A new codon adaptation metric predicts vertebrate body size and tendency to protein disorder

Catherine A. Weibel[1,2,3], Andrew L. Wheeler[4], Jennifer E. James[5,6], Sara M. Willis[5,7], Joanna Masel[5]

[1]Department of Mathematics, University of Arizona, Tucson, Arizona 85721, USA

[2]Department of Physics, University of Arizona, Tucson, Arizona 85721, USA

[3]present address: Department of Applied Physics, Stanford University, California, USA

[4]Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, Arizona 85721, USA

[5]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

[6]present address: Department of Ecology and Genetics, Evolutionary Biology Center, Uppsala University, Sweden

[7]present address: University Information Technology Services, University of Arizona, Tucson, Arizona 85721, USA

Corresponding author: Joanna Masel

Email: masel@arizona.edu

Abstract

The nearly neutral theory of molecular evolution posits variation among species in the effectiveness of selection. In an idealized model, the census population size determines both this minimum magnitude of the selection coefficient required for deleterious variants to be reliably purged, and the amount of neutral diversity. Empirically, an "effective population size" is often estimated from the amount of putatively neutral genetic diversity, and is assumed to also capture a species' effectiveness of selection. The degree to which selection maintains preferred codons has the potential to more directly quantify the effectiveness of selection. However, past metrics that compare codon bias across species are confounded by among-species variation in %GC content and/or amino acid composition. Here we propose a new Codon Adaptation Index of Species (CAIS) that corrects for both confounders. Unlike previous metrics of codon bias, CAIS yields the expected relationship with adult vertebrate body mass. We demonstrate the use of CAIS correlations to show that the protein domains of more highly adapted vertebrate species evolve higher intrinsic structural disorder.

Introduction

Species differ from each other in many ways, including mating system, ploidy, spatial distribution, life history, size, lifespan, and population size. These differences make the process of purifying selection more efficient in some species than others. Our understanding of both the causes and consequences of these differences is limited in part by a reliable metric with which to measure them. In the long-term, the probability that a gene is fixed for one allele rather than another allele is given by the ratio of fixation and counter-fixation probabilities (Bulmer 1991). In an idealized population of constant population size and no selection at linked sites, a mutation-selection-drift model describes how this ratio of fixation probabilities depends on the census population size *N* (Kimura 1962), and hence gives the fraction of sites expected to be found in preferred vs. non-preferred states (Figure 1).

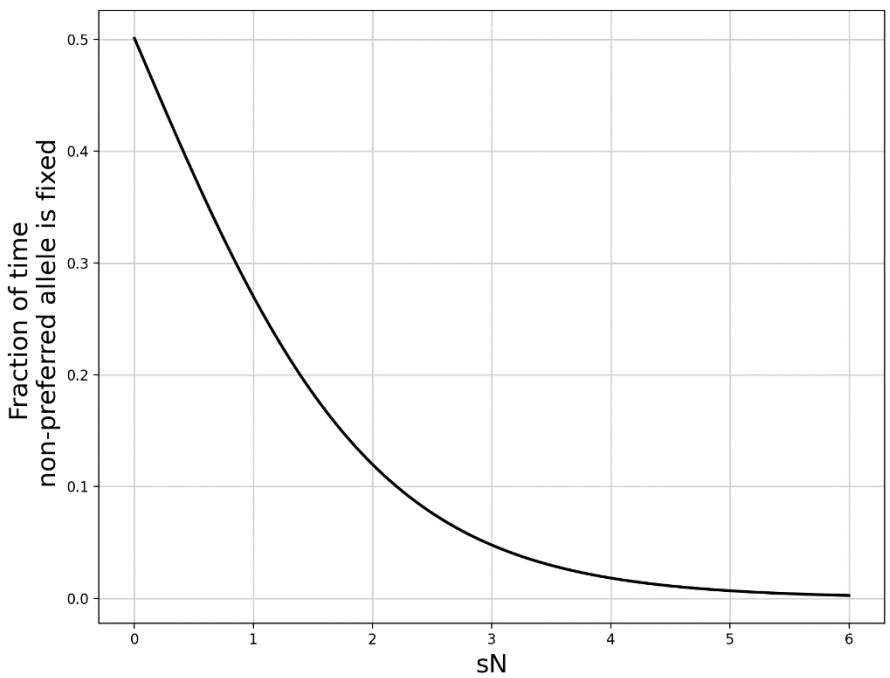


Figure 1: The effectiveness of selection, calculated as the long-term ratio of time spent in fixed deleterious: fixed beneficial allele states given symmetric mutation rates, is a function of the product *Ns*. Assuming a diploid Wright-Fisher population with s <<1, the probability of fixation of a new mutation $\pi(N,s) = \frac{1-e^{-\frac{s}{2}}}{1-e^{-Ns}}$, and the y-axis is calculated as $\pi(N,-s)/(\pi(N,-s)+\pi(N,s))$. *s* is held constant at a value of 0.001 and *N* is varied. Results for other small magnitude values of *s* are superimposable.

This reasoning has been extended to real populations by positing that species have an "effective" population size, N_e (Ohta 1973). N_e is the census population size of an idealized population that reproduces a property of interest in the focal population. N_e is therefore not a single quantity per population, but instead depends on which property is of interest.

The amount of neutral polymorphism is the usual property used to empirically estimate N_e (Charlesworth 2009; Doyle et al. 2015; Lynch et al. 2016). However, the property of most relevance to nearly neutral theory is instead the inflection point *s* at which non-preferred alleles become common enough to matter (Figure 1), and hence the degree to which highly exquisite adaptation can be maintained in the face of ongoing mutation and genetic drift (Kimura 1962; Ohta 1972; Ohta 1992). While genetic diversity has been found to reflect some aspects of life history strategy (Romiguier et al. 2014), there remain concerns about whether neutral genetic diversity and the limits to weak selection always remain closely coupled. One source of concern is that the classic one-locus models that underpin N_e calculations assume that neutral diversity is shaped by genetic drift, but it can instead be shaped instead by some combination of background selection and hitchhiking (Masel 2011; Kern & Hahn 2018).

As a practical matter, N_e is usually calculated by taking some measure of the amount of putatively neutral (often synonymous) polymorphism segregating in a population, and dividing by that species' mutation rate (Charlesworth 2009). As a result, N_e values are only available for species that have both polymorphism data and accurate mutation rate estimates, limiting its use. Worse, N_e is not a robust statistic. In the absence of a clear species definition, polymorphism is sometimes calculated across too broad a range of genomes, substantially inflating N_e (Daubin & Moran 2004). Similarly, a recent bottleneck or poor sampling scheme will deflate estimated values of N_e, even if the degree of fine-tuned adaptation remains high. Transient hypermutation (Plotkin et al. 2006), which is common in microbes, causes further short-term inconsistencies in polymorphism levels.

An alternative approach to measure the efficiency of selection exploits codon usage bias, which is influenced by weak selection for factors such as translational speed and accuracy (Hershberg

& Petrov 2008; Plotkin & Kudla 2011; Hunt et al. 2014). The degree of bias in synonymous codon usage that is driven by selective preference offers a more direct way to assess how effective selection is at the molecular level in a given species (Li 1987; Bulmer 1991; Akashi 1996; Subramanian 2008). Conveniently, it can be estimated from only a single genome, i.e., without polymorphism or mutation rate data for that species.

One commonly used metric, the Codon Adaptation Index (CAI) (Sharp & Li 1987; Sharp et al. 2010) takes the average of Relative Synonymous Codon Usage (RSCU) scores, which quantify how often a codon is used, relative to the codon that is most frequently used to encode that amino acid in that species. While this works well for comparing genes within the same species, it unfortunately means that the species-wide strength of codon bias appears in the normalizing denominator (see equation 4 below and Supplementary Figure 1A). Paradoxically, this can make more exquisitely adapted species have lower rather than higher species-averaged CAI scores (as we will later find in Figure 3B). A handful of publications (Jansen et al. 2003; Botzman & Margalit 2011; LaBella et al. 2019) have inappropriately used mean CAI as a metric of codon adaptation across species, presumably generating inverse results.

To compare species using CAI, it has been suggested that instead of taking a genome-wide average, one should consider a set of highly expressed reference genes (Sharp et al. 2005; Vicario et al. 2007; dos Reis & Wernisch 2008; Subramanian 2008). This approach assumes that the relative strength of selection on those reference genes (often a function of gene expression) remains approximately constant across the set of species considered (red boxes in Figure 2 indicating same quartiles). Its use also requires careful attention to the length of reference genes (Urrutia & Hurst 2001; Doherty & McInerney 2013).

If more highly adapted species have a higher proportion of their genes subject to effective selection on codon bias (Galtier et al. 2018), this is best detected using a proteome-wide approach (Figure 2). Averaging across the entire proteome provides robustness to shifts in the expression level of or strength of selection on particular genes. The proteome-wide average depends on the fraction of sites whose selection coefficients exceed the “drift barrier” for that particular species (Figure 2, blue threshold).

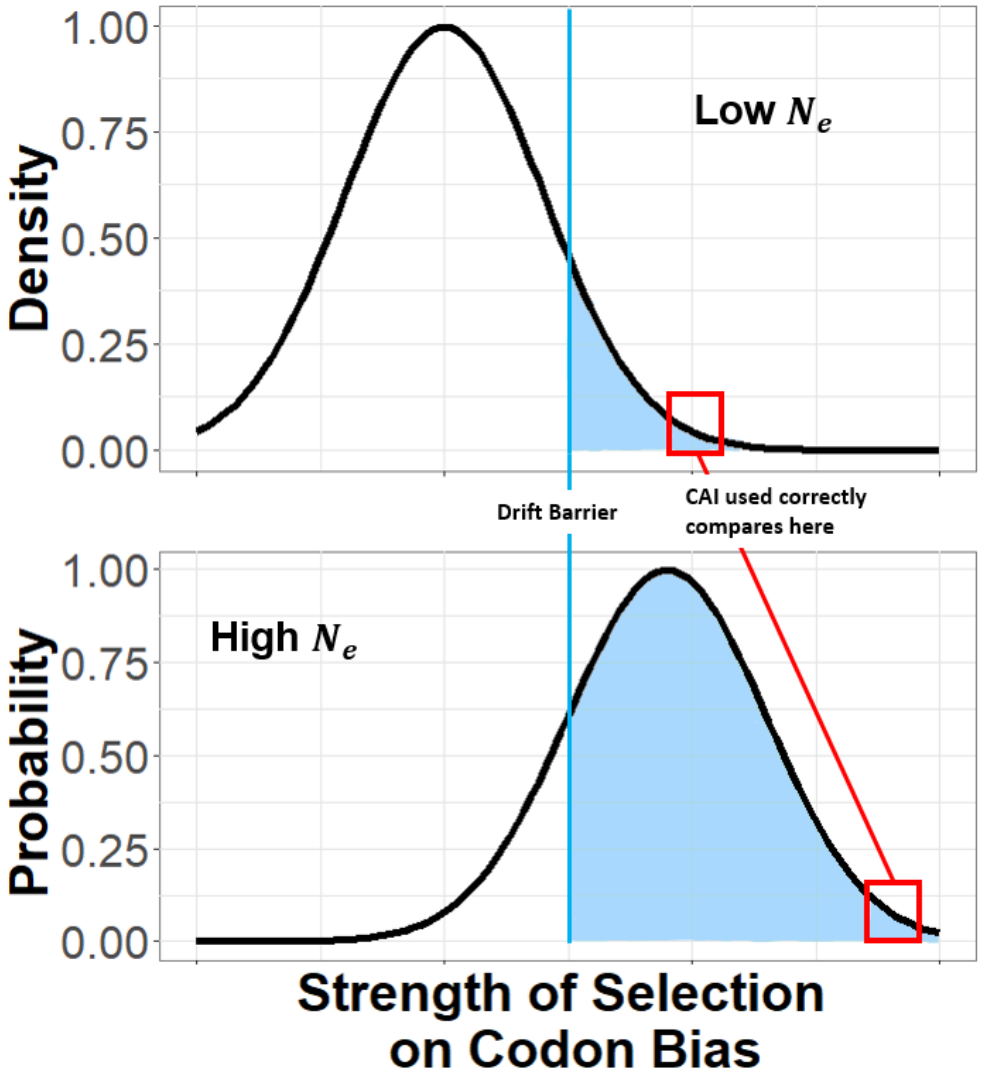


Figure 2: More highly adapted species (bottom) have a higher proportion of their genes subject to effective selection on codon bias (blue area). The CAI, when used correctly, compares the intensity of selection in comparable regions of the right tail (red boxes). Our new metric captures both this, and differences in the proportion of the genome subject to substantial selection.

However, any whole-proteome comparisons must be corrected for differences in GC content among species. Genomic GC content is a major driver of codon usage differences among species. Species differ in their mutational bias with respect to GC and in their frequency of GC-biased gene conversion (Urrutia & Hurst 2001; Duret & Galtier 2009; Doherty & McInerney 2013; Figuet et al. 2014). Here, we control for both by simply replacing RSCU scores with the ratio of how frequently a codon is used to encode a particular amino acid relative to how

117 frequently we expect it to be used given the %GC content. There are also GC-related
118 differences in which codons are preferred (Hershberg & Petrov 2009); we expect our method to
119 be more robust to these than CAI is.

120 An alternative metric, the Effective Number of Codons (ENC) originally quantified how far the
121 codon usage of a sequence departs from equal usage of synonymous codons (Wright 1990),
122 creating a complex relationship with GC content (Fuglsang 2008). The ENC was later modified to
123 correct for GC content, in a similar manner to the new method we present here (Novembre
124 2002). However, a remaining issue with this modified ENC is that differences among species in
125 amino acid composition might act as a confounding factor, even after controlling for GC
126 content. Specifically, species that make more use of an amino acid for which there is stronger
127 selection among codons (which is sometimes the case (Vicario et al. 2007)) would have higher
128 codon bias, even if each amino acid, considered on its own, had identical codon bias
129 irrespective of which species it is in. Neither ENC (Fuglsang 2004; Fuglsang 2008) nor the CAI
130 (Sharp & Li 1987) adequately control for differences in amino acid composition when applied
131 across species. Despite early claims to the contrary (Wright 1990), this problem is not easy to fix
132 for ENC (Fuglsang 2004; Fuglsang 2008).

133 Here, we extend the CAI so that it corrects for both GC and amino acid composition (see
134 Methods) to create a new Codon Adaptation Index of Species (CAIS), and compare its
135 performance to that of the ENC. The availability of complete genomes allows both metrics to be
136 readily calculated without data on polymorphism or mutation rate, without selecting reference
137 genes, and without concerns about demographic history. Our purpose is to find an accessible
138 metric that can quantify the limits to weak selection important to nearly neutral theory; this
139 differs from past evaluations focused on comparing different genes of the same species and
140 recapitulating "ground truth" simulations thereof (Sun et al. 2012; Zhang et al. 2012; Liu et al.
141 2018). Our first measure of success is to produce the expected correlation with body size (used
142 here as a proxy for census population size and hence the expected exquisiteness of adaptation).
143 We use vertebrates as a challenging test case, given relatively weak selection on codon usage
144 (Doherty & McInerney 2013; Galtier et al. 2018). Correlations between codon adaptation and
145 body size across vertebrates have previously been elusive, which is surprising given that the

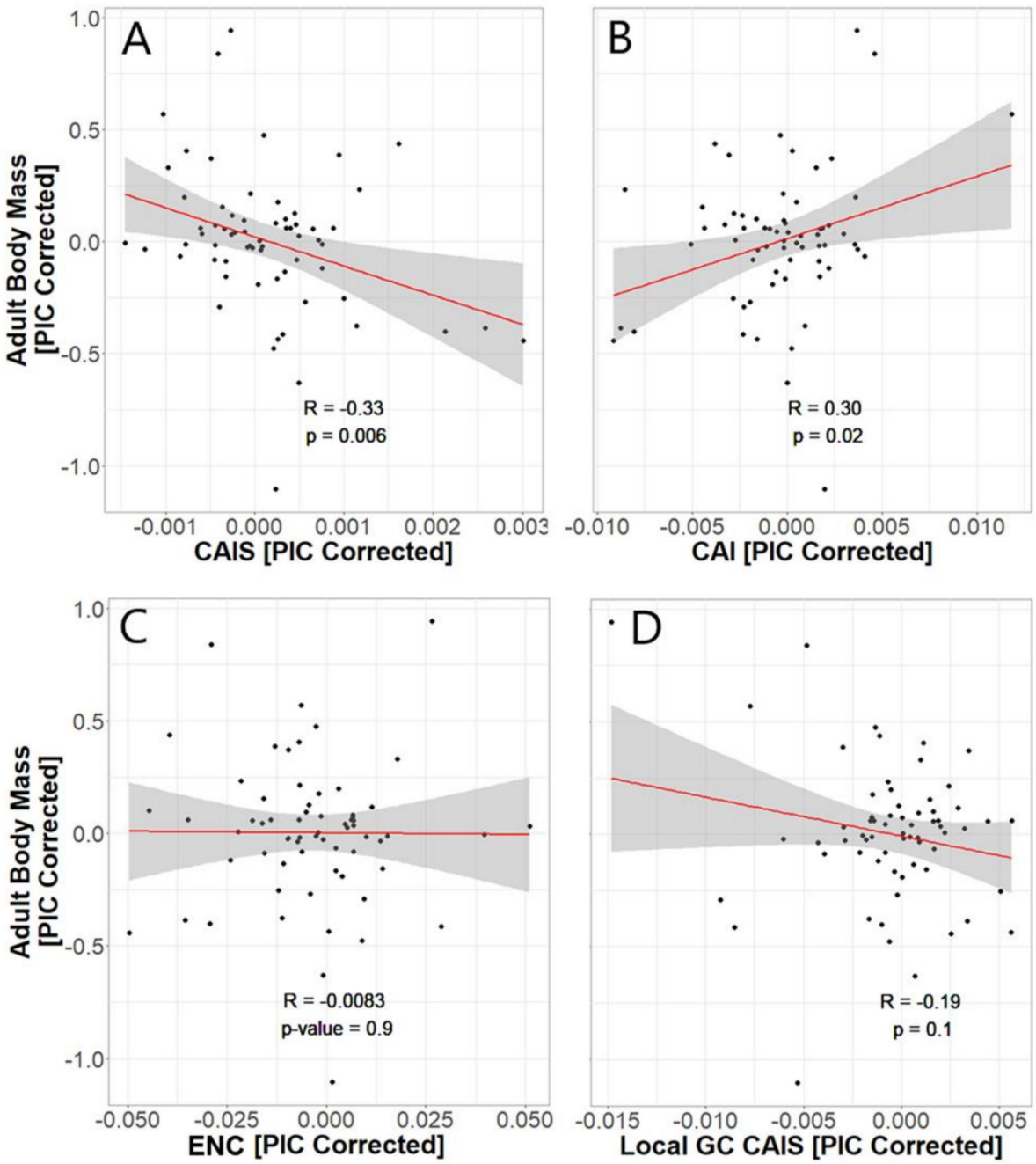
146 expected differences in the efficiency of selection between X-chromosomes vs. autosomes and
147 between high-recombination vs. low recombination regions are found (Kessler & Dean 2014).
148 Our second measure of success is a demonstration of how our metric can be used to find novel
149 correlations pointing to what else might be subject to weak selective preferences at the
150 molecular level.

151 Results

152 Species with larger body mass have smaller CAIS

153 CAI scores averaged across the proteome of each species would paradoxically suggest that
154 larger species have more effective selection (Figure 3B; note that we control for phylogenetic
155 non-independence using Phylogenetic Independent Contrasts (PIC) (Felsenstein 1985)). This is
156 because the behavior of the CAI is driven by the normalization term on its denominator, rather
157 than by its numerator (Supplementary Figure 1A). This problem is removed in our improved
158 CAIS measure (Supplementary Figure 1B), which yields the expected result that species with
159 more codon adaptation according to the CAIS tend to be smaller (Figure 3A).

160 In contrast, the ENC is independent of adult vertebrate body mass (Figure 3C), consistent with
161 past reports that ENC does not predict body mass or generation time in mammals (Kessler &
162 Dean 2014) and that genome-wide ENC does not predict generation time in eukaryotes more
163 broadly (Subramanian 2008). This difference between ENC and CAIS is surprising because the
164 two methods are conceptually similar (see Methods). One difference is that CAIS corrects for
165 amino acid composition differences across the dataset while ENC does not (see Methods).
166 However, when we remove the amino acid composition correction from CAIS, we retain the
167 relationship that smaller-bodied species tend to have more codon adaptation (Supplementary
168 Figure 2), ruling this out as the cause of their different behaviors.



169

170 Figure 3: CAIS reflects the expected relationship between effectiveness of selection and body size, while
171 CAI, ENC, and CAIS calculated using local GC content do not. Note that a high ENC value means more
172 codons are being used in the genome of the given species, so that the given species is less codon
173 adapted, while a high CAIS value means that a species is more codon adapted. Body size data are from
174 PanTHERIA database, originally in $\log_{10}$(mass) in grams prior to PIC correction. Data are shown for 62
175 species in common between PanTHERIA and our own dataset of 118 vertebrate species that have both

176 “Complete” genome sequence available for calculating %GC, and TimeTree divergence dates available
177 for use in PIC correction. P-values shown are for Pearson’s correlation. Red line shows unweighted
178 lm(y~x) with grey region indicated 95% confidence interval.

179 Another difference between ENC and CAIS is that CAIS is linear in observed codon frequencies
180 (equation 7), while ENC has a quadratic term (equation 14). The quadratic term in ENC
181 magnifies the differences in more extreme deviations of observed frequencies from the GC-
182 informed expected frequencies than CAIS. We speculate that this might somehow swamp more
183 informative differences. Other ENC-like approaches that use a quadratic term have also
184 struggled to reliably detect selection (Urrutia & Hurst 2001; Kessler & Dean 2014).

185 Surprisingly, we achieved a successful version of CAIS by using genome-wide GC content to
186 calculate expected codon use. Different parts of the genome have different GC contents
187 (Bernardi 2000; Eyre-Walker & Hurst 2001; Lander et al. 2001), primarily because the extent to
188 which GC-biased gene conversion increases GC content depends on the local rate of
189 recombination (Galtier et al. 2001; Meunier & Duret 2004; Duret et al. 2006; Duret & Galtier
190 2009). We therefore also calculated a version of CAIS whose codon frequency expectations are
191 based on local intergenic GC content. This performed worse (Fig. 3D) than our simple use of
192 genome-wide GC content (Fig. 3A) with respect to the strength of correlation between CAIS and
193 body size. We therefore use the simpler, genome-wide metric from here on. It is possible that
194 local GC content evolves more rapidly than codon usage, and that genome-wide GC serves as
195 an appropriately time-averaged proxy. It is also possible that the local non-coding sequences
196 we used were too short (at 3000bp or more), creating excessive noise that obscured the signal.
197 Given the body size results, we advocate for the use of CAIS as a metric of species’ effectiveness
198 of selection.

199 Proteins in better adapted species evolve more structural disorder

200 As an example of how correlations with CAIS can be used to identify weak selective
201 preferences, we investigate protein intrinsic structural disorder (ISD). Disordered proteins are
202 more likely to be harmful when overexpressed (Vavouri et al. 2009), and ISD is more abundant

203 in eukaryotic than prokaryotic proteins (Schad et al. 2011; Xue et al. 2012; Basile et al. 2019),
204 suggesting that low ISD might be favored by more effective selection.

205 However, compositional differences among proteomes might be driven by the recent birth of
206 ISD-rich proteins in animals (James et al. 2021), and/or by differences among sequences in their
207 subsequent tendency to proliferate into many different genes (James et al. 2022), rather than
208 by differences in how a given protein sequence evolves as a function of the effectiveness of
209 selection. To focus only on the latter, we use a linear mixed model, with each species having a
210 fixed effect on ISD, while controlling for Pfam domain identity as a random effect. We note that
211 once GC is controlled for, codon adaptation can be assessed similarly in intrinsically disordered
212 vs. ordered proteins (Gossmann et al. 2012). Controlling for Pfam identity is supported, with
213 standard deviation in ISD among Pfams estimated as 0.178 in comparison to residual standard
214 deviation of 0.058 and a p-value on the significance of the Pfam random effect term of 3×10^{-13}.
215 We then ask whether the fixed species effects on ISD are correlated with CAIS.

216 Surprisingly, more exquisitely adapted species have more disordered protein domains (Figure
217 4). Results are similar using ENC instead of CAIS (Supplementary Figure 3), despite the
218 weakness of ENC in predicting body size; the correlation coefficient drops only slightly from
219 0.54 to 0.5, and the p-value rises by less than 2 orders of magnitude.

220

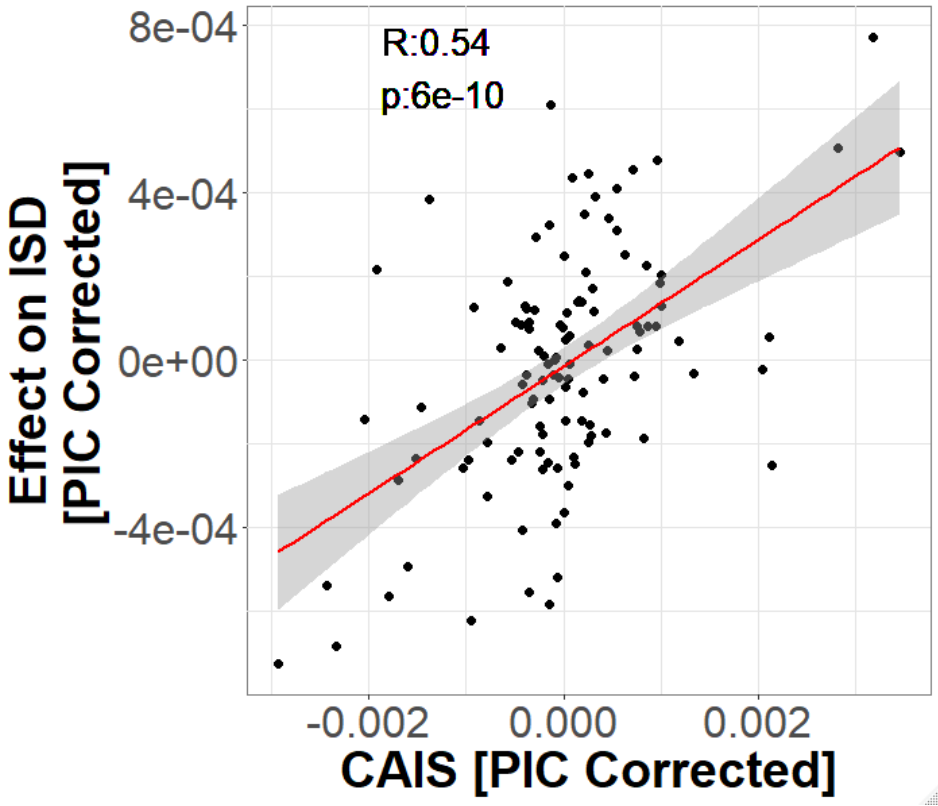


Figure 4: Protein domains have higher ISD when found in more exquisitely adapted species. Each datapoint is one of 118 vertebrate species with "complete" intergenic genomic sequence available (allowing for %GC correction) and TimeTree divergence dates (allowing for PIC correction). "Effects" on ISD shown on the y-axis are fixed effects of species identity in our linear mixed model, after PIC correction. Red line shows unweighted lm(y~x) with grey region as 95% confidence interval.

Younger animal-specific protein domains have higher ISD (James et al. 2021). We therefore hypothesize that selection in favor of high ISD might be strongest in young domains, which might use more primitive methods to avoid aggregation (Foy et al. 2019; Bertram & Masel 2020). To test this, we analyze two subsets of our data: those that emerged prior to the last eukaryotic common ancestor (LECA), here referred to as "old" protein domains, and "young" protein domains that emerged after the divergence of animals and fungi from plants. Young and old domains both show a trend of increasing disorder with species' adaptedness (Figure 5). Because PIC analysis makes the units incommensurable, we quantitatively compare the slopes of non-PIC-corrected weighted regressions. As predicted, the slope is stronger among young protein domains than among old domains (0.223 +/- 0.021 versus 0.161 +/- 0.014, respectively;

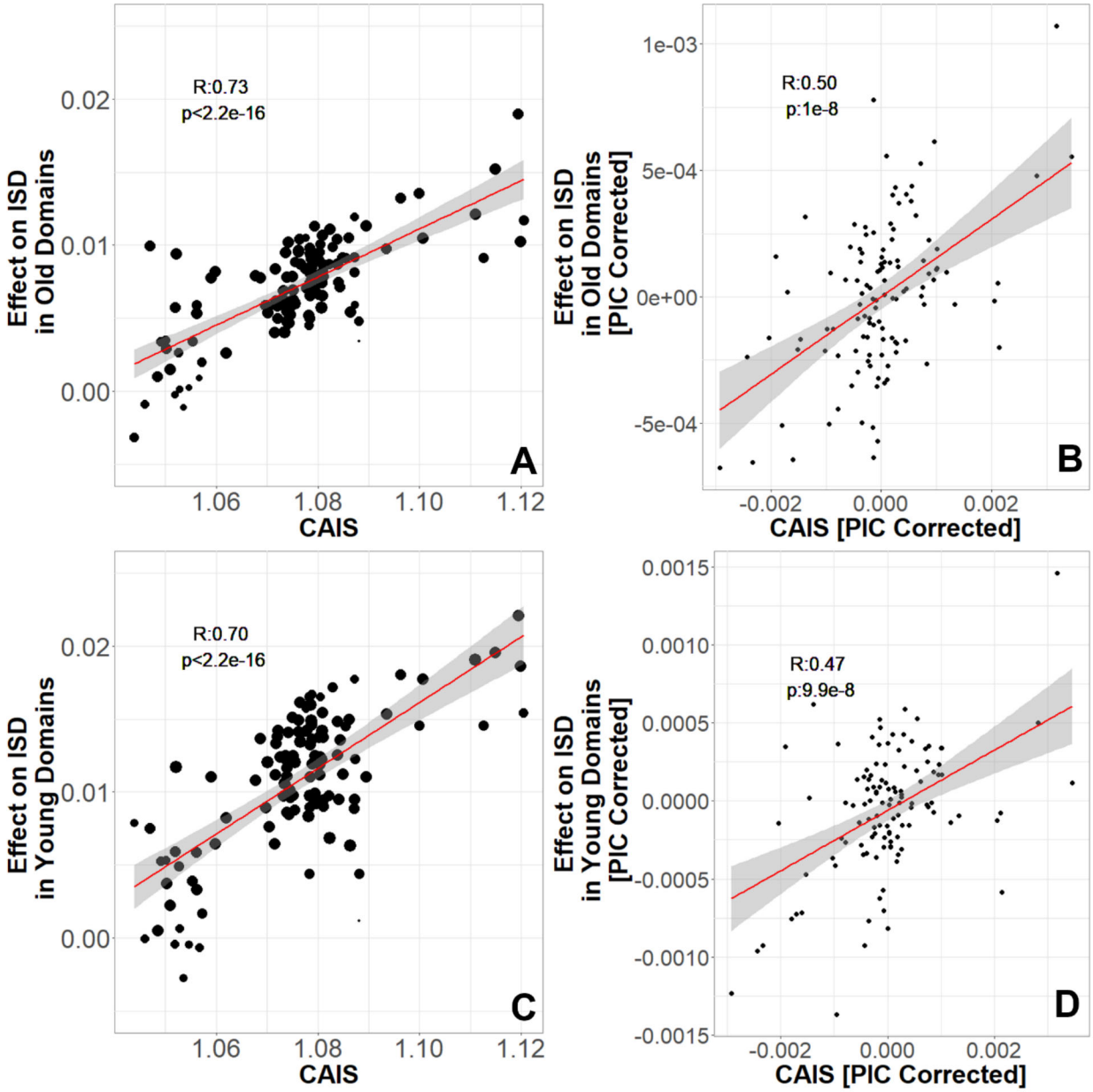


Figure 5: More exquisitely adapted species have higher ISD in both ancient and recent protein domains. Age assignments are taken from James et al. (2021), with vertebrate protein domains that emerged prior to LECA classified as "old", and vertebrate protein domains that emerged after the divergence of animals and fungi from plants as "young". "Effects" on ISD shown on the y-axis are fixed effects of species identity in our linear mixed model. The same n=118 datapoints are shown as in Figures 3 and 4. P-values shown are for Pearson's correlation. Red line shows lm(y~x), with grey region as 95% confidence interval, using a weighted model for non-PIC-corrected figures and unweighted for PIC-corrected figures. Weighted models make the quantitative estimation of slopes more accurate. Dot size on left indicates weight.

247 units of proportion ISD/ CAIS), although we are struck by the degree to which even ancient
248 domains prefer higher ISD. We find similar relationships of modestly lower strength and
249 statistical significance using ENC (Supplementary Figure 4).

250 Discussion

251 Here we propose CAIS as a new metric to quantify how species differ in the effectiveness of
252 selection. CAIS corrects codon bias both for total genomic GC content and for amino acid
253 composition, to extract a measure of codon adaptation. Surprisingly, we obtained better results
254 correcting for total GC content than for local GC content in the vicinity of each coding region.
255 Unlike the ENC, CAIS is sensitive enough to show the expected relationship with adult
256 vertebrate body mass.

257 CAIS can be used to detect which properties are under selection. To accomplish this, we note
258 that when different properties are each causally affected by a species' exquisiteness of
259 adaptation, this will create a correlation between the properties. We use codon adaptation as a
260 reference property, such that correlations with codon adaptation indicate selection. We
261 operationalize this using a linear mixed model approach that controls for Pfam identity as a
262 random effect, and then apply phylogenetic correction to the resulting fixed effects of species
263 on ISD. This approach shows that the same Pfam domain tends to be more disordered when
264 found in a well-adapted species (i.e. a species with a higher CAIS). This is true for both ancient
265 and recently emerged protein domains, albeit of slightly larger magnitude for the latter.

266 It is important that no additional variable such as GC content creates a spurious correlation by
267 affecting both CAIS and our property of interest. For this reason, we control for GC content
268 during the construction of CAIS. GC content is the product of many different processes, most
269 notably the genome-wide forces of mutation bias and gene conversion (Romiguier & Roux
270 2017), but potentially also selection on individual nucleotide substitutions that is hypothesized
271 to favor higher %GC (Long et al. 2018). By controlling for %GC, we exclude all these forces from
272 influencing CAIS. CAIS thus captures the extent of adaptation in codon bias, including
273 translational speed, accuracy, and any intrinsic preference for GC over AT that is specific to
274 coding regions. These remaining codon-adaptive factors do not create a statistically significant

275 correlation between CAIS and GC (Supplementary Figures 5A, 5B), nor between ENC and GC
276 (Supplementary Figures 5C, 5D), although CAI is correlated with GC (Supplementary Figures 5E,
277 F).

278 A direct effect of ISD on fitness agrees with studies of random ORFs in *Escherichia coli*, where
279 fitness was driven more by amino acid composition than %GC content, after controlling for the
280 intrinsic correlation between the two (Kosinski et al. 2022). However, we have not ruled out a
281 role for selection for higher %GC in ways that are general rather than restricted to coding
282 regions, whether in shaping mutational biases (Smith & Eyre-Walker 2001; Hershberg & Petrov
283 2009; Hildebrand et al. 2010; Novoa et al. 2019; Forcelloni & Giansanti 2020) or the extent of
284 gene conversion, or even at the single nucleotide level in a manner shared between coding
285 regions and intergenic regions (Long et al. 2018).

286 A more complex metric could control for more than just GC content and amino acid
287 frequencies. First vs. second vs. third codon positions have different nucleotide usage on
288 average, but while correcting for this might be useful for comparing genes (Zhang et al. 2012),
289 the issue in correcting for it while comparing species is that one might be removing the effect of
290 interest. Similarly, while it might be useful to control for dinucleotide and trinucleotide
291 frequencies (Brbić et al. 2015), to avoid circularity these would need to be taken from
292 intergenic sequences, with care needed to avoid influence from unannotated protein-coding
293 genes or even highly decayed pseudogenes.

294 Note that if a species were to experience a sudden reduction in population size, e.g. due to
295 habitat loss, leading to less effective selection, it would take some time for CAIS to adjust. CAIS
296 thus represents a relatively long-term historical pattern of adaptation. The timescales setting
297 neutral polymorphism based N_e estimates are likely shorter (Gossmann et al. 2012). It is
298 possible that the reason that we obtained better results when we control for genome-wide GC
299 content than when we control for local GC content is also that codon adaptation adjusts slowly
300 relative to the timescale of fluctuations in local GC content.

301 Here we developed a new metric of species adaptedness at the codon level, capable of
302 quantifying degrees of codon adaptation even among vertebrates. We chose vertebrates partly

303 due to the abundance of suitable data, and partly as a stringent test case, given past studies
304 finding limited evidence for codon adaptation (Kessler & Dean 2014). We restricted our analysis
305 to only the best annotated genomes, in part to ensure the quality of intergenic %GC estimates,
306 and in part limited by the feasibility of running linear mixed models with six million data points.
307 The phylogenetic tree is well resolved for vertebrate species, with an overrepresentation of
308 mammalian species. Despite the focus on vertebrates, we see a remarkably strong signal for a
309 subtle codon adaptation effect across closely related species, to the point where we can
310 comfortably detect the ISD signal across subsets of domains.

311 An alternative approach to measuring effective population size focuses on d_N/d_S, the ratio of
312 nonsynonymous to synonymous substitutions (Weber et al. 2014; Lefébure et al. 2017; Payne &
313 Alvarez-Ponce 2018; Lin et al. 2019; Luzuriaga-Neira & Alvarez-Ponce 2022). A potential issue
314 with this approach is that it is assumes that differences among species are driven by differences
315 in dN and not by differences in dS. The success of our codon adaptation metric suggests that
316 this assumption does not hold across animals, because the proportion of synonymous
317 substitutions that is effectively neutral varies across species. This may explain why d_N/d_S is not
318 consistently related to life history traits across different taxonomic groups (Figuet et al. 2016).

319 Our finding that weak selection prefers higher ISD was unexpected, given that lower-N_e
320 eukaryotes have more disordered proteins than higher-N_e prokaryotes (Ahrens et al. 2017;
321 Basile et al. 2019). However, this can be reconciled in a model in which highly disordered
322 sequences are less likely to be found in high-N_e species, but the sequences that are present
323 tend to have slightly higher disorder than their low-N_e homologs. High ISD might help mitigate
324 the trade-off between affinity and specificity (Dunker et al. 1998; Huang & Liu 2013; Lazar et al.
325 2022); non-specific interactions might be short-lived due to the high entropy associated with
326 disorder, which specific interactions are robust to.

327 Our new CAIS metric more directly quantifies how species vary in their exquisiteness of
328 adaptation than estimates of effective population size that are based on neutral polymorphism.
329 It can also be estimated for far more species because it does not require polymorphism or
330 mutation rate data, but only a single complete genome. We therefore expect CAIS to have

many uses as a new tool for exploring nearly neutral theory, of which our worked example of ISD could be just the start.

Methods

Species

Pfam sequences and IUPRED2 estimates of Intrinsic Structural Disorder (ISD) predictions were taken from James et al. (2021), who studied species marked as "Complete" in the GOLD database, with divergence dates available in TimeTree (Kumar et al. 2017). James et al. (2021) applied a variety of quality controls to exclude contaminants from the set of Pfams and assign accurate dates of Pfam emergence. Pfams that emerged prior to LECA are classified here as "old", and Pfams that emerged after the divergence of animals and fungi from plants are classified as "young", following annotation by James et al. (2021).

Codon Adaptation Index

Sharp and Li (1987) quantified codon bias through the Codon Adaptation Index (CAI), a normalized geometric mean of synonymous codon usage bias across sites, excluding stop and start codons. We modify this to calculate CAI including stop and start codons, because of documented preferences among stop codons in mammals (Wangen & Green 2020). While usually used to compare genes within a species, among-species comparisons can be made using a reference set of genes that are highly expressed in yeast genome (Sharp & Li 1987). Each codon i is assigned a Relative Synonymous Codon Usage value:

$$RSCU_i = \frac{N_i}{\frac{1}{n_a}\sum_{j=1}^{n_a} N_j} \tag{1}$$

where N_i denotes the number of times that codon i is used, and the denominator sums over all n_a codons that code for that specific amino acid. RSCU values are normalized to produce a relative adaptiveness values w_i for each codon, relative to the best adapted codon for that amino acid:

$$w_i \equiv \frac{RSCU_i}{RSCU_{max}} \tag{2}$$

Let L be the number of codons across all protein-coding sequences considered. Then

$$CAI = [\prod_{i=1}^{L} w_i]^{\frac{1}{L}} \quad (3)$$

To understand the effects of normalization, it is useful to rewrite this as:

$$CAI = \left[\prod_{i=1}^{L} \frac{RSCU_i}{RSCU_{max}}\right]^{\frac{1}{L}} = \frac{CAI_{raw}}{CAI_{max}} \quad (4)$$

where CAI_{raw} is the geometric mean of the "unnormalized" or observed synonymous codon usages, and CAI_{max} is the maximum possible CAI given the observed codon frequencies.

GC Content

We calculated total %GC content (intergenic and genic) during a scan of all six reading frames across genic and intergenic sequences available from NCBI with access dates between May and July 2019 (described in James et al. (2022)). Of the 170 vertebrates meeting the quality criteria of James et al. (2021), 118 had annotated intergenic sequences within NCBI, so we restricted the dataset further to keep only the 118 species for which total GC content was available.

Codon Adaptation Index of Species (CAIS)

Controlling for GC bias in Synonymous Codon Usage

Consider a sequence region r within species s where each nucleotide has an expected probability of being G or C = g_r. For our main analysis, we consider just one region r encompassing the entire genome of a species s. In a secondary analysis, we break the genome up and use local values of g_r in the non-coding regions within and surrounding a gene or set of overlapping genes. To annotate the boundaries of these local regions, we first selected 1500 base pairs flanking each side of every coding sequence identified by NCBI annotations. Coding sequence annotations are broken up according to exon by NCBI. When coding sequences of the same gene did not fall within 3000 base pairs of each other, they were treated as different regions. When two coding sequences, whether from the same gene or from different genes,

had overlapping 1500bp catchment areas, we merged them together. g_r was then calculated based on the non-coding sites within each region, including both genic regions such as promoters and non-genic regions such as introns and intergenic sequences.

With no bias between C vs. G, nor between A vs. T, nor patterns beyond the overall composition taken one nucleotide at a time, the expected probability of seeing codon i in a triplet within r is

$$p_{i,r} = {g_r}^{k_{GC}}(1 - g_r)^{k_{AT}} \tag{5}$$

where $k_{GC} + k_{AT} = 3$ total positions in codon i. The expected probability that amino acid a in region r is encoded by codon i is

$$E_{i,r} = \frac{p_{i,r}}{\sum_{j=1}^{n_a} p_{j,r}} \ . \tag{6}$$

The Relative Synonymous Codon Usage (RSCU) value used by the CAI measures the degree to which a codon's relative frequency differs from the null expectation that all synonymous codons are used equally. Using equations 5 and 6, we replace this by one or a set of Relative Synonymous Codon Usage of Species (RSCUS) values for each region:

$$RSCUS_{i,r,s} = \frac{O_{i,r}}{E_{i,r}} \tag{7}$$

where O_i is the observed frequency with which amino acid a is encoded by codon i within region r of species s.

Controlling for Amino Acid Composition

Some amino acids may be more intrinsically prone to codon bias. We want a metric that quantifies effectiveness of selection (not amino acid frequency), so we re-weight CAIS on the basis of a standardized amino acid composition, to remove the effect of variation among species in amino acid frequencies.

Let F_a be the frequency of amino acid a across the entire dataset of 118 vertebrate genomes, and $f_{i,s}$ the frequency of codon i in given species s. A candidate CAIS that controls for global GC content but not for amino acid composition can be written as

$$\prod_{i=1}^{64} RSCUS_{i,global,s}{}^{f_{i,s}} \tag{8}$$

We want to re-weight $f_{i,s}$ on the basis of F_a to ensure that differences in amino acid frequencies among species do not affect CAIS, while preserving relative codon frequencies for the same amino acid. We do this by solving for $\alpha_{a,s}$ so that

$$F_a = \alpha_{a,s} \sum_{j=1}^{n_a} f_{j,s}. \tag{9}$$

We then define $f'_{j,s} = \alpha_{a,s} f_{j,s}$ to obtain an amino acid frequency adjusted CAIS:

$$CAIS(s) = \prod_{i=1}^{64} RSCUS_{i,global,s}{}^{f'_{i,s}} \tag{10}$$

For convenient implementation in code, we used the following form:

$$CAIS(s) = e^{\sum_{i=1}^{64} f'_{i,s} \ln(RSCUS_{i,global,s})} \tag{11}$$

The F_a values for our species set are at https://github.com/MaselLab/Codon-Adaptation-Index-of-Species/blob/main/CAIS_ENC_calculation/Total_amino_acid_frequency_vertebrates.txt. Similarly, CAIS corrected for local intergenic GC content but not species-wide amino acid composition is

$$CAIS_{localGC}(s) = \left(\prod_{r=1}^{G} \prod_{i=1}^{64} RSCUS_{i,r,s}{}^{n_{i,r}}\right)^{1/L} \tag{12}$$

where $n_{i,r}$ is the number of times codon i appears in region r of species s, G is the number of regions, and $L = \sum_{r=1}^{G} \sum_{i=1}^{64} n_{i,r}$ is the total number of codons in the genome. Rewritten for greater computational ease:

$$CAIS_{localGC}(s) = e^{\frac{1}{L} \sum_{r=1}^{G} \sum_{i=1}^{64} n_{i,r} \ln(RSCUS_{i,r,s})}. \tag{13}$$

Novembre's Effective Number of Codons (ENC) controlled for total GC Content

Following from equations 5 and 6, the X_a^2 value representing the deviation of the frequencies of the codons for amino acid a from null expectations is

$$X_a^2 = \sum_{i=1}^{n_a} \frac{N_a(O_i - E_i)^2}{E_i} \tag{14}$$

where N_a is the total number of times that amino acid a appears. Novembre (2002) defines the corrected "F value" of amino acid a as

$$\widehat{F'}_a = \frac{X_a^2 + N_a - n_a}{n_a(N_a - 1)} \tag{15}$$

and the Effective Number of Codons as

$$ENC = 2 + \frac{9}{\widehat{F'}_2} + \frac{1}{\hat{F'}_3} + \frac{5}{\widehat{F'}_4} + \frac{3}{\widehat{F'}_6} \tag{16}$$

where each $\widehat{F'}_{n_a}$ is the average of the "F values" for amino acids with n_a synonymous codons. Past measures of ENC do not contain stop or start codons (Wright 1990; Novembre 2002; Fuglsang 2004), but as we did for CAI and CAIS above, we include stop codons as an "amino acid" and therefore amend (19) to

$$ENC = 2 + \frac{9}{\widehat{F'}_2} + \frac{2}{\widehat{F'}_3} + \frac{5}{\widehat{F'}_4} + \frac{3}{\widehat{F'}_6} \tag{17}$$

Statistical Analysis

All statistical modelling was done in R 3.5.1. Scripts for calculating CAI and CAIS were written in Python 3.7.

Phylogenetic Independent Contrasts

Spurious phylogenetically confounded correlations can occur when closely related species share similar values of both metrics. One danger of such pseudoreplication is Simpson's paradox, where there are negative slopes within taxonomic groups and a positive slope among them might combine to yield an overall positive slope. We avoid pseudoreplication by using Phylogenetic Independent Contrasts (PIC) (Felsenstein 1985) to assess correlation. PIC analysis was done using the R package "ape" (Paradis & Schliep 2019).

Data Availability

454 All data and code underlying this article are available in the public repository at

455 https://github.com/MaselLab/Codon-Adaptation-Index-of-Species.

456 Acknowledgements

457 This work was supported by the National Institutes of Health (R01 GM104040 and T32

458 GM132008), the John Templeton Foundation (60814), the Arnold and Mabel Beckman

459 Foundation Scholars Program, the Western Alliance to Expand Student Opportunities (WAESO)

460 Louis Stokes Alliance for Minority Participation (LSAMP) National Science Foundation (NSF)

461 Cooperative Agreement (HRD-1101728), and UA/NASA Space Grant Undergraduate Research

462 Internship program. We thank Luke Kosinski, Hanon McShea, David Liberles, and Sawsan Wehbi

463 for helpful discussions, Paul Nelson for providing the genome-wide GC contents, and the

464 University of Arizona Undergraduate Biology Research Program for training.

465 Literature Cited

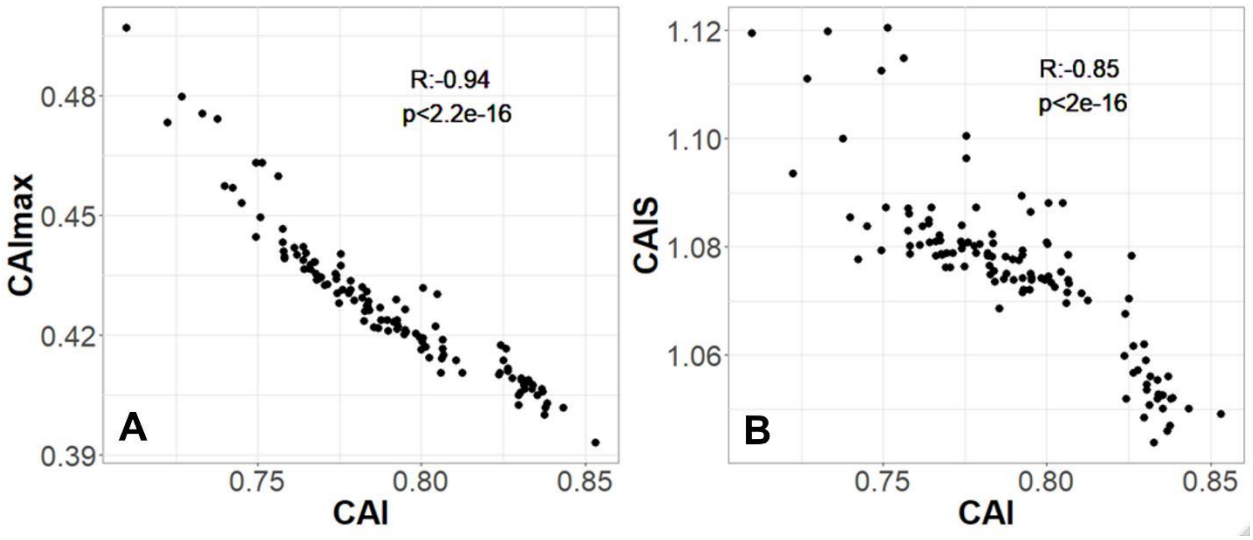
466 Ahrens, JB, Nunez-Castilla, J, and Siltberg-Liberles, J. 2017. Evolution of intrinsic disorder in eukaryotic
467 proteins. *Cellular and Molecular Life Sciences* 74:3163-3174.
468 Akashi, H. 1996. Molecular Evolution Between *Drosophila melanogaster* and *D. simulans* Reduced Codon
469 Bias, Faster Rates of Amino Acid Substitution, and Larger Proteins in *D. melanogaster*. *Genetics*
470 144:1297-1307.
471 Basile, W, Salvatore, M, Bassot, C, and Elofsson, A. 2019. Why do eukaryotic proteins contain more
472 intrinsically disordered regions? *PLoS Computational Biology* 15:e1007186.
473 Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3-17.
474 Bertram, J, and Masel, J. 2020. Evolution rapidly optimizes stability and aggregation in lattice proteins
475 despite pervasive landscape valleys and mazes. *Genetics* 214:1047-1057.
476 Botzman, M, and Margalit, H. 2011. Variation in global codon usage bias among prokaryotic organisms is
477 associated with their lifestyles. *Genome Biology* 12:R109.
478 Brbić, M, Warnecke, T, Kriško, A, and Supek, F. 2015. Global Shifts in Genome and Proteome
479 Composition Are Very Tightly Coupled. *Genome Biology and Evolution* 7:1519-1532.
480 Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897-
481 907.
482 Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation.
483 *Nature Reviews Genetics* 10:195-205.
484 Daubin, V, and Moran, NA. 2004. Comment on "The Origins of Genome Complexity". *Science* 306:978-
485 978.
486 Doherty, A, and McInerney, JO. 2013. Translational Selection Frequently Overcomes Genetic Drift in
487 Shaping Synonymous Codon Usage Patterns in Vertebrates. *Molecular Biology and Evolution*
488 30:2263-2267.

489 dos Reis, M, and Wernisch, L. 2008. Estimating Translational Selection in Eukaryotic Genomes. *Molecular*
490 *Biology and Evolution* 26:451-461.
491 Doyle, JM, Hacking, CC, Willoughby, JR, Sundaram, M, and DeWoody, JA. 2015. Mammalian genetic
492 diversity as a function of habitat, body size, trophic class, and conservation status. *Journal of*
493 *Mammalogy* 96:564-572.
494 Dunker, AK, Garner, E, Guilliot, S, Romero, P, Albrecht, K, Hart, J, Obradovic, Z, Kissinger, C, and
495 Villafranca, JE. 1998. *Protein disorder and the evolution of molecular recognition: theory,*
496 *predictions and observations*. Pp. 473-484. Pac Symp Biocomput.
497 Duret, L, Eyre-Walker, A, and Galtier, N. 2006. A new perspective on isochore evolution. *Gene* 385:71-74.
498 Duret, L, and Galtier, N. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic
499 Landscapes. *Annual Review of Genomics and Human Genetics* 10:285-311.
500 Eyre-Walker, A, and Hurst, LD. 2001. The evolution of isochores. *Nature Reviews Genetics* 2:549-555.
501 Felsenstein, J. 1985. Phylogenies and the Comparative Method. *The American Naturalist* 125:1-15.
502 Figuet, E, Ballenghien, M, Romiguier, J, and Galtier, N. 2014. Biased Gene Conversion and GC-Content
503 Evolution in the Coding Sequences of Reptiles and Vertebrates. *Genome Biology and Evolution*
504 7:240-250.
505 Figuet, E, Nabholz, B, Bonneau, M, Mas Carrio, E, Nadachowska-Brzyska, K, Ellegren, H, and Galtier, N.
506 2016. Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes.
507 *Molecular Biology and Evolution* 33:1517-1527.
508 Forcelloni, S, and Giansanti, A. 2020. Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic
509 Disorder in the Human Proteome. *Journal of Molecular Evolution* 88:164-178.
510 Foy, SG, Wilson, BA, Bertram, J, Cordes, MHJ, and Masel, J. 2019. A Shift in Aggregation Avoidance
511 Strategy Marks a Long-Term Direction to Protein Evolution. *Genetics* 211:1345-1355.
512 Fuglsang, A. 2004. The 'effective number of codons' revisited. *Biochemical and Biophysical Research*
513 *Communications* 317:957-964.
514 Fuglsang, A. 2008. Impact of bias discrepancy and amino acid usage on estimates of the effective
515 number of codons used in a gene, and a test for selection on codon usage. *Gene* 410:82-88.
516 Galtier, N, Piganeau, G, Mouchiroud, D, and Duret, L. 2001. GC-Content Evolution in Mammalian
517 Genomes: The Biased Gene Conversion Hypothesis. *Genetics* 159:907-911.
518 Galtier, N, Roux, C, Rousselle, M, Romiguier, J, Figuet, E, Glémin, S, Bierne, N, and Duret, L. 2018. Codon
519 Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size,
520 and GC-Biased Gene Conversion. *Molecular Biology and Evolution* 35:1092-1103.
521 Gossmann, TI, Keightley, PD, and Eyre-Walker, A. 2012. The Effect of Variation in the Effective
522 Population Size on the Rate of Adaptive Molecular Evolution in Eukaryotes. *Genome Biology and*
523 *Evolution* 4:658-667.
524 Hershberg, R, and Petrov, DA. 2008. Selection on Codon Bias. *Annual Review of Genetics* 42:287-299.
525 Hershberg, R, and Petrov, DA. 2009. General Rules for Optimal Codon Choice. *PLoS Genetics* 5:e1000556.
526 Hildebrand, F, Meyer, A, and Eyre-Walker, A. 2010. Evidence of Selection upon Genomic GC-Content in
527 Bacteria. *PLoS Genetics* 6:e1001107.
528 Huang, Y, and Liu, Z. 2013. Do Intrinsically Disordered Proteins Possess High Specificity in Protein–
529 Protein Interactions? *Chemistry – A European Journal* 19:4462-4467.
530 Hunt, RC, Simhadri, VL, Iandoli, M, Sauna, ZE, and Kimchi-Sarfaty, C. 2014. Exposing synonymous
531 mutations. *Trends In Genetics* 30:308-321.
532 James, J, Nelson, P, and Masel, J. 2022. Differential retention of Pfam domains creates long-term
533 evolutionary trends. *bioRxiv*:2022.10.27.514087.
534 James, JE, Willis, SM, Nelson, PG, Weibel, C, Kosinski, LJ, and Masel, J. 2021. Universal and taxon-specific
535 trends in protein sequences as a function of age. *eLife* 10:e57347.

536 Jansen, R, Bussemaker, HJ, and Gerstein, M. 2003. Revisiting the codon adaptation index from a whole-
537 genome perspective: analyzing the relationship between gene expression and codon occurrence
538 in yeast using a variety of models. *Nucleic Acids Research* 31:2242-2251.
539 Kern, AD, and Hahn, MW. 2018. The Neutral Theory in Light of Natural Selection. *Molecular Biology and*
540 *Evolution* 35:1366-1371.
541 Kessler, MD, and Dean, MD. 2014. Effective population size does not predict codon usage bias in
542 mammals. *Ecology and Evolution* 4:3887-3900.
543 Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713-719.
544 Kosinski, LJ, Aviles, NR, Gomez, K, and Masel, J. 2022. Random Peptides Rich in Small and Disorder-
545 Promoting Amino Acids Are Less Likely to Be Harmful. *Genome Biology and Evolution*
546 14:evac085.
547 Kumar, S, Stecher, G, Suleski, M, and Hedges, SB. 2017. TimeTree: A Resource for timelines, timetrees,
548 and divergence times. *Molecular Biology and Evolution* 34:1812-1819.
549 LaBella, AL, Opulente, DA, Steenwyk, JL, Hittinger, CT, and Rokas, A. 2019. Variation and selection on
550 codon usage bias across an entire subphylum. *PLoS Genetics* 15:e1008304.
551 Lander, ES, Linton, LM, Birren, B, Nusbaum, C, Zody, MC, Baldwin, J, Devon, K, Dewar, K, Doyle, M,
552 FitzHugh, W, Funke, R, Gage, D, Harris, K, Heaford, A, Howland, J, Kann, L, Lehoczky, J, LeVine, R,
553 McEwan, P, McKernan, K, Meldrim, J, Mesirov, JP, Miranda, C, Morris, W, Naylor, J, Raymond, C,
554 Rosetti, M, Santos, R, Sheridan, A, Sougnez, C, Stange-Thomann, N, Stojanovic, N, Subramanian,
555 A, Wyman, D, Rogers, J, Sulston, J, Ainscough, R, Beck, S, Bentley, D, Burton, J, Clee, C, Carter, N,
556 Coulson, A, Deadman, R, Deloukas, P, Dunham, A, Dunham, I, Durbin, R, French, L, Grafham, D,
557 Gregory, S, Hubbard, T, Humphray, S, Hunt, A, Jones, M, Lloyd, C, McMurray, A, Matthews, L,
558 Mercer, S, Milne, S, Mullikin, JC, Mungall, A, Plumb, R, Ross, M, Shownkeen, R, Sims, S,
559 Waterston, RH, Wilson, RK, Hillier, LW, McPherson, JD, Marra, MA, Mardis, ER, Fulton, LA,
560 Chinwalla, AT, Pepin, KH, Gish, WR, Chissoe, SL, Wendl, MC, Delehaunty, KD, Miner, TL,
561 Delehaunty, A, Kramer, JB, Cook, LL, Fulton, RS, Johnson, DL, Minx, PJ, Clifton, SW, Hawkins, T,
562 Branscomb, E, Predki, P, Richardson, P, Wenning, S, Slezak, T, Doggett, N, Cheng, J-F, Olsen, A,
563 Lucas, S, Elkin, C, Uberbacher, E, Frazier, M, Gibbs, RA, Muzny, DM, Scherer, SE, Bouck, JB,
564 Sodergren, EJ, Worley, KC, Rives, CM, Gorrell, JH, Metzker, ML, Naylor, SL, Kucherlapati, RS,
565 Nelson, DL, Weinstock, GM, Sakaki, Y, Fujiyama, A, Hattori, M, Yada, T, Toyoda, A, Itoh, T,
566 Kawagoe, C, Watanabe, H, Totoki, Y, Taylor, T, Weissenbach, J, Heilig, R, Saurin, W, Artiguenave,
567 F, Brottier, P, Bruls, T, Pelletier, E, Robert, C, Wincker, P, Rosenthal, A, Platzer, M, Nyakatura, G,
568 Taudien, S, Rump, A, Smith, DR, Doucette-Stamm, L, Rubenfield, M, Weinstock, K, Lee, HM,
569 Dubois, J, Yang, H, Yu, J, Wang, J, Huang, G, Gu, J, Hood, L, Rowen, L, Madan, A, Qin, S, Davis,
570 RW, Federspiel, NA, Abola, AP, Proctor, MJ, Roe, BA, Chen, F, Pan, H, Ramser, J, Lehrach, H,
571 Reinhardt, R, McCombie, WR, de la Bastide, M, Dedhia, N, Blöcker, H, Hornischer, K, Nordsiek, G,
572 Agarwala, R, Aravind, L, Bailey, JA, Bateman, A, Batzoglou, S, Birney, E, Bork, P, Brown, DG,
573 Burge, CB, Cerutti, L, Chen, H-C, Church, D, Clamp, M, Copley, RR, Doerks, T, Eddy, SR, Eichler,
574 EE, Furey, TS, Galagan, J, Gilbert, JGR, Harmon, C, Hayashizaki, Y, Haussler, D, Hermjakob, H,
575 Hokamp, K, Jang, W, Johnson, LS, Jones, TA, Kasif, S, Kaspryzk, A, Kennedy, S, Kent, WJ, Kitts, P,
576 Koonin, EV, Korf, I, Kulp, D, Lancet, D, Lowe, TM, McLysaght, A, Mikkelsen, T, Moran, JV, Mulder,
577 N, Pollara, VJ, Ponting, CP, Schuler, G, Schultz, J, Slater, G, Smit, AFA, Stupka, E, Szustakowki, J,
578 Thierry-Mieg, D, Thierry-Mieg, J, Wagner, L, Wallis, J, Wheeler, R, Williams, A, Wolf, YI, Wolfe,
579 KH, Yang, S-P, Yeh, R-F, Collins, F, Guyer, MS, Peterson, J, Felsenfeld, A, Wetterstrand, KA,
580 Myers, RM, Schmutz, J, Dickson, M, Grimwood, J, Cox, DR, Olson, MV, Kaul, R, Raymond, C,
581 Shimizu, N, Kawasaki, K, Minoshima, S, Evans, GA, Athanasiou, M, Schultz, R, Patrinos, A,
582 Morgan, MJ, International Human Genome Sequencing, C, Whitehead Institute for Biomedical
583 Research, CfGR, The Sanger, C, Washington University Genome Sequencing, C, Institute, UDJG,

584 Baylor College of Medicine Human Genome Sequencing, C, Center, RGS, Genoscope, Cnrs, UMR,
585 Department of Genome Analysis, IoMB, Center, GTCS, Beijing Genomics Institute/Human
586 Genome, C, Multimegabase Sequencing Center, TIfSB, Stanford Genome Technology, C,
587 University of Oklahoma's Advanced Center for Genome, T, Max Planck Institute for Molecular,
588 G, Cold Spring Harbor Laboratory, LAHGC, Biotechnology, GBGRCf, *Genome Analysis, G,
589 Scientific management: National Human Genome Research Institute, USNIoH, Stanford Human
590 Genome, C, University of Washington Genome, C, Department of Molecular Biology, KUSoM,
591 University of Texas Southwestern Medical Center at, D, Office of Science, USDoE, and The
592 Wellcome, T. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
593 Lazar, T, Tantos, A, Tompa, P, and Schad, E. 2022. Intrinsic protein disorder uncouples affinity from
594 binding specificity. *Protein Science* 31:e4455.
595 Lefébure, T, Morvan, C, Malard, F, François, C, Konecny-Dupré, L, Guéguen, L, Weiss-Gayet, M, Seguin-
596 Orlando, A, Ermini, L, Sarkissian, CD, Charrier, NP, Eme, D, Mermillod-Blondin, F, Duret, L, Vieira,
597 C, Orlando, L, and Douady, CJ. 2017. Less effective selection leads to larger genomes. *Genome*
598 *Research* 27:1016-1028.
599 Li, W-H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of
600 synonymous codons. *Journal of Molecular Evolution* 24:337-345.
601 Lin, J-J, Bhattacharjee, MJ, Yu, C-P, Tseng, YY, and Li, W-H. 2019. Many human RNA viruses show
602 extraordinarily stringent selective constraints on protein evolution. *Proceedings of the National*
603 *Academy of Sciences* 116:19009-19018.
604 Liu, SS, Hockenberry, AJ, Jewett, MC, and Amaral, LAN. 2018. A novel framework for evaluating the
605 performance of codon usage bias metrics. *Journal of The Royal Society Interface* 15:20170667.
606 Long, H, Sung, W, Kucukyildirim, S, Williams, E, Miller, SF, Guo, W, Patterson, C, Gregory, C, Strauss, C,
607 Stone, C, Berne, C, Kysela, D, Shoemaker, WR, Muscarella, ME, Luo, H, Lennon, JT, Brun, YV, and
608 Lynch, M. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nature*
609 *Ecology & Evolution* 2:237-240.
610 Luzuriaga-Neira, AR, and Alvarez-Ponce, D. 2022. Rates of Protein Evolution across the Marsupial
611 Phylogeny: Heterogeneity and Link to Life-History Traits. *Genome Biology and Evolution*
612 14:evab277.
613 Lynch, M, Ackerman, MS, Gout, J-F, Long, H, Sung, W, Thomas, WK, and Foster, PL. 2016. Genetic drift,
614 selection and the evolution of the mutation rate. *Nature Reviews Genetics* 17:704-714.
615 Masel, J. 2011. Genetic drift. *Current Biology* 21:R837-R838.
616 Meunier, J, and Duret, L. 2004. Recombination Drives the Evolution of GC-Content in the Human
617 Genome. *Molecular Biology and Evolution* 21:984-990.
618 Novembre, JA. 2002. Accounting for Background Nucleotide Composition When Measuring Codon Usage
619 Bias. *Molecular Biology and Evolution* 19:1390-1394.
620 Novoa, EM, Jungreis, I, Jaillon, O, and Kellis, M. 2019. Elucidation of Codon Usage Signatures across the
621 Domains of Life. *Molecular Biology and Evolution* 36:2328-2339.
622 Ohta, T. 1972. Population size and rate of evolution. *Journal of Molecular Evolution* 1:305-314.
623 Ohta, T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* 246:96-98.
624 Ohta, T. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and*
625 *Systematics* 23:263-286.
626 Paradis, E, and Schliep, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary
627 analyses in R. *Bioinformatics* 35:526-528.
628 Payne, BL, and Alvarez-Ponce, D. 2018. Higher Rates of Protein Evolution in the Self-Fertilizing Plant
629 *Arabidopsis thaliana* than in the Out-Crossers *Arabidopsis lyrata* and *Arabidopsis halleri*.
630 *Genome Biology and Evolution* 10:895-900.

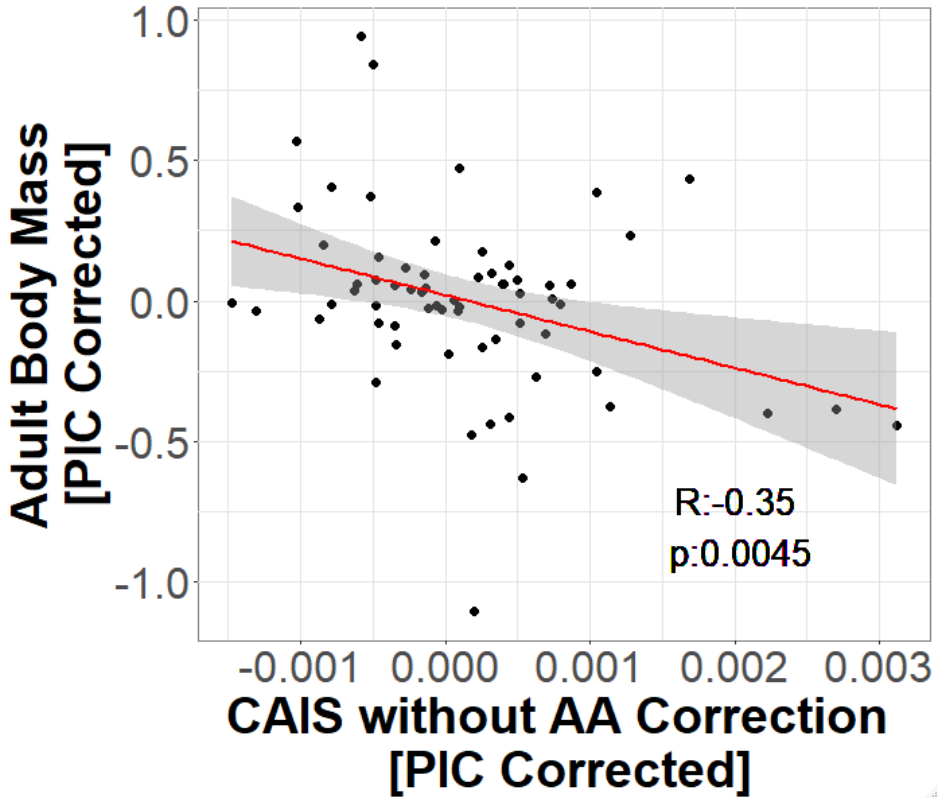
631 Plotkin, JB, Dushoff, J, Desai, MM, and Fraser, HB. 2006. Codon Usage and Selection on Proteins. *Journal*
632 *of Molecular Evolution* 63:635-653.
633 Plotkin, JB, and Kudla, G. 2011. Synonymous but not the same: the causes and consequences of codon
634 bias. *Nature Reviews Genetics* 12:32-42.
635 Romiguier, J, Gayral, P, Ballenghien, M, Bernard, A, Cahais, V, Chenuil, A, Chiari, Y, Dernat, R, Duret, L,
636 Faivre, N, Loire, E, Lourenco, JM, Nabholz, B, Roux, C, Tsagkogeorga, G, Weber, AAT, Weinert,
637 LA, Belkhir, K, Bierne, N, Glémin, S, and Galtier, N. 2014. Comparative population genomics in
638 animals uncovers the determinants of genetic diversity. *Nature* 515:261-263.
639 Romiguier, J, and Roux, C. 2017. Analytical Biases Associated with GC-Content in Molecular Evolution.
640 *Frontiers in Genetics* 8:16.
641 Schad, E, Tompa, P, and Hegyi, H. 2011. The relationship between proteome size, structural disorder and
642 organism complexity. *Genome Biology* 12:R120.
643 Sharp, PM, Bailes, E, Grocock, RJ, Peden, JF, and Sockett, RE. 2005. Variation in the strength of selected
644 codon usage bias among bacteria. *Nucleic Acids Research* 33:1141-1153.
645 Sharp, PM, Emery, LR, and Zeng, K. 2010. Forces that influence the evolution of codon bias. *Philosophical*
646 *Transactions of the Royal Society B: Biological Sciences* 365:1203-1212.
647 Sharp, PM, and Li, W-H. 1987. The codon adaptation index-a measure of directional synonymous codon
648 usage bias, and its potential applications. *Nucleic Acids Research* 15:1281-1295.
649 Smith, NGC, and Eyre-Walker, A. 2001. Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C-
650 Rich Genes in Humans. *Molecular Biology and Evolution* 18:982-986.
651 Subramanian, S. 2008. Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes.
652 *Genetics* 178:2429-2432.
653 Sun, X, Yang, Q, and Xia, X. 2012. An Improved Implementation of Effective Number of Codons (Nc).
654 *Molecular Biology and Evolution* 30:191-196.
655 Urrutia, AO, and Hurst, LD. 2001. Codon Usage Bias Covaries With Expression Breadth and the Rate of
656 Synonymous Evolution in Humans, but This Is Not Evidence for Selection. *Genetics* 159:1191-
657 1199.
658 Vavouri, T, Semple, JI, Garcia-Verdugo, R, and Lehner, B. 2009. Intrinsic Protein Disorder and Interaction
659 Promiscuity Are Widely Associated with Dosage Sensitivity. *Cell* 138:198-208.
660 Vicario, S, Moriyama, EN, and Powell, JR. 2007. Codon usage in twelve species of Drosophila. *BMC*
661 *Evolutionary Biology* 7:226.
662 Wangen, JR, and Green, R. 2020. Stop codon context influences genome-wide stimulation of termination
663 codon readthrough by aminoglycosides. *eLife* 9:e52611.
664 Weber, CC, Nabholz, B, Romiguier, J, and Ellegren, H. 2014. Kr/Kc but not dN/dS correlates positively
665 with body mass in birds, raising implications for inferring lineage-specific selection. *Genome*
666 *Biology* 15:542.
667 Wright, F. 1990. The 'effective number of codons' used in a gene. *Gene* 87:23-29.
668 Xue, B, Dunker, AK, and Uversky, VN. 2012. Orderly order in protein intrinsic disorder distribution:
669 disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular*
670 *Structure and Dynamics* 30:137-149.
671 Zhang, Z, Li, J, Cui, P, Ding, F, Li, A, Townsend, JP, and Yu, J. 2012. Codon Deviation Coefficient: a novel
672 measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics*
673 13:43.
674



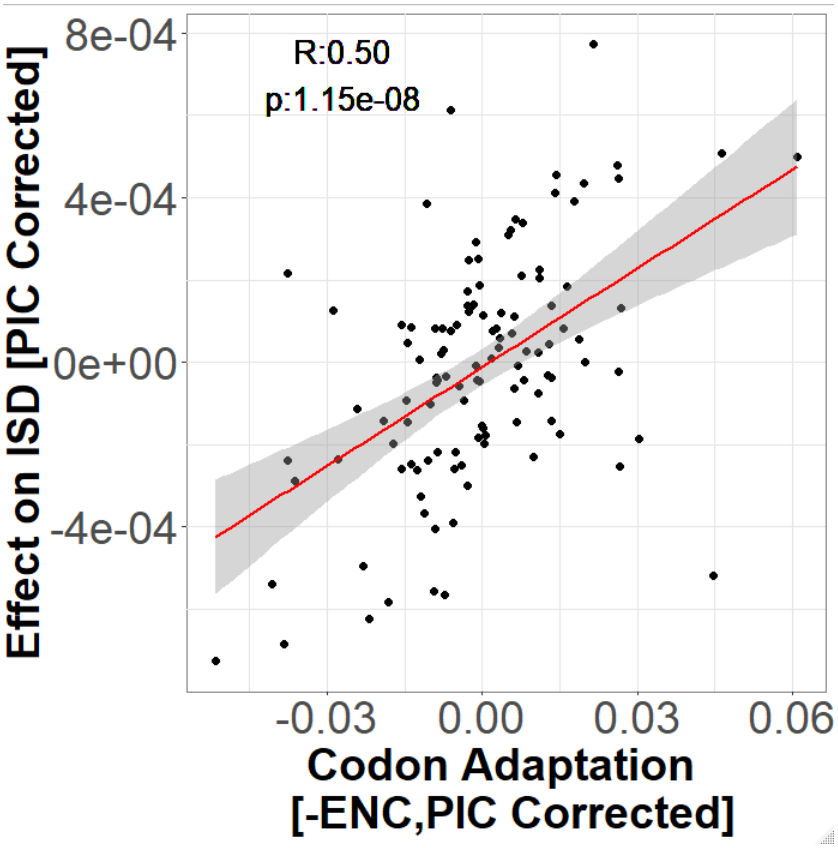
675

676 Supplementary Figure 1: Codon Adaptation Index is not appropriate for species-wide
677 effectiveness of selection measurements. Each CAI value shown is averaged over an entire
678 species' proteome. A) The value of CAI is driven by its normalizing denominator term, CAI_{max}. B)
679 As a result, CAI is inversely proportional to CAIS. Each datapoint is one of 118 vertebrate species
680 with "Complete" intergenic genomic sequence available (allowing for %GC correction) and
681 TimeTree divergence dates (allowing for PIC correction). P-values shown are for Pearson's
682 correlation.

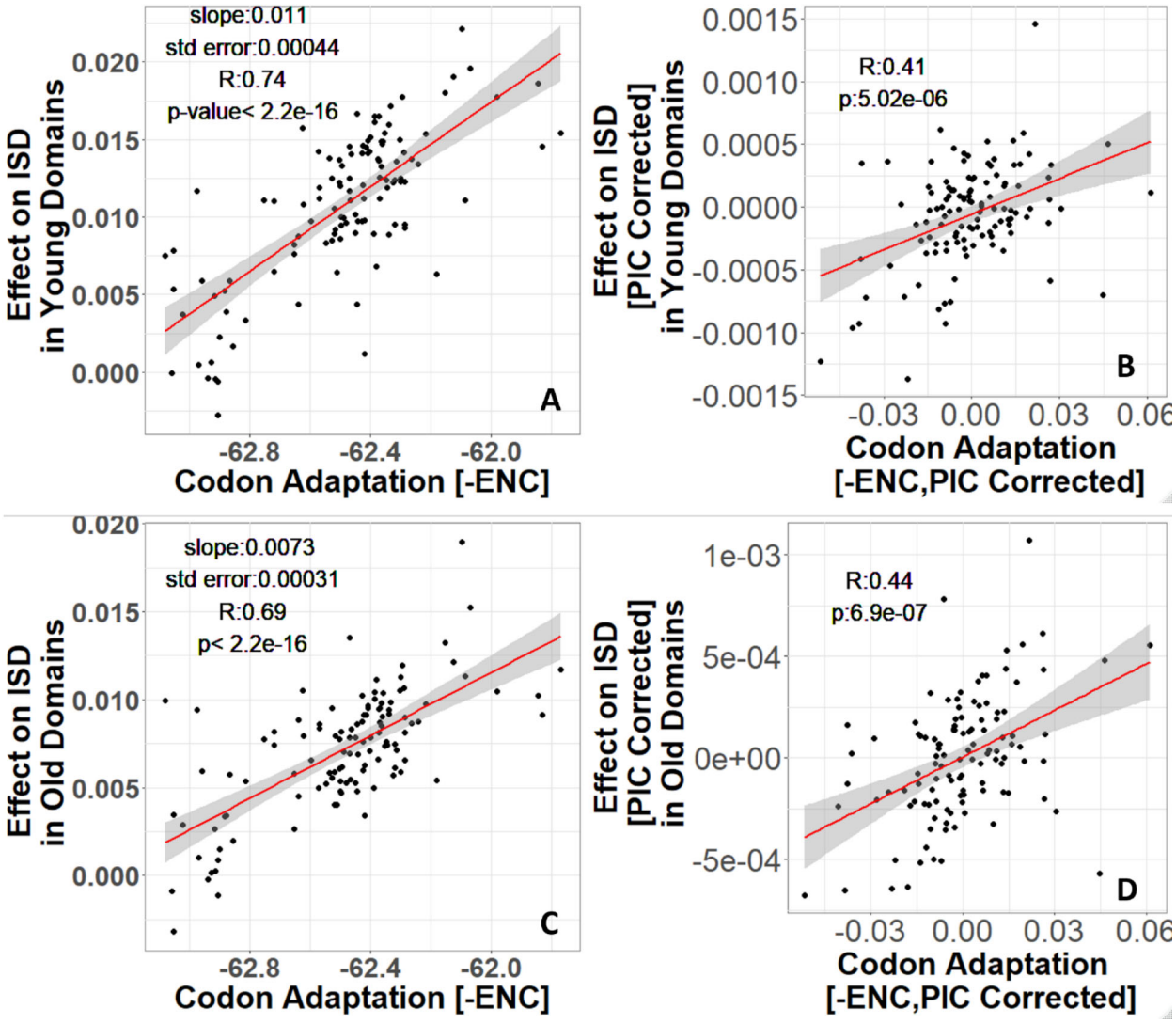
683



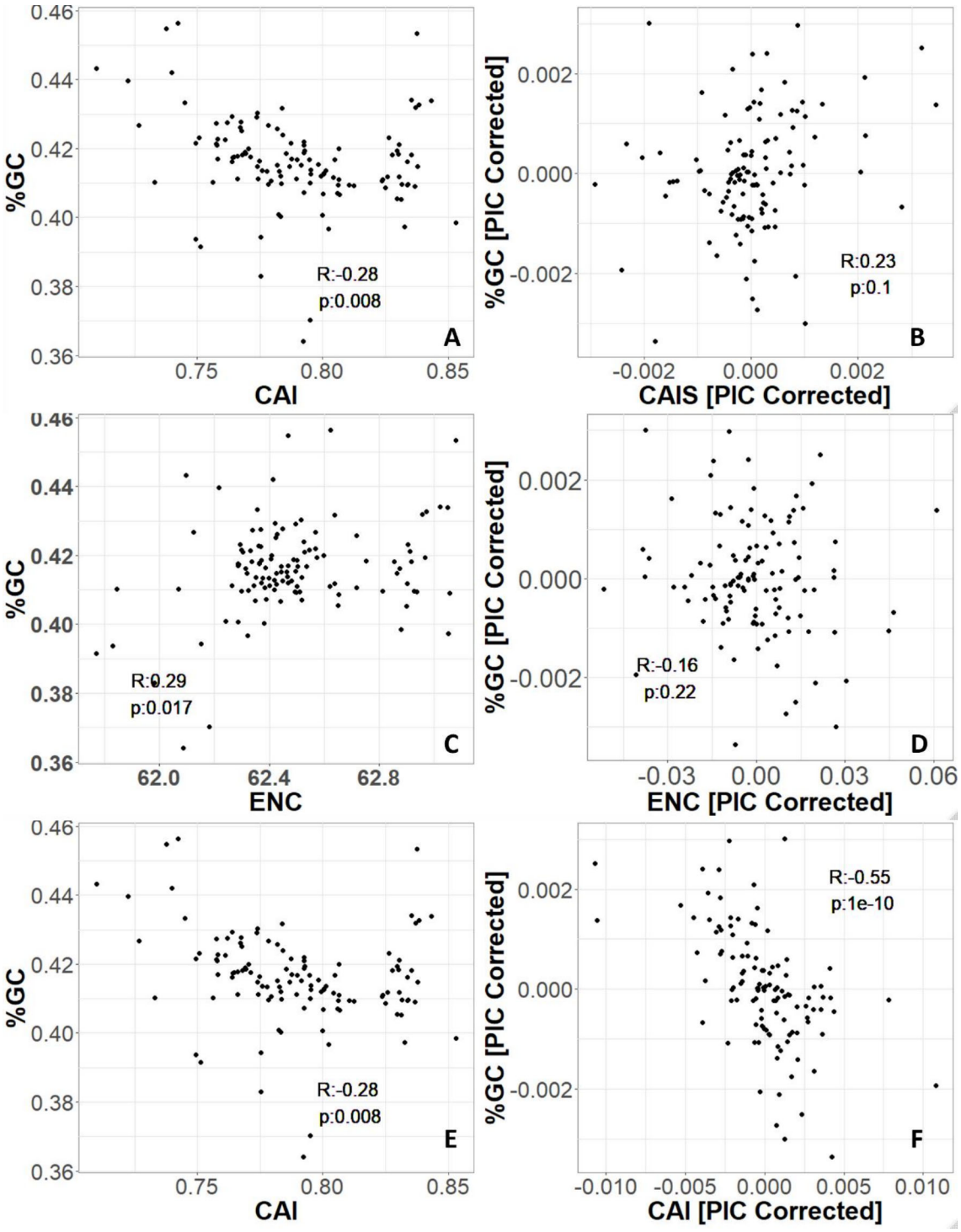
Supplementary Figure 2: CAIS without correction for amino acid composition still reflects the expected relationship between effectiveness of selection and body. Body size data from PanTHERIA database, originally in $\log_{10}$(mass) in grams; data shown for 62 species in common between PANTHERIA and our own dataset of 118 vertebrate species with a "Complete" intergenic genomic sequence and TimeTree divergence dates. P-value is shown for Pearson's correlation. Red line shows unweighted lm(y~x) with grey region as 95% confidence interval.



Supplementary Figure 3: Our Figure 4 finding that more exquisitely adapted species have protein domains with higher ISD is confirmed by ENC. The same n=118 datapoints are shown as in Figure 4. P-values shown are for Pearson's correlation. Red line shows unweighted lm(y~x) with grey region as 95% confidence interval.



Supplementary Figure 4: More exquisitely adapted species have higher ISD in both ancient and recent protein domains, as confirmed by ENC. Protein domains that emerged prior to LECA are identified as “old”, and protein domains that emerged after the divergence of animals and fungi from plants and found in vertebrates are identified as “young”. Age assignments are taken from James et al. (2021). The same n=118 datapoints are shown as in Figure 3. P-values shown are for Pearson’s correlation. Red line shows lm(y~x), with grey region as 95% confidence interval, using a weighted model for non-PIC-corrected figures and unweighted for PIC-corrected figures.



Supplementary Figure 5: Effective Number of Codons (ENC) and Codon Adaptation Index of Species (CAIS) are uncorrelated with total genomic GC Content. Each datapoint is one of 118 vertebrate species with “Complete” intergenic genomic sequence and TimeTree divergence dates. Statistical significance is only supported if it survives controlling for phylogenetic confounding via Phylogenetic Independent Contrasts (PIC). P-values shown are for Pearson’s correlation.