



The Effects of Ecological Traits on the Rate of Molecular Evolution in Ray-Finned Fishes: A Multivariable Approach

Jacqueline A. May¹ · Zeny Feng² · Matthew G. Orton¹ · Sarah J. Adamowicz¹

Received: 31 January 2019 / Accepted: 22 September 2020 / Published online: 3 October 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Myriad environmental and biological traits have been investigated for their roles in influencing the rate of molecular evolution across various taxonomic groups. However, most studies have focused on a single trait, while controlling for additional factors in an informal way, generally by excluding taxa. This study utilized a dataset of cytochrome *c* oxidase subunit I (COI) barcode sequences from over 7000 ray-finned fish species to test the effects of 27 traits on molecular evolutionary rates. Environmental traits such as temperature were considered, as were traits associated with effective population size including body size and age at maturity. It was hypothesized that these traits would demonstrate significant correlations with substitution rate in a multivariable analysis due to their associations with mutation and fixation rates, respectively. A bioinformatics pipeline was developed to assemble and analyze sequence data retrieved from the Barcode of Life Data System (BOLD) and trait data obtained from FishBase. For use in phylogenetic regression analyses, a maximum likelihood tree was constructed from the COI sequence data using a multi-gene backbone constraint tree covering 71% of the species. A variable selection method that included both single- and multivariable analyses was used to identify traits that contribute to rate heterogeneity estimated from different codon positions. Our analyses revealed that molecular rates associated most significantly with latitude, body size, and habitat type. Overall, this study presents a novel and systematic approach for integrative data assembly and variable selection methodology in a phylogenetic framework.

Keywords Molecular clock · Molecular evolutionary rates · DNA barcoding · Mitochondrial DNA · Bioinformatics pipeline · Multivariable statistical analysis

Introduction

Heterogeneity in molecular evolutionary rates has been observed in a variety of taxonomic groups, challenging the assumptions of a strict and universal molecular clock (Li and

Tanumura 1987; Martin and Palumbi 1993; Nabholz et al. 2008). Identifying the sources of molecular rate heterogeneity is important for informing clock-based molecular dating techniques that are used to estimate phylogenetic relationships and the ages of important evolutionary events (Lanfear et al. 2010). Previous endeavors to identify correlates of molecular rates have implicated a plethora of ecological traits. These traits include the biological attributes of a species such as body size (Martin and Palumbi 1993), metabolic rate (Gillooly et al. 2005), and generation time (Ohta 1993; Thomas et al. 2010) as well as environmental factors such as latitude (Wright et al. 2011) and UV exposure (Smith et al. 1992; Hebert et al. 2002). These traits are theorized to impact the rate of mutation over time. Alternatively, factors that influence the effective population size of a species, such as habitat (Woolfit 2009) and dispersal ability (Fujisawa et al. 2015), are suggested to affect the fixation rate and thus the substitution rate. However, these factors are often found to be highly correlated amongst themselves in vertebrate

Handling Editor: Joana Projecto-Garcia.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00239-020-09967-9>) contains supplementary material, which is available to authorized users.

✉ Jacqueline A. May
mayj@uoguelph.ca

¹ Department of Integrative Biology and Biodiversity Institute of Ontario, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada

² Department of Mathematics and Statistics, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada

groups. For instance, smaller organisms often have larger effective population sizes, accelerated metabolic rates, and shorter generation times (Thomas et al. 2010). These correlations render it difficult to separate out genuine correlates from spurious associations, particularly if only a single trait is considered at a time in the analysis. Despite this, only a limited number of molecular rates studies have considered the joint effects of multiple traits (e.g. Nabholz et al. 2008; Bromham et al. 2015; Fujisawa et al. 2015), largely due to the challenge of assembling a species-level trait dataset that does not contain a disproportionate amount of missing data (Goolsby et al. 2017). The amount of available trait data has increased substantially in recent years (Dececchi et al. 2016) and provides the opportunity for an investigation of multiple potential correlates in a concurrent fashion. This type of approach would allow one to account for the effects of confounding variables and facilitate identification of important rate correlates.

Actinopterygii or ray-finned fishes comprise the most diverse vertebrate group and display a considerable amount of trait variation across multiple levels of taxonomy. Due to their ectothermic nature, the metabolic rates of fishes are positively correlated with environmental temperature (Clarke and Johnston 1999); as a result, fishes are expected to exhibit higher molecular rates in warmer environments (Rohde 1992; Wright et al. 2011). Indeed, a previous study by Wright et al. (2011) implicated latitude as a correlate of molecular rate in marine teleost fishes. Other studies found evidence for biological correlates of rates in certain fish groups, including lifestyle (i.e. high-performance vs. more sedentary species) (Strohm et al. 2015), longevity (Hua et al. 2015), and body size (Qui et al. 2014). Considering the number of different traits that have been associated with rate heterogeneity in fishes, a broader-scale comparative analysis is warranted that takes advantage of the large amounts of curated sequence and trait data available online for fishes. Such a study could also test whether trends observed in specific fish groups are prevalent at a more general phylogenetic scale.

An approach must be used to account for the shared evolutionary history among related species as sequence and trait data are considered to be statistically non-independent. A common approach is the sister pair method, which pairs together sister taxa that contrast in a certain trait and quantifies the relative distances of each sister taxon to an outgroup species (Lanfear et al. 2010). This approach is often used in studies that examine the effects of a single trait (e.g. latitude; Gillman et al. 2009; Wright et al. 2011) and is convenient as it does not require a full phylogenetic tree. As taxa that are not successfully paired are discarded using this method, however, the sample size is limited (Lanfear et al. 2010). Phylogenetic generalized least squares (PGLS) is an alternative method for

accounting for phylogeny. PGLS models incorporate the evolutionary history of a group of species into the estimation of the parameters of a regression model (Grafen 1989; Freckleton et al. 2002). PGLS has been increasingly used in the realm of molecular rate studies as its design allows for the analysis of both larger sample sizes and a larger number of explanatory variables or traits (Lanfear et al. 2010).

The aim of the current study is to investigate a broad array of biological traits and environmental factors and quantify their relative contribution towards molecular rate variation in fishes using a multivariable approach. To achieve this task, we constructed a bioinformatics pipeline that assembles both sequence and trait data from the Barcode of Life Data System (BOLD) (Ratnasingham and Hebert 2007) and FishBase (Froese and Pauly 2019) and performs multiple PGLS analyses using a multivariable approach. Previously investigated traits, such as body size, latitude, and longevity as well as traits whose influence on molecular evolution in fishes are unknown or uncertain, including diet and morphologic features, were included in the analysis. Rates were estimated using both the whole alignment (i.e. all three codon positions) and an alignment including only the third codon positions. Observations of rate variability across different codon positions may have different implications regarding the underlying evolutionary mechanisms. Rate variability among lineages at the first and second codon positions is most influenced by factors that affect the rate of fixation of new mutations (e.g. strength of positive selection or purifying selection) as changes at these positions may alter the encoded amino acid. Changes at the third codon position, however, are evolutionarily silent (i.e. do not alter the encoded amino acid). Molecular evolutionary rates estimated using the third codon position can therefore provide a measure of the relative rate of mutation among lineages. In general, it was expected that traits that influence either the mutation rate or the fixation rate can be revealed by their correlations with estimated substitution rates along lineages over time (see Supplementary Table S1 for hypotheses/expectations for each trait). More specifically, factors that influence the fixation rate, such as life history traits related to effective population size, were expected to correlate significantly with substitution rates estimated using the whole alignment (i.e. first and second codon positions will be affected). Alternatively, factors that influence the mutation rate, such as those related to DNA replication or oxygen radical production, should correlate with both measures of substitution rate (i.e. third codon positions will also be affected). This work illustrates the utility of using a systematic multivariable approach to pinpoint traits that play a role in generating molecular rate variation. In addition, our bioinformatics integrates data assembly and model-selection methods in a manner that may be easily adapted by other

researchers to investigate new taxonomic groups, ecological traits, or genomic regions.

Materials and Methods

A bioinformatics pipeline was constructed in the R programming language to assemble DNA barcode and trait data from different online data sources and perform multiple PGLS analyses using a multivariable approach. Figure 1 presents an overview that briefly outlines our pipeline procedures.

Barcode Sequence Data

BOLD provides researchers access to standardized cytochrome *c* oxidase subunit I (COI) sequence data for hundreds of thousands of species. As of May 4th, 2020, over 8.2 million COI barcode sequences are publicly available on BOLD, and over 221,000 animal species are represented. Due to its comprehensive collection of barcode records and accessibility, BOLD was utilized as the source for sequence data in the current study. BOLD uses a Barcode Index Number (BIN) system to cluster COI barcode sequences into groups of sequences, or molecular operational taxonomic units (MOTU, Floyd et al. 2002), that strongly resemble species (Ratnasingham and Hebert 2013). BINs were used as a proxy for species in the initial stages of the current study, as doing so enabled us to include geographic information for specimens currently lacking taxonomic identifications

on BOLD, but which could be identified on the basis of their sequences.

180,091 Actinopterygii barcode sequence records were downloaded directly into R version 3.3.2 (R Core Team 2019) using BOLD's API tool. These sequences comprised data for 16,862 BINs. Several filters were applied to the BOLD records for quality control purposes. Records with sequences lacking a BIN URI, a form of BIN identification, were omitted. In addition, only sequences from the 5' terminus region of the COI gene (COI-5P) were retained. To facilitate the multiple sequence alignment that is performed in a later section of the pipeline, sequences with N content and/or gaps found on their terminal ends were identified and trimmed. Filters were also used to remove sequences that contained over 1% internal N content or gaps over their total sequence length and to retain only those sequences whose lengths were between 640 and 1000 bp. A filter was also used to remove BINs that had no species-level information (e.g. BINs that only contained sequences sampled from unidentified or unknown species). This was necessary as the trait data are attributed to the scientific name of the species, and the species name was therefore required for matching purposes. Each BIN was subsequently assigned a species label based on those sequence records with species-level identifications within the BIN (see Supplementary Information File for details on this process). After this, species were subsequently used as the taxonomic units for further analysis.

Trait Data

Latitude data were obtained for each barcode record, grouped by species, and the median value for each species calculated. Prior to taking the median, the absolute values of latitude were used such that the values would represent biologically relevant information (e.g. equatorial vs. polar). To ensure that the BOLD latitudinal data were not biased and that they correlate well with another known information about species distributions, the median absolute latitude values were validated against data downloaded from the Global Biodiversity Information Facility (GBIF 2019) (<https://www.gbif.org>). Occurrence data for Actinopterygii were downloaded from GBIF using the R package “rgbif” version 0.9.9 (Chamberlain 2017) and used to calculate the median of the absolute latitude values for each species that had latitudinal data available in both databases. A regression analysis determined that the latitudinal data obtained from BOLD largely agreed with the GBIF data ($R^2=0.76$, p value $<2 \times 10^{-16}$) (Supplementary Fig. S1).

FishBase is a publicly accessible online database that holds ecological, geographical, life history, and morphometric trait information for more than 30,000 fish species from around the world (Froese and Pauly 2019). BOLD species

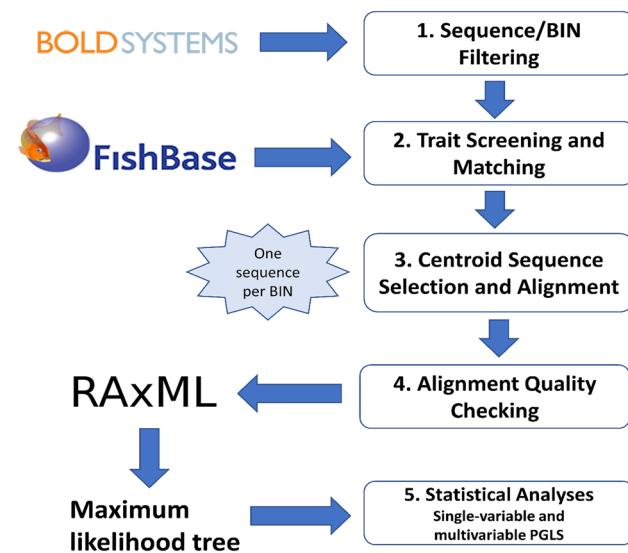


Fig. 1 Visualization of the five major components of the R programming pipeline. The components work in a sequential manner and may be adapted by users to analyze additional sequence data (i.e. from organisms other than ray-finned fishes). COI maximum likelihood tree estimated using multi-gene constraint. *BIN* barcode index number, *PGLS* phylogenetic generalized least squares

names from the previously filtered BINs were obtained and matched against trait information from the Species, Stocks, Ecology, Morphology, Reproduction, and Maturity data tables on FishBase (Boettiger et al. 2012; Froese and Pauly 2019). Species-level data for 27 traits were selected for inclusion in downstream analyses based on sample size ($n > 300$), data variability, and use of a controlled vocabulary to facilitate data parsing (see Supplementary File 1 for trait matrix and Supplementary Information File for details on this process).

Sequence Alignment and Quality Control

The next component of the pipeline focuses on the selection of the “centroid” sequence of each species (Orton et al. 2019). Selecting one representative sequence per species significantly reduces the computation time required for estimating a phylogenetic tree. To achieve this task, the sequence data were grouped by species, and a multiple sequence alignment was performed for each species using the R package “muscle” version 3.28.0 (Edgar 2004; Kalinka 2019). A gap penalty of – 3000 was specified (Orton et al. 2019), and default settings were used for all other parameters. Pairwise distance matrices were then constructed from each species alignment, and the TN93 model of DNA evolution was specified (Tamura and Nei 1993) (for specifics on distance matrix parameter settings, see Supplementary Information File). The centroid sequence for each species was defined as the sequence with the lowest sum of pairwise distances in the distance matrix.

A “reference” sequence of the Actinopterygii class was used to standardize the lengths of the centroid sequences. The reference sequence in this case was a published yellow perch *Perca flavescens* 658 bp COI-5P sequence (BOLD:AAA4391: BNAF367-09.COI-5P; April et al. 2011, 2013). This sequence was chosen as an amino acid alignment with the other Actinopterygii sequences used in this study indicated a typical sequence among fishes. The reference sequence was symmetrically trimmed down to a standard sequence length of 620 bp, as many sequences in an alignment may fall short of the standard barcode length. Base-calling errors are also more common at the terminal ends of sequences (Athey 2013). The R package “muscle” was used to perform a second multiple sequence alignment that included the centroid sequences in addition to the reference sequence. A gap penalty of – 3000 was again specified, and default settings were used for all other parameters. The alignment was trimmed according to the start and end positions of the 620 bp reference sequence. The alignment was inspected to ensure it was in the correct reading frame by translating the alignment in MEGA6 (Tamura et al. 2013) and checking for the presence of stop codons. Additional quality control tests were performed on the alignment to

check for potentially deviant and/or contaminated sequences that made it through the previous filters (see Supplementary Information File for details). DAMBE7 (Xia 2018) was also used to assess whether saturation occurred at the third codon position.

Phylogenetic Tree Construction and Rate Estimation

The centroid sequences were used to construct a phylogeny for the downstream PGLS analyses. The function *modelTest* from the R package “phangorn” version 2.2.0 (Schliep 2011) was used to identify the best model of nucleotide substitution for the centroid sequences. The GTR + Γ + I model (Lanave et al. 1984) was subsequently selected as it produced the lowest BIC values of all the models considered. A maximum likelihood (ML) tree search was performed in RAxML version 8.0 (Stamatakis 2014) on the centroid sequence alignment to construct a COI gene tree. A published multi-gene tree of Actinopterygii (Rabosky et al. 2018) based on 27 genetic markers was used as a backbone constraint to preserve the monophyly of the ingroup and previously established phylogenetic relationships in Actinopterygii. 5073 out of 7094 species or 71% of the COI ML gene tree was constrained using the backbone, which accounted for 67 out of the 69 orders (97%), 372 out of the 398 families (93%), and 2062 out of the 2526 genera (82%) in the COI tree. Consequently, the backbone constraint was useful for resolving deeper evolutionary relationships within the COI ML gene tree. Erroneous phylogenetic relationships that arose from the use of COI would therefore be confined towards the tips of the trees and were less likely to introduce significant error into the estimation of PGLS parameters. To ensure results were robust to choice of phylogeny, we re-ran single-variable PGLS analyses for the three traits with the largest sample sizes using a different multi-gene tree published by Betancur-R et al. (2017) as a constraint for our gene tree.

The *chronos* function in the “ape” R package version 5.1 (Paradis et al. 2004) was used to estimate absolute molecular evolutionary rates using the penalized likelihood method (Sanderson, 2002; Paradis 2013). The tree was calibrated using known fish fossil dates obtained from Rabosky et al. (2018) (see Supplementary Information File for details). Cross-validation was used to select the smoothing parameter, and a correlated clock model was specified. Rates estimated using all three codon positions are a product of the processes that impact fixation as well as mutation rates. By considering rates that were estimated using only third codon positions, we gain access to a more specific measure of the relative mutation rate. Thus, each species had an absolute terminal rate measurement for both whole codon (RATE_{WHOLE}) and third codon (RATE_{THIRD}) positions. Rate estimates for terminal branches were used because saturation is expected for comparisons of distantly related species, particularly at third

codon positions of COI. Terminal rates were also deemed appropriate for the current analysis as recommended by Lanfear et al. (2010) for analyses of contemporary trait data. The rate data were further scrutinized for potential outliers by identifying observations that exceeded the upper (quartile $3 + 1.5 \times \text{IQR}$) or lower thresholds (quartile $1 - 1.5 \times \text{IQR}$) of the data. If a rate estimate for a species was flagged as an outlier, the original COI sequence for that species was checked for possible sequence contamination using BLASTn (Basic Local Alignment Search Tool; Altschul et al. 1990). An expected value threshold of $1e-6$ was used and the search was optimized for highly similar sequences. If the sequence matched most closely (percent identity greater than 95%) with other sequences from the same or congeneric species, it was retained for analysis. These steps ensured that the estimated rate data were representative of the assigned fish species and the potential for human error reduced.

Phylogenetic Signal Estimation

To measure the amount of phylogenetic signal for each continuous variable, Pagel's lambda (λ) (Pagel 1999) values were approximated using the *phylosig* function from the R package "phytools" version 0.6.2 (Revell 2012). λ is estimated through maximum likelihood and provides a value that best transforms the phylogenetic variance–covariance matrix of a trait to fit the observed data (Pagel 1999). λ ranges from 0 (the trait evolved independently from phylogeny) to 1 (the data adhere to a Brownian motion model of evolution, and the value of the trait is dependent on phylogeny) (Pagel 1999). A likelihood ratio test was also performed via the *phylosig* function (the argument "test" was set to TRUE) to evaluate if the λ value indicated a significant amount of phylogenetic signal when compared to a null model ($\lambda=0$) (Revell 2012). To measure the amount of phylogenetic signal for each categorical variable, Fritz and Purvis (2010)'s *D* metrics were calculated using the *phylo.d* function from the R package "caper" version 0.5.2 (Orme et al. 2013). *D* tests whether the number of transitions of a binary trait differs from the number that would be expected if the evolution of the trait followed a Brownian motion model (Fritz and Purvis 2010). A *D* value of 0 suggests that the trait data support a Brownian motion model of evolution, and a value of 1 indicates no relationship between the trait and species relatedness (random). A *D* value greater than 1 indicates phylogenetic overdispersion (trait values are not similar in related species), whereas a negative *D* value indicates the trait is more conserved than would be expected under a Brownian motion model (Fritz and Purvis 2010). A hypothesis test was also performed via the *phylo.d* function to evaluate if the *D* value varied significantly from the *D* values of a simulated null model (*D*=1) (Orme et al. 2013).

Two-step Multivariable Model Selection Process

Traits varied greatly in terms of sample size, ranging from $n=436$ for age at maturity to $n=6032$ for body shape. Inclusion of all 27 traits was therefore not feasible in a multivariable analysis, as no species had data available for every trait. To mitigate this problem, a screening process was invoked that involved running a PGLS regression analysis for each trait. In these analyses, each trait was considered as the sole explanatory variable in the PGLS regression model, and RATE_{WHOLE} and RATE_{THIRD} were fitted as response variables in two separate models. The function *ppls* from the "caper" R package (Orme et al. 2013) was used for each PGLS analysis. The results of a PGLS analysis include estimated slopes for the regression equations (*B*) of the explanatory variables and the associated standard errors (SE_B). The λ parameter was estimated using maximum likelihood for each trait model. A multiple-comparison correction of the p values was made using the Benjamini and Hochberg (1995) adjustment.

For inclusion in the multivariable model selection process, traits that had a *p* value below 0.15 in their respective single-variable analysis were retained (i.e. "candidate" traits). A *p* value of 0.15 was chosen as this value was not too stringent as to exclude potentially important variables in the final multivariable model (cut-off values of 0.10 and 0.05 were also considered, see Supplementary Information File). A data subset was found for species that had data available for all candidate traits. However, when all candidate traits were included in the subset, the sample size (number of species) was greatly reduced due to the sparseness of the trait data available among all species. To overcome this problem, candidate traits with inadequate data overlap ($n < 50$) with other traits were removed from the analysis at this stage. In addition, categorical traits that no longer exhibited variation in this data subset were also removed. A check for multicollinearity was performed, as highly correlated explanatory variables can lead to erroneous estimations of regression coefficients. To control for multicollinearity, only candidate traits with variance inflation factors (VIFs) lower than 5 were retained for analysis. A "global" multivariable model was built that included all of the candidate traits in the final subset. A backwards selection process was then invoked that involved (1) removing the trait with the largest *p* value from the model and (2) examining the Bayesian information criterion (BIC) of the model. The set of traits that produced significant terms and the lowest BIC value was selected as the "best-fit model". This process was performed separately for RATE_{WHOLE} and RATE_{THIRD}. Although AIC is generally more conservative than BIC, AIC does not account for situations wherein a larger sample size allows for selection of a larger model than necessary; given the range of sample sizes used in this study, BIC was deemed a more appropriate

choice for variable selection. A *p* value of 0.05 was used to test for significance in the final multivariable model. Supplementary multivariable analyses were also performed for distantly related fish orders separately (Scorpaeniformes and Cypriniformes) to assess whether effects were consistent across evolutionarily independent groups.

Results

Data Filtering

Data for 16,862 BINs (180,091 sequences) were initially downloaded from BOLD into R (Fig. 2). A filter removed 352 BINs that contained sequences with internal N and/or gap content over 1% of their total sequence length. Sequences with lengths longer than 1000 bp and shorter than 640 bp were then filtered out, resulting in the removal of 2619 BINs. Following this, 3424 BINs were removed that did not contain any records with a species-level identification. Filters used to resolve taxonomic conflicts within BINs resulted in the removal of 2910 BINs. 309 BINs were

subsequently removed as they did not match to any trait data from FishBase and did not contain latitudinal information in their BOLD records. Finally, checks for extremely “gappy” centroid sequences and close neighbor taxonomic conflicts resulted in the removal of 154 BINs (refer to Supplementary Information File for definition of “gappy”). The final alignment contained centroid sequences for 7094 species upon application of these quality control and taxonomic conflict filters (see Supplementary File 2 for multiple sequence alignment).

Gene Tree Phylogeny and Molecular Rate Variation

Fish species from 69 orders, 398 families, and 2526 genera were represented by the COI ML gene tree (Supplementary File 3). The taxonomic composition of the dataset corresponded well with known species counts in Actinopterygii. For example, among all species in the dataset, the orders that were represented by the largest numbers of species were: Perciformes ($n=898$), Cypriniformes ($n=882$), Scorpaeniformes ($n=489$), Gobiiformes ($n=478$), and Siluriformes ($n=416$).

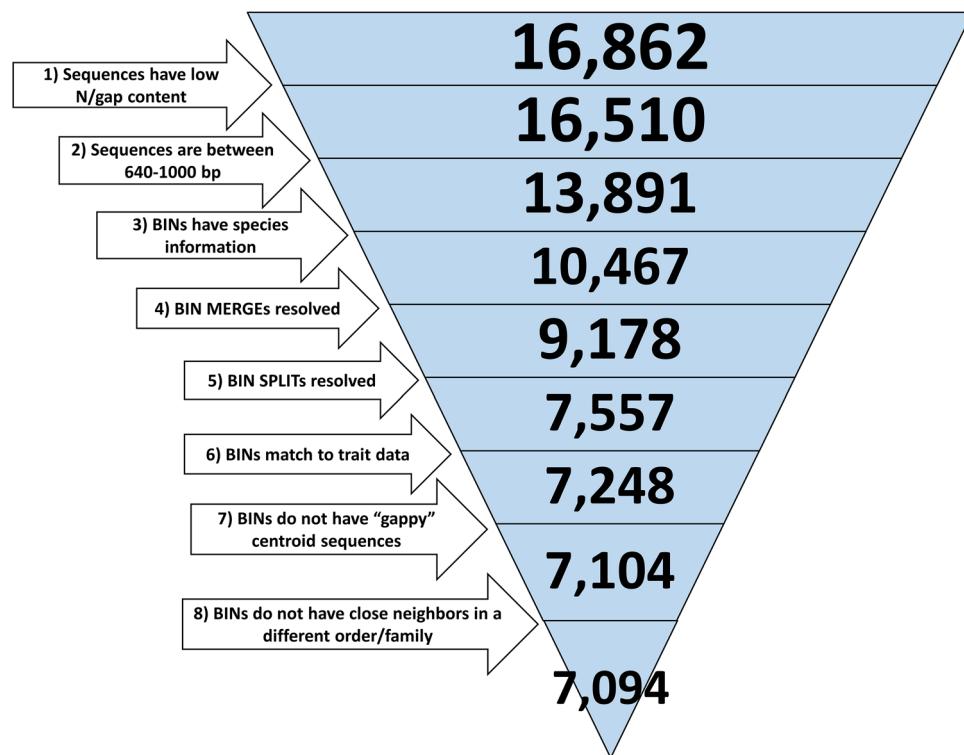


Fig. 2 The number of BINs removed at each filtering step. The filters included (1) removing sequences with a high internal N or gap content (greater than 1%), (2) removing sequences longer than 1000 bp and shorter than 640 bp, (3) ensuring that each BIN has at least one sequence with a species-level taxonomic identification, (4) resolving within-BIN taxonomy conflicts (MERGEs), (5) checking that there was only one BIN per species (i.e. SPLITs; when more than one

BIN shared the same species name, the BIN with the highest sample size was retained), (6) ensuring that each BIN had data available for at least one trait, (7) checking that the centroid sequence for the BIN did not have a disproportionate number of gaps, and (8) checking that the BIN did not have a close neighbor (<5% divergence) that was in a different order or family. Note: no sequences were removed due to exceeding the upper threshold of the IQR of the divergence values

The median values were 0.015 substitutions/site/Myr for RATE_{WHOLE} and 0.19 substitutions/site/Myr for RATE_{THIRD}, respectively. 95% of rate values fell within a range of 0.0018–0.10 for RATE_{WHOLE} and 0.011–1.48 for RATE_{THIRD}. These ranges are indicative of variation in molecular evolutionary rates among COI sequences in fishes. Both rate measures were ln transformed for all analyses and residuals checked for normality and homogeneity to ensure adherence to PGLS assumptions. Each biological and environmental variable exhibits some degree of phylogenetic signal, and all of the metrics were significantly different from either $\lambda=0$ for continuous traits or $D=1$ for discrete traits (i.e. phylogenetic signal was detected in our traits) (for visualization of rate and trait data for selected families see Figs. 3 and 4; see Supplementary Tables S2 and S3 for phylogenetic signal values).

PGLS Analyses

There were 20 candidate traits for RATE_{WHOLE} and 17 candidate traits for RATE_{THIRD} that were selected based on the results of the single-variable analyses (p value < 0.15)

(Fig. 5; see Supplementary Tables S2 and S3 for results of all single-variable PGLS analyses). It should be noted that the traits age at maturity, longevity, minimum water temperature, and diet trophic level were not included in the multivariable model selection process due to inadequate data overlap ($n < 50$). PGLS analyses that were performed to test for robustness to choice of phylogeny indicated consistent results (see Supplementary Table S4 and Supplementary File 4 for alternative gene tree).

Following the backward selection process, the best-fit model for RATE_{WHOLE} included median latitude, maximum body length, minimum water depth, schooling, coral reefs, and body shape as explanatory variables (Table 1a). A data subset that contained complete information for all variables was used for this analysis, which included observations for 844 species. There was a significant and negative association between median latitude and RATE_{WHOLE} ($B_{LATITUDE} = -0.0083$, p value = 0.0042). Maximum body length also had a significant negative effect on RATE_{WHOLE} ($B_{LENGTH} = -0.0021$; p value = 0.0060), as did minimum water depth ($B_{DEPTH} = -0.0030$; p value = 0.0096). Species that exhibit schooling behavior had significantly lower

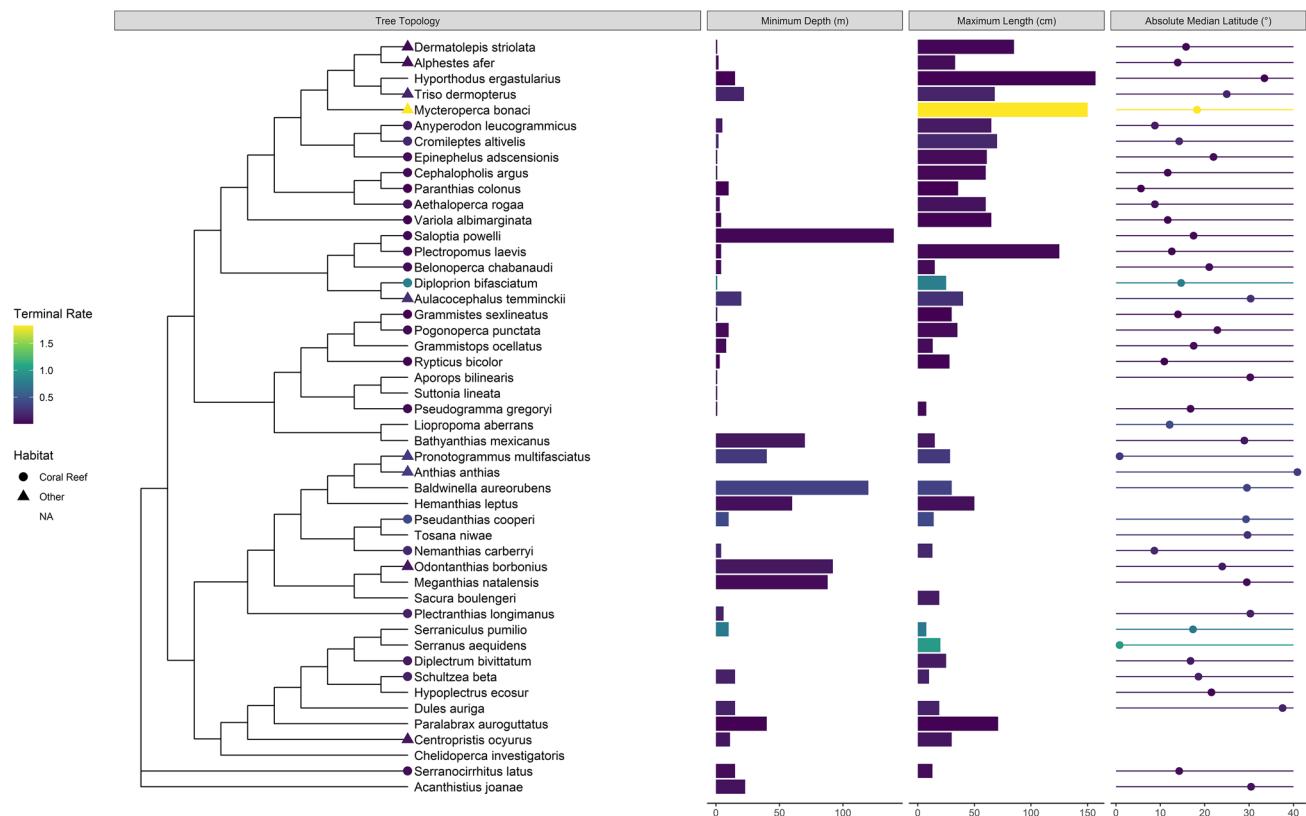


Fig. 3 Visualization of the distribution of molecular rate (RATE_{WHOLE}) and trait data (coral reef, minimum water depth, maximum body length, median latitude) across the topology of the Actinopterygii family Serranidae. The units for latitude represent the median of the absolute latitude values for each species or distance

away from the equator (°). Each genus in the family is represented by a randomly selected species. This family was chosen for plot readability and availability of trait data. RATE_{WHOLE}: terminal rates estimated using all three codon positions

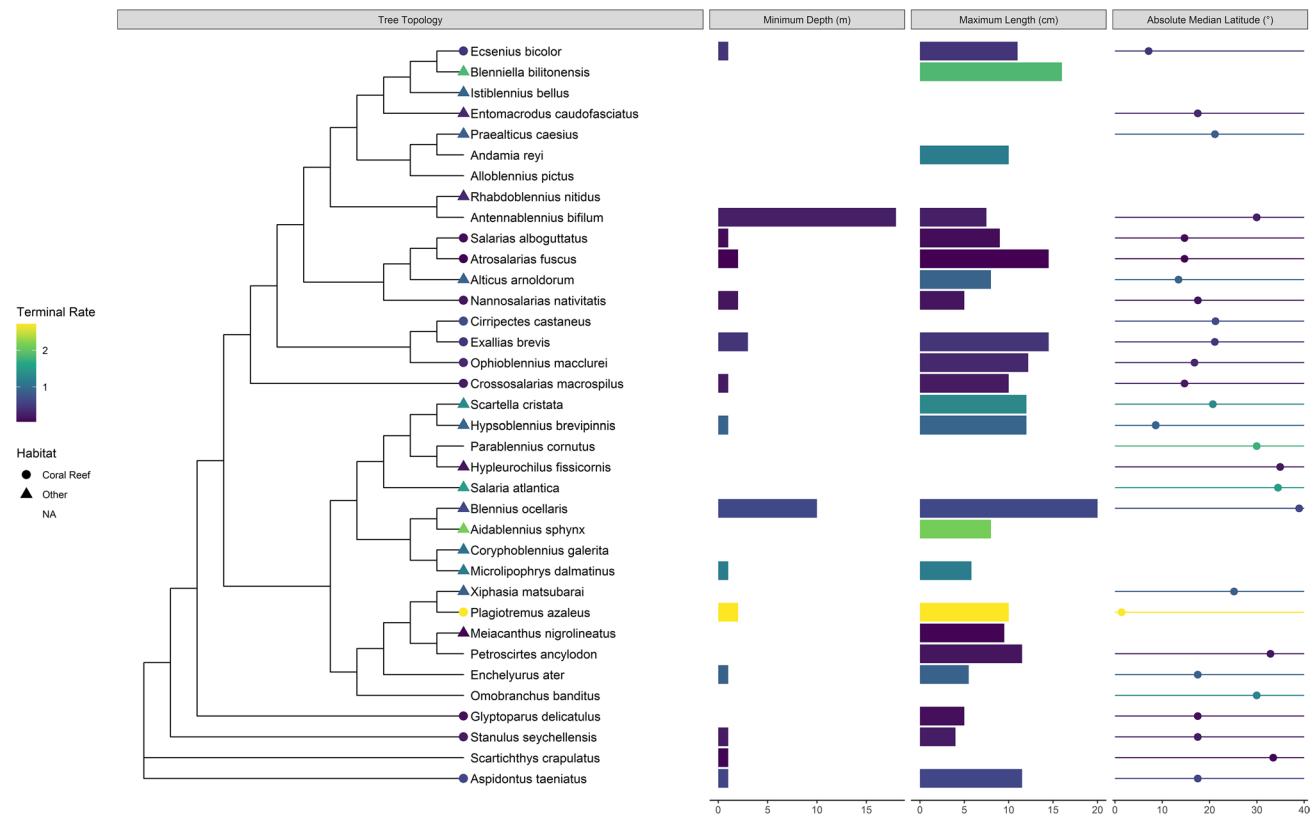


Fig. 4 Visualization of the distribution of molecular rate ($\text{RATE}_{\text{WHOLE}}$) and trait data (coral reef, minimum water depth, maximum body length, median latitude) across the topology of the Actinopterygii family Blenniidae. The units for latitude represent the median of the absolute latitude values for each species or distance

$\text{RATE}_{\text{WHOLE}}$ values ($B_{\text{SCHOOL}} = -0.26$, p value = 0.015) on average in comparison to non-schooling species. Species that inhabit coral reefs had significantly higher $\text{RATE}_{\text{WHOLE}}$ values ($B_{\text{REEFS}} = 0.15$, p value = 0.017) on average when compared to species that do not reside in such habitats. Finally, when considering body shape, species with short and/or deep body shapes had significantly lower $\text{RATE}_{\text{WHOLE}}$ values ($B_{\text{SHAPE}} = -0.23$, p value = 0.014) on average in comparison to species with fusiform/normal body shapes. The R^2 value for the $\text{RATE}_{\text{WHOLE}}$ model was 0.052, suggesting that 5.2% of the observed variation in $\text{RATE}_{\text{WHOLE}}$ may be explained by the traits included in the best-fit model.

The best-fit model for $\text{RATE}_{\text{THIRD}}$ included median latitude, neritic, coral reefs, body shape, and feeding type as explanatory variables (Table 1b). A data subset was used for this analysis that included observations for 653 species. Median latitude had a significant and negative effect on $\text{RATE}_{\text{THIRD}}$ ($B_{\text{LATITUDE}} = -0.0095$; p value = 0.00062). Species found in neritic environments had significantly higher $\text{RATE}_{\text{THIRD}}$ values ($B_{\text{NERITIC}} = 0.13$, p value = 0.0055) on average in comparison to species inhabiting other types of environments. Species that inhabit coral reefs also had significantly higher

$\text{RATE}_{\text{WHOLE}}$ values ($B_{\text{REEFS}} = 0.36$, p value = 2.24×10^{-9}) on average when compared to species that do not inhabit coral reef environments. Species with elongated body shapes had significantly lower $\text{RATE}_{\text{THIRD}}$ values ($B_{\text{SHAPE}} = -0.33$, p value = 0.00011) on average in comparison to species with fusiform/normal body shapes. Finally, when compared to species that hunt macrofauna, species with substrate browsing ($B_{\text{BROWSE}} = 0.48$, p value = 0.0069), aquatic plant grazing ($B_{\text{GRAZE}} = 0.47$, p value = 0.0015), selective plankton feeding ($B_{\text{PLANKTON}} = 0.39$, p value = 0.0032), and variable ($B_{\text{VARIABLE}} = 0.31$, p value = 0.0091) feeding types had on average significantly higher $\text{RATE}_{\text{THIRD}}$ values. The R^2 value for the $\text{RATE}_{\text{THIRD}}$ model was 0.16, suggesting that 16% of the observed variation in $\text{RATE}_{\text{THIRD}}$ may be explained by the traits included in the best-fit model.

Discussion

Identifying the sources of variation in molecular evolutionary rates is a complicated undertaking as trait data are often correlated, and comprehensive species-level datasets are

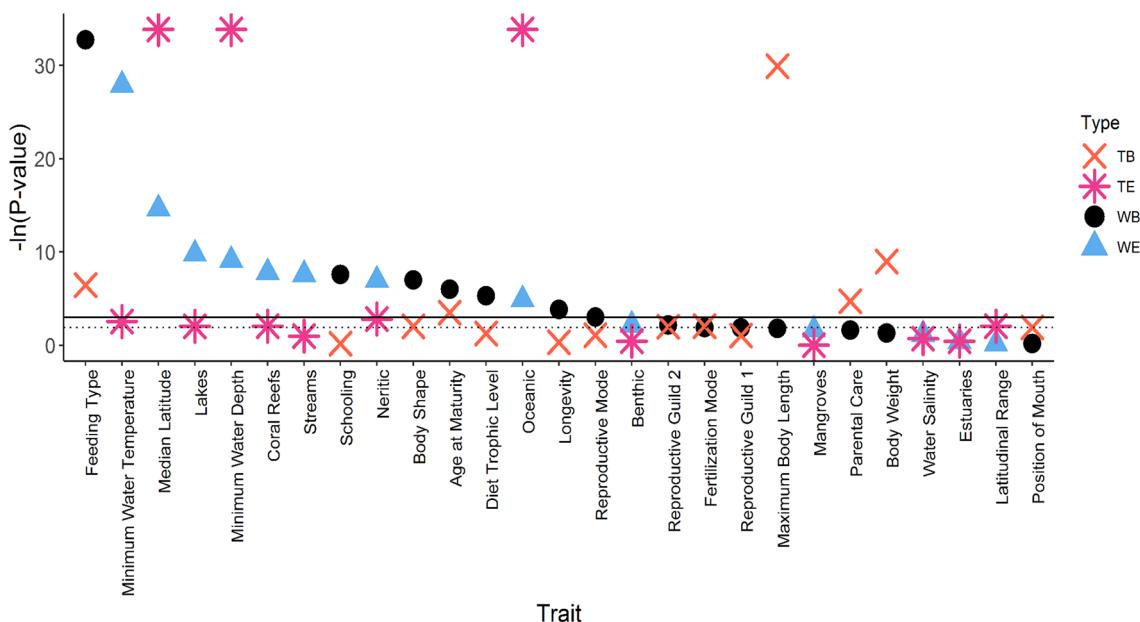


Fig. 5 A comparison of the negative natural logarithm (\ln) of Benjamini and Hochberg (1995) adjusted p values from the single-variable analyses. The solid line indicates significance at the 0.05 level after adjustment. Traits are ordered by p -values from the RATE_{WHOLE} analyses. The dotted line indicates the cut-off for inclusion in the multivariable model selection process (p value < 0.15). *TB* Bio-

logical trait with rates estimated using only the third codon position (RATE_{THIRD}) as the response variable, *TE* Environmental trait with RATE_{THIRD} as the response variable, *WB* Biological trait with RATE_{WHOLE} as the response variable, *WE* Environmental trait with RATE_{WHOLE} as the response variable

scarce. However, comparative phylogenetic analyses and contemporary bioinformatics tools offer much promise for performing this task. The results presented here offer new insights into the most important correlates of COI molecular rates in a group of organisms. In addition, the bioinformatics pipeline developed in this study combines data assembly tools together with model-selection techniques in a manner that may be easily tailored to meet the needs of other researchers.

Molecular Rate Correlates

In contrast to the constant and internally regulated body temperatures of endotherms, the body temperatures of ectotherms are directly related to environmental temperature. Latitude and environmental temperature are, therefore, theorized to correlate with the rate of molecular evolution in ectothermic organisms, stemming from an increase in the metabolic rates of those species that inhabit warmer environments (Rohde 1992; Martin and Palumbi 1993). The results of our multivariable PGLS analyses support this, as median absolute latitude, or distance away from the equator, exhibited a significant and negative effect on rates. The negative effect found on RATE_{THIRD} (i.e. mutation rates) is particularly supportive of this hypothesis, as accelerated metabolic rates are thought to result in an increase in the number of DNA-damaging oxygen radicals and mutagenic events

(Martin and Palumbi 1993; Gillooly et al. 2005). This is corroborated by previous studies on ectothermic groups that found environmental factors such as latitude to be significant correlates of substitution rates in ectothermic groups (e.g. plants, amphibians, turtles; Davies et al. 2004; Wright et al. 2010; Lourenço et al. 2013). More recently, using a sample size of 435 Actinopterygii species pairs, Orton et al. (2019) found that species residing closer to the equator have higher evolutionary rates compared to their northern counterparts; the effect was moderate, however, as only 56% of the pairs displayed a longer branch length in the more tropical species relative to its polar sister species. These findings are supported by the results of the current study, whereby median latitude had a significantly negative effect on RATE_{WHOLE} and RATE_{THIRD}, but also low explanatory power in the single-variable analyses (R^2 values = 0.0052 and 0.033, respectively). This pattern may be mirrored at the order-level, as a latitudinal effect was observed in Scorpaeniformes but not in Cypriniformes (Supplementary Table S5).

The water depth at which species reside is expected to correlate with environmental temperature, and therefore, molecular rates in these organisms. This association is supported by the results of the current study, as both minimum water depth and living in a neritic (near shore) environment had significant effects on RATE_{WHOLE} and RATE_{THIRD}, respectively. It is worth noting that, in addition to environmental temperature, the depth of a fish's habitat is also

Table 1 Best-fit multivariable PGLS models testing for associations between biological traits or environmental variables and rates of molecular evolution for (a) RATE_{WHOLE} (terminal molecular evolutionary rates calculated for all nucleotides in the COI DNA barcode sequences) and (b) RATE_{THIRD} (terminal rates of molecular evolution calculated using third codon positions in COI only)

Trait	Categories ^a	Regression coefficient (<i>B</i>)	Standard error (SE _B)	<i>p</i> value
(a) RATE _{WHOLE} (<i>n</i> =844, <i>R</i> ² =0.052, Pagel's λ =1.00)				
Median latitude	-	-0.0083	0.0029	0.0042
Maximum body length	-	-0.0021	0.00077	0.0060
Minimum water depth	-	-0.0030	0.0012	0.0096
Schooling	Does not school	-	-	-
	Schools	-0.26	0.11	0.015
Coral reefs	Does not inhabit coral reefs	-	-	-
	Inhabits coral reefs	0.15	0.063	0.017
Body shape	Fusiform/normal	-	-	-
	Eel-like	0.57	0.31	0.065
	Elongated	0.19	0.12	0.12
	Short and/or deep	-0.23	0.093	0.014
(b) RATE _{THIRD} (<i>n</i> =653, <i>R</i> ² =0.16, Pagel's λ =1.00)				
Median latitude	-	-0.0095	0.0028	0.00062
Neritic	Not neritic	-	-	-
	Neritic	0.13	0.048	0.0055
Coral reefs	Does not inhabit coral reefs	-	-	-
	Inhabits coral reefs	0.36	0.059	2.24×10 ⁻⁹
Body shape	Fusiform/normal	-	-	-
	Eel-like	0.072	0.66	0.91
	Elongated	-0.33	0.084	0.00011
	Short and/or deep	-0.082	0.097	0.40
Feeding type	Hunting macrofauna	-	-	-
	Browsing on substrate	0.48	0.18	0.0069
	Grazing on aquatic plants	0.47	0.15	0.0015
	Selective plankton feeding	0.39	0.13	0.0032
	Variable	0.31	0.12	0.0091

Bold rows indicate significance at the 0.05 level. RATE_{WHOLE} and RATE_{THIRD} values were log transformed prior to analysis

^aCategories with hyphens (-) under the *B*, SE_B, and *p* value columns indicate the reference category for that trait analysis

inversely related to the amount of UV radiation to which it is exposed (Zagarese and Williamson 2001), and many fish eggs are transparent and demonstrate sensitivity to UV damage (Häder et al. 2007). Exposure to elevated levels of UV radiation has been associated with a rise in mutagenic events in both vertebrates (Ziegler et al. 1993) and aquatic invertebrates (Hebert et al. 2002; Smith et al. 1992). Moreover, both latitude and depth were found to associate negatively with substitution rates in a previous study completed on marine fish mtDNA (Wright et al. 2011). Thus, it is possible that species residing in shallower waters are exposed to higher levels of UV radiation and, as a consequence, exhibit higher rates of mutation compared to their deeper-dwelling counterparts. A future multivariable study that incorporates direct measurements of UV radiation levels as a predictor variable may shed further light on its relative importance in the context of fish molecular evolution rates. Species that live in coral reef environments also exhibited elevated rates

for both RATE_{WHOLE} and RATE_{THIRD}. Coral reefs are found in shallow waters and are home to a diverse array of fish species. Population subdivision, for example, due to lower dispersal rates in reef species (e.g. Jones et al. 2005), may play a role in this observed association through increased population structure and fixation of nonsynonymous substitutions (Woolfit 2009; Bromham 2011). It is also possible that coral reef species exhibit accelerated mutation rates due to warmer environmental temperatures, or some other underlying factor associated with reef environments. Regardless, this finding warrants further investigation into the correlates of molecular evolutionary rates in coral reef fish species; this may be particularly intriguing in the context of a potential link between molecular evolutionary rates and tropical biodiversity (Cowman 2014).

Maximum body length and body shape both had significant effects on rates. This finding is in agreement with previous studies that found a link between body size and substitution

rates (Martin and Palumbi 1993; Welch et al. 2008; Nabholz et al. 2016). Body size has previously been implicated as a correlate of life span and generation time in ectothermic vertebrates (Blueweiss et al. 1978; Blanco and Sherman 2005; Speakman 2005; Hua et al. 2015). While significant in our single-variable analyses, age at maturity and longevity could not be included in our multivariable model selection process due to a lack of data overlap with other traits (i.e. missing data). Differences in population sizes have also been shown to correlate with body size estimates (Dunham and Vinyard 1997; Lewis et al. 2008), and can contribute to the fixation rate of non-synonymous mutations (i.e. smaller populations may have a higher fixation rate of slightly deleterious mutations due to a stronger role of genetic drift) (Woolfit 2009). It is possible that the observed negative effect of maximum body length on molecular rates is a manifestation of its underlying correlations with these other life history traits. Conversely, in addition to temperature, metabolic rate also scales with body size in fishes (Clarke and Johnston 1999). As metabolic rate was not directly tested in the current study, it is difficult to rule it out as a driving factor behind the observed body size effect. However, previous studies that considered ectotherms and that accounted for the covarying effect of body size did not reveal any significant correlation between basal metabolic rate and molecular rate (e.g. Lanfear et al. 2007; Qui et al. 2014), while molecular rates may be correlated with active metabolic rate (Santos 2012).

Traits that have not been previously implicated as correlates of molecular rates in fishes were shown to exhibit effects in the current study, including schooling behavior and feeding type. Fish species that exhibit schooling behavior had lower RATE_{WHOLE} values on average compared to non-schooling species. This signal indicates an association between social group structure and substitution rate in fishes. As this effect was only found on RATE_{WHOLE}, this may relate to differences in population dynamics and fixation rate in schooling and non-schooling species. Intriguingly, all feeding types (browsing on substrate, grazing on aquatic plants, selective plankton feeding, variable) exhibited elevated RATE_{THIRD} observations when compared to the hunting macrofauna feeding type. As this association was only observed when RATE_{THIRD} was specified as the response variable, this finding may be indicative of an effect on mutation rate. As our analyses are limited to the barcoding region of COI and studies on diet and molecular rates are scarce, further investigations are warranted that consider this potential link in fishes and ectotherms at large.

Limitations

A caveat of utilizing online trait databases is that there is often a considerable level of disparity in data availability for different traits. For instance, traits that may be more easily

quantified, such as length measurements, offer a much larger subset of data compared to traits for which data collection is often invasive or difficult, such as those related to age and reproduction (Jarić and Gačić 2012). This disparity in data collection and recording often results in incomplete or patchy datasets. This is evident in the current study, as the sample size of species dropped substantially ($n < 19$) when age at maturity was added as a term in the multivariable analyses. As researchers become more accustomed to sharing their raw data in a public format, opportunities for future studies to utilize more complete datasets will inevitably arise. Another limitation of our dataset may stem from saturation at the third codon position (Supplementary Fig. S2). It is possible that rate heterogeneity across the phylogeny was underestimated when considering RATE_{THIRD}. However, as we used terminal rate estimates for our research, it is unlikely that this had a major impact on our results. Finally, comparative methods such as PGLS have been criticized for their susceptibility to pseudoreplication in the context of singular evolutionary events (Maddison and FitzJohn 2015; Uyeda et al. 2018). Our analyses showed largely consistent results for distantly related fish orders (Supplementary Table S5). However, it is possible that some of the trait states included in this study arose from a single evolutionary event, and the significant associations we detected between traits and rates are inflated as a result. Probabilistic graphical models that more adequately account for this type of phylogenetic structure are currently under development (Uyeda et al. 2018) and researchers should consider implementing these models in future studies.

Conclusion

Our analyses revealed that molecular rates associate most significantly with latitude, body size, and habitat type. Results also indicate that behavioral traits such as schooling and feeding type play a role in generating molecular rate variation in this ectothermic group. It is evident from these results that numerous traits contribute to COI rate heterogeneity in a synergistic fashion by impacting both the fixation and mutation rates. Ultimately, by investigating correlates of molecular rates, we acquire a more thorough understanding of the evolutionary processes that are tied to molecular change. Phylogenetic tree dating methods may then be improved by correcting for factors that can affect the generality of a molecular clock across lineages. Knowledge of rate correlates is particularly useful for mtDNA regions that are often used as markers of biodiversity and for species-level phylogenetic inference (i.e. DNA barcoding projects (Hebert et al. 2003)) such as using COI. This study highlights not only the usefulness of a multi-parameter, comparative approach for identifying molecular rate correlates but

also the utility of bioinformatics tools for processing and analyzing complex biodiversity data.

Acknowledgements We thank Dr. Robert Hanner for his advice regarding the processing and analyzing of fish barcode sequence data. We would like to thank the many contributors of sequence and trait data to BOLD and FishBase, respectively. Finally, we would like to thank three anonymous reviewers for their invaluable insight and suggestions for improving this manuscript.

Funding This work was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) (2016–06199 to S.J.A and 400095 to Z.F) and by a grant in Bioinformatics and Computational Biology (15401) from the Government of Canada through Genome Canada and Ontario Genomics (to S.J.A., Z.F., et al.). Additionally, this study represents a contribution to the “Food from Thought” research program led by the University of Guelph and supported through the Canada First Research Excellence Fund.

Code Availability The source code used for this research is available at <https://github.com/jmay29/phylo>.

Data Availability All sequence and trait data used are available as supplementary material for this article.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- April J, Mayden RL, Hanner RH, Bernatchez L (2011) Genetic calibration of species diversity among North America’s freshwater fishes. *Proc Natl Acad Sci USA* 108:10602–10607. <https://doi.org/10.1073/pnas.1016437108>
- April J, Hanner RH, Mayden RL, Bernatchez L (2013) Metabolic rate and climatic fluctuations shape continental wide pattern of genetic divergence and biodiversity in fishes. *PLoS ONE* 8:e70296. <https://doi.org/10.1371/journal.pone.0070296>
- Athey TBT (2013) Assessing errors in DNA barcode sequence records. Master’s thesis, University of Guelph
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 57:289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Betancur-R R, Wiley EO, Arratia G, Acero A, Bailly N, Miya M, Lecointre G, Ortí G (2017) Phylogenetic classification of bony fishes. *BMC Evol Biol* 17:162. <https://doi.org/10.1186/s12862-017-0958-3>
- Blanco MA, Sherman PW (2005) Maximum longevities of chemically protected and non-protected fishes, reptiles, and amphibians support evolutionary hypotheses of aging. *Mech Ageing Dev* 126:794–803. <https://doi.org/10.1016/j.mad.2005.02.006>
- Blueweiss L, Fox H, Kudzma V, Nakashima D, Peters R, Sams S (1978) Relationships between body size and some life history parameters. *Oecologia* 37:257–272
- Boettiger C, Lang DT, Wainwright P (2012) rfishbase: Exploring, manipulating and visualizing FishBase data from R. *J Fish Biol* 81:2030–2039. <https://doi.org/10.1111/j.1095-8649.2012.03464.x>
- Bromham L (2011) The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philos Trans R Soc Lond B* 366:2503–2513. <https://doi.org/10.1098/rstb.2011.0014>
- Bromham L, Hua X, Lanfear R, Cowman PF (2015) Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *Am Nat* 185:507–524. <https://doi.org/10.1086/680052>
- Chamberlain S (2017) rgbif: interface to the global ‘biodiversity’ information facility API. R package version 0.9.9
- Clarke A, Johnston NM (1999) Scaling of metabolic rate with body mass and temperature in teleost fish. *J Anim Ecol* 68:893–905. <https://doi.org/10.1046/j.1365-2656.1999.00337.x>
- Cowman PF (2014) Historical factors that have shaped the evolution of tropical reef fishes: a review of phylogenies, biogeography, and remaining questions. *Front Genet* 5:394. <https://doi.org/10.3389/fgene.2014.00394>
- Davies TJ, Savolainen V, Chase MW, Moat J, Barraclough TG (2004) Environmental energy and evolutionary rates in flowering plants. *Proc R Soc Lond Ser B* 271:2195–2200. <https://doi.org/10.1098/rspb.2004.2849>
- Dececchi TA, Mabee PM, Blackburn DC (2016) Data sources for trait databases: comparing the phenomic content of monographs and evolutionary matrices. *PLoS ONE* 11:e0155680. <https://doi.org/10.1371/journal.pone.0155680>
- Dunham JB, Vinyard GL (1997) Relationships between body mass, population density, and the self-thinning rule in stream-living salmonids. *Can J Fish Aquat Sci* 54:1025–1030. <https://doi.org/10.1139/f97-012>
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 19:1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Floyd R, Abebe E, Papert A, Blaxter M (2002) Molecular barcodes for soil nematode identification. *Mol Ecol* 11:839–850. <https://doi.org/10.1046/j.1365-294X.2002.01485.x>
- Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *Am Nat* 160:712–726. <https://doi.org/10.1086/343873>
- Fritz SA, Purvis A (2010) Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic strength in binary traits. *Conserv Biol* 24:1042–1051. <https://doi.org/10.1111/j.1523-1739.2010.01455.x>
- Froese R, Pauly D (eds) (2019) FishBase. World Wide Web electronic publication. www.fishbase.org
- Fujisawa T, Vogler AP, Barraclough TG (2015) Ecology has contrasting effects on genetic variation within species versus rates of molecular evolution across species in water beetles. *Proc R Soc B Biol Sci* 282:1–9. <https://doi.org/10.1098/rspb.2014.2476>
- GBIF: The Global Biodiversity Information Facility (2019) What is GBIF? <https://www.gbif.org/what-is-gbif>
- Gillman LN, Keeling DJ, Ross HA, Wright SD (2009) Latitude, elevation and the tempo of molecular evolution in mammals. *Proc R Soc B* 276:3353–3359. <https://doi.org/10.1098/rspb.2009.0674>
- Gillooly JF, Allen AP, West GB, Brown JH (2005) The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc Natl Acad Sci USA* 102:140–145. <https://doi.org/10.1073/pnas.0407735101>
- Goolsby EW, Bruggeman J, Ané C (2017) Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods Ecol Evol* 8:22–27. <https://doi.org/10.1111/2041-210X.12612>
- Grafen A (1989) The phylogenetic regression. *Philos Trans R Soc Lond B* 326:119–154. <https://doi.org/10.1098/rstb.1989.0106>

- Häder DP, Kumar HD, Smith RC, Worrest RC (2007) Effects of solar radiation on aquatic ecosystems and interactions with climate change. *Photochem Photobiol Sci* 6:267–285. <https://doi.org/10.1039/C0PP90036B>
- Hebert PDN, Remigio EA, Colbourne JK, Taylor DJ, Wilson CC (2002) Accelerated molecular evolution in halophilic crustaceans. *Evolution* 56:909–926. <https://doi.org/10.1111/j.0014-3820.2002.tb01404.x>
- Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proc R Soc B* 270:313–321. <https://doi.org/10.1098/rspb.2002.2218>
- Hua X, Cowman P, Warren D, Bromham L (2015) Longevity is linked to mitochondrial mutation rates in rockfish: a test using Poisson regression. *Mol Biol Evol* 32:2633–2645. <https://doi.org/10.1093/molbev/msv137>
- Jarić I, Gačić Z (2012) Relationship between the longevity and the age at maturity in long-lived fish: Rikhter/Efanov's and Hoenig's methods. *Fish Res* 129–130:61–63. <https://doi.org/10.1016/j.fishres.2012.06.010>
- Jones GP, Planes S, Thorrold SR (2005) Coral reef fish larvae settle close to home. *Curr Biol* 15:131401318. <https://doi.org/10.1016/j.cub.2005.06.061>
- Kalinka AT (2019) Multiple sequence alignment with MUSCLE. R package version 3.28.0
- Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. *J Mol Evol* 20:86–93. <https://doi.org/10.1007/BF02101990>
- Lanfear R, Thomas JA, Welch JJ, Brey T, Bromham L (2007) Metabolic rate does not calibrate the molecular clock. *Proc Natl Acad Sci USA* 104:15388–15393. <https://doi.org/10.1073/pnas.0703359104>
- Lanfear R, Welch JJ, Bromham L (2010) Watching the clock: Studying variation in rates of molecular evolution between species. *Trends Ecol Evol* 25:495–503. <https://doi.org/10.1016/Zj.tree.2010.06.007>
- Lewis HM, Law R, McKane AJ (2008) Abundance-body size relationships: the roles of metabolism and population dynamics. *J Anim Ecol* 77:1056–1062. <https://doi.org/10.1111/j.1365-2656.2008.01405.x>
- Li WH, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. *Nature* 326:93–96. <https://doi.org/10.1038/326093a0>
- Lourenço JM, Glémén S, Chiari Y, Galtier N (2013) The determinants of the molecular substitution process in turtles. *J Evol Biol* 26:38–50. <https://doi.org/10.1111/jeb.12031>
- Maddison WP, FitzJohn RG (2015) The unresolved challenge to phylogenetic correlation tests for categorical characters. *Syst Biol* 64:127–136. <https://doi.org/10.1093/sysbio/syu070>
- Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci USA* 90:4087–4091. <https://doi.org/10.1073/pnas.90.9.4087>
- Nabholz B, Glemin S, Galtier N (2008) Strong variations of mitochondrial mutation rate across mammals—the longevity hypothesis. *Mol Biol Evol* 25:120–130. <https://doi.org/10.1093/molbev/msm248>
- Nabholz B, Lanfear R, Fuchs J (2016) Body mass-corrected molecular rate for bird mitochondrial DNA. *Mol Ecol* 25:4438–4449. <https://doi.org/10.1111/mec.13780>
- Ohta T (1993) An examination of the generation-time effect on molecular evolution. *Proc Natl Acad Sci USA* 90:10676–10680. <https://doi.org/10.1073/pnas.90.22.10676>
- Orme CDL, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W (2013) caper: comparative analyses of phylogenetic and evolution in R. R package version 0.5.2
- Orton MG, May JA, Ly W, Lee DJ, Adamowicz SJ (2019) Is molecular evolution faster in the tropics? *Heredity* 122:513–524. <https://doi.org/10.1038/s41437-018-0141-7>
- Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature* 401:877–884. <https://doi.org/10.1038/44766>
- Paradis E (2013) Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion. *Mol Phylogenet Evol* 67:436–444. <https://doi.org/10.1016/j.ympev.2013.02.008>
- Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290. <https://doi.org/10.1093/bioinformatics/btg412>
- Qui F, Kitchen A, Burleigh JG, Miyamoto MM (2014) Scombrid fishes provide novel insights into the trait/rate associations of molecular evolution. *J Mol Evol* 78:338–348. <https://doi.org/10.1007/s00239-014-9621-4>
- R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Rabosky DL, Chang J, Title PO, Cowman PF, Sallan L, Friedman M, Kaschner K et al (2018) An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* 559:392–395. <https://doi.org/10.1038/s41586-018-0273-1>
- Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system (www.barcodinglife.org). *Mol Ecol Notes* 7:355–364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>
- Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) System. *PLoS ONE* 8:e66213. <https://doi.org/10.1371/journal.pone.0066213>
- Revell LJ (2012) Phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223. <https://doi.org/10.1111/j.2041-210X.2011.00169.x>
- Rohde K (1992) Latitudinal gradients in species-diversity—the search for the primary cause. *Oikos* 65:514–527. <https://doi.org/10.2307/3545569>
- Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol Biol Evol* 19:101–109. <https://doi.org/10.1093/oxfordjournals.molbeva.003974>
- Santos JC (2012) Fast molecular evolution associated with high active metabolic rates in poison frogs. *Mol Biol Evol* 29:2001–2018. <https://doi.org/10.1093/molbev/mss069>
- Schliep KP (2011) phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593
- Smith AB, Lafay B, Christen R (1992) Comparative variation of morphological and molecular evolution through geologic time: 28S ribosomal RNA versus morphology in echinoids. *Philos Trans R Soc Lond B* 338:365–382. <https://doi.org/10.1098/rstb.1992.0155>
- Speakman JR (2005) Body size, energy metabolism and lifespan. *J Exp Biol* 208:1717–1730. <https://doi.org/10.1242/jeb.01556>
- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Strohm JHT, Gwiadzowski RA, Hanner R (2015) Fast fish face fewer mitochondrial mutations: patterns of dN/dS across fish mitogenomes. *Gene* 572:27–34. <https://doi.org/10.1016/j.gene.2015.06.074>
- Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526. <https://doi.org/10.1093/oxfordjournals.molbev.a040023>
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis (MEGA) software version 6.0. *Mol Biol Evol* 30:2725–2729. <https://doi.org/10.1093/molbev/mst197>

- Thomas JA, Welch JJ, Lanfear R, Bromham L (2010) A generation time effect on the rate of molecular evolution in invertebrates. *Mol Biol Evol* 27:1173–1180. <https://doi.org/10.1093/molbev/msq009>
- Uyeda JC, Zenil-Ferguson R, Pennell MW (2018) Rethinking phylogenetic comparative methods. *Syst Biol* 67:1091–1109. <https://doi.org/10.1093/sysbio/syy031>
- Welch JJ, Bininda-Emonds ORP, Bromham L (2008) Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol Biol* 8:53. <https://doi.org/10.1186/1471-2148-8-53>
- Woolfit M (2009) Effective population size and the rate and pattern of nucleotide substitutions. *Biol Lett* 5:417–420. <https://doi.org/10.1098/rsbl.2009.0155>
- Wright SD, Gillman LN, Ross HA, Keeling DJ (2010) Energy and the tempo of evolution in amphibians. *Global Ecol Biogeogr* 19:733–740. <https://doi.org/10.1111/j.1466-8238.2010.00549.x>
- Wright SD, Ross HA, Keeling DK, McBride P, Gillman LN (2011) Thermal energy and the rate of genetic evolution in marine fishes. *Evol Ecol* 25:525–530. <https://doi.org/10.1007/s10682-010-9416-z>
- Xia X (2018) DAMBE7: New and improved tools for data analysis in molecular biology and evolution. *Mol Biol Evol* 35:1550–1552. <https://doi.org/10.1093/molbev/msy073>
- Zagarese HE, Williamson CE (2001) The implications of solar UV radiation exposure for fish and fisheries. *Fish Fish* 2:250–260. <https://doi.org/10.1046/j.1467-2960.2001.00048.x>
- Ziegler AD, Leffel DJ, Kunala S, Sharma HW, Gailani M, Simon JA, Halperin AJ et al (1993) Mutation hotspots due to sunlight in the p53 gene of nonmelanoma skin cancers. *Proc Natl Acad Sci USA* 90:4216–4220. <https://doi.org/10.1073/pnas.90.9.4216>