

Lifespan prediction in protostomes from trinucleotide composition

Yuiki Kondo

kondo-yuiki746@g.ecc.u-tokyo.ac.jp

The University of Tokyo

Simon Jarman

Curtin University

Alyssa M. Budd

Curtin University

Research Article

Keywords: longevity, invertebrates, nucleotide composition bias, CpG density, LASSO regression, k-mer analysis

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11265326/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Title

Lifespan prediction in protostomes from trinucleotide composition

Yuiki Kondo^{1,2,3}, Simon Jarman⁴ & Alyssa M. Budd^{3,4,5}

¹Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Chiba, Japan

²Department of Biological Sciences, Graduate School of Science, The University of Tokyo,
Hongo, Tokyo, Japan

³Environomics Future Science Platform, Indian Ocean Marine Research Centre, Commonwealth
Scientific and Industrial Research Organisation (CSIRO), Crawley, Western Australia, Australia

⁴School of Molecular and Life Sciences, Curtin University, Perth, Western Australia, Australia

⁵UWA Oceans Institute, The University of Western Australia, Perth, Western Australia,
Australia

Corresponding author: Yuiki Kondo

Atmosphere and Ocean Research Institute, The University of Tokyo

5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8564, Japan

Email: kondo-yuiki746@g.ecc.u-tokyo.ac.jp

Keywords

longevity, invertebrates, nucleotide composition bias, CpG density, LASSO regression, k-mer
analysis

Abstract

Background

Promoter CpG composition has been proposed to predict vertebrate lifespan, with higher CpG density hypothesized to buffer age-related DNA methylation drift. Recent comparative analyses in protostomes suggest that lifespan associations extend to other short motifs across multiple genomic regions. We tested whether these sequence features can predict lifespan in protostomes.

Results

Using lifespan and genomic data for 466 species spanning eight protostome phyla, we compared di-, tri-, and tetranucleotide lifespan predictors across different sets of genomic regions. Models were fitted using least absolute shrinkage and selection operator (LASSO) regression with 327 species for model development and evaluated in a fixed-holdout set of 139 species. Conserved genes were identified with Benchmarking Universal Single-Copy Orthologs (BUSCO). The best-performing model combined relative frequencies (Frequency) and observed/expected ratios (O/E) of 64 trinucleotides across six BUSCO-associated regions: exons, introns, and two regions each around the start and stop codons. Fixed-holdout predictions achieved a predictive R^2 of 0.594 for $\ln(\text{lifespan})$, a mean absolute error of 5.24 years, and a median fold error of 1.91. The model outperformed models based on phylogenetic relatedness or genome size, as well as an individual-gene upstream CpG O/E model adapted from previous promoter-based approaches. Frequency and O/E contributed complementary regional information, with Frequency coefficient magnitudes most prominent around translation initiation and O/E coefficient magnitudes in upstream and downstream flanking regions. In leave-one-taxon-out validation, models were

trained excluding any species from the target taxon. At higher taxonomic ranks (phylum to order), the models captured relative lifespan differences within excluded groups, although absolute lifespan estimates were inaccurate for phyla and classes. At lower ranks (family and genus), the models captured little within-group variation but produced lower prediction errors, suggesting that their higher accuracy largely arose from correctly estimating the average lifespan of each family or genus, rather than predicting variation among its species.

Conclusions

Regional trinucleotide composition supports lifespan prediction across multiple invertebrate phyla and provides information beyond that captured by the phylogenetic predictors. Its ability to recover lifespan associations in taxa absent from training supports the potential applicability to genomic lifespan estimation in any animal group.

Background

Animal life histories vary extensively as a result of adaptation to different environments and diversification of niche exploitation strategies within them (Healy et al., 2014; Jones et al., 2014). Lifespan is a key component of species life history and covaries with a range of physiological and demographic traits; for example, larger animals generally mature later and live longer (Healy et al., 2019). Lifespan is also strongly linked to reproductive strategies, with longer-lived species often having lower mass-specific annual reproductive output over extended periods (Jones et al., 2014; Healy et al., 2019). The adaptive significance of lifespan makes it

important in predicting population biology parameters such as population growth rate and viability (Wang et al., 2019). However, obtaining lifespan estimates for many animal species is difficult. Estimates often require prolonged observation, reliable age determination, or inference from age-specific mortality rates in intensively studied populations (Baylis et al., 2014; Reinke et al., 2020). The lifespan records that are currently available cover only a small fraction of animal diversity, often focusing on abundant and commercially harvested species (de Magalhães et al., 2024; La Mesa & Eastman, 2024).

Genome sequence analyses offer an alternative source of lifespan estimates (Mayne et al., 2019; Budd et al., 2025). Previous work has linked promoter CpG composition to vertebrate longevity (Sheldon et al., 2023). Positive associations in conserved mammalian promoters led to the hypothesis that higher CpG density could buffer age-related changes in DNA methylation and thereby help maintain gene regulation in longer-lived species (McLain & Faulk, 2018). Promoter CpG density has subsequently been used to predict lifespan across vertebrates (Mayne et al., 2019), and CpG observed/expected ratios have supported a fish-specific predictor (Budd et al., 2025). However, negative associations at many promoters show that a uniformly positive CpG–lifespan relationship does not explain all of these patterns (Mayne et al., 2019; Budd et al., 2025). Comparative methylation data provide partial support for CpG buffering, but CpG density does not fully explain interspecific differences in epigenetic drift (Bertucci-Richter & Parrott, 2023). The substantial evolutionary variation in invertebrate gene-body methylation further limits the assumption that the same relationships apply across animals (Sarda et al., 2012; Engelhardt et al., 2022). The vertebrate predictor also showed limited accuracy when applied to non-vertebrates (Mayne et al., 2019). Evaluating other short motifs and gene-associated regions could therefore reveal additional sources of lifespan-related information in protostome genomes.

Associations between sequence composition and lifespan that extend beyond promoter CpG to numerous di-, tri-, and tetranucleotide motifs across multiple genomic regions have been identified in protostomes (Kondo & Budd, 2026). Using sequence counts pooled across conserved genes, they detected widespread lifespan associations in gene-associated non-coding regions, suggesting that regional sequence summaries could predict species lifespan. Comparative associations alone, however, do not establish how accurately lifespan can be predicted for an unobserved species. Shared ancestry presents an additional challenge because sequence composition and lifespan may both be conserved within lineages (Felsenstein, 1985). Evaluating predictions for species from taxonomic groups represented in or absent from the training data therefore addresses distinct aspects of generalization (Roberts et al., 2017).

Here, we used lifespan and genomic data for 466 species from eight protostome phyla to address three questions:

- (1) how accurately can short nucleotide motifs predict lifespan in protostomes?
- (2) does motif composition provide lifespan-related information beyond that captured by phylogenetic relatedness and genome size?
- (3) can these relationships support prediction for taxonomic groups absent from model training?

We addressed these questions by comparing sequence-based models and evaluating predictions for species not used in training, both from taxonomic groups represented in the training data and from groups entirely absent from it.

Methods

Lifespan and genomic data collection

We analyzed 466 protostome species spanning eight phyla from the data set of Kondo & Budd (2026), which describes lifespan curation, gene annotation, and nucleotide counting in detail. Briefly, the response variable was the median published maximum lifespan for each species, and genome assemblies were obtained from the National Center for Biotechnology Information. Lifespan in days was natural-log transformed for analysis. Species names, assembly accessions, lifespan values, taxonomy, and data-set assignments are provided in Supplementary Table S1.

Sequence regions and nucleotide predictors

Sequence regions followed Kondo & Budd (2026). Conserved metazoan orthologs were annotated with BUSCO v5.8.2 (Manni et al., 2021) and Miniprot (Li, 2023) using metazoa_odb10. Coding segments were treated as exons and the intervals between successive coding segments as introns. Four boundary regions were defined relative to coding boundaries: Upstream 1, -300 to -100 bp from the start codon; Upstream 2, -100 to +100 bp from the start codon; Downstream 1, -100 to 0 bp from the stop codon; and Downstream 2, 0 to +200 bp from the stop codon. These four regions, pooled exons, and pooled introns retained gene orientation. Candidate models also included whole-genome and intergenic sequence. Intergenic nucleotide counts were calculated using the full gene set predicted by AUGUSTUS (Stanke et al., 2006). For each assembly, counts from extended gene regions from 2 kb upstream of the start codon to 1 kb downstream of the stop codon were subtracted from whole-genome counts. We examined

each of the eight regions individually and five combined region sets, giving 13 prespecified region sets. As a gene-set sensitivity analysis, candidate-model performance was also evaluated using all AUGUSTUS-predicted genes instead of Miniprot BUSCO genes (Supplementary Fig. S1).

Overlapping di-, tri-, and tetranucleotides were counted with a one-base step after excluding windows containing non-ATGC characters. Counts were pooled by species and region. For motif j , Frequency was its observed frequency,

$$\text{Frequency}_j = f_j = \frac{n_j}{\sum_{k=1}^K n_k},$$

where n_j is the count of motif j , K is the number of possible motifs of the same length, and the denominator is the total count of valid motifs of that length in the same species and region.

Expected frequencies were calculated from lower-order composition in the same species and region:

$$E(AB) = f(A)f(B), \quad E(ABC) = \frac{f(AB)f(BC)}{f(B)}, \quad E(ABCD) = \frac{f(ABC)f(BCD)}{f(BC)}.$$

For motif j , observed/expected enrichment was

$$O/E_j = \frac{f_j}{E_j},$$

where f_j and E_j are its observed and expected frequencies, respectively.

Because strand orientation outside annotated genes is arbitrary, each motif was combined with its reverse complement in whole-genome and intergenic sequence (e.g., AAC with GTT).

For paired motifs, observed frequencies were summed and O/E was calculated by dividing the summed observed frequencies by the summed expectations; self-reverse-complementary motifs remained singletons. This reduced the 16 dinucleotides, 64 trinucleotides, and 256

tetranucleotides to 10, 32, and 136 reverse-complement classes, respectively. Reverse-complement collapsing was not applied to the six oriented gene-associated regions, which retained 16, 64, and 256 motifs. Trinucleotide counts and the corresponding Frequency and O/E measures for the final model are provided in Supplementary Table S2.

Model training and selection

Species were assigned once to a development set ($n = 327$) or fixed-holdout set ($n = 139$), and the same assignment was used throughout. The fixed holdout was not used for model selection or tuning. Within the development set, we compared 117 predictor configurations defined by three motif lengths, three composition measure sets (Frequency, O/E, and Frequency + O/E), and 13 region sets.

For each predictor configuration, the development set was repeatedly divided into an outer-training subset (70%) and an outer-test subset (30%) for 1,000 outer resampling iterations. A linear LASSO regression model with an intercept was fitted to each outer-training subset (Tibshirani, 1996). The coefficients were chosen to minimize the following objective, which balances squared prediction error against an absolute-coefficient penalty:

$$\mathcal{L}(\beta_0, \boldsymbol{\beta}; \lambda) = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|.$$

Here, n is the number of outer-training species, p is the number of predictors, y_i is known $\ln(\text{lifespan days})$, $\mathbf{x}_i$ is the predictor vector for species i , β_0 is the intercept, $\boldsymbol{\beta}$ is the vector of predictor coefficients, and λ controls the strength of the penalty. Larger values of λ produce greater coefficient shrinkage and more coefficients equal to zero. This is the Gaussian family in glmnet (Friedman et al., 2010) with the mixing parameter $\alpha = 1$, corresponding to the LASSO

penalty. Predictors were internally mean-centered and scaled by *glmnet* using the outer-training data; the same transformations were applied when predicting the corresponding outer-test species.

Within each outer-training subset, λ was evaluated over 100 values by tenfold inner cross-validation. We selected λ_{1se} , the largest value of λ whose mean cross-validation error was within one standard error of the minimum, thereby favoring a more strongly regularized model with cross-validation error close to the minimum. The fitted model was then applied without refitting to the outer-test subset. Predictor configurations were ranked by mean outer-test mean squared error (MSE) across the 1,000 iterations. The lowest-error configuration—trinucleotide Frequency + O/E from the six gene-associated regions—was fixed for all subsequent sequence-model analyses. It contained $64 \text{ trinucleotides} \times 2 \text{ composition measures} \times 6 \text{ regions} = 768 \text{ predictors}$.

Prediction generation, performance, and error analysis

For each development species, predictions were averaged only from splits in which that species was in the outer-test subset, typically about 300 of the 1,000 splits. This produced one cross-fitted prediction: an average based only on models that had not been trained on that species. For each fixed-holdout species, predictions were averaged across all 1,000 outer models.

Performance was summarized separately for development and fixed-holdout species by MSE and the predictive coefficient of determination (R^2_{pred}):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad R^2_{\text{pred}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Here, n is the number of evaluated species; y_i and $\hat{y}_i$ are the known and predicted ln(lifespan days) of species i ; $\bar{y}$ is the mean known value in the evaluated subset; SSE is the sum of squared prediction errors; and SST is the total sum of squares around $\bar{y}$. Thus, $R^2_{\text{pred}} = 0$ indicates no reduction in squared error relative to predicting the subset mean, and negative values indicate worse performance.

Prediction errors were summarized on three complementary scales: absolute error (AE, years) on the original time scale, fold error (FE) on the ratio scale, and signed log error (SLE) on the log scale:

$$AE_i = \left| \frac{e^{\hat{y}_i}}{365} - \frac{e^{y_i}}{365} \right|, \quad FE_i = e^{|\hat{y}_i - y_i|}, \quad SLE_i = \hat{y}_i - y_i.$$

Positive and negative SLE values denote overprediction and underprediction, respectively (Supplementary Fig. S2).

Comparison with alternative predictors and control models

The selected trinucleotide predictor set (Motifs, M) was compared with 12 phylogenetic eigenvectors (Phylogeny, P) and ln-transformed genome size (Genome size, G) on the same fixed holdout. Pairwise and three-way combinations of these predictor sets were also evaluated, giving seven models in total. The 12 phylogenetic eigenvectors were obtained by eigen decomposition of the species-by-species phylogenetic correlation matrix derived from a protostome tree downloaded from the Open Tree of Life (Hinchliff et al., 2015). Species absent from the tree were assigned to the terminal branch of a closely related species. The species-to-tree-tip assignments and the 12 eigenvector values are provided in Supplementary Table S3.

Three additional benchmark models were prespecified. The individual-BUSCO CpG model used one Upstream 1 CpG O/E value per BUSCO gene, corresponding to the promoter-

CpG approaches of Mayne et al. (2019) and Budd et al. (2025). The BUSCO presence/absence model encoded recovery of each *metazoa_odbl0* ortholog as 1 or 0. The randomized-predictor control used the same 768 trinucleotide Frequency/O/E predictor columns after their values had been shuffled among species before model fitting. All benchmark models used the same development/fixed-holdout assignment as the selected model.

For M, P, and G, we applied the inclusion–exclusion equations of three-set commonality analysis (Nimon et al., 2008, 2017) to partition the fixed-holdout predictive R^2 of the complete M + P + G model into unique and shared components. The total contribution of a predictor set was the sum of every component containing that set, expressed relative to the complete-model predictive R^2 .

Feature selection and coefficient summaries

Feature summaries used the 1,000 outer fits and no fixed-holdout outcomes. For predictor j , selection frequency was the proportion of fits with a nonzero coefficient. Unselected predictors were assigned coefficients of zero when calculating ensemble means. Because *glmnet* returns coefficients on the original predictor scale, coefficients were standardized using the standard deviation s_j of predictor j across all 327 development species:

$$\bar{\beta}_j^{\text{std}} = s_j \left(\frac{1}{R} \sum_{r=1}^R \beta_{jr} \right), \quad \overline{|\beta|}_j^{\text{std}} = s_j \left(\frac{1}{R} \sum_{r=1}^R |\beta_{jr}| \right).$$

Here, β_{jr} is the coefficient of predictor j in outer fit r , $R = 1,000$, $\bar{\beta}_j^{\text{std}}$ is the mean signed standardized coefficient, and $\overline{|\beta|}_j^{\text{std}}$ is the mean absolute standardized coefficient. The latter was summed across predictors to obtain the total standardized coefficient magnitude of a region, composition measure, or motif category. For a group g , its coefficient-magnitude share was

247

$$S_g = \frac{\sum_{j \in g} \overline{|\beta|_j}^{\text{std}}}{\sum_{j=1}^p \overline{|\beta|_j}^{\text{std}}}.$$

248

Here, g denotes the predictor group and $p = 768$ is the total number of predictors. For matched

249

Frequency and O/E predictors, selection frequencies were compared by Spearman's ρ within

250

each region, with false-discovery-rate (FDR) adjustment across the six tests using the

251

Benjamini–Hochberg method (Supplementary Fig. S3).

252

253 *Motif category enrichment*

254

We tested whether coefficient magnitude was concentrated in three simple trinucleotide

255

sequence categories—AT-rich motifs (at least two A/T bases), homopolymers (three consecutive

256

identical bases), and purine-rich motifs (at least two A/G bases)—or in motifs containing each of

257

the 16 possible dinucleotides in either adjacent position. Category membership could overlap.

258

Each test corresponded to one category within one of six regions and one of two composition

259

measures, giving $19 \times 6 \times 2 = 228$ tests.

260

For a category group g , coefficient-magnitude enrichment was calculated as

261

$$Q_g = \frac{S_g}{n_g/p} = \frac{\left(\sum_{j \in g} \overline{|\beta|_j}^{\text{std}}\right)/n_g}{\left(\sum_{j=1}^p \overline{|\beta|_j}^{\text{std}}\right)/p},$$

262

where S_g is the observed coefficient-magnitude share, n_g is the number of predictors in the

263

category group, and $p = 768$. Thus, $Q_g > 1$ indicates greater mean coefficient magnitude than

264

expected from the number of predictors in the group. For each category containing n_g predictors,

265

we sampled 10,000 random sets of n_g distinct predictors from the full set of 768 predictors and

266

calculated the mean absolute standardized coefficient for each set. The two-sided

empirical p value was calculated as $(b + 1)/(10,000 + 1)$, where b is the number of random-set means whose absolute deviation from the overall predictor mean was at least as large as that of the observed category mean (Phipson & Smyth, 2010). The resulting p values were FDR-adjusted across the 228 tests using the Benjamini–Hochberg method (Supplementary Fig. S4).

Within-taxon analyses

We tested whether predictions tracked lifespan among species within named phyla, classes, orders, families, and genera. Development species contributed cross-fitted predictions and fixed-holdout species contributed 1,000-model averages. Taxa with fewer than three species were excluded. We first estimated a slope for each taxon and summarized the proportion above zero with a Wilson 95% confidence interval (Wilson, 1927).

At each rank, a common within-taxon slope was estimated with the additive fixed-effects model

$$\hat{y}_{it} = \alpha_t + \beta x_{it}^{\text{within}} + \varepsilon_{it},$$

where $\hat{y}_{it}$ is predicted $\ln(\text{lifespan})$ for species i in taxon t , α_t is the taxon-specific intercept, x_{it}^{within} is known $\ln(\text{lifespan})$ centered on the mean of taxon t , β is the common within-taxon slope, and ε_{it} is the residual. A second model allowed this slope to vary with within-taxon lifespan range and taxon size:

$$\hat{y}_{it} = \alpha_t + (\beta + \gamma_R R_t + \gamma_N N_t) x_{it}^{\text{within}} + \varepsilon_{it},$$

where R_t is the z-standardized range of known $\ln(\text{lifespan})$ in taxon t , N_t is z-standardized $\ln(\text{number of species})$, and γ_R and γ_N quantify their associations with the within-taxon slope. Z-standardization was performed across eligible taxa separately at each taxonomic rank.

Coefficients were estimated by ordinary least squares. Species in the same taxon may share unmodeled evolutionary and genomic characteristics, so their residuals need not be independent, and residual variance may differ among taxa. We therefore used taxon-clustered standard errors with the CR2 small-sample correction (Pustejovsky & Tipton, 2018). Satterthwaite-adjusted degrees of freedom were used separately to obtain confidence intervals and p values from the CR2 variance estimates. P values were FDR-adjusted across the five taxonomic ranks separately for the additive slope, lifespan-range interaction, and group-size interaction.

Leave-one-taxon-out validation

To evaluate model generalization to taxa absent from training, leave-one-taxon-out (LOTO) validation was conducted separately at the phylum, class, order, family, and genus levels. For each target taxon represented in the development set, all development species belonging to that taxon were withheld. A new LASSO model using the fixed final configuration of 768 trinucleotide Frequency and O/E predictors from the six gene-associated regions was fitted to the remaining development species and used to predict both development and fixed-holdout species in the withheld taxon. For taxa represented only in the fixed holdout, the complete development set served as the training set.

For each LOTO training set, λ was selected from 100 candidate values using up to 10-fold taxon-blocked inner cross-validation and the one-standard-error rule. Among the taxa remaining after the outer exclusion, all species belonging to the same taxon at the evaluated rank were assigned to the same inner fold. Thus, taxa assigned to an inner-validation fold were absent from the corresponding inner-training subset. This prevented λ selection from benefiting from

within-taxon similarity between training and validation species and provided a more stringent tuning criterion for generalization across taxonomic groups. After λ was selected, the model was refitted using all remaining development species. Predictive R^2 was calculated at each rank for all species, development species, and fixed-holdout species. To test recovery of relative lifespan within unseen taxa, we fitted the additive model above using LOTO predictions among taxa represented by at least three species. The common slope was tested against zero with the same taxon-clustered CR2 and Satterthwaite procedure, and p values were FDR-adjusted across the five ranks.

Code availability

The analysis code is available in the Zenodo repository specified in the Availability of data and materials section. Portions of the code were developed with assistance from ChatGPT (OpenAI) and subsequently reviewed and validated by the authors.

Results

Model selection and fixed-holdout evaluation

Across the 117 candidate sequence models, trinucleotide Frequency + O/E predictors from the six gene-associated regions had the lowest mean outer-test MSE (1.404 on the $\ln[\text{lifespan-days}]$ scale; Fig. 1a and Supplementary Table S4). This 768-predictor configuration was therefore fixed for all subsequent analyses (Supplementary Table S5). When subsequently evaluated on the independent fixed holdout, it ranked second (MSE = 1.255), only 0.00026

above the holdout-best model, which additionally included intergenic sequence. Candidate rankings were highly concordant between outer testing and fixed-holdout evaluation (Spearman's $\rho = 0.963$), and the AUGUSTUS all-gene sensitivity analysis showed a similar performance landscape (Supplementary Fig. S1 and Supplementary Table S4).

On the fixed holdout, the trinucleotide model (M) outperformed models based only on phylogenetic eigenvectors (P), genome size (G), BUSCO presence/absence, randomized predictors, or individual-BUSCO CpG O/E (MSE: M, 1.25; P, 2.00; G, 2.87; BUSCO presence/absence, 2.28; randomized predictors, 3.11; CpG O/E, 1.91; Fig. 1b and Supplementary Table S6). Adding phylogeny modestly improved performance (M + P, 1.19), whereas adding genome size did not (M + G, 1.26); the complete M + P + G model also had an MSE of 1.19.

The predictive R^2 decomposition similarly showed that M captured approximately 97% of the predictive R^2 of the complete model, compared with 58% for P and 12% for G (Fig. 1c and Supplementary Table S7). These overlapping totals indicate that motifs contained most of the available predictive information, although part of their signal was shared with phylogeny.

Out-of-sample performance and prediction errors

Species-level cross-fitted predictions explained 56.6% of the variation in $\ln(\text{lifespan})$ among the 327 development species (MSE = 1.36; Fig. 2a). Predictions for the 139 fixed-holdout species explained 59.4% (MSE = 1.25; Fig. 2a). Mean absolute error was 5.36 years in the development set and 5.24 years in the fixed holdout; median fold errors were 2.18 and 1.91, respectively (Fig. 2b,c and Supplementary Table S8). Thus, performance was similar for cross-fitted development predictions and for species excluded from all model fitting. Errors showed regression toward the center of the lifespan distribution in both data sets: short-lived species

tended to be overpredicted and long-lived species underpredicted (Supplementary Fig. S2). The model therefore recovered broad lifespan variation but was less accurate at the short- and long-lived extremes.

Distributed contributions of Frequency and O/E predictors

Repeatedly selected predictors generally had stronger correlations with lifespan and larger ensemble-averaged standardized coefficients (Fig. 3a,b and Supplementary Table S9). Selection frequency and coefficient direction depended on motif, composition measure, and genomic region. O/E predictors accounted for 65.6% of the total standardized coefficient magnitude and Frequency predictors for 34.4% (Fig. 3c). Their contributions differed by region: O/E contribution was greatest in Downstream 2 and Upstream 1, whereas Frequency contribution was greatest in Upstream 2. Matched Frequency and O/E selection frequencies had little correlation in most regions (Spearman's $\rho = 0.05\text{--}0.20$) and were moderately correlated only in Downstream 1 ($\rho = 0.45$, FDR-adjusted $p = 0.0006$; Supplementary Fig. S3). Frequency and O/E therefore contributed complementary regional information.

Several motif categories accounted for relatively large shares of the total coefficient magnitude, including purine-rich Frequency predictors in Upstream 2 and GC-containing O/E predictors in Downstream 2. However, none differed significantly from the share expected from category size (FDR-adjusted $p > 0.05$; Supplementary Fig. S4 and Supplementary Table S10), providing no evidence that the predictive signal was concentrated in a few simple sequence categories.

Prediction within taxa

Taxon-specific slopes were positive in 5 of 5 eligible phyla, 12 of 15 classes, 28 of 46 orders, 23 of 51 families, and 9 of 22 genera (Fig. 4a,b and Supplementary Table S11). The proportion of positive slopes therefore declined from 100% at the phylum level to 41% at the genus level. In the additive models accounting for taxon-specific intercepts, the common within-taxon slope also decreased toward finer taxonomic ranks. It was positive after FDR adjustment at the phylum level ($\beta = 0.440$, 95% CI = 0.396–0.484, FDR-adjusted $p = 0.00066$) and class level ($\beta = 0.364$, 95% CI = 0.251–0.478, FDR-adjusted $p = 0.00066$). The order-level slope was also positive ($\beta = 0.258$, 95% CI = 0.026–0.491, unadjusted $p = 0.035$) but did not remain significant after correction across ranks (FDR-adjusted $p = 0.059$). The family- and genus-level slopes did not differ from zero ($\beta = -0.019$ and -0.031 , respectively) (Supplementary Table S12).

At the order level, a one-standard-deviation increase in the range of known $\ln(\text{lifespan})$ was associated with a 0.119 increase in the within-order slope (95% CI = 0.042–0.196, FDR-adjusted $p = 0.030$; Fig. 4c and Supplementary Table S13). Lifespan range was not associated with the slope at the other ranks, and there was no evidence that the number of species modified the within-taxon slope at any rank.

Prediction after taxon exclusion

After entire target taxa were excluded from training, LOTO predictions retained positive within-taxon associations at the three broader taxonomic ranks. The common slopes were positive after phylum exclusion ($\beta = 0.153$, 95% CI = 0.073–0.234, FDR-adjusted $p = 0.020$), class exclusion ($\beta = 0.242$, 95% CI = 0.136–0.348, FDR-adjusted $p = 0.007$), and order exclusion ($\beta = 0.232$, 95% CI = 0.107–0.357, FDR-adjusted $p = 0.011$; Supplementary Fig. S5a and Supplementary Table S14). In contrast, the slopes did not differ from zero after family exclusion

($\beta = 0.023$, 95% CI = -0.026 – 0.071 , FDR-adjusted $p = 0.424$) or genus exclusion ($\beta = -0.005$, 95% CI = -0.094 – 0.084 , FDR-adjusted $p = 0.901$).

Predictive R^2 across all species showed a contrasting pattern. It was low after phylum and class exclusion (0.137 and 0.016, respectively) but substantially higher after order, family, and genus exclusion (0.483, 0.536, and 0.562, respectively; Supplementary Fig. S5b and Supplementary Table S15). Development and fixed-holdout species followed the same general pattern. Thus, overall species-level predictive accuracy was generally higher at finer taxonomic ranks, whereas positive within-group relationships were retained only after exclusion at broader ranks. This contrast suggests that the higher R^2 after family or genus exclusion primarily reflected differences in mean lifespan among families or genera, rather than accurate recovery of lifespan variation among species within those groups.

Discussion

Lifespan was predicted in a range of protostome species by regional trinucleotide composition. To our knowledge, this is the first model developed to predict species lifespan from genomic sequence composition across multiple invertebrate phyla. The selected model achieved a predictive R^2 of 0.594 for $\ln(\text{lifespan})$ in the 139 fixed-holdout species, with a median fold error of 1.91. Short motifs provided lifespan-related information beyond that captured by the phylogenetic predictors, and models fitted after taxon exclusion retained predictive associations within excluded phyla, classes, and orders. These results support genomic lifespan prediction

across a broad taxonomic range of protostomes, while recognizing difficulties in resolving lifespan differences among closely related species.

Sequence predictors and gene sets

The results presented here support expanding genomic lifespan predictors beyond CpG composition in upstream regions. The trinucleotide model had lower fixed-holdout error than the individual-BUSCO CpG O/E model evaluated under the same framework (Fig. 1). Our CpG predictors were calculated from fixed intervals 300–100 bp upstream of the annotated translation start in conserved BUSCO genes. By contrast, Mayne et al. (2019) and Budd et al. (2025) identified homologous promoters using experimentally defined human and zebrafish transcription start sites from the Eukaryotic Promoter Database (EPD; Meylan et al., 2020). The comparison therefore favors the broader motif model for our protostome dataset, although differences in gene sets and promoter definitions limit direct comparison with those published models. Whether the same advantage extends to vertebrates remains to be tested.

Gene-set choice can itself influence prediction. Miniprot BUSCO and all-AUGUSTUS models showed broadly similar performance patterns, but the combined-region models generally performed better with BUSCO genes (Supplementary Fig. S1). In phylogenetic comparative analyses, Kondo & Budd (2026) likewise found stronger lifespan associations for AUGUSTUS-predicted BUSCO genes than for the full AUGUSTUS gene set. This indicates that the difference is not restricted to comparisons between annotation tools. Several explanations remain possible. More accurate annotation of conserved genes may reduce errors in the sequences and coding boundaries used to calculate motif compositions (Waterhouse et al., 2018). Given that shared ancestry can shape both genomic and phenotypic similarities (Dewar et al., 2025),

BUSCO-associated composition may capture this signal more strongly and thereby improve prediction of phylogenetically conserved lifespan differences (Kondo & Budd, 2026). Conserved genes may also experience functional constraints on gene expression and protein homeostasis that connect sequence composition to life-history evolution (Drummond & Wilke, 2008; Ke et al., 2017; see below). The relative contributions of annotation accuracy, phylogenetic signal, and functional constraints remain unresolved by these comparisons.

Phylogenetic signal and prediction across taxa

The model comparisons indicate that short motifs contain lifespan-related information beyond the phylogenetic structure captured by the eigenvector predictors. The trinucleotide model recovered approximately 97% of the predictive R^2 of the model combining motifs, phylogeny, and genome size, including information shared among these predictor sets. Adding phylogenetic eigenvectors modestly improved prediction, whereas adding genome size did not (Fig. 1). Motif composition therefore captures much of the information needed for prediction while complementing the phylogenetic predictors. These predictors summarize a topology-based Open Tree of Life phylogeny (Hinchliff et al., 2015); its branch lengths do not represent empirical estimates of divergence times or sequence divergence. The 12 eigenvectors consequently provide an approximate description of evolutionary relatedness. A phylogeny with empirically estimated branch lengths would allow a more detailed assessment of the information shared between sequence composition and evolutionary history.

The analyses within taxa and after taxon exclusion clarify where this predictive information is useful. Positive associations within phyla and classes show that the original model recovered lifespan differences within broad groups, whereas slopes near zero within families and

genera indicate limited discrimination among closely related species (Fig. 4). After exclusion of entire phyla, classes, or orders, positive within-taxon associations remained, showing that models fitted and tuned using other taxa could still recover lifespan differences within the excluded groups (Supplementary Fig. S5). The low overall predictive R^2 after phylum or class exclusion nevertheless indicates that absolute prediction remained difficult. Conversely, the higher overall R^2 after family or genus exclusion, together with within-group slopes near zero, is consistent with predictions mainly capturing differences in mean lifespan among families or genera. Thus, predictions can remain informative for species from taxonomic groups absent from training, but their ability to resolve lifespan variation depends on the taxonomic scale.

Poor discrimination within families and genera may reflect limited variation in both regional motif composition and lifespan. Closely related species may retain similar motif frequencies despite substantial divergence in lifespan. For example, in the fish CpG model, *Sebastes* species varied considerably in lifespan but relatively little in CpG O/E (Budd et al., 2025). Limited lifespan variation may pose an additional constraint, as suggested by the positive association between within-order slopes and the range of known $\ln(\text{lifespan})$ (Fig. 4c). At lower taxonomic levels, where true lifespan differences among species are often small, measurement error in lifespan could obscure the available signal. Analyses of closely related species with substantial lifespan divergence would test whether regional composition tracks these changes or whether more specific regulatory and functional sequence features are needed.

Biological interpretation of selected motifs

Selected motif–region combinations can reflect several evolutionary processes. Similarity inherited from common ancestors may make motif composition informative because lifespan is

495 phylogenetically conserved (Zhu et al., 2023). Motif composition may also evolve in association
496 with lifespan through functional constraints on gene expression or through life-history-related
497 differences in mutation and fixation biases (Romiguier et al., 2010). Feature selection alone
498 cannot determine which processes account for the predictive value of an individual motif.
499 Correlated motifs can also substitute for one another or share predictive information, so a large
500 mean absolute standardized coefficient does not necessarily identify an independent biological
501 effect (Zou & Hastie, 2005). Moreover, none of the tested motif categories showed significant
502 enrichment after FDR adjustment (Supplementary Fig. S4). Individual motifs are therefore best
503 discussed as candidates for functional investigation within the broader pattern of regional
504 sequence variation.

505 Frequency coefficients were most prominent in Upstream 2, whereas O/E coefficients
506 were prominent in Downstream 2 and Upstream 1 (Fig. 3). Frequency captures observed
507 frequency, while trinucleotide O/E measures departures from dinucleotide-based expectations.
508 Upstream 1 and Downstream 2 were also among the regions with the highest proportions of
509 lifespan-associated motifs in the phylogenetic comparative analyses of Kondo & Budd (2026).
510 One interpretation of the present model is that prominent Upstream 2 Frequency features partly
511 reflect functional constraints around translation initiation, while O/E features in flanking regions
512 capture regulatory sequence variation beyond lower-order composition. ATG and TGG in
513 Upstream 2 provide a specific example: their repeated selection may partly reflect initiation
514 contexts such as ATGG. Start-codon context affects translation initiation, with preferences
515 varying among eukaryotic lineages (Kozak, 1986; Nakagawa et al., 2008). Higher translation
516 fidelity is associated with longer lifespan across rodents (Ke et al., 2017), and increasing fidelity
517 extended lifespan in yeast, worms, and flies (Martinez-Miguel et al., 2021). Reducing

translation-initiation factor levels can also extend nematode lifespan (Hansen et al., 2007). These findings motivate testing whether translation-related sequence variation contributes to the lifespan relationships detected here. The prominence of Downstream 2 O/E may partly reflect sequence-dependent regulation of mRNA stability and transcript turnover in 3' UTR sequences (Litterman et al., 2019). Reported associations between shorter lifespan and faster basal protein turnover in mammals (Swovick et al., 2021) suggest a possible link between these sequence features and lifespan-related variation in molecular turnover and cellular metabolism.

Codon usage may contribute further information from exons and the coding portion of Upstream 2. Synonymous codon choice can affect mRNA stability (Presnyak et al., 2015; Bazzini et al., 2016), and selection on codon usage varies across animals. Bénitière et al. (2025) found that species with the strongest selection on translationally optimal codons tended to have short lifespans and other traits associated with large effective population sizes, although associations with these proxies were weak. Lifespan-related codon variation could therefore reflect differences in the efficacy of selection as well as functional demands on gene expression. Mutation bias and fixation processes also shape nucleotide composition (Long et al., 2018). Analyses that preserve reading frame and distinguish synonymous codon usage from amino-acid composition could help resolve their contributions to the trinucleotide predictors.

Limitations and future directions

Lifespan prediction accuracy depends on the quality and distribution of lifespan observations used for model calibration. Here, short-lived species tended to be overpredicted and long-lived species underpredicted, indicating limited recovery of the extremes of the observed range (Supplementary Fig. S2). Published maximum lifespan estimates vary with observation

conditions and sampling effort (Moorad et al., 2012), and taxonomic coverage remains uneven. Genome assembly quality may also influence the sequence features used for prediction. Previous analyses found that estimated intergenic proportions covaried with assembly-quality metrics, although associations between intergenic dinucleotide composition and lifespan were largely unchanged after adjustment for these metrics (Kondo & Budd, 2026). Better-supported lifespan records, additional observations at the extremes of the lifespan distribution as well as from poorly represented lineages, and improved genome assemblies would strengthen the basis for model calibration and independent evaluation.

Further gains may come from refining the design of sequence-based predictors. Within the present dataset and modelling framework, tetranucleotide models offered no clear improvement over the selected trinucleotide model despite their larger predictor sets (Fig. 1). Regional differences in predictor contributions and the gene-set comparisons also suggest that the choice of sequences matters. Future work may therefore be better directed toward identifying additional informative genomic regions, defining regulatory boundaries more precisely, and refining gene sets used to construct predictors. Functional analyses of selected motif–region combinations could then test how the underlying sequences influence gene expression, translation, and lifespan. More broadly, assembling comparable lifespan and genomic data across a wider range of animal lineages would enable taxon-exclusion tests to clarify where a common model predicts lifespan adequately and where lineage-specific models are needed. Such models could extend lifespan estimation to species with limited demographic information and support comparative studies of lifespan evolution across animals.

Conclusions

Our findings address the three questions posed in this study. (1) Short nucleotide motifs predicted substantial lifespan variation across multiple invertebrate phyla, as demonstrated by our regional trinucleotide model. (2) Model comparisons suggest that motif composition provides lifespan-related information beyond that captured by the phylogenetic and genome-size predictors used here. (3) Validation supported the model's applicability to species from taxonomic groups absent from training, although accurately predicting lifespans at the short- and long-lived extremes and resolving differences among closely related species remain challenging. These findings extend the scope of genomic lifespan prediction beyond vertebrate promoter-CpG models and provide a foundation for its broader application across animal diversity.

References

- Baylis SM, de Lisle M, Hauber ME. Inferring maximum lifespan from maximum recorded longevity in the wild carries substantial risk of estimation bias. *Ecography*. 2014;37:770-80.
- Bazzini AA, del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, et al. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J*. 2016;35:2087-103.
- Bénitière F, Lefébure T, Duret L. Variation in the fitness impact of translationally optimal codons among animals. *Genome Res*. 2025;35:446-58.

Bertucci-Richter EM, Parrott BB. The rate of epigenetic drift scales with maximum lifespan
 across mammals. *Nat Commun.* 2023;14:7731.

Budd AM, Mayne B, Berry O, Jarman S. Fish species lifespan prediction from promoter
 cytosine-phosphate-guanine density. *Mol Ecol Resour.* 2025;25:e13774.

de Magalhães JP, Abidi Z, dos Santos GA, Avelar RA, Barardo D, Chatsirisupachai K, et al.
 Human Ageing Genomic Resources: updates on key databases in ageing research. *Nucleic
 Acids Res.* 2024;52:D900-D908.

Dewar AE, Belcher LJ, West SA. A phylogenetic approach to comparative genomics. *Nat Rev
 Genet.* 2025;26:395-405.

Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint
 on coding-sequence evolution. *Cell.* 2008;134:341-52.

Engelhardt J, Scheer O, Stadler PF, Prohaska SJ. Evolution of DNA methylation across
 Ecdysozoa. *J Mol Evol.* 2022;90:56-72.

Felsenstein J. Phylogenies and the comparative method. *Am Nat.* 1985;125:1-15.

Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via
 coordinate descent. *J Stat Softw.* 2010;33(1):1-22.

Hansen M, Taubert S, Crawford D, Libina N, Lee SJ, Kenyon C. Lifespan extension by
 conditions that inhibit translation in *Caenorhabditis elegans*. *Aging Cell.* 2007;6:95-110.

Healy K, Ezard THG, Jones OR, Salguero-Gómez R, Buckley YM. Animal life history is shaped
 by the pace of life and the distribution of age-specific mortality and reproduction. *Nat Ecol
 Evol.* 2019;3:1217-24.

Healy K, Guillaume T, Finlay S, Kane A, Kelly SBA, McClean D, et al. Ecology and mode-of-
 life explain lifespan variation in birds and mammals. *Proc Biol Sci.* 2014;281:20140298.

609 Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, et al. Synthesis of
 610 phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci U S A*.
 611 2015;112:12764-9.
 612 Jones OR, Scheuerlein A, Salguero-Gómez R, Camarda CG, Schaible R, Casper BB, et al.
 613 Diversity of ageing across the tree of life. *Nature*. 2014;505:169-73.
 614 Ke Z, Mallik P, Johnson AB, Luna F, Nevo E, Zhang ZD, et al. Translation fidelity coevolves
 615 with longevity. *Aging Cell*. 2017;16:988-93.
 616 Kondo Y, Budd A. Di-, tri-, and tetranucleotide frequencies covary with lifespan and genome
 617 size across protostome invertebrates. *J Evol Biol*. 2026. doi:10.1093/jeb/voag086.
 618 Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates
 619 translation by eukaryotic ribosomes. *Cell*. 1986;44:283-92.
 620 La Mesa M, Eastman JT. Assessing current knowledge and future challenges of age
 621 determination, life span and growth performance in notothenioid fishes: a review. *Rev Fish*
 622 *Biol Fish*. 2024;34:575-96.
 623 Li H. Protein-to-genome alignment with miniprot. *Bioinformatics*. 2023;39:btad014.
 624 Litterman AJ, Kageyama R, Le Tonqueze O, Zhao W, Gagnon JD, Goodarzi H, et al. A
 625 massively parallel 3' UTR reporter assay reveals relationships between nucleotide content,
 626 sequence conservation, and mRNA destabilization. *Genome Res*. 2019;29:896-906.
 627 Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, et al. Evolutionary
 628 determinants of genome-wide nucleotide composition. *Nat Ecol Evol*. 2018;2:237-40.
 629 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and
 630 streamlined workflows along with broader and deeper phylogenetic coverage for scoring of
 631 eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021;38:4647-54.

632 Martinez-Miguel VE, Lujan C, Espie-Caullet T, Martinez-Martinez D, Moore S, Backes C, et al.
 633 Increased fidelity of protein synthesis extends lifespan. *Cell Metab.* 2021;33:2288-2300.e12.
 634 Mayne B, Berry O, Davies C, Farley J, Jarman S. A genomic predictor of lifespan in vertebrates.
 635 *Sci Rep.* 2019;9:17866.
 636 McLain AT, Faulk C. The evolution of CpG density and lifespan in conserved primate and
 637 mammalian promoters. *Aging (Albany NY).* 2018;10:561-72.
 638 Meylan P, Dreos R, Ambrosini G, Groux R, Bucher P. EPD in 2020: enhanced data visualization
 639 and extension to ncRNA promoters. *Nucleic Acids Res.* 2020;48:D65-D69.
 640 Moorad JA, Promislow DEL, Flesness N, Miller RA. A comparative assessment of univariate
 641 longevity measures using zoological animal records. *Aging Cell.* 2012;11:940-8.
 642 Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K. Diversity of preferred nucleotide
 643 sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.*
 644 2008;36:861-71.
 645 Nimon K, Lewis M, Kane R, Haynes RM. An R package to compute commonality coefficients
 646 in the multiple regression case: an introduction to the package and a practical example.
 647 *Behav Res Methods.* 2008;40:457-66.
 648 Nimon K, Lewis M, Kane R, Haynes RM. Erratum to: An R package to compute commonality
 649 coefficients in the multiple regression case: an introduction to the package and a practical
 650 example. *Behav Res Methods.* 2017;49:2275.
 651 Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values
 652 when permutations are randomly drawn. *Stat Appl Genet Mol Biol.* 2010;9:Article 39.
 653 Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a
 654 major determinant of mRNA stability. *Cell.* 2015;160:1111-24.

655 Pustejovsky JE, Tipton E. Small-sample methods for cluster-robust variance estimation and
 656 hypothesis testing in fixed effects models. *J Bus Econ Stat.* 2018;36:672-83.
 657 Reinke BA, Hoekstra L, Bronikowski AM, Janzen FJ, Miller D. Joint estimation of growth and
 658 survival from mark–recapture data to improve estimates of senescence in wild populations.
 659 *Ecology.* 2020;101:e02877.
 660 Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation
 661 strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography.*
 662 2017;40:913-29.
 663 Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33
 664 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome*
 665 *Res.* 2010;20:1001-9.
 666 Sarda S, Zeng J, Hunt BG, Yi SV. The evolution of invertebrate gene body methylation. *Mol*
 667 *Biol Evol.* 2012;29:1907-16.
 668 Sheldon EL, Schrey AW, Lauer ME, Martin LB. Epigenetic potential: Promoter CpG content
 669 positively covaries with lifespan and is dependent on gene function among vertebrates. *J*
 670 *Hered.* 2023;114:207-18.
 671 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio
 672 prediction of alternative transcripts. *Nucleic Acids Res.* 2006;34:W435-9.
 673 Swovick K, Firsanov D, Welle KA, Hryhorenko JR, Wise JP Sr, George C, et al. Interspecies
 674 differences in proteome turnover kinetics are correlated with life spans and energetic
 675 demands. *Mol Cell Proteomics.* 2021;20:100041.
 676 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat*
 677 *Methodol.* 1996;58:267-88.

Wang T, Fujiwara M, Gao X, Liu H. Minimum viable population size and population growth rate of freshwater fishes and their relationships with life history traits. *Sci Rep.* 2019;9:3612.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol.* 2018;35:543-8.

Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc.* 1927;22:209-12.

Zhu P, Liu W, Zhang X, Li M, Liu G, Yu Y, et al. Correlated evolution of social organization and lifespan in mammals. *Nat Commun.* 2023;14:372.

Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol.* 2005;67:301-20.

Supplementary Information

Additional file 1 (.xlsx): Supplementary Tables S1–S15. Species data, sequence composition, model predictions and coefficients, and statistical analysis results.

Additional file 2 (.pdf): Supplementary Figures S1–S5. Additional model comparisons, prediction-error and feature analyses, and taxon-exclusion validation, with figure legends.

Declarations

701 ***Ethics approval and consent to participate***

702 Not applicable.

703

704 ***Consent for publication***

705 Not applicable.

706

707 ***Availability of data and materials***

708 Genome assemblies were downloaded from NCBI, with the accession numbers provided in

709 Supplementary Table S1. The supplementary tables primarily present data and results for the

710 main trinucleotide model. Detailed input data and model outputs, including those for alternative

711 models, together with the analysis code, have been uploaded to Zenodo (DOI:

712 10.5281/zenodo.23034047) and will be made publicly available upon acceptance of the

713 manuscript. During peer review, these materials can be accessed via the following private link:

714 [https://zenodo.org/records/23034047?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImRiZDhjYTQ3LTA0Y2QtNDA3ZC1iOTVhLTRhMTIxODBhZjYwOSIsImRhdGEiOnt9LCJyYW5kb20iOiIzMtM0OGQ2OWJlNmUyNjU3N2QwODliMWRlZjUyOTYzYyJ9.Lt0UP3ikojNEa_G5n077sqFhlDVFL2g7JHTLJ0HzfW0UYOZfmCd4Iol7Ty5VwFv9c-1EK9KS4jtWi0RqkMfAzg].

715 RiZDhjYTQ3LTA0Y2QtNDA3ZC1iOTVhLTRhMTIxODBhZjYwOSIsImRhdGEiOnt9LCJyY

716 W5kb20iOiIzMtM0OGQ2OWJlNmUyNjU3N2QwODliMWRlZjUyOTYzYyJ9.Lt0UP3ikojNE

717 a_G5n077sqFhlDVFL2g7JHTLJ0HzfW0UYOZfmCd4Iol7Ty5VwFv9c-

718 1EK9KS4jtWi0RqkMfAzg].

719

720 ***Competing interests***

721 The authors declare that they have no competing interests.

722

723 ***Funding***

This research was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [grant number JP24KJ0854 to Y.K.].

Authors' contributions

YK and AMB conceived and designed the study. YK compiled the data, performed the analyses, and drafted the manuscript. AMB and SJ contributed to the interpretation of the results and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Olly Berry and members of the Environomics Future Science Platform, Commonwealth Scientific and Industrial Research Organisation (CSIRO) for their encouragement during this project. We are also grateful to the World-leading Innovative Graduate Study Program in Proactive Environmental Studies (WINGS-PES), The University of Tokyo, for general support on research activities, and to the librarians at the Atmosphere and Ocean Research Institute (AORI), for their assistance in obtaining numerous articles through copying and interlibrary loan. Finally, we gratefully acknowledge the immense effort of all those involved in the creation, submission, and curation of publicly available data that made this study possible.

Figures and their legends

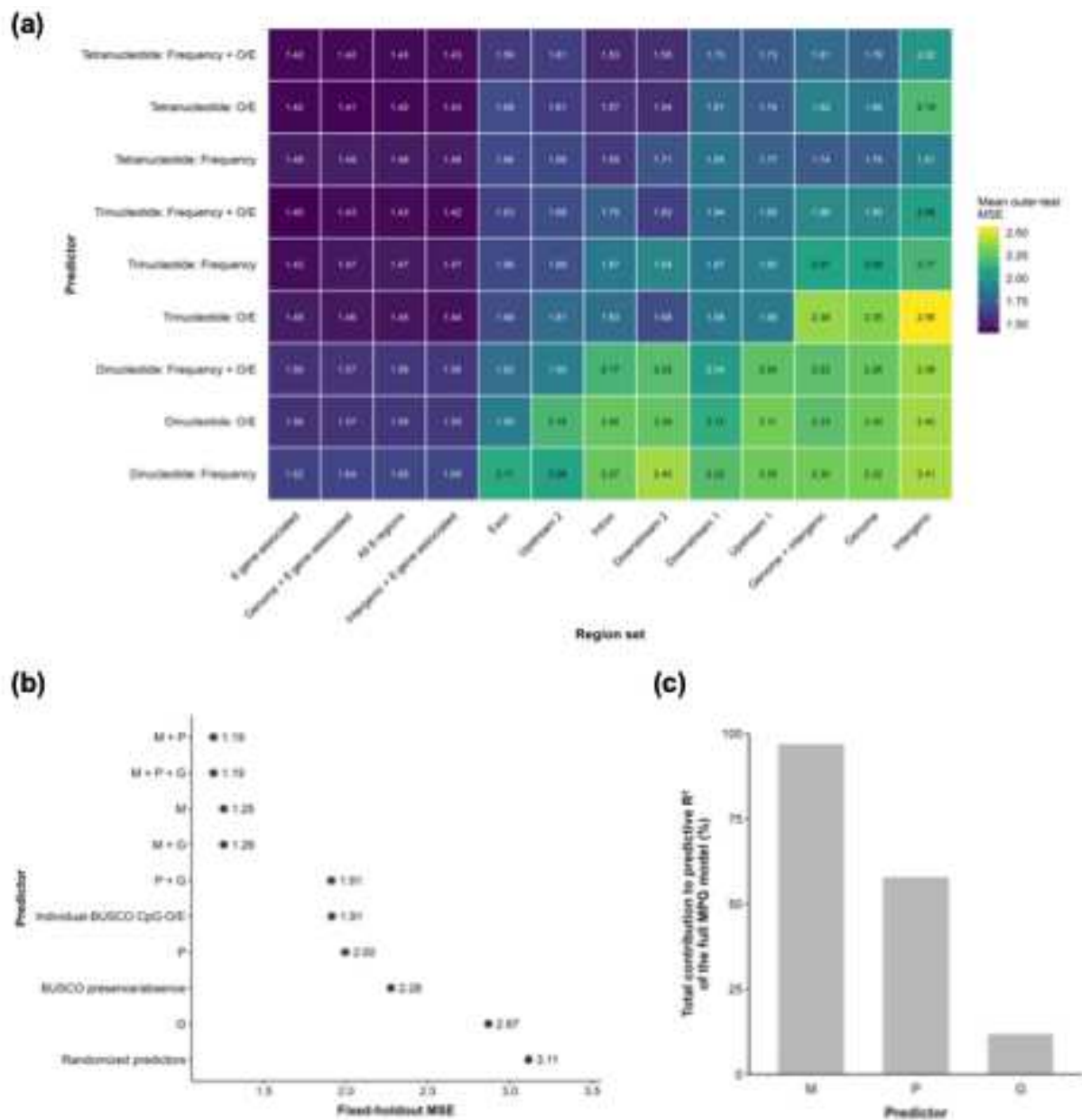


Figure 1. Model selection, fixed-holdout evaluation, and predictor contributions.

(a) Heatmap of mean outer-test mean squared error (MSE) across 1,000 resampling iterations of the development set ($n = 327$ species). The 117 candidate configurations combined three motif lengths (di-, tri-, or tetranucleotide), three composition measure sets (Frequency, O/E, or Frequency + O/E), and 13 region sets. O/E denotes the observed-to-expected ratio. Lower MSE indicates better prediction of ln-transformed lifespan. Rows and columns are ordered independently by mean MSE. The six gene-associated regions comprise two upstream regions, exon and intron regions, and two downstream regions; the eight-region set additionally includes whole-genome and intergenic sequences. Trinucleotide Frequency + O/E predictors from the six gene-associated regions were selected as the final motif predictor set (M). (b) Fixed-holdout MSE for ten models evaluated on 139 species excluded from model selection and fitting. P denotes 12 phylogenetic eigenvectors, and G denotes ln-transformed genome size. The full MPG model combines all three predictor sets. M, P, G, and their pairwise and three-way combinations were compared with individual-BUSCO CpG O/E, BUSCO presence/absence, and randomized-predictor models. The CpG model used Upstream 1 O/E ratios for individual BUSCO genes. Randomized predictors were generated by shuffling each motif predictor among species. Predictions were averaged across 1,000 outer models. (c) Total contributions of M, P, and G to the fixed-holdout predictive coefficient of determination (R^2) of the full MPG model. Each bar includes the contribution unique to that predictor and contributions shared with other predictors, expressed as a percentage of the full-model R^2 .

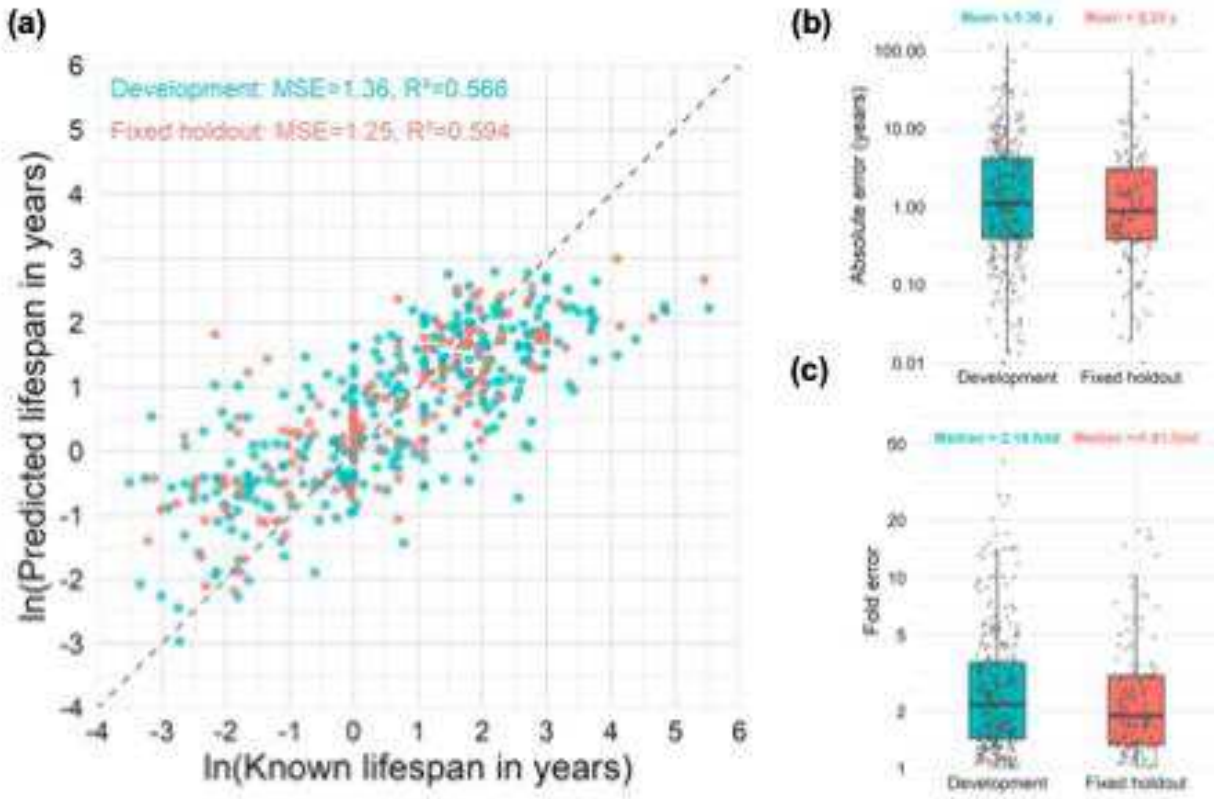


Figure 2. Performance of the final lifespan prediction model for protostomes.

Results are shown for the selected 768-predictor trinucleotide model. Predictions for development species were cross-fitted by averaging only predictions from outer iterations in which each species belonged to the outer-test subset. Predictions for fixed-holdout species were averaged across all 1,000 outer models. (a) Relationship between known and predicted maximum lifespans. Cyan points represent the 70% development species ($n = 327$), while coral points represent the 30% fixed-holdout species ($n = 139$). Mean squared error (MSE) and predictive R^2 are shown separately for each dataset. The dashed line indicates equality between known and predicted lifespan. (b) Absolute prediction error in years, calculated as $|\text{predicted lifespan} - \text{known lifespan}|$. Colored text indicates the mean absolute error for each dataset. (c) Fold error, calculated as $\exp(|\ln(\text{predicted lifespan}) - \ln(\text{known lifespan})|) = \max(\text{predicted/known}, \text{known/predicted})$. A fold error of 1 indicates an exact prediction, and a fold error of 2 indicates a twofold over- or underprediction. Colored text indicates the median fold error for each dataset. Across all panels, each point represents one species.

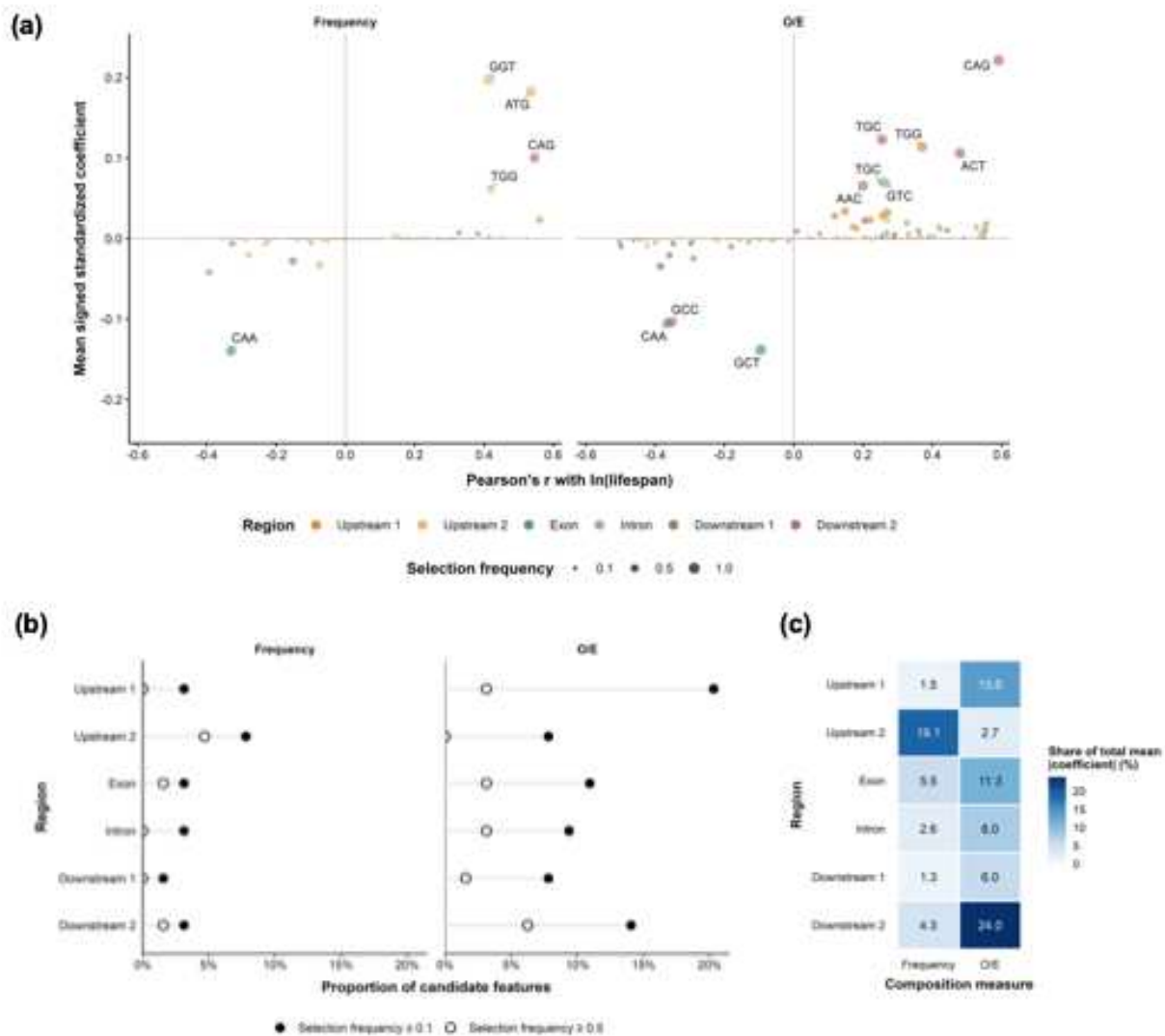


Figure 3. Feature selection and coefficient patterns in the final trinucleotide model.

Feature selection was summarized across the 1,000 outer models using the 327 development species; fixed-holdout outcomes were not used. The model included trinucleotide frequencies (Frequency) and observed-to-expected ratios (O/E) from six BUSCO gene-associated regions: two upstream regions, exons, introns and two downstream regions. Selection frequency was the proportion of models in which a predictor had a nonzero coefficient. Coefficient means were calculated across all 1,000 models, including zeros for unselected predictors, and coefficients were standardized to represent a one-standard-deviation increase in the predictor. **(a)** Relationship between each predictor's Pearson correlation with $\ln(\text{lifespan})$ and its mean signed standardized coefficient. Point color indicates region, and point size indicates selection frequency. Labels identify predictors that accounted for at least 2% of total coefficient magnitude. **(b)** Proportions of the 64 trinucleotides selected in at least 10% (filled points) or 50% (open points) of the 1,000 models for each region and composition measure. **(c)** Coefficient-magnitude share for each genomic region and composition measure, calculated by summing the mean absolute standardized coefficients and dividing by the total across all 768 predictors.

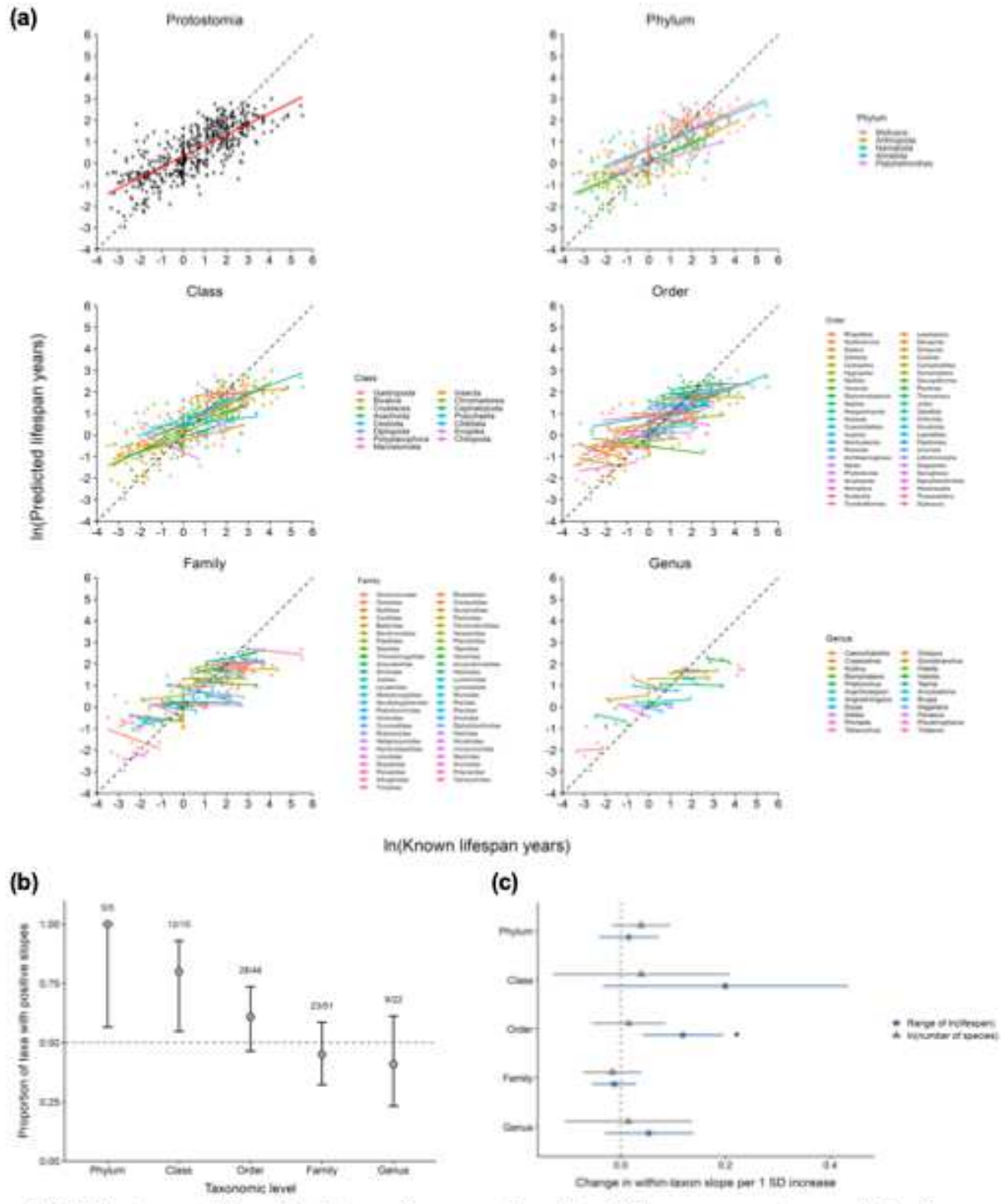


Figure 4. Within-taxon relationships between known and predicted lifespans across taxonomic levels.

807 Within-taxon slopes were estimated by regressing predicted lifespan on known lifespan for
808 taxonomic groups represented by at least three species. Predictions were cross-fitted for
809 development species and averaged across 1,000 outer models for fixed-holdout
810 species. (a) Scatterplots of known and predicted lifespans for six taxonomic levels. The
811 Protostomia panel includes all 466 species; the remaining panels include only taxa represented
812 by at least three species, with points and fitted regression lines colored by taxon. The dashed line
813 indicates equality between known and predicted lifespan. (b) Proportion of taxa with a positive
814 taxon-specific slope. Points show the observed proportions, vertical lines show Wilson 95%
815 confidence intervals, and labels give the number of positive slopes relative to the total number of
816 taxa analyzed. The dashed line at 0.5 indicates equal proportions of positive and non-positive
817 slopes. (c) Associations of the within-taxon range of known $\ln(\text{lifespan})$ and $\ln(\text{number of}$
818 $\text{species})$ with the within-taxon slope. Points show the estimated change in slope per one-
819 standard-deviation increase in each variable, and horizontal lines show taxon-clustered CR2 95%
820 confidence intervals. Positive values indicate that larger lifespan ranges or more species are
821 associated with steeper within-taxon slopes. The dashed vertical line indicates no association. An
822 asterisk indicates FDR-adjusted $p < 0.05$, with adjustment performed separately for each variable
823 across the five taxonomic ranks.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1Suppltables.xlsx](#)
- [Additionalfile2Supplfigs.pdf](#)