | 0    | 0     | 0.69 | 0.31 |
| R18Gen tree        | 200 m        | 1370.71| 1292.87| 636.85| 637.89| 0    | 0     | 0.63 | 0.37 |
| R18Gen tree        | 600 m        | 1370.71| 1261.98| 636.85| 638.84| 0    | 0     | 0.72 | 0.28 |

Figure 2. Time-calibrated B17+GB phylogeny containing the 648 species with genome size in Megabases (Mb; black bars) and maximum depth (m) data. Taxonomy annotations highlight the 11 subclades used for expanded dataset analyses (black labels, red circles at nodes), as well as the four clades examined by EEA71 (gray labels, gray circles at nodes) that failed to meet the minimum species criteria for further analysis (see Methods). For visualization purposes only, branch colors denote ancestral depth reconstructions based on the phytools function “contmap” (Revell 2012). A complete phylogeny with tip labels and taxonomic annotations is available in Figure S2.

being favored most often by the R17+GB tree (seven clades vs. five using the R18Gen tree), and PGLS-BM being favored most often by the R18Gen tree (four clades vs. two using the R17+GB tree; Figs. 3 and 4; Table S4).

The pANCOVA analyses largely failed to identify consistent and significant differences in slopes and/or intercepts for any of the clades examined (Table S7). Although for some clades the P-values obtained with this analysis were significant or marginal (e.g., Ovalentaria, Labriformes, Spariformes; see Table S7), in all cases different trees produced ambiguous results. Failure to identify any clear slope/intercept differences across clades may be a reflection of the high scatter of residuals around the trendlines.

Finally, analyses based on all 708 species in our genome size/habitat database using the set of 100 R18Sim trees also identified similar patterns of correlation for all species in 92% of the trees, and for Carangaria, Ovalentaria, and Labriformes in >50% of the trees based on PGLS-BM (Fig. S3; Tables S8 and S9). However, with PGLS-OU (the best-fit model in all these cases; mean AICw = 0.59–1.0) regressions, the percentage of significant correlations decreased to 59% for all species, and only Labriformes produced significant results in 30% of the trees. Additional details on these set of analyses are reported in the Supporting Information (Extended Results and Fig. S3; Tables S8 and S10). The pANCOVA analyses using one of the
Figure 3. Summary of (a) Pearson’s $r$ and (b) slope values for regressions of genome size and depth for all species with phylogenetic placement in the R18Gen and B17+GB trees. Each cluster of five bars represents an OLS regression analysis without accounting for phylogeny (OLS), and phylogenetic regression analyses (PGLS-BM and PGLS-OU) using both the B17+GB and R18Gen trees. Sample size ($n$) is indicated for each group/clade. Pearson’s $r$ and slope values above and below zero indicate positive and negative correlations, respectively. The level of significance is indicated as follows: $^*P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$. Akaike weights (AICw) based on model comparisons between OLS, PGLS-BM, and PGLS-OU (using the same tree per comparison) are denoted as follows: $\dagger$AICw $> 0.5$, $\ddagger$AICw $> 0.75$, and $\ddagger\ddagger$AICw $> 0.95$. For scatterplots, coefficient of determination ($r^2$), standard error (SE), significance ($p$), and phylogenetic signal values, see Figure 4 and Table S4. For analyses based on the 100 R18Sim trees, see Figure S3 and Tables S8–S10.
Figure 4. Scatterplots between genome size and depth for OLS regression analysis without accounting for phylogeny (OLS) and phylogenetic regression analyses (PGLS-BM and PGLS-OU) using the (a) B17+GB and (b) R18Gen trees (see also Fig. 3 for a summary). For analyses based on the 100 R18Sim trees, see Tables S8–S10.

100 R18Sim trees produced similar results to those based on B17+GB and R18Gen trees (Table S11).

Discussion

By conducting phylogenetic comparative analyses using data for genome size and habitat obtained from hundreds of fish species, we revisited and re-tested Ebeling et al.’s (1971) hypothesis regarding a positive relationship between genome size and depth. Our initial reanalysis of the EEA71 dataset using sampling with replacement shows that the pattern the authors identified was exclusively driven by one of the three clades examined (Argentiniformes). Phylogenetic ANOVAs resulted in positive and significant comparisons for the EEA71 data, but not for any of the resampled datasets that replaced argentiniforms with species from four other clades with similar depth properties. Furthermore, after assigning random depth categories for argentiniforms, only a small fraction of the phylogenetic ANOVAs was significant, although statistical power for this clade may be limited due to small sample size (n = 5).

Subsequent analyses based on data from 594–708 species using both inferred (B17+GB, R18Gen) and taxonomically imputed (R18Sim) trees identified modest evidence for an increase in genome size with depth for all species in the corresponding datasets (Figs. 3 and 4), although with high dispersion of residuals around the regression lines (Fig. 4). We also examined 11 representative subclades individually with results showing that the pattern identified with the complete datasets may be restricted to a limited number of groups, as only one clade (Labriformes) consistently produced significant phylogenetic correlations (Figs. 3 and 4). Although most analyses based on different sets of trees are similar, for some the best-fit models (PGLS-BM or
Building upon half a century of research on genome size, our analyses offer modest corroboration for emerging trends showing that genome size can be affected by ecological factors and can also evolve due to adaptive modulation (Ebeling et al. 1971; Brainerd et al. 2001; Smith and Gregory 2009). Different mechanisms may explain the influence of environmental constraints in genome size, including gene duplications (Kondrashov 2012; Johnson et al. 2018; Musilova et al. 2019; Roddy et al. 2020), ploidy shifts influencing cell and nucleus size (Hardie and Hebert 2003; but see Gardner et al. 2020), and egg diameter suggesting linkages with parental care (Hardie and Hebert 2004). Although depth is an important factor that is negatively correlated with metabolic rate in vertebrates (e.g., Ikeda 2016), studies have failed to identify a direct functional connection between genome size and basal metabolic rate (see Gardner et al. 2020).

Ebeling et al. (1971) suggested that larger genomes observed among deep-sea argentiniforms may have originated as a result of passive accumulations of DNA over time. A recent study using whole-genome sequence data and phylogenetic placement for 101 fish species (Musilova et al. 2019) demonstrated the independent expansion of the rhodopsin 1 gene repertoires through single-gene duplications associated with vision enhancement in multiple deep-sea fish lineages. Therefore, two contrasting hypotheses may explain a trend of increased genome size with depth. On the one hand, large genomes could simply be a byproduct of passive accumulation over time (Gregory and Hebert 1999). On the other hand, differences found in genome size along the depth gradient may be the result of adaptive evolution (Ohno 1972; Cavalier-Smith 1982; Innan and Kondrashov 2010; Kondrashov 2012).

We find that genome sizes among marine fish clades are much more often correlated positively than negatively with depth (Fig. 3). Although these positive correlations have a weak fit overall, they may indeed reflect potential interactions between evolutionary processes and ecological factors (Hardie and Hebert 2003, 2004; Smith and Gregory 2009; Nayfach and Pollard 2015; Musilova et al. 2019), facilitated by genome-level mechanisms that affect genome size (e.g., genome duplications, transposable elements; Brainerd et al. 2001). As demonstrated previously in other groups, larger genomes are associated with rapid adaptability (Price 1988; Mowforth and Grime 1989; Minelli et al. 1996; Nayfach and Pollard 2015; Roddy et al. 2020), which can be advantageous in hostile habitats such as the deep sea. For example, the expansion of the rod opsin gene repertoire in three lineages of deep-sea fishes may be linked to increased spectral sensitivity in dim-light conditions (Musilova et al. 2019). Another possibility that may explain weak positive correlations identified for some of the clades without invoking adaptability (e.g., Zeiogadaria, Perciformes, Gobiaria) is that deep habitats are environmentally more stable than shallower ones, thus allowing the accumulation of redundant or slightly deleterious genome elements that are not being purged by selection. Thus, although passive accumulations should be decoupled from any ecological factors, resulting in no obvious deviations from random correlation patterns, the actual evolutionary history of individual clades could be more nuanced, with both passive and adaptive trends interacting to shape genome size evolution.
It is also possible that other indirect life-history traits (not examined here) correlated with depth, such as water temperature (Hardie and Hebert 2003, 2004), have a more direct effect on genome size. By analyzing the relationship between genome size and niche breadth in pomacentrid fishes, dos Santos et al. (2021) also found that water depth has a low explanatory power. As mentioned above, species with larger eggs have larger genomes (Hardie and Hebert 2004), a finding that is in line with the nucleoskeletal theory that suggests a link between nuclear and cell volumes—that is, shifts in cellular volume modulate changes in nuclear volume (Gregory 2001). Deep-sea species are characterized by having larger eggs than their shallower counterparts, suggesting that deep-sea lineages are investing higher amounts of energy into the production of offspring (Fernandez-Arcaya et al. 2016). Thus, an increase in genome size could be associated with the modulation of cellular parameters that play a role in the selection of life-history traits, such as parental care.

In summary, we find that although Ebeling et al. (1971) were able to unveil a trend of genome size increasing along the depth axis in some marine teleosts, the reasons argued for such observations would be clade specific, as our reanalysis of their dataset failed to identify any positive relationships outside of argentiniforms. With more data and analytical tools available, we were able to analyze up to 708 species and gauge the relationship between genome size and depth in a phylogenetic comparative framework. Our analyses identified a similar trend showing modest evidence for genome size increasing with depth, a pattern that is nonetheless restricted to a few marine fish clades. Thus, although a positive relationship between genome size and depth may exist, this trend seems to be clade specific.

By providing a macroevolutionary perspective on genome size evolution across different orders of ray-finned fishes, we highlight important insights into the relationship between environmental factors (e.g., depth) and genome size. Although our results are based on the largest number of species with genome size data available to date, it is still likely that small sample sizes are limiting statistical power, in particular at the intra-clade level. It is also important to emphasize that even significantly positive correlations have weak coefficient of correlation values (e.g., Labriformes), implying that depth may have low explanatory power (see also dos Santos et al. 2021). Other ecological variables (e.g., cold-water environments; Hardie and Hebert 2003), habitat type (e.g., reef-associated or pelagic; Smith and Gregory 2009), or life-history traits (e.g., parental care and longevity; Griffith 2003; Hardie and Hebert 2004) may better explain the great disparity in genome size across the fish diversity (Fig. 2). As data on genome size become available for many more fish species, future phylogenetic comparative analyses may reveal a stronger link between genome size and depth (or other ecological factors).

**AUTHOR CONTRIBUTIONS**
APMM collected data, ran statistical analyses, made figures and tables, and drafted the manuscript. BAS supervised work, discussed ideas, and edited the manuscript. RBR conceived the study, proposed experimental design, collected and curated data, supervised work, and drafted the manuscript.

**ACKNOWLEDGMENTS**
We are grateful to Associate Editor J. Clavel and four anonymous reviewers for helpful comments that on the manuscript. This research was supported by National Science Foundation (NSF) grants DEB-1932759 and DEB-1929248 to RBR. APMM received scholarship support by the Programa Institucional de Internacionalização (PrInt) of CAPES (Brazilian Federal Agency for Support and Evaluation of Graduate Education, Ministry of Education of Brazil), the Fundação de Apoio à Pesquisa do Estado da Paraíba (Fapesq-PB), and the PADI Foundation (application #32777). BAS received funding by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; grant number 312178/2019-0).

**CONFLICT OF INTEREST**
The authors declare no conflict of interest.

**DATA ARCHIVING**
Data are available in Dryad Digital repository at https://doi.org/10.5061/dryad.gzht76hj6.

**LITERATURE CITED**
Betancur-R., Wiley, E.O., Arratia, G., Acero, A., Bailly, N., Miya, M., Lecointre, G. & Ortí, G. (2017) Phylogenetic classification of bony fishes. *Bmc Evolutionary Biology* 17, 162.
Boettiger, C., Lang, D.T. & Wainwright, P.C. (2012) fishbase: exploring, manipulating and visualizing FishBase data from R. *Journal of Fish Biology*, 81, 2030–2039.
Brainerd, E.L., Slutz, S.S., Hall, E.K. & Phillips, R.W. (2001) Patterns of genome size evolution in tetraodontiform fishes. *Evolution*, 55, 2363–2368.
Cavalier-Smith, T. (1982) Skeletal DNA and the evolution of genome size. *Annual Review of Biophysics and Bioengineering*, 11, 273–302.
Clavel, J., Escarguel, G. & Merceron, G. (2015) mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, 6, 1311–1319.
Dolezel, J., Bartos, J., Voglmayr, H. & Greilhuber, J. (2003) Nuclear DNA content and genome size of trout and human. *Cytometry*, 51A, 127–128.
Dos Santos, E.V., Martinez, P.A., Souza, G. & Jacobina, U.P. (2021) Genome size drives ecological breadth in Pomacentridae reef fishes. *J. Exp. Mar. Bio. Ecol.*, 540, 151544.
Eastman, J.M., Harmon, L.J. & Tank, D.C. (2013) Congruification: support for time scaling large phylogenetic trees. *Methods in Ecology and Evolution*, 4, 688–691.
Ebeling, A.W., Atkin, N.B. & Setzer, P.Y. (1971) Genome sizes of teleostean fishes: increases in some deep-sea species. *American Naturalist*, 105, 549–561.
Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics [Electronic Resource]*, 5, 1–19.
Felsenstein, J. (1985) Phylogenies and the comparative method. *American Naturalist*, 125, 1–15.
Somero, G.N. & Hochachka, P.W. (1969) Isoenzymes and short-term temperature compensation in poikilotherms: activation of lactate dehydrogenase isoenzymes by temperature decreases. Nature, 223, 194–195.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22, 2688–2690.

Stamatakis, A., Hoover, P. & Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML web servers. Systematic Biology, 57, 758–771.

Stone, G.N., Nee, S. & Felsenstein, J. (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Philos. Trans. R. Soc. B Biol. Sci., 366, 1410–1424.

Sun, C. & Mueller, R.L. (2014) Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. Genome Biol. Evol., 6, 1818–1829.

Takuno, S., Ran, J.H. & Gaut, B.S. (2016) Evolutionary patterns of genomic DNA methylation vary across land plants. Nat. Plants, 2, 1–7.

Tavera, J., Acero P, A. & Wainwright, P.C. (2018) Multilocus phylogeny, divergence times, and a major role for the benthic-to-pelagic axis in the diversification of grunts (Haemulidae). Molecular Phylogenetics and Evolution, 121, 212–223.

Tung Ho, L.S. & Ané, C. (2014) A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Systematic Biology, 63, 397–408.

Vitturi, R., Libertini, A., Campolmi, M., Calderazzos, F. & Mazzola, A. (1998) Conventional karyotype, nucleolar organizer regions and genome size in five Mediterranean species of Syngnathidae (Pisces, Syngnathiformes). Journal of Fish Biology, 52, 677–687.

Volff, J.N. (2005) Genome evolution and biodiversity in teleost fish. Heredity, 94, 280–294.

Supporting Information
Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Information

Associate Editor: S. Edie
Handling Editor: A. McAdam