The protein domains of vertebrate species in which selection is more effective have greater intrinsic structural disorder

Catherine A. Weibel[1,2,3,*], Andrew L. Wheeler[4,*], Jennifer E. James[5,6], Sara M. Willis[5,7], Hanon McShea[8], 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

[8]Department of Earth System Science, Stanford University

*these authors contributed equally

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. A potentially more direct measure of the effectiveness of selection is the degree to which selection maintains preferred codons. 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), based on Kullback-Leibler divergence, that corrects for both confounders. We demonstrate the use of CAIS correlations, as well as the Effective Number of Codons, 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, genome size, mutation rate, selective pressure, 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).

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.

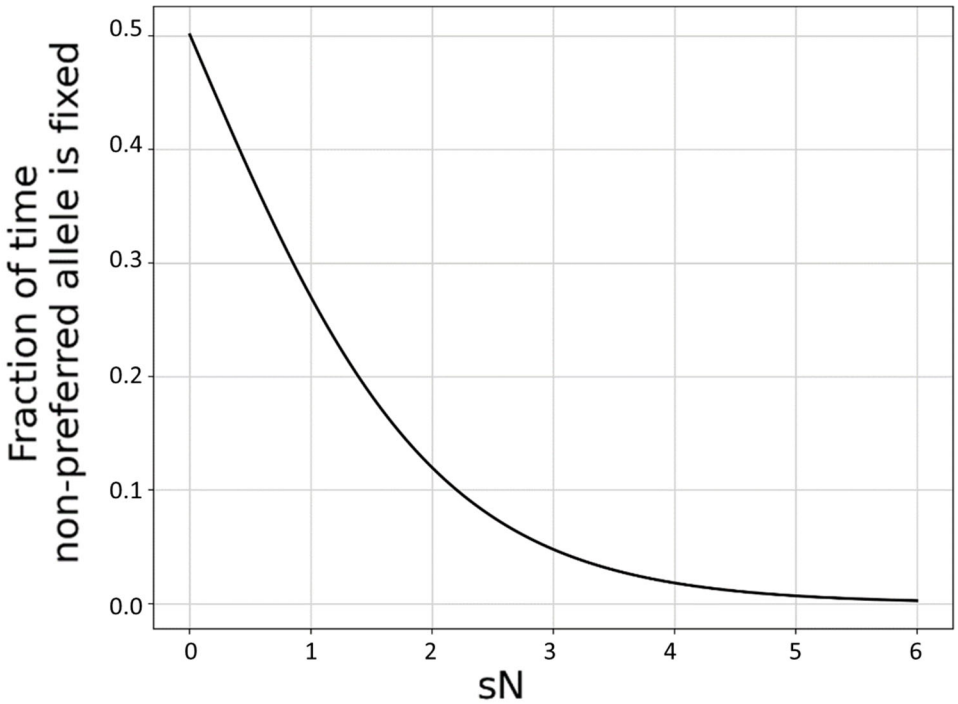


Figure 1: The effectiveness of selection, calculated as the long-term ratio of time spent in fixed deleterious : fixed beneficial allele states given symmetric mutation rates, is a function of the product *sN*. Assuming a diploid Wright-Fisher population with s <<1, the probability of fixation of a new 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. For small sN, selection is ineffective at producing codon bias. For large sN, selection is highly effective. For only a relatively narrow range of intermediate values of sN, the degree of codon bias depends quantitatively on sN.

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 in non-equilibrium settings.

As a practical matter, N_e is usually calculated by dividing some measure of the amount of putatively neutral (often synonymous) polymorphism segregating in a population 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 their 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); a poor sampling scheme can have the converse effect of deflating genetic diversity. Transient hypermutation (Plotkin et al. 2006), which is common in microbes, causes further short-term inconsistencies in polymorphism levels. Perhaps most importantly, a recent bottleneck will deflate N_e based on the coalescence, even if too brief to lead to significant erosion of fine-tuned adaptations. But drift barrier theory concerns the level with which adaptation is fine-tuned, and so a better metric would capture that directly, rather than indirectly rely on neutral diversity.

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

93 denominator (see equation 4 below and Supplementary Figure 1A). Paradoxically, this can
94 make more exquisitely adapted species have lower rather than higher species-averaged CAI
95 scores (Supplementary Figure 1B; Rocha 2004; Botzman & Margalit 2011).

96 To compare species using CAI, it has been suggested that instead of taking a genome-wide
97 average, one should consider a set of highly expressed reference genes (Sharp et al. 2005;
98 Vicario et al. 2007; Subramanian 2008; dos Reis & Wernisch 2009). This approach assumes that
99 the relative strength of selection on those reference genes (often a function of gene
100 expression) remains approximately constant across the set of species considered (red
101 distributions in Figure 2). Its use also requires careful attention to the length of reference genes
102 (Urrutia & Hurst 2001; Doherty & McInerney 2013), and some approaches also require
103 information about tRNA gene copy numbers and abundances (dos Reis & Wernisch 2009).

104 Since codon bias varies quantitatively within only a small range of sN (Figure 1), a promising
105 approach is to measure the proportion of sites at which codon adaptation is effective. We posit
106 that more highly adapted species have a higher proportion of both genes and sites subject to
107 effective selection on codon bias (Figure 2; Galtier et al. 2018). Indeed, CAI might also rely in
108 part on variation in the fraction of sites within the reference genes that is subject to effective
109 selection as a function of species (Figure 2, red). Here we take this logic further, considering all
110 sites in a proteome-wide approach. Averaging across the entire proteome provides robustness
111 to shifts in the expression level of or strength of selection on particular genes. The proteome-
112 wide average depends on the fraction of sites whose selection coefficients exceed the "drift
113 barrier" for that particular species (Figure 2, blue threshold).

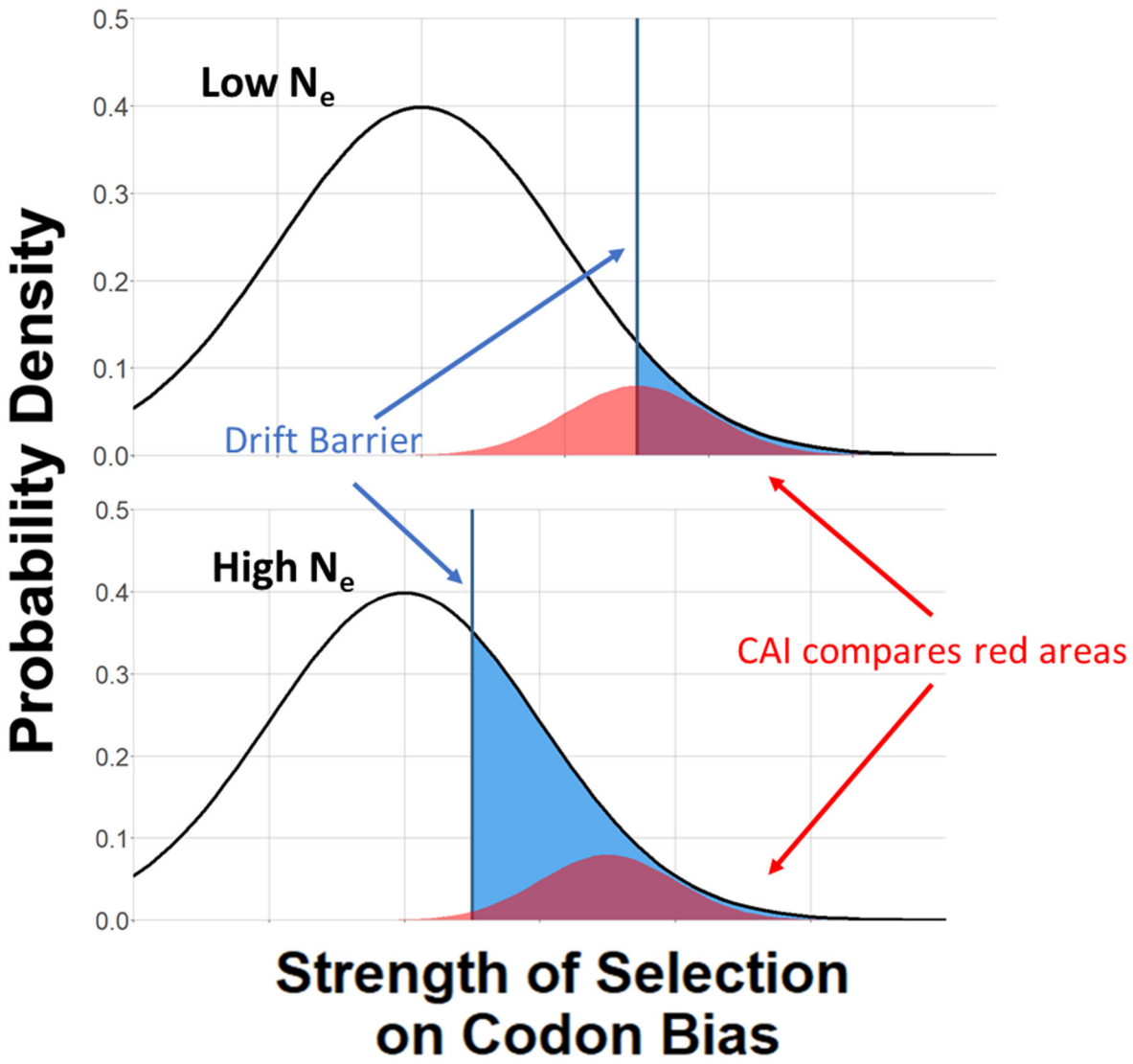


Figure 2: More highly adapted species (bottom) have a higher proportion of their sites subject to effective selection on codon bias (blue area). The CAI attempts to compare the intensity of selection (Figure 1, x-axis) in a subset of genes under strong selection (red areas). Given the narrow range of quantitative dependence of codon bias on sN shown in Figure 1, our new metric is intended to capture differences in the proportion of the proteome subject to substantial selection (blue areas).

In estimating the effects of selection, it is critical to control for other causes of codon bias. In particular, species differ in their mutational bias with respect to %GC, and in the frequency of GC-biased gene conversion (Urrutia & Hurst 2001; Duret & Galtier 2009; Doherty & McInerney

124 2013; Figuet et al. 2014). Here, we control for %GC, capturing species differences both in
125 mutation and in gene conversion, by calculating the Kullback-Leibler divergence of the observed
126 codon frequencies away from the codon frequencies that we would expect to see given the
127 genomic %GC content of the species. Kullback-Leibler divergence measures the distance of an
128 observed probability distribution from an expected reference distribution given %GC, capturing
129 a measure of surprise (Kullback & Leibler 1951). This method does not require us to specify
130 preferred vs. non-preferred codons, and can thus also accommodate situations in which
131 different genes have different codon preferences (Gingold et al. 2014; Cope et al. 2018).

132 An alternative metric, the Effective Number of Codons (ENC) originally quantified how far the
133 codon usage of a sequence departs from equal usage of synonymous codons (Wright 1990),
134 with lower ENC values indicating greater departure. This approach creates a complex
135 relationship with GC content (Fuglsang 2008), and so ENC was later modified to correct for GC
136 content (Novembre 2002). However, a remaining issue with this modified ENC is that
137 differences among species in amino acid composition might act as a confounding factor, even
138 after controlling for GC content. Specifically, species that make more use of an amino acid for
139 which there is stronger selection among codons (which is sometimes the case (Vicario et al.
140 2007)) would have higher codon bias, even if each amino acid considered on its own had
141 identical codon bias irrespective of which species it is in. Confounding with amino acid
142 frequencies has been shown to be a problem at the individual protein level (Cope et al. 2018).
143 Neither ENC (Fuglsang 2004; Fuglsang 2008) nor the CAI (Sharp & Li 1987) adequately control
144 for differences in amino acid composition when applied across species. Despite early claims to
145 the contrary (Wright 1990), this problem is not easy to fix for ENC (Fuglsang 2004; Fuglsang
146 2008).

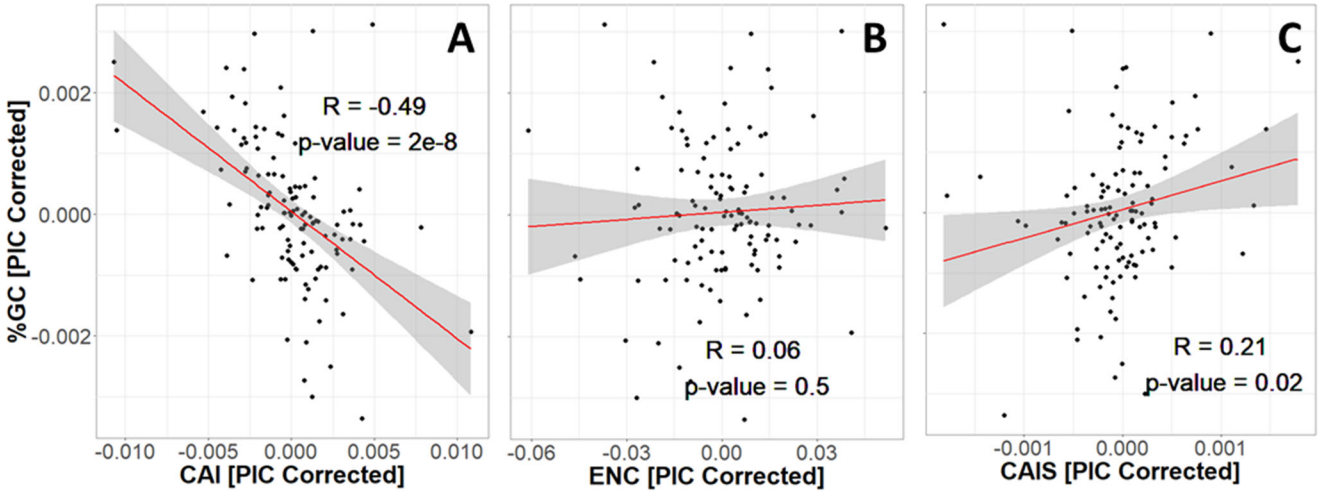
147 Here, we extend the CAI, using the information-theory based Kullback-Leibler divergence, so
148 that it corrects for both GC and amino acid composition (see Methods) to create a new Codon
149 Adaptation Index of Species (CAIS). The availability of complete genomes allows both metrics to
150 be readily calculated without data on polymorphism or mutation rate, without selecting
151 reference genes, and without concerns about demographic history. Our purpose is to find an
152 accessible metric that can quantify the limits to weak selection important to nearly neutral

153 theory; this differs from past evaluations focused on comparing different genes of the same
154 species and recapitulating "ground truth" simulations thereof (Sun et al. 2012; Zhang et al.
155 2012; Liu et al. 2018). To demonstrate the usefulness of our method, we identify a novel
156 correlation with intrinsic structural disorder (ISD), pointing to what else might be subject to
157 weak selective preferences at the molecular level. While ENC can also identify subtle selection
158 on ISD, CAIS can do so without the risk of confounding with amino acid frequencies.

159 Results

160 Both ENC and CAIS solve the GC confounding problem that plagues CAI

161 CAI is seriously confounded with GC content (Figure 3A). ENC is not confounded with GC
162 content (Figure 3B), while CAIS has only a very weak correlation that is not significant after
163 correction for multiple comparisons (Figure 3C).

164 

165 Figure 3: CAI is seriously confounded with GC content, while ENC and CAIS are not. We control
166 for phylogenetic confounding via Phylogenetic Independent Contrasts (PIC) (Felsenstein 1985);
167 this yields an unbiased R^2 estimate (Rohlf 2006). Each datapoint is one of 118 vertebrate
168 species with "Complete" intergenic genomic sequence (allowing for %GC correction) and
169 TimeTree divergence dates (allowing for PIC correction). Red line shows unweighted lm(y~x)
170 with grey region as 95% confidence interval. Plots without PIC correction are shown in
171 Supplementary Figure 2.

172

173 ## Proteins in better adapted species evolve more structural disorder

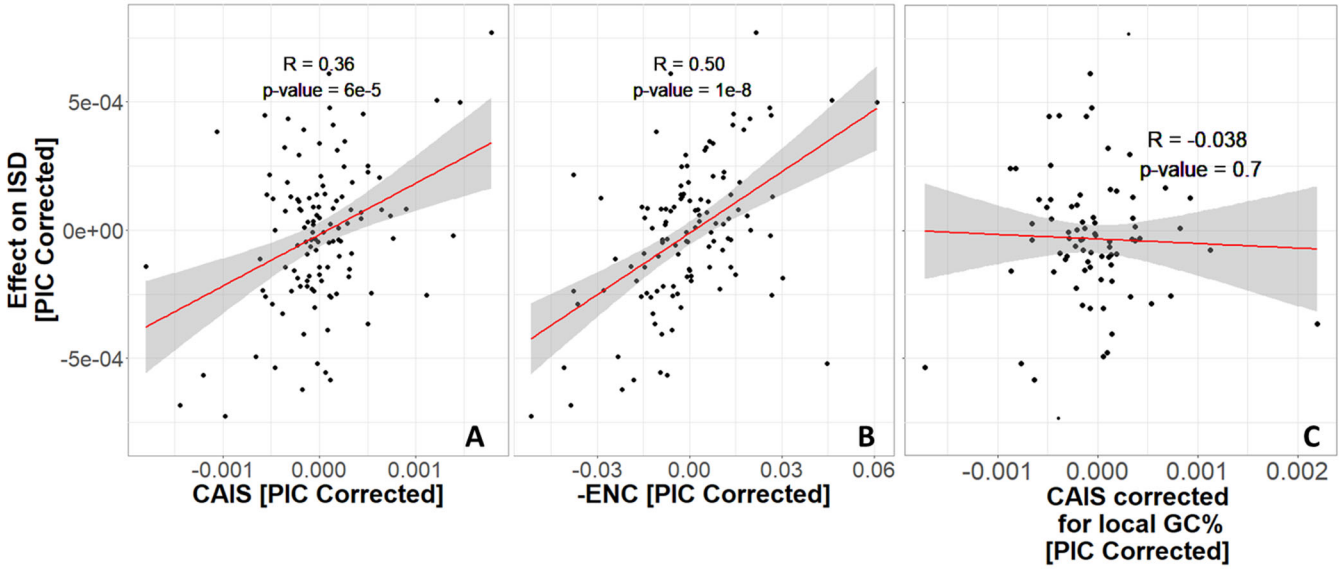
174 As an example of how correlations with codon adaptation metrics can be used to identify weak
175 selective preferences, we investigate protein ISD. Disordered proteins are more likely to be
176 harmful when overexpressed (Vavouri et al. 2009), and ISD is more abundant in eukaryotic than
177 prokaryotic proteins (Schad et al. 2011; Xue et al. 2012; Basile et al. 2019), suggesting that low
178 ISD might be favored by more effective selection.

179 However, compositional differences among proteomes might not be driven by differences in
180 how a given protein sequence evolves as a function of the effectiveness of selection. Instead,
181 they might be driven by the recent birth of ISD-rich proteins in animals (James et al. 2021),
182 and/or by differences among sequences in their subsequent tendency to proliferate into many
183 different genes (James et al. 2023). To focus only on the effects of descent with modification,
184 we use a linear mixed model, with each species having a fixed effect on ISD, while controlling
185 for Pfam domain identity as a random effect. We note that once GC is controlled for, codon
186 adaptation can be assessed similarly in intrinsically disordered vs. ordered proteins (Gossmann
187 et al. 2012). Controlling for Pfam identity is supported, with standard deviation in ISD of 0.178
188 among Pfams compared to residual standard deviation of 0.058, and a p-value on the
189 significance of the Pfam random effect term of 3×10^{-13}. Controlling in this way for Pfam identity,
190 we then ask whether the fixed species effects on ISD are correlated with CAIS and with ENC.

191 Surprisingly, more exquisitely adapted species have more disordered protein domains (Figure
192 4). Results using ENC and CAIS are similar, with ENC having higher power; the correlation
193 coefficient is 0.36 for CAIS compared to 0.50 for ENC, and the p-value for ENC is 3 orders of
194 magnitude lower. We note, however, that amino acid frequencies strongly influence ISD
195 (Theillet et al. 2013). The CAIS correlation is more reliable than the ENC correlation because by
196 construction, CAIS controls for differences in amino acid frequencies among species.

197 Different parts of the genome have different GC contents (Bernardi 2000; Eyre-Walker & Hurst
198 2001; International Human Genome Sequencing Consortium 2001), primarily because the
199 extent to which GC-biased gene conversion increases GC content depends on the local rate of
200 recombination (Galtier et al. 2001; Meunier & Duret 2004; Duret et al. 2006; Duret & Galtier

201 2009). We therefore also calculated a version of CAIS whose codon frequency expectations are
202 based on local intergenic GC content. This performed worse (Figure 4C) than our simple use of
203 genome-wide GC content (Figure 4A) with respect to the strength of correlation between CAIS
204 and ISD. If GC-biased gene conversion is a more powerful force than weak selective
205 preferences among codons, then local GC content will evolve more rapidly than codon usage
206 (Kondrashov et al. 2010). In this case, genome-wide GC may serve as an appropriately time-
207 averaged proxy. It is also possible that the local non-coding sequences we used were too short
208 (at 3000bp or more), creating excessive noise that obscured the signal.



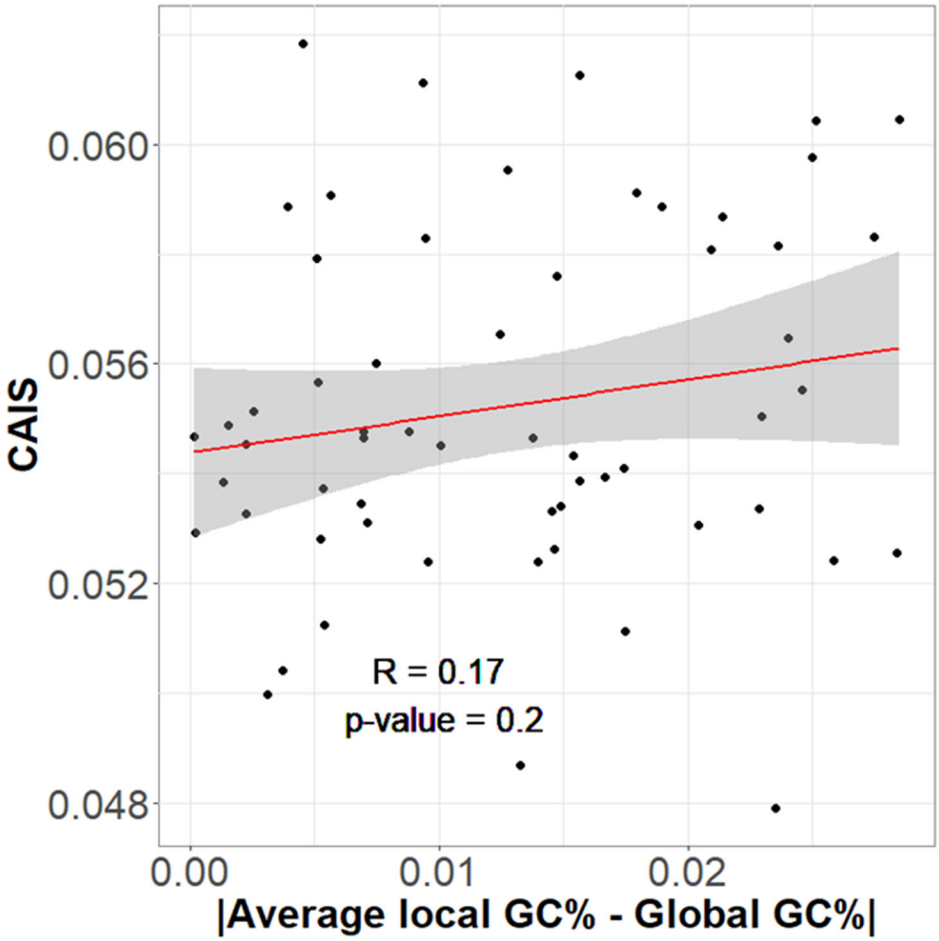
209

210 Figure 4: Protein domains have higher ISD when found in more exquisitely adapted species. We show -
211 ENC instead of ENC to more easily compare with CAIS. Each datapoint is one of 118 vertebrate species
212 with "complete" intergenic genomic sequence available (allowing for %GC correction), and TimeTree
213 divergence dates (allowing for PIC correction). "Effects" on ISD shown on the y-axis are fixed effects of
214 species identity in our linear mixed model, after PIC correction. Red line shows unweighted lm(y~x) with
215 grey region as 95% confidence interval. Panels without PIC correction are presented in Supplementary
216 Figure 3.

217

218 Many vertebrates have higher recombination rates and hence GC-biased gene conversion near
219 genes; in this case genome-wide GC content would misestimate the codon usage expected
220 from the combination of mutation bias and GC-biased gene conversion in the vicinity of genes.

221 If GC-biased gene conversion drove CAIS, we expect high $|\overline{\text{local GC}}$ - global GC| to predict high
222 CAIS. We do not see this relationship (Figure 5), suggesting that gene conversion strength is not
223 a confounding factor impacting CAIS.



224

225 Figure 5: CAIS is not correlated with the degree to which local genomic regions differ in their GC
226 content from global GC content. If CAIS were driven by GC biased gene conversion, genomes
227 with more heterogeneous %GC distributions should have higher CAIS scores.

228 Younger animal-specific protein domains have higher ISD (James et al. 2021). It is possible that
229 selection in favor of high ISD is strongest in young domains, which might use more primitive
230 methods to avoid aggregation (Foy et al. 2019; Bertram & Masel 2020). To test this, we analyze
231 two subsets of our data: those that emerged prior to the last eukaryotic common ancestor
232 (LECA), here referred to as “old” protein domains, and “young” protein domains that emerged
233 after the divergence of animals and fungi from plants. Young and old domains show equally
234 strong trends of increasing disorder with species’ adaptedness (Figure 6).

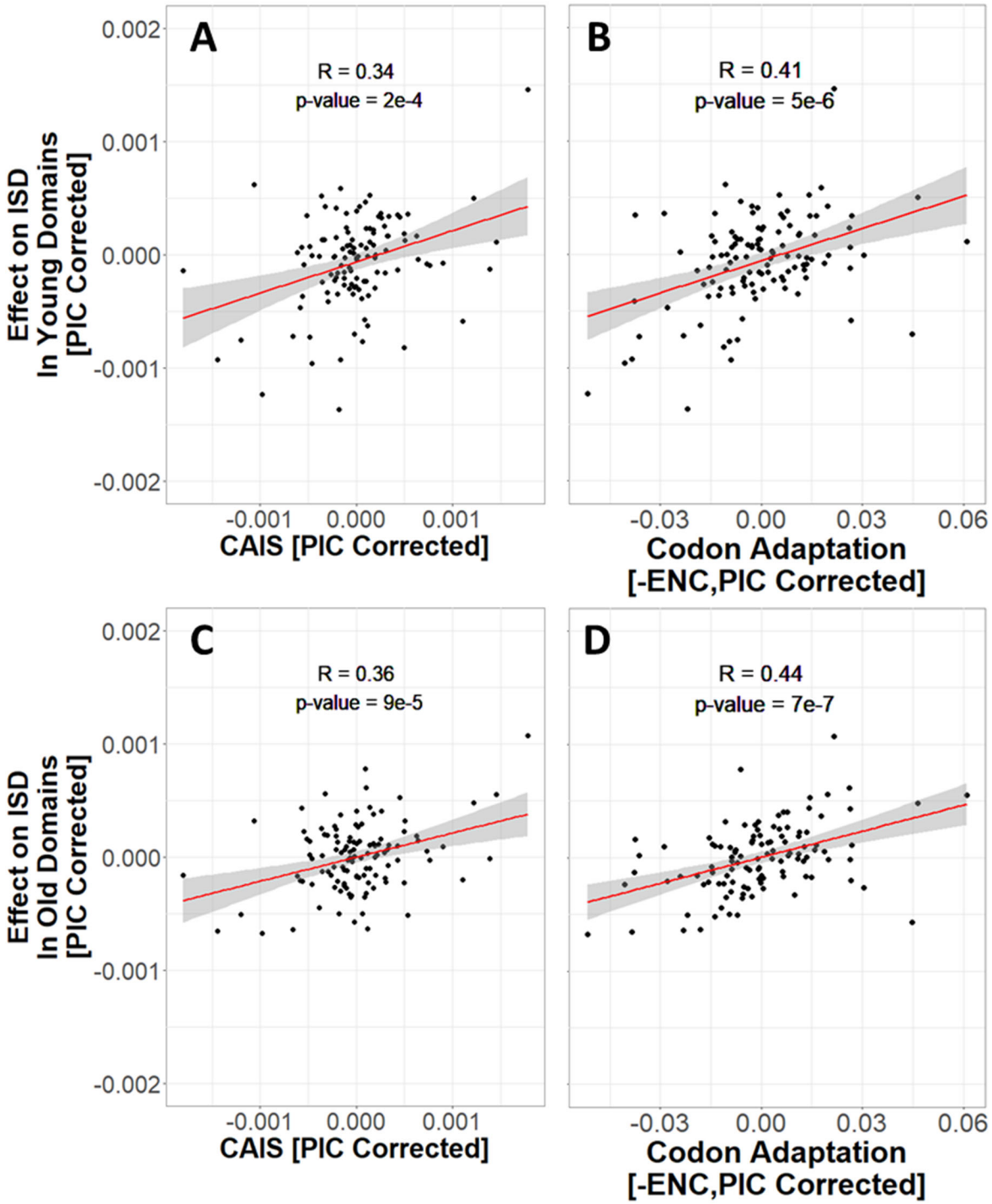


Figure 6: More exquisitely adapted species have higher ISD in both ancient and recent protein domains. Age assignments are taken from James et al. (2021), with vertebrate protein domains that emerged prior to LECA classified as “old”, and vertebrate protein domains that emerged after the divergence of 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.

241 Red line shows lm(y~x), with grey region as 95% confidence interval. Panels without PIC correction are
242 shown in Supplementary Figure 4. We show -ENC instead of ENC to more easily compare with CAIS.

243

244 Discussion

245 When different properties are each causally affected by a species' exquisiteness of adaptation,
246 this will create a correlation between the properties. We use codon adaptation as a reference
247 property, such that correlations with codon adaptation indicate selection. To detect ISD as a
248 novel property under selection, we used a linear mixed model approach that controls for Pfam
249 identity as a random effect. This approach shows that the same Pfam domain tends to be more
250 disordered when found in a well-adapted species (i.e. a species with a higher CAIS or ENC). This
251 is true for both ancient and recently emerged protein domains.

252 It is important that no additional variable such as GC content or amino acid frequencies creates
253 a spurious correlation by affecting both CAIS and our property of interest. For this reason, we
254 define CAIS as the observed Kullback-Leibler divergence (Kullback & Leibler 1951) from the
255 codon usage expected given the GC content. The GC content pertinent to this expectation is
256 primarily the product of mutation bias and GC-biased gene conversion (Romiguier & Roux
257 2017), but potentially also of selection on individual nucleotide substitutions that is
258 hypothesized to favor higher %GC (Long et al. 2018). By controlling for %GC, we exclude all
259 these forces from influencing CAIS or ENC. We thus capture the extent of adaptation in codon
260 bias, including translational speed, accuracy, and any intrinsic preference for GC over AT that is
261 specific to coding regions. These remaining codon-adaptive factors do not create a statistically
262 convincing correlation between CAIS and GC (Figures 3C), nor between ENC and GC (Figure 3B),
263 although CAI is strongly correlated with GC (Figure 3A). Notably, our new CAIS metric of codon
264 adaptation controls for amino acid frequencies, rather than, like ENC, only GC content.

265 A direct effect of ISD on fitness agrees with studies of random ORFs in *Escherichia coli*, where
266 fitness was driven more by amino acid composition than %GC content, after controlling for the
267 intrinsic correlation between the two (Kosinski et al. 2022). However, we have not ruled out a
268 role for selection for higher %GC in ways that are general rather than restricted to coding

269 regions, whether in shaping mutational biases (Smith & Eyre-Walker 2001; Hershberg & Petrov
270 2009; Hildebrand et al. 2010; Novoa et al. 2019; Forcelloni & Giansanti 2020) or the extent of
271 gene conversion, or even at the single nucleotide level in a manner shared between coding
272 regions and intergenic regions (Long et al. 2018).

273 A more complex metric could control for more than just GC content and amino acid
274 frequencies. First vs. second vs. third codon positions have different nucleotide usage on
275 average, but while correcting for this might be useful for comparing genes (Zhang et al. 2012),
276 correcting for it while comparing species might remove the effect of interest. Similarly, while it
277 might be useful to control for dinucleotide and trinucleotide frequencies (Brbić et al. 2015), to
278 avoid circularity these would need to be taken from intergenic sequences, with care needed to
279 avoid influence from unannotated protein-coding genes or even pseudogenes.

280 Note that if a species were to experience a sudden reduction in census population size, e.g. due
281 to habitat loss, leading to less effective selection, it would take some multiple of the neutral
282 coalescent time for CAIS to fully adjust. CAIS thus represents a relatively long-term historical
283 pattern of adaptation. The timescales setting neutral polymorphism based N_e estimates are
284 likely shorter, based on a single round of coalescence. It is possible that the reason that we
285 obtained correlations when we controlled for genome-wide GC content, but not when we
286 controlled for local GC content, is also that codon adaptation adjusts slowly relative to the
287 timescale of fluctuations in local GC content.

288 Here we developed a new metric of species adaptedness at the codon level, capable of
289 quantifying degrees of codon adaptation even among vertebrates. We chose vertebrates partly
290 due to the abundance of suitable data, and partly as a stringent test case, given past studies
291 finding limited evidence for codon adaptation (Kessler & Dean 2014). It remains to be seen how
292 CAIS scales among species with stronger codon adaptation. We restricted our analysis to only
293 the best annotated genomes, in part to ensure the quality of intergenic %GC estimates, and in
294 part limited by the feasibility of running linear mixed models with six million data points. The
295 phylogenetic tree is well resolved for vertebrate species, with an overrepresentation of

296 mammalian species. Despite the focus on vertebrates, we were able to discover new results
297 regarding selection on ISD.

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

306 Codon adaptation metrics more directly quantify how species vary in their exquisiteness of
307 adaptation, than do estimates of effective population size that are based on neutral
308 polymorphism. Both CAIS and ENC can also be estimated for far more species because they do
309 not require polymorphism or mutation rate data, nor tRNA gene copy numbers and
310 abundances, but only a single complete genome. CAIS has the additional advantage of not
311 being confounded with amino acid frequencies. This makes CAIS a useful tools for applying
312 nearly neutral theory to protein evolution, as shown by our worked example of ISD.

313 Methods

314 Species

315 Pfam sequences and IUPRED2 estimates of Intrinsic Structural Disorder (ISD) predictions were
316 taken from James et al. (2021), who studied species marked as "Complete" in the GOLD
317 database, with divergence dates available in TimeTree (Kumar et al. 2017). James et al. (2021)
318 applied a variety of quality controls to exclude contaminants from the set of Pfams and assign
319 accurate dates of Pfam emergence. Pfams that emerged prior to LECA are classified here as
320 "old", and Pfams that emerged after the divergence of animals and fungi from plants are
321 classified as "young", following annotation by James et al. (2021). Species list and other
322 information can be found at https://github.com/MaselLab/Codon-Adaptation-Index-of-Species.
323

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 (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

$$CAI = [\prod_{i=1}^{L} w_i]^{\frac{1}{L}}. \tag{3}$$

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

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

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. (2023)). 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

$$p_{i,r} = \frac{g_r}{2}^{k_{GC}} \left(1 - \frac{g_r}{2}\right)^{k_{AT}}, \tag{5}$$

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

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

We can then measure the degree to which the observed codon frequencies diverge from these expected probabilities using the Kullback–Leibler divergence. This gives a CAIS metric for a species *s* where $O_{i,s}$ is the observed frequency of codon *i*:

$$CAIS(s) = \sum_{i=1}^{64} O_{i,s} \log\left(\frac{O_{i,s}}{E_{i,s}}\right). \quad (7)$$

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. We want to re-weight $O_{i,s}$ 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

$$F_a = \alpha_{a,s} \sum_{j=1}^{n_a} O_{j,s}. \quad (8)$$

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

$$CAIS(S) = \sum_{i=1}^{64} f'_{i,s} \log\left(\frac{O_{i,s}}{E_{i,s}}\right). \quad (9)$$

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. Use of the standardized set of amino acid frequencies F_a has only a small effect on computed CAIS values relative to using each vertebrate species' own amino acid frequencies (Supplementary Figure 5).

CAIS corrected for local intergenic GC content but not species-wide amino acid composition is

$$CAIS_{localGC}(s) = \left(\prod_{r=1}^{G} \sum_{i=1}^{64} O_{i,r} \log\left(\frac{O_{i,r}}{E_{i,r}}\right)\right)^{\frac{1}{L}}, \quad (10)$$

where $O_{i,r}$ is the number of times codon i appears in region r of species s, $E_{i,r}$ is the expected number of times codon i would appear in region r of species s given the local intergenic GC content, G is the number of regions, and $L = \sum_{r=1}^{G} \sum_{i=1}^{64} O_{i,r}$ is the total number of codons in the genome. Rewritten for greater computational ease:

$$CAIS_{localGC}(s) = e^{\frac{1}{L}\sum_{r=1}^{G} \ln\left(\sum_{i=1}^{64} O_{i,r} \log\left(\frac{O_{i,r}}{E_{i,r}}\right)\right)}. \quad (11)$$

Given the limited impact of amino acid frequency correction, we used Equation 11 for the local GC results, but we could correct for amino acid composition by replacing the $O_{i,r}$ prefactor with $f'_{i,s}$, or even $f'_{i,r}$.

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

The expected number of codons is based on the squared deviations X_a^2 of the frequencies of the codons for each amino acid a from null expectations:

$$X_a^2 = \sum_{i=1}^{n_a} \frac{N_a (O_i - E_i)^2}{E_i}, \quad (12)$$

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

$$\widehat{F'}_a = \frac{X_a^2 + N_a - n_a}{n_a(N_a - 1)} \quad (13)$$

and

$$ENC = 2 + \frac{9}{\widehat{F'}_2} + \frac{1}{\widehat{F'}_3} + \frac{5}{\widehat{F'}_4} + \frac{3}{\widehat{F'}_6}, \quad (14)$$

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 (14) to

$$ENC = 2 + \frac{9}{\widehat{F'}_2} + \frac{2}{\widehat{F'}_3} + \frac{5}{\widehat{F'}_4} + \frac{3}{\widehat{F'}_6}. \quad (15)$$

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, but 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

All data and code underlying this article are available in the public repository at https://github.com/MaselLab/Codon-Adaptation-Index-of-Species.

Acknowledgements

This work was supported by the National Institutes of Health (R01 GM104040 and T32 GM132008), the John Templeton Foundation (60814), the Arnold and Mabel Beckman Foundation Scholars Program, the Western Alliance to Expand Student Opportunities (WAESO) Louis Stokes Alliance for Minority Participation (LSAMP) National Science Foundation (NSF) Cooperative Agreement (HRD-1101728), and UA/NASA Space Grant Undergraduate Research Internship program. We thank Luke Kosinski, David Liberles, and Sawsan Wehbi for helpful discussions, Paul Nelson for providing the genome-wide GC contents, and the University of Arizona Undergraduate Biology Research Program for training. We thank Gavin Douglas for writing a convenient end-to-end implementation of CAIS on the basis of our preprint, which can be found at https://github.com/gavinmdouglas/handy_pop_gen/blob/main/CAIS.py, and for catching a minor bug in our code in time for us to correct it in the version of record. We thank

the anonymous reviewers for constructive feedback, and Laurent Duret for helpful elaboration on the concerns of reviewer 1.

Literature Cited

Ahrens, JB, Nunez-Castilla, J, and Siltberg-Liberles, J. 2017. Evolution of intrinsic disorder in eukaryotic proteins. *Cellular and Molecular Life Sciences* 74:3163-3174.

Akashi, H. 1996. Molecular Evolution Between *Drosophila melanogaster* and *D. simulans* Reduced Codon Bias, Faster Rates of Amino Acid Substitution, and Larger Proteins in *D. melanogaster*. *Genetics* 144:1297-1307.

Basile, W, Salvatore, M, Bassot, C, and Elofsson, A. 2019. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Computational Biology* 15:e1007186.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3-17.

Bertram, J, and Masel, J. 2020. Evolution rapidly optimizes stability and aggregation in lattice proteins despite pervasive landscape valleys and mazes. *Genetics* 214:1047-1057.

Botzman, M, and Margalit, H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biology* 12:R109.

Brbić, M, Warnecke, T, Kriško, A, and Supek, F. 2015. Global Shifts in Genome and Proteome Composition Are Very Tightly Coupled. *Genome Biology and Evolution* 7:1519-1532.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897-907.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10:195-205.

Cope, AL, Hettich, RL, and Gilchrist, MA. 2018. Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1860:2479-2485.

Daubin, V, and Moran, NA. 2004. Comment on "The Origins of Genome Complexity". *Science* 306:978-978.

Doherty, A, and McInerney, JO. 2013. Translational Selection Frequently Overcomes Genetic Drift in Shaping Synonymous Codon Usage Patterns in Vertebrates. *Molecular Biology and Evolution* 30:2263-2267.

dos Reis, M, and Wernisch, L. 2009. Estimating Translational Selection in Eukaryotic Genomes. *Molecular Biology and Evolution* 26:451-461.

Doyle, JM, Hacking, CC, Willoughby, JR, Sundaram, M, and DeWoody, JA. 2015. Mammalian genetic diversity as a function of habitat, body size, trophic class, and conservation status. *Journal of Mammalogy* 96:564-572.

Dunker, AK, Garner, E, Guilliot, S, Romero, P, Albrecht, K, Hart, J, Obradovic, Z, Kissinger, C, and Villafranca, JE. 1998. *Protein disorder and the evolution of molecular recognition: theory, predictions and observations*. Pp. 473-484. Pac Symp Biocomput.

Duret, L, Eyre-Walker, A, and Galtier, N. 2006. A new perspective on isochore evolution. *Gene* 385:71-74.

Duret, L, and Galtier, N. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics* 10:285-311.

Eyre-Walker, A, and Hurst, LD. 2001. The evolution of isochores. *Nature Reviews Genetics* 2:549-555.

Felsenstein, J. 1985. Phylogenies and the Comparative Method. *The American Naturalist* 125:1-15.

Figuet, E, Ballenghien, M, Romiguier, J, and Galtier, N. 2014. Biased Gene Conversion and GC-Content Evolution in the Coding Sequences of Reptiles and Vertebrates. *Genome Biology and Evolution* 7:240-250.

Forcelloni, S, and Giansanti, A. 2020. Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome. *Journal of Molecular Evolution* 88:164-178.

Foy, SG, Wilson, BA, Bertram, J, Cordes, MHJ, and Masel, J. 2019. A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution. *Genetics* 211:1345-1355.

Fuglsang, A. 2004. The 'effective number of codons' revisited. *Biochemical and Biophysical Research Communications* 317:957-964.

Fuglsang, A. 2008. Impact of bias discrepancy and amino acid usage on estimates of the effective number of codons used in a gene, and a test for selection on codon usage. *Gene* 410:82-88.

Galtier, N, Piganeau, G, Mouchiroud, D, and Duret, L. 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* 159:907-911.

Galtier, N, Roux, C, Rousselle, M, Romiguier, J, Figuet, E, Glémin, S, Bierne, N, and Duret, L. 2018. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution* 35:1092-1103.

Gingold, H, Tehler, D, Christoffersen, Nanna R, Nielsen, Morten M, Asmar, F, Kooistra, Susanne M, Christophersen, Nicolaj S, Christensen, LL, Borre, M, Sørensen, Karina D, Andersen, Lars D, Andersen, Claus L, Hulleman, E, Wurdinger, T, Ralfkiær, E, Helin, K, Grønbæk, K, Ørntoft, T, Waszak, Sebastian M, Dahan, O, . . . Pilpel, Y. 2014. A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. *Cell* 158:1281-1292.

Gossmann, TI, Keightley, PD, and Eyre-Walker, A. 2012. The Effect of Variation in the Effective Population Size on the Rate of Adaptive Molecular Evolution in Eukaryotes. *Genome Biology and Evolution* 4:658-667.

Hershberg, R, and Petrov, DA. 2008. Selection on Codon Bias. *Annual Review of Genetics* 42:287-299.

Hershberg, R, and Petrov, DA. 2009. General Rules for Optimal Codon Choice. *PLoS Genetics* 5:e1000556.

Hildebrand, F, Meyer, A, and Eyre-Walker, A. 2010. Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genetics* 6:e1001107.

Huang, Y, and Liu, Z. 2013. Do Intrinsically Disordered Proteins Possess High Specificity in Protein–Protein Interactions? *Chemistry – A European Journal* 19:4462-4467.

Hunt, RC, Simhadri, VL, Iandoli, M, Sauna, ZE, and Kimchi-Sarfaty, C. 2014. Exposing synonymous mutations. *Trends In Genetics* 30:308-321.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

James, J, Nelson, P, and Masel, J. 2023. Differential retention of Pfam domains contributes to long-term evolutionary trends. *Molecular Biology and Evolution*:msad073.

James, JE, Willis, SM, Nelson, PG, Weibel, C, Kosinski, LJ, and Masel, J. 2021. Universal and taxon-specific trends in protein sequences as a function of age. *eLife* 10:e57347.

Kessler, MD, and Dean, MD. 2014. Effective population size does not predict codon usage bias in mammals. *Ecology and Evolution* 4:3887-3900.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713-719.

Kondrashov, AS, Povolotskaya, IS, Ivankov, DN, and Kondrashov, FA. 2010. Rate of sequence divergence under constant selection. *Biology Direct* 5:5.

Kosinski, LJ, Aviles, NR, Gomez, K, and Masel, J. 2022. Random Peptides Rich in Small and Disorder-Promoting Amino Acids Are Less Likely to Be Harmful. *Genome Biology and Evolution* 14:evac085.

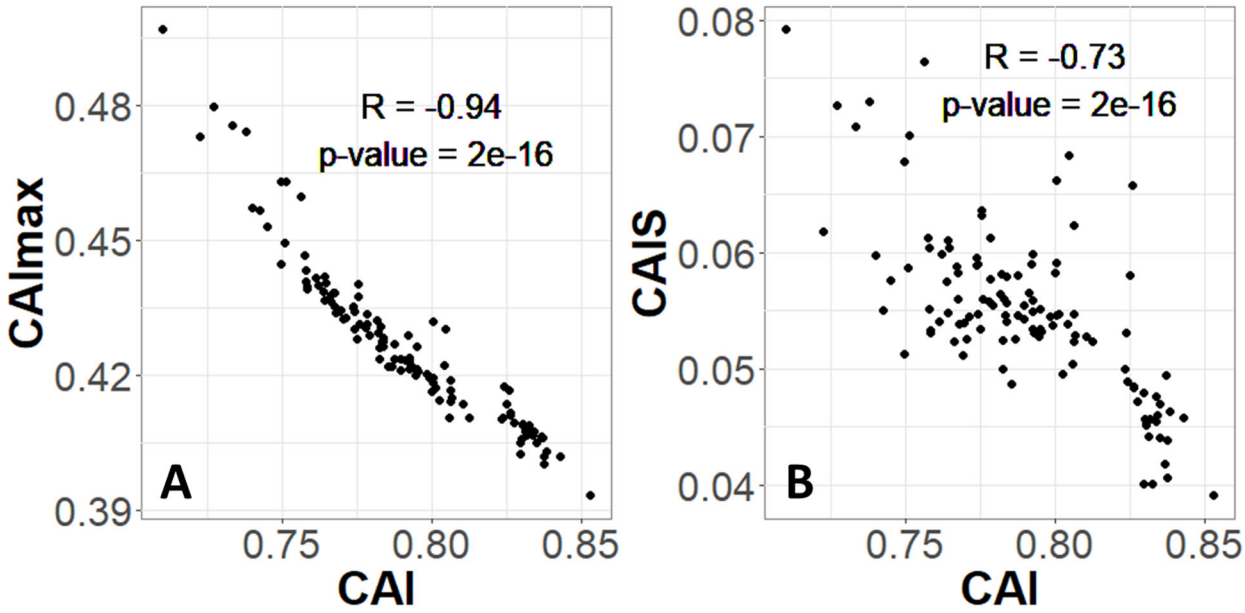
Kullback, S, and Leibler, RA. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22:79-86, 8.

539 Kumar, S, Stecher, G, Suleski, M, and Hedges, SB. 2017. TimeTree: A Resource for timelines, timetrees,
540 and divergence times. *Molecular Biology and Evolution* 34:1812-1819.
541 Lazar, T, Tantos, A, Tompa, P, and Schad, E. 2022. Intrinsic protein disorder uncouples affinity from
542 binding specificity. *Protein Science* 31:e4455.
543 Li, W-H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of
544 synonymous codons. *Journal of Molecular Evolution* 24:337-345.
545 Liu, SS, Hockenberry, AJ, Jewett, MC, and Amaral, LAN. 2018. A novel framework for evaluating the
546 performance of codon usage bias metrics. *Journal of The Royal Society Interface* 15:20170667.
547 Long, H, Sung, W, Kucukyildirim, S, Williams, E, Miller, SF, Guo, W, Patterson, C, Gregory, C, Strauss, C,
548 Stone, C, Berne, C, Kysela, D, Shoemaker, WR, Muscarella, ME, Luo, H, Lennon, JT, Brun, YV, and
549 Lynch, M. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nature*
550 *Ecology & Evolution* 2:237-240.
551 Lynch, M, Ackerman, MS, Gout, J-F, Long, H, Sung, W, Thomas, WK, and Foster, PL. 2016. Genetic drift,
552 selection and the evolution of the mutation rate. *Nature Reviews Genetics* 17:704-714.
553 Meunier, J, and Duret, L. 2004. Recombination Drives the Evolution of GC-Content in the Human
554 Genome. *Molecular Biology and Evolution* 21:984-990.
555 Novembre, JA. 2002. Accounting for Background Nucleotide Composition When Measuring Codon Usage
556 Bias. *Molecular Biology and Evolution* 19:1390-1394.
557 Novoa, EM, Jungreis, I, Jaillon, O, and Kellis, M. 2019. Elucidation of Codon Usage Signatures across the
558 Domains of Life. *Molecular Biology and Evolution* 36:2328-2339.
559 Ohta, T. 1972. Population size and rate of evolution. *Journal of Molecular Evolution* 1:305-314.
560 Ohta, T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* 246:96-98.
561 Ohta, T. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and*
562 *Systematics* 23:263-286.
563 Paradis, E, and Schliep, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary
564 analyses in R. *Bioinformatics* 35:526-528.
565 Plotkin, JB, Dushoff, J, Desai, MM, and Fraser, HB. 2006. Codon Usage and Selection on Proteins. *Journal*
566 *of Molecular Evolution* 63:635-653.
567 Plotkin, JB, and Kudla, G. 2011. Synonymous but not the same: the causes and consequences of codon
568 bias. *Nature Reviews Genetics* 12:32-42.
569 Rocha, EPC. 2004. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient
570 decoding for translation optimization. *Genome Research* 14:2279-2286.
571 Rohlf, FJ. 2006. A Comment on Phylogenetic Correction. *Evolution* 60:1509-1515.
572 Romiguier, J, Gayral, P, Ballenghien, M, Bernard, A, Cahais, V, Chenuil, A, Chiari, Y, Dernat, R, Duret, L,
573 Faivre, N, Loire, E, Lourenco, JM, Nabholz, B, Roux, C, Tsagkogeorga, G, Weber, AAT, Weinert,
574 LA, Belkhir, K, Bierne, N, Glémin, S, . . . Galtier, N. 2014. Comparative population genomics in
575 animals uncovers the determinants of genetic diversity. *Nature* 515:261-263.
576 Romiguier, J, and Roux, C. 2017. Analytical Biases Associated with GC-Content in Molecular Evolution.
577 *Frontiers in Genetics* 8:16.
578 Schad, E, Tompa, P, and Hegyi, H. 2011. The relationship between proteome size, structural disorder and
579 organism complexity. *Genome Biology* 12:R120.
580 Sharp, PM, Bailes, E, Grocock, RJ, Peden, JF, and Sockett, RE. 2005. Variation in the strength of selected
581 codon usage bias among bacteria. *Nucleic Acids Research* 33:1141-1153.
582 Sharp, PM, Emery, LR, and Zeng, K. 2010. Forces that influence the evolution of codon bias. *Philosophical*
583 *Transactions of the Royal Society B: Biological Sciences* 365:1203-1212.
584 Sharp, PM, and Li, W-H. 1987. The codon adaptation index-a measure of directional synonymous codon
585 usage bias, and its potential applications. *Nucleic Acids Research* 15:1281-1295.

586 Smith, NGC, and Eyre-Walker, A. 2001. Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C-
587 Rich Genes in Humans. *Molecular Biology and Evolution* 18:982-986.
588 Subramanian, S. 2008. Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes.
589 *Genetics* 178:2429-2432.
590 Sun, X, Yang, Q, and Xia, X. 2012. An Improved Implementation of Effective Number of Codons (Nc).
591 *Molecular Biology and Evolution* 30:191-196.
592 Theillet, F-X, Kalmar, L, Tompa, P, Han, K-H, Selenko, P, Dunker, AK, Daughdrill, GW, and Uversky, VN.
593 2013. The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins* 1:e24360.
594 Urrutia, AO, and Hurst, LD. 2001. Codon Usage Bias Covaries With Expression Breadth and the Rate of
595 Synonymous Evolution in Humans, but This Is Not Evidence for Selection. *Genetics* 159:1191-
596 1199.
597 Vavouri, T, Semple, JI, Garcia-Verdugo, R, and Lehner, B. 2009. Intrinsic Protein Disorder and Interaction
598 Promiscuity Are Widely Associated with Dosage Sensitivity. *Cell* 138:198-208.
599 Vicario, S, Moriyama, EN, and Powell, JR. 2007. Codon usage in twelve species of Drosophila. *BMC*
600 *Evolutionary Biology* 7:226.
601 Wangen, JR, and Green, R. 2020. Stop codon context influences genome-wide stimulation of termination
602 codon readthrough by aminoglycosides. *eLife* 9:e52611.
603 Wright, F. 1990. The 'effective number of codons' used in a gene. *Gene* 87:23-29.
604 Xue, B, Dunker, AK, and Uversky, VN. 2012. Orderly order in protein intrinsic disorder distribution:
605 disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular*
606 *Structure and Dynamics* 30:137-149.
607 Zhang, Z, Li, J, Cui, P, Ding, F, Li, A, Townsend, JP, and Yu, J. 2012. Codon Deviation Coefficient: a novel
608 measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics*
609 13:43.
610

611

612



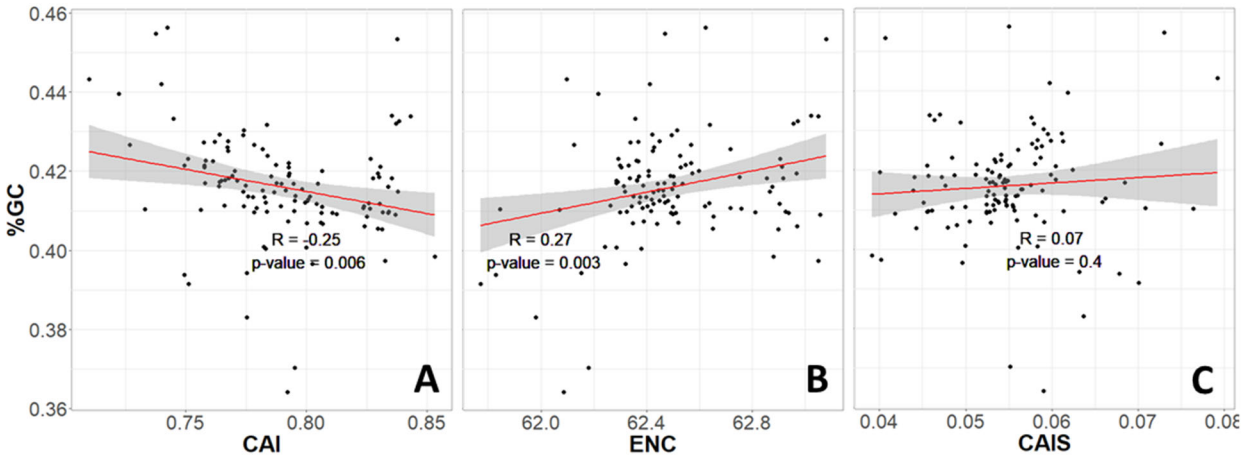
613

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

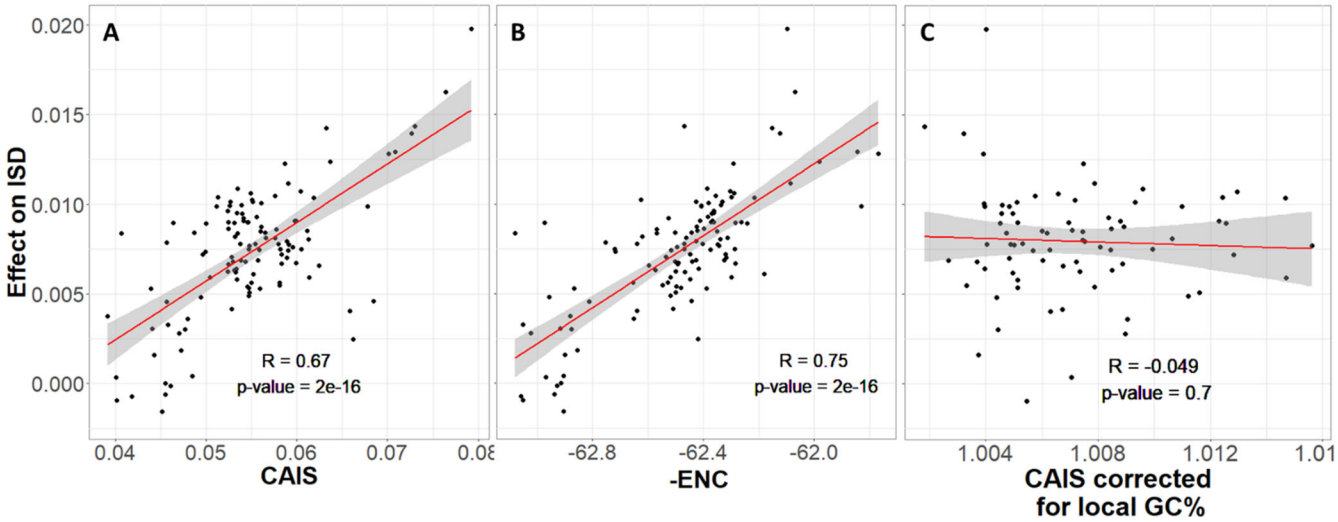
621

622

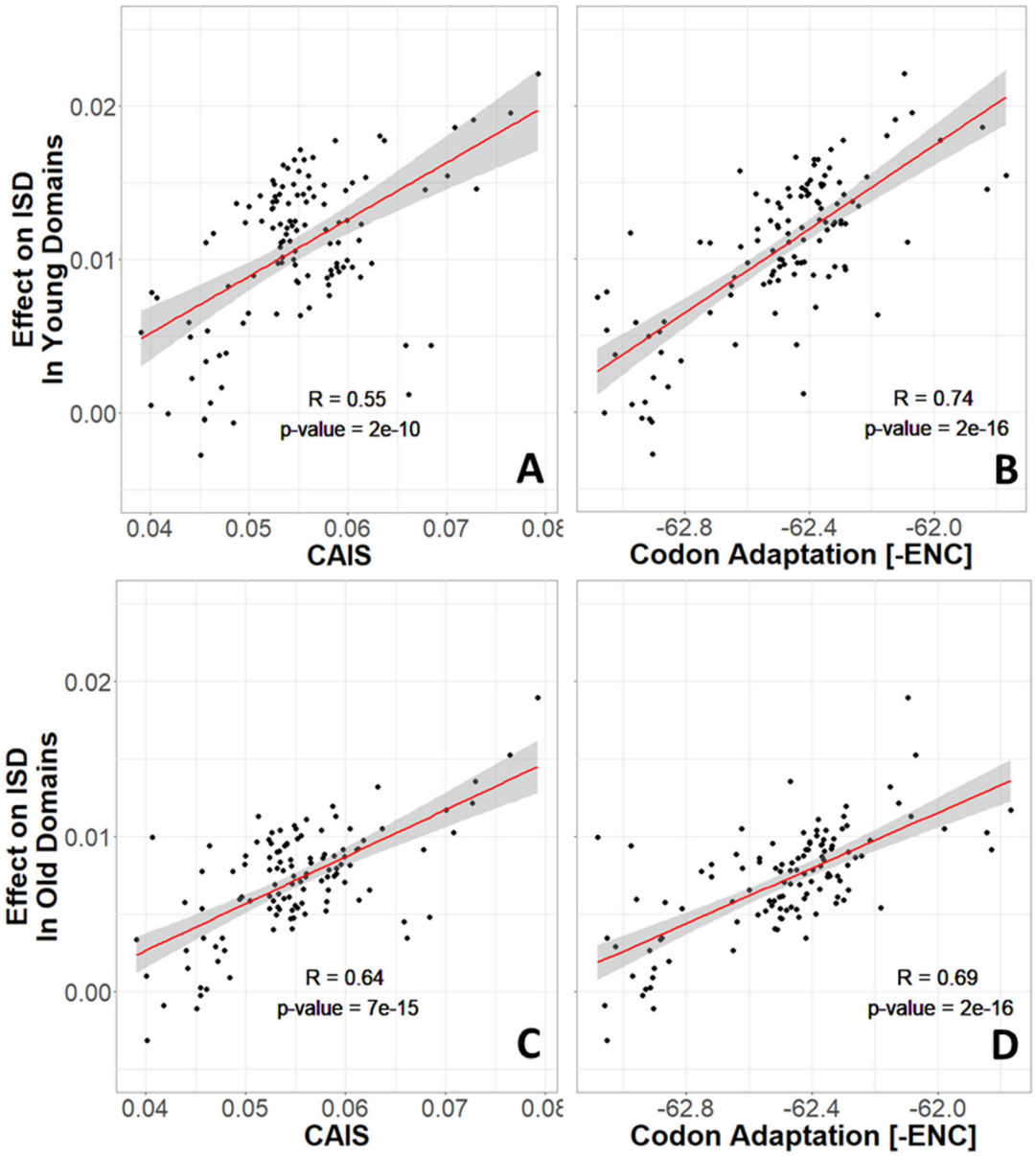
623



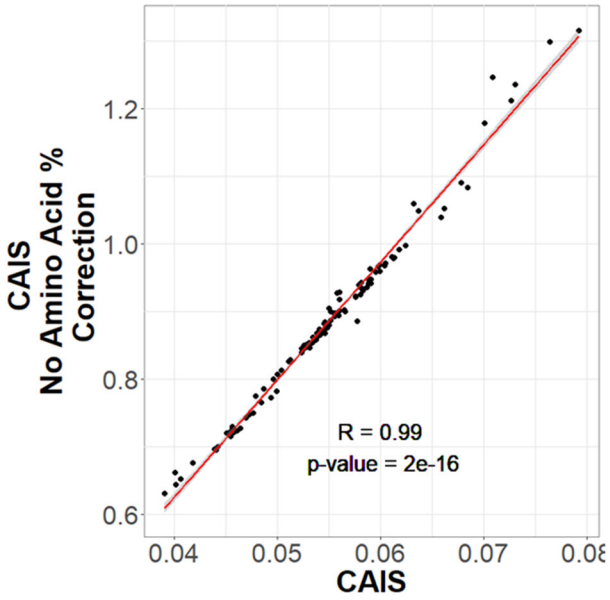
Supplementary Figure 2: The same relationships are shown as in Figure 3, here without correction for phylogenetic confounding. CAI and ENC both corelate with genomic GC, but CAIS does not. Red line shows lm(y~x), with grey region as 95% confidence interval. We use PIC corrected results rather than these results because PIC correction removes non-independent errors to produce an unbiased R^2 estimate (Rohlf 2006).



Supplementary Figure 3: The same relationships are shown as in Figure 4, here without correction for phylogenetic confounding. As in Figure 4, ISD of protein domains is higher in more highly adapted species, as measured by CAIS and ENC, but not by CAIS calculated with local GC% rather than genome-wide GC%. ISD is calculated as in Figure 4. Red line shows lm(y~x), with grey region as 95% confidence interval.



Supplementary Figure 4: Without correction for phylogenetic confounding, more highly adapted species have higher ISD in both young and old protein domains. Age assignments and ISD effects are calculated as in Figure 6. same n=118 datapoints are shown as in Figures 3-5. Red line shows lm(y~x), with grey region as 95% confidence interval.



Supplementary Figure 5: Vertebrate CAIS values are not greatly affected by computation for a standardized amino acid composition vs. computation for the amino acid frequencies in the species in question.