Evolutionary Biology
Genetics and Genomics

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

Catherine A. Weibel, Andrew L. Wheeler, Jennifer E. James, Sara M. Willis, Joanna Masel

Department of Mathematics, University of Arizona, Tucson, Arizona 85721, USA • Department of Physics, University of Arizona, Tucson, Arizona 85721, USA • Department of Applied Physics, Stanford University, California, USA • Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, Arizona 85721, USA • Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA • Department of Ecology and Genetics, Evolutionary Biology Center, Uppsala University, Sweden • University Information Technology Services, University of Arizona, Tucson, Arizona 85721, USA

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted
July 26, 2023 (this version)

Sent for peer review
March 15, 2023

Posted to bioRxiv
March 3, 2023

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.

eLife assessment

In this intriguing study, the authors offer a new metric, Codon Adaptation Index of Species (CAIS), which allows one to determine the strength of selection and effective population sizes using data on GC content and amino acid composition. This could be of broad use in molecular evolution, as it could be applied across species. The study offers **important** findings that may aid in our ability to compute key population genomic parameters from genomic data, and the conclusions are based on **solid** evidence.

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**).

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

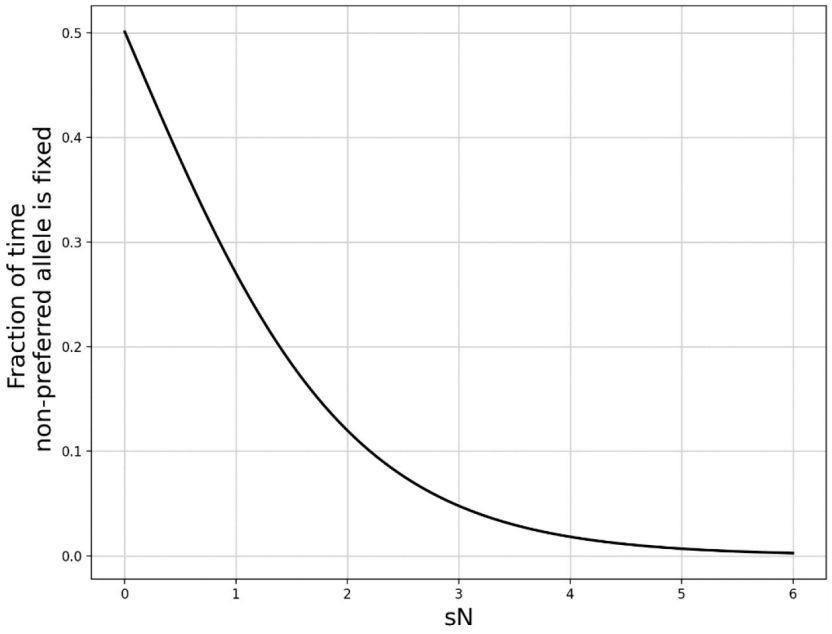


Figure 1

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 $\pi(N,s)=\frac{1-e^{-2}}{1-e^{-4N}}$, and the y-axis is calculated as *π(N, —s)/(π(N, —s) + π(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.

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).

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

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

Here, we extend the CAI so that it corrects for both GC and amino acid composition (see Methods) to create a new Codon Adaptation Index of Species (CAIS), and compare its performance to that of the ENC. The availability of complete genomes allows both metrics to be readily calculated without data on polymorphism or mutation rate, without selecting reference genes, and without concerns about demographic history. Our purpose is to find an accessible metric that can quantify the limits to weak selection important to nearly neutral theory; this differs from past evaluations focused on comparing different genes of the same species and recapitulating “ground truth” simulations thereof (Sun et al. 2012; Zhang et al. 2012; Liu et al. 2018). Our first measure of success is to produce the expected correlation with body size (used here as a proxy for census population size

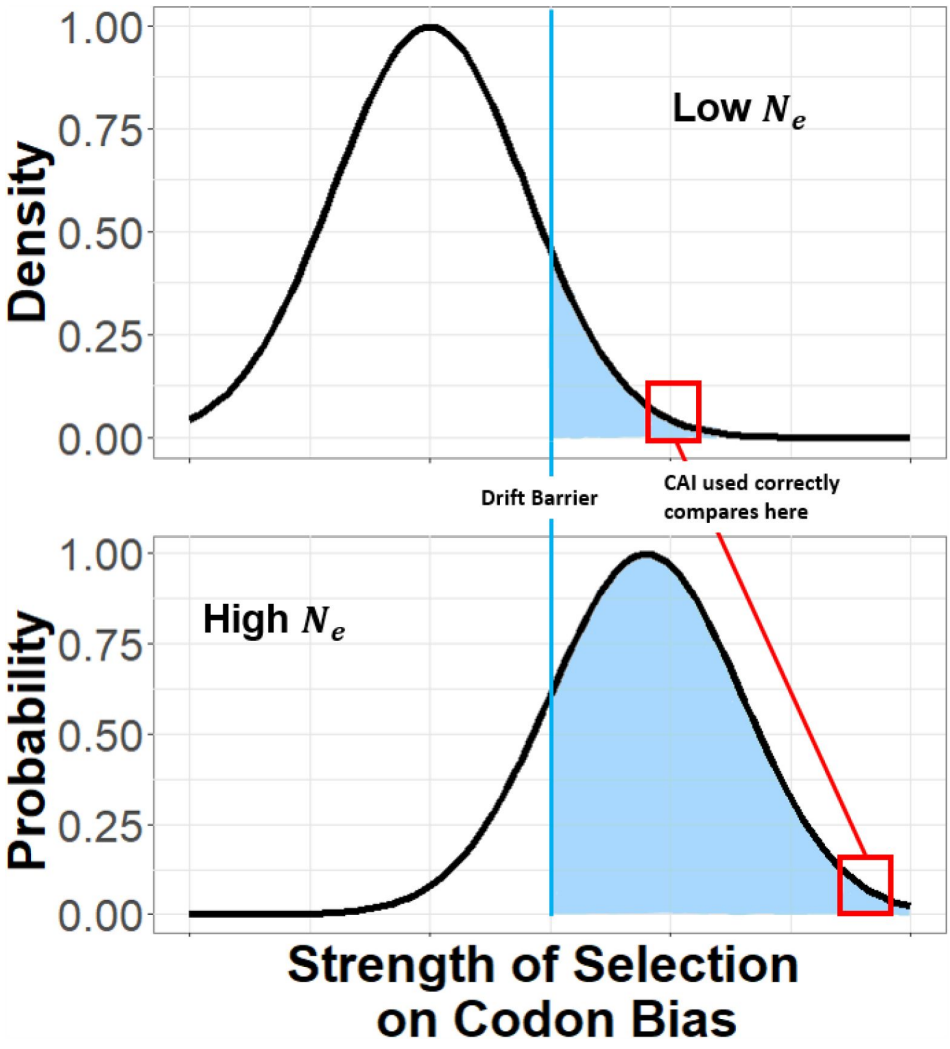


Figure 2

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.

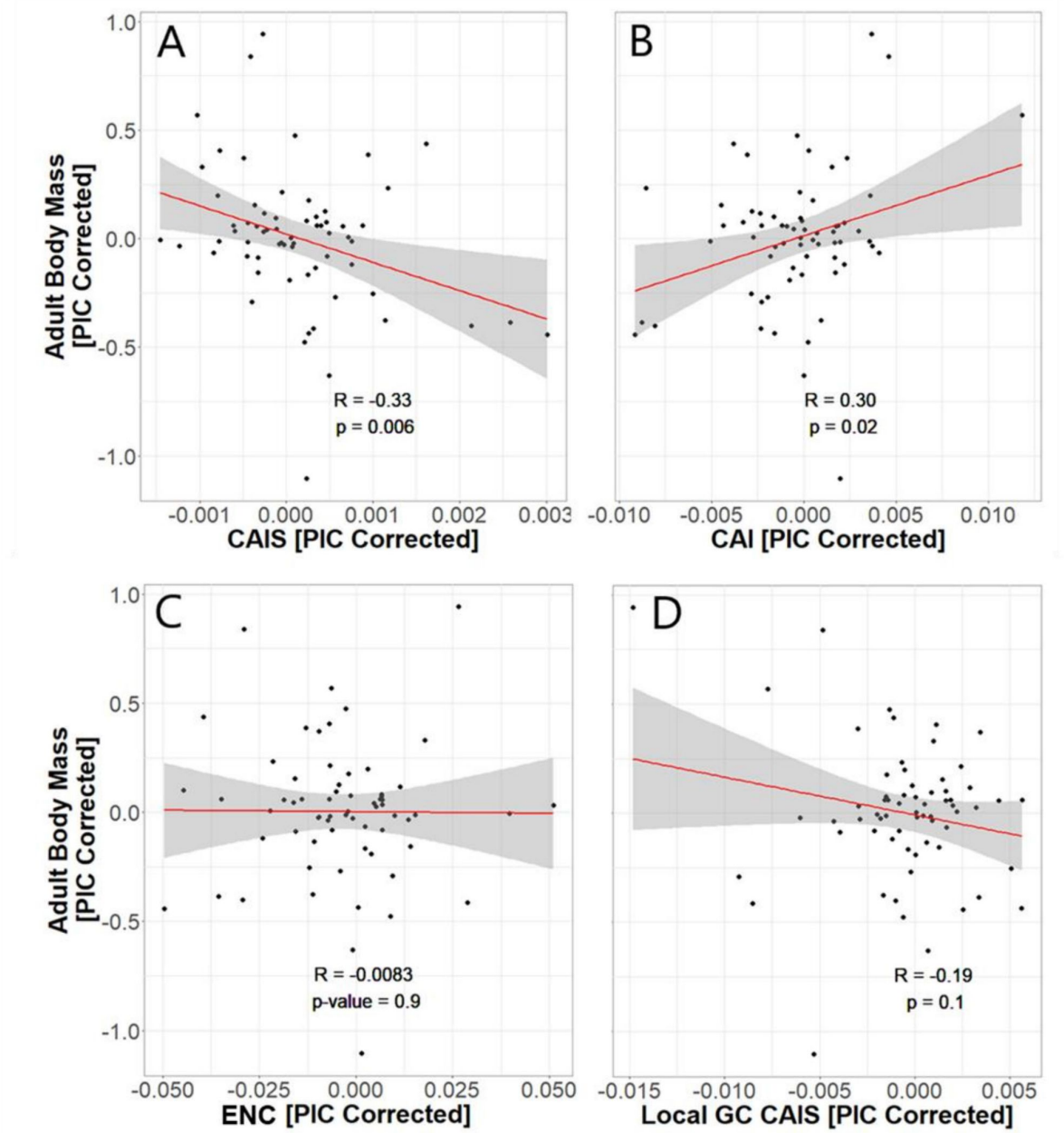


Figure 3

CAIS reflects the expected relationship between effectiveness of selection and body size, while CAI, ENC, and CAIS calculated using local GC content do not. Note that a high ENC value means more codons are being used in the genome of the given species, so that the given species is less codon adapted, while a high CAIS value means that a species is more codon adapted. Body size data are from PanTHERIA database, originally in $\log_{10}$(mass) in grams prior to PIC correction. Data are shown for 62 species in common between PanTHERIA and our own dataset of 118 vertebrate species that have both "Complete" genome sequence available for calculating %GC, and TimeTree divergence dates available for use in PIC correction. P-values shown are for Pearson's correlation. Red line shows unweighted Im(y~x) with grey region indicated 95% confidence interval.

and hence the expected exquisiteness of adaptation). We use vertebrates as a challenging test case, given relatively weak selection on codon usage (Doherty & McInerney 2013; Galtier et al. 2018). Correlations between codon adaptation and body size across vertebrates have previously been elusive, which is surprising given that the expected differences in the efficiency of selection between X-chromosomes vs. autosomes and between high-recombination vs. low recombination regions are found (Kessler & Dean 2014). Our second measure of success is a demonstration of how our metric can be used to find novel correlations pointing to what else might be subject to weak selective preferences at the molecular level.

Results

Species with larger body mass have smaller CAIS

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

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

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

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

Proteins in better adapted species evolve more structural disorder

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

However, compositional differences among proteomes might be driven by the recent birth of ISD-rich proteins in animals (James et al. 2021), and/or by differences among sequences in their subsequent tendency to proliferate into many different genes (James et al. 2022), rather than by differences in how a given protein sequence evolves as a function of the effectiveness of selection. To focus only on the latter, we use a linear mixed model, with each species having a fixed effect on ISD, while controlling for Pfam domain identity as a random effect. We note that once GC is controlled for, codon adaptation can be assessed similarly in intrinsically disordered vs. ordered proteins (Gossmann et al. 2012). Controlling for Pfam identity is supported, with standard deviation in ISD among Pfams estimated as 0.178 in comparison to residual standard deviation of 0.058 and a p-value on the significance of the Pfam random effect term of $3\times10'^{13}$. We then ask whether the fixed species effects on ISD are correlated with CAIS.

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

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

Discussion

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

CAIS can be used to detect which properties are under selection. To accomplish this, we note that when different properties are each causally affected by a species’ exquisiteness of adaptation, this will create a correlation between the properties. We use codon adaptation as a reference property, such that correlations with codon adaptation indicate selection. We operationalize this using a linear mixed model approach that controls for Pfam identity as a random effect, and then apply

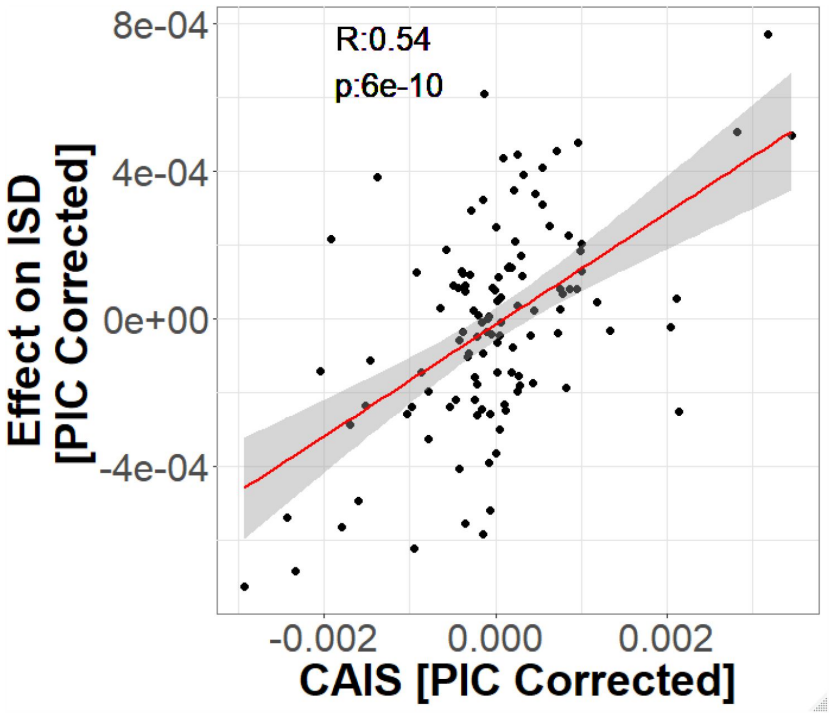


Figure 4

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.

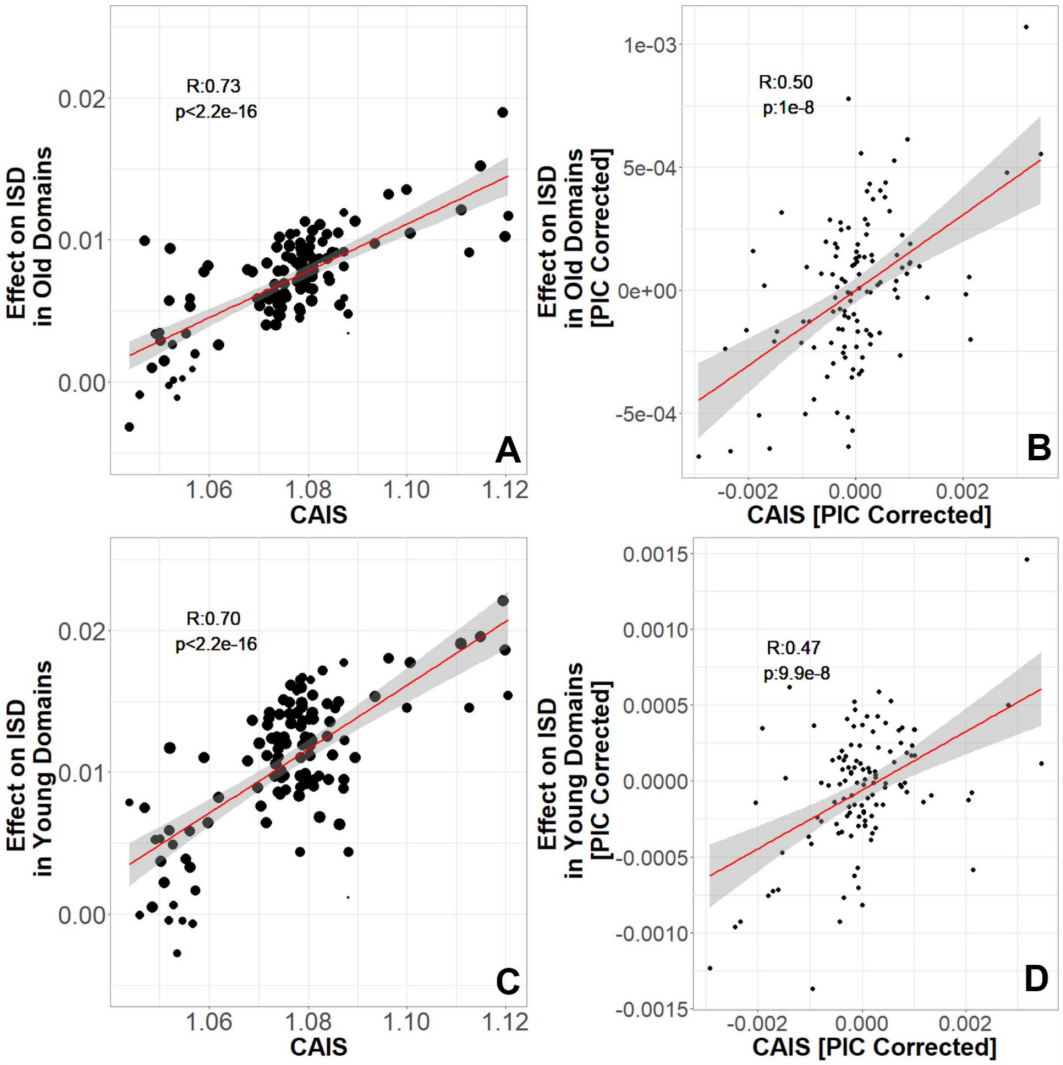


Figure 5

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-a×is are fixed effects of species identity in our linear mixed model. The same n=H8 datapoints are shown as in Figures 3 and 4. P-values shown are for Pearson's correlation. Red line shows Im(y~×), 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.

phylogenetic correction to the resulting fixed effects of species on ISD. This approach shows that the same Pfam domain tends to be more disordered when found in a well-adapted species (i.e. a species with a higher CAIS). This is true for both ancient and recently emerged protein domains, albeit of slightly larger magnitude for the latter.

It is important that no additional variable such as GC content creates a spurious correlation by affecting both CAIS and our property of interest. For this reason, we control for GC content during the construction of CAIS. GC content is the product of many different processes, most notably the genome-wide forces of mutation bias and gene conversion (Romiguier & Roux 2017), but potentially also selection on individual nucleotide substitutions that is hypothesized to favor higher %GC (Long et al. 2018). By controlling for %GC, we exclude all these forces from influencing CAIS. CAIS thus captures the extent of adaptation in codon bias, including translational speed, accuracy, and any intrinsic preference for GC over AT that is specific to coding regions. These remaining codon-adaptive factors do not create a statistically significant correlation between CAIS and GC (**Supplementary Figures 5A, 5B**), nor between ENC and GC (**Supplementary Figures 5C, 5D**), although CAI is correlated with GC (**Supplementary Figures 5E, F**).

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

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

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

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

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

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

Our new CAIS metric more directly quantifies how species vary in their exquisiteness of adaptation than estimates of effective population size that are based on neutral polymorphism. It can also be estimated for far more species because it does not require polymorphism or 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:

$$RSCU_i = \frac{N_i}{\frac{1}{n_a}\sum_{j=1}^{n_a} N_j} \quad (1)$$

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_t for each codon, relative to the best adapted codon for that amino acid:

$$w_i \equiv \frac{RSCU_i}{RSCU_{max}} \quad (2)$$

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 l5OObp 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

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

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

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

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:

$$RSCUS_{i,r,s} = \frac{O_{i,r}}{E_{i,r}} \quad (7)$$

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

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

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 $a_{a,s}$ so that

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

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

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

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

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

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

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

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:

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

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

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

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

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

and the Effective Number of Codons as

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

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

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

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).

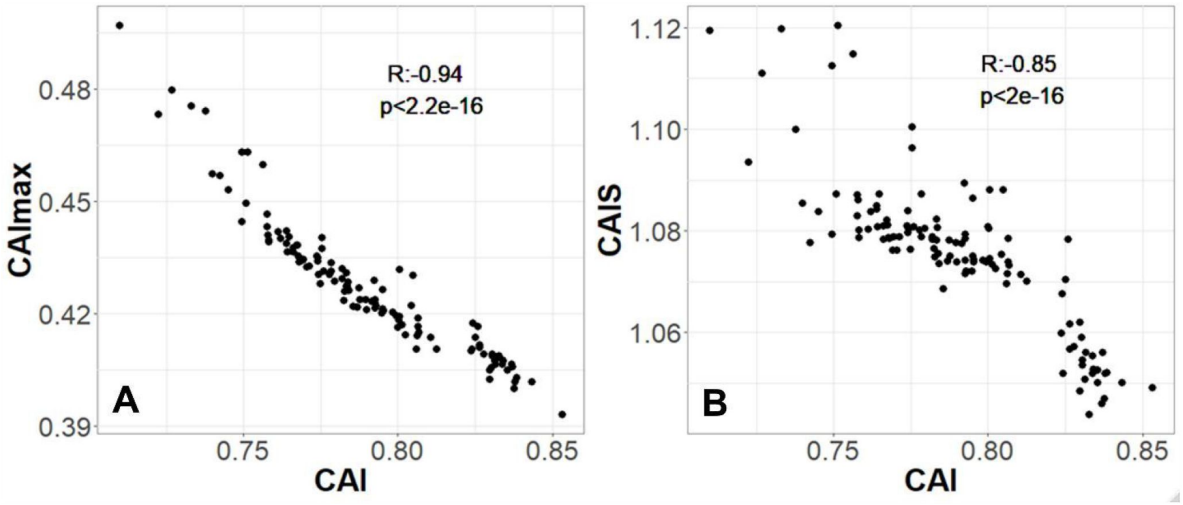
Data Availability

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

Acknowledgements

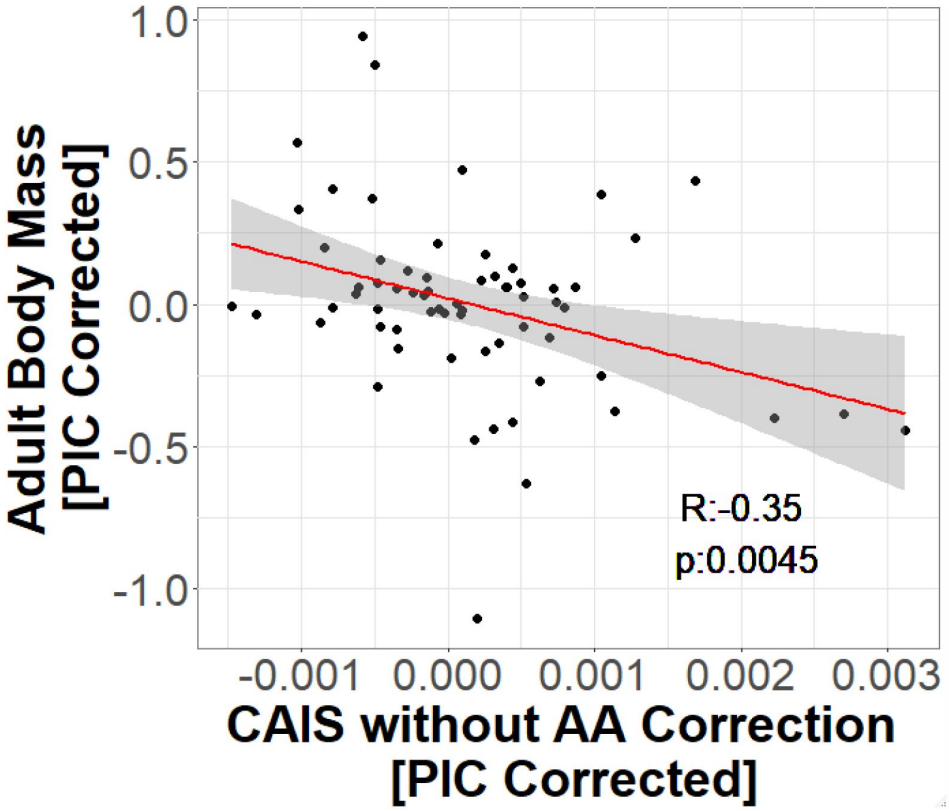
This work was supported by the National Institutes of Health (R01 GM104040 and T32 GM132008), the John Templeton Foundation (60814), the Arnold and Mabel Beckman Foundation Scholars Program, the Western Alliance to Expand Student Opportunities (WAESO) Louis Stokes Alliance for Minority Participation (LSAMP) National Science Foundation (NSF) Cooperative Agreement (HRD-

1101728), and UA/NASA Space Grant Undergraduate Research Internship program. Wethank Luke Kosinski, Hanon McShea, David Liberies, 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.



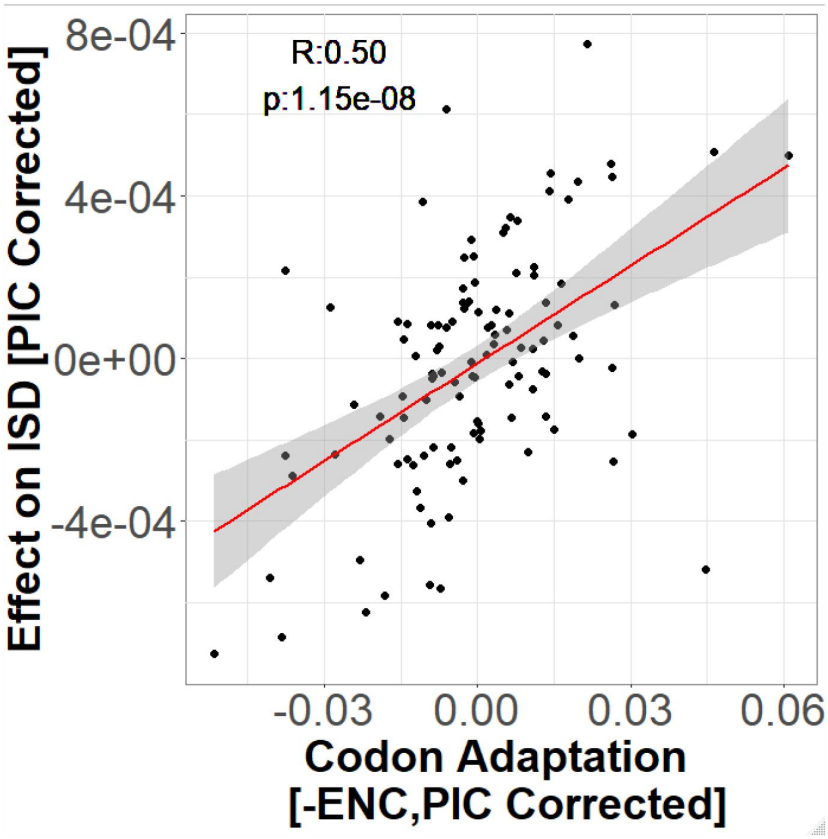
Supplementary Figure 1

Codon Adaptation Index is not appropriate for species-wide effectiveness of selection measurements. Each CAI value shown is averaged over an entire species' proteome. A) The value of CAI is driven by its normalizing denominator term, CAI_{max}. B) As a result, CAI is inversely proportional to CAIS. 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). P-values shown are for Pearson's correlation.



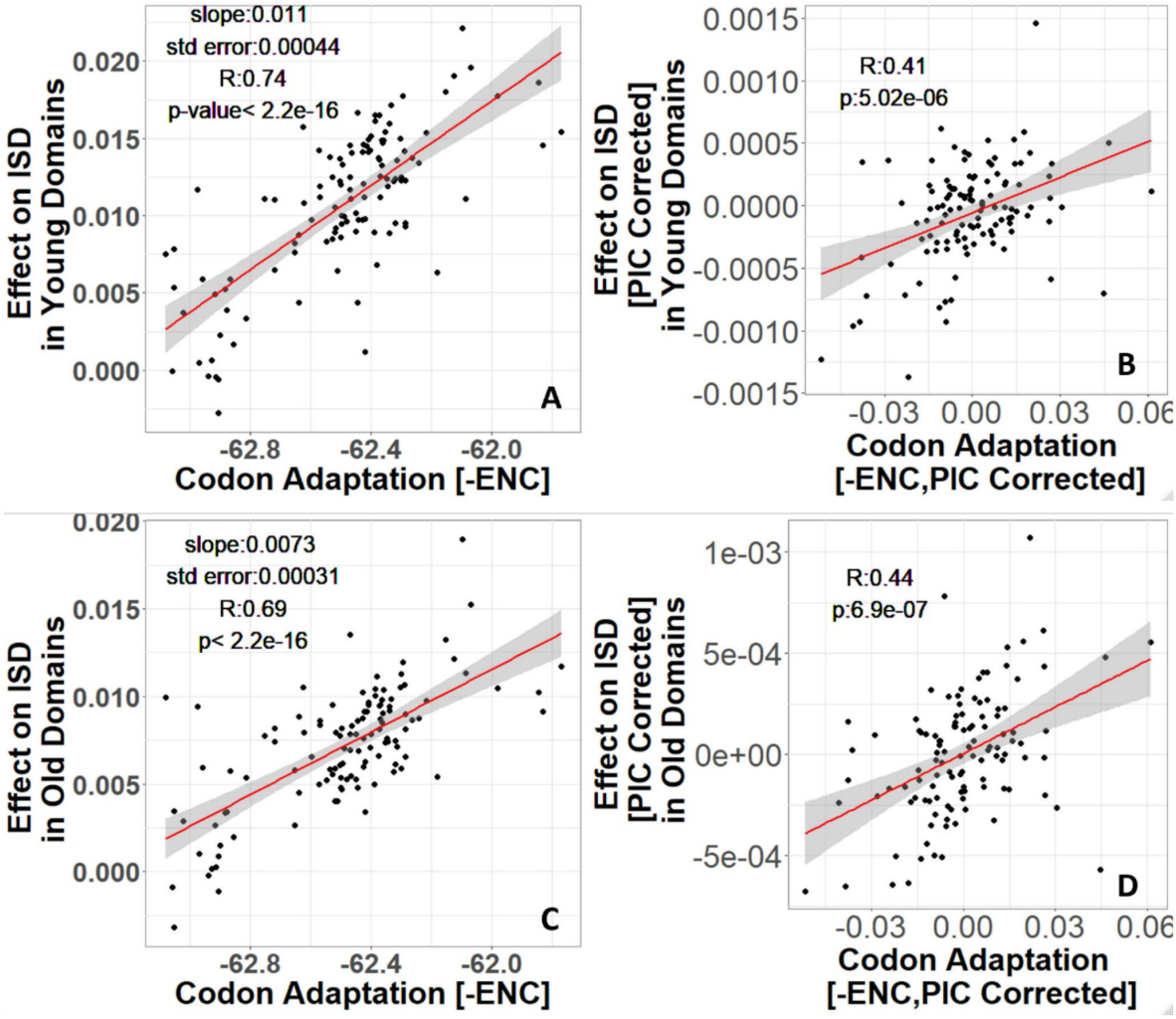
Supplementary Figure 2

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 Im(y~x) with grey region as 95% confidence interval.



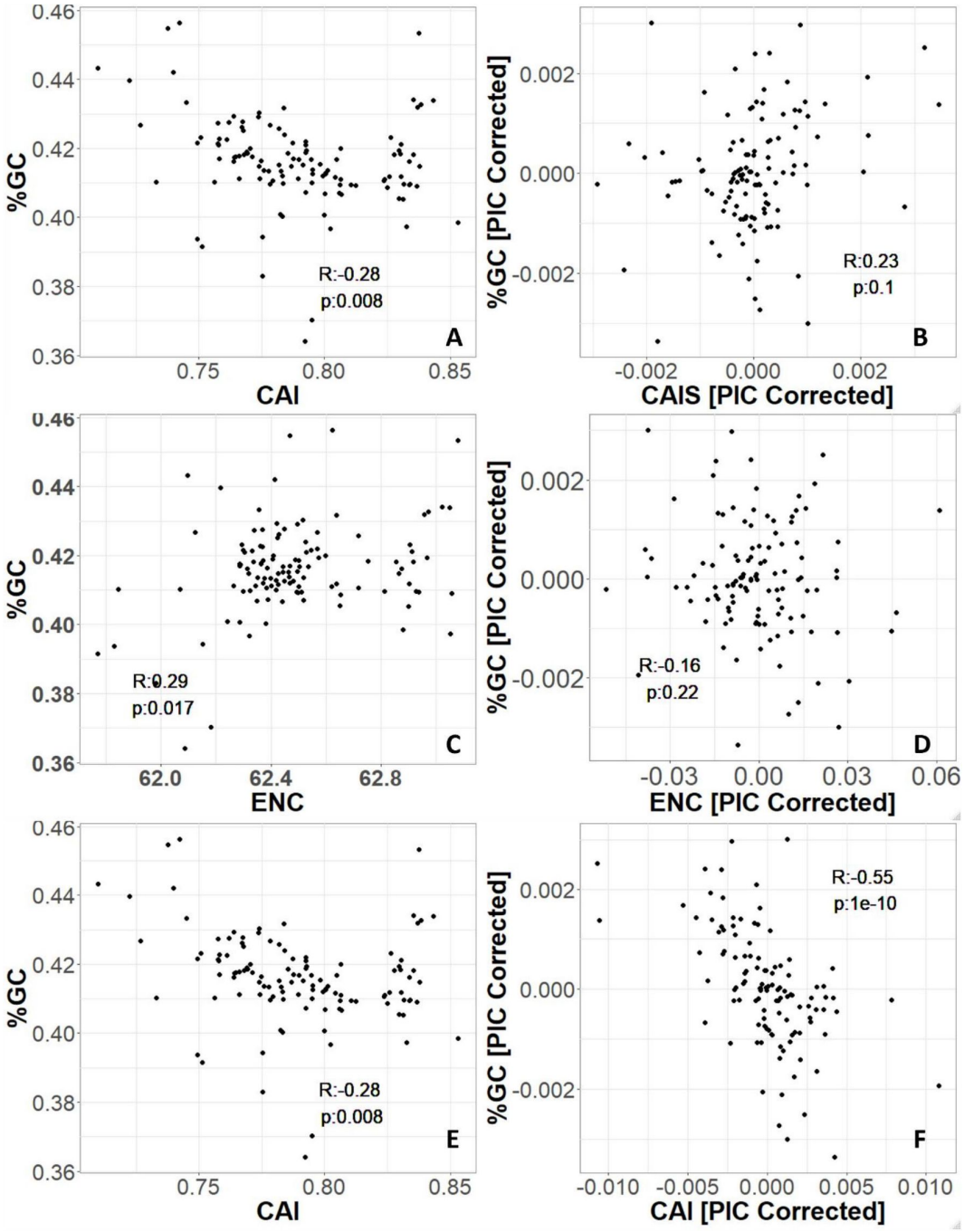
Supplementary Figure 3

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 Im(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=H8 datapoints are shown as in **Figure 3**. P-values shown are for Pearson's correlation. Red line shows Im(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.

References

Ahrens JB, Nunez-Castilla J, 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, Elofsson A. (2019) **Why do eukaryotic proteins contain more intrinsically disordered regions?** *PLoS Computational Biology*

Bernardi G. (2000) **Isochores and the evolutionary genomics of vertebrates** *Gene* **241**:3–17

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

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

Brbić M, Warnecke T, Kriško A, 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**

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

Daubin V, Moran NA (2004) **Comment on "The Origins of Genome Complexity"** *Science* **306**

Doherty A, 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., Wernisch L. (2008) **Estimating Translational Selection in Eukaryotic Genomes** *Molecular Biology and Evolution* **26**:451–461

Doyle JM, Hacking CC, Willoughby JR, Sundaram M, 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, Villafranca JE (1998) **Protein disorder and the evolution of molecular recognition: theory, predictions and observations** *Pac Symp Biocomput* :473–484

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

Duret L, 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, 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, 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

Figuet E, Nabholz B, Bonneau M, Mas Carrio E., Nadachowska-Brzyska K, Ellegren H, Galtier N. (2016) **Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes** *Molecular Biology and Evolution* **33**:1517–1527

Forcelloni S, 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, 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, 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, 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

Gossmann TI, Keightley PD, 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, Petrov DA (2008) **Selection on Codon Bias** *Annual Review of Genetics* **42**:287–299

Hershberg R, Petrov DA (2009) **General Rules for Optimal Codon Choice** *PLoS Genetics* **5**

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

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

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

James J, Nelson P, Masel J. (2022) **Differential retention of Pfam domains creates long-term evolutionary trends** *bioRxiv:2O22* **10**

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

Jansen R, Bussemaker HJ, Gerstein M. (2003) **Revisiting the codon adaptation index from a wholegenome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models** *Nucleic Acids Research* **31**:2242–2251

Kern AD, Hahn MW (2018) **The Neutral Theory in Light of Natural Selection** *Molecular Biology and Evolution* **35**:1366–1371

Kessler MD, 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

Kosinski LJ, Aviles NR, Gomez K, Masel J. (2022) **Random Peptides Rich in Small and DisorderPromoting Amino Acids Are Less Likely to Be Harmful** *Genome Biology and Evolution*

Kumar S, Stecher G, Suleski M, Hedges SB (2017) **TimeTree: A Resource for timelines, timetrees, and divergence times** *Molecular Biology and Evolution* **34**:1812–1819

LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. (2019) **Variation and selection on codon usage bias across an entire subphylum** *PLoS Genetics*

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendi MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J-F, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen H-C, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang S-P, Yeh R-F, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Patrinos A, Morgan MJ, International Human Genome Sequencing, C, Whitehead Institute for Biomedical Research, CfGR, The Sanger, C. Washington University Genome Sequencing, C, Institute, UDJG, Baylor College of Medicine Human Genome Sequencing, C, Center, RGS, Genoscope, Cnrs, UMR, Department of Genome Analysis, loMB, Center, GTCS, Beijing Genomics Institute/Human Genome, C. Multimegabase Sequencing Center, TlfSB, Stanford Genome Technology, C, University of Oklahoma's Advanced Center for Genome, T, Max Planck Institute for Molecular, G, Cold Spring Harbor Laboratory, LAHGC, Biotechnology, GBGRCf, *Genome Analysis, G, Scientific management: National Human Genome Research Institute, USNIoH, Stanford Human Genome, C, University of Washington Genome, C, Department of Molecular Biology, KUSoM, University of Texas Southwestern Medical Center at, D, Office of Science, USDoE, and The Wellcome, T (2001) **Initial sequencing and analysis of the human genome** *Nature* **409**:860–921

Lazar T, Tantos A, Tompa P, Schad E. (2022) **Intrinsic protein disorder uncouples affinity from binding specificity** *Protein Science* **3**

Lefébure T, Morvan C, Malard F, François C, Konecny-Dupré L, Guéguen L, Weiss-Gayet M, Seguin-Orlando A, Ermini L, Sarkissian CD, Charrier NP, Erne D, Mermillod-Blondin F, Duret L, Vieira C, Orlando L, Douady CJ. (2017) **Less effective selection leads to larger genomes** *Genome Research* **27**:1016–1028

Li W-H. (1987) **Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons** *Journal of Molecular Evolution* **24**:337–345

Lin J-J, Bhattacharjee MJ, Yu C-P, Tseng YY, Li W-H. (2019) **Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution** *Proceedings of the National Academy of Sciences* **116**:19009–19018

Liu SS, Hockenberry AJ, Jewett MC, Amaral LAN. (2018) **A novel framework for evaluating the performance of codon usage bias metrics** *Journal of The Royal Society Interface* **15**

Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, Berne C, Kysela D, Shoemaker WR, Muscarella ME, Luo H, Lennon JT, Brun YV, Lynch M. (2018) **Evolutionary determinants of genome-wide nucleotide composition** *Nature Ecology & Evolution* **2**:237–240

Luzuriaga-Neira AR, Alvarez-Ponce D. (2022) **Rates of Protein Evolution across the Marsupial Phylogeny: Heterogeneity and Link to Life-History Traits** *Genome Biology and Evolution*

Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. (2016) **Genetic drift, selection and the evolution of the mutation rate** *Nature Reviews Genetics* **17**:704–714

Masel J. (2011) **Genetic drift** *Current Biology* **2**:R837–R838

Meunier J, Duret L. (2004) **Recombination Drives the Evolution of GC-Content in the Human Genome** *Molecular Biology and Evolution* **21**:984–990

Novembre JA. (2002) **Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias** *Molecular Biology and Evolution* **19**:1390–1394

Novoa EM, Jungreis I, Jail Ion O., Kellis M. (2019) **Elucidation of Codon Usage Signatures across the Domains of Life** *Molecular Biology and Evolution* **36**:2328–2339

Ohta T. (1972) **Population size and rate of evolution** *Journal of Molecular Evolution* **1**:305–314

Ohta T. (1973) **Slightly Deleterious Mutant Substitutions in Evolution** *Nature* **246**:96–98

Ohta T. (1992) **The Nearly Neutral Theory of Molecular Evolution** *Annual Review of Ecology and Systematics* **23**:263–286

Paradis E, Schliep K. (2019) **ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R** *Bioinformatics* **35**:526–528

Payne BL, Alvarez-Ponce D. (2018) **Higher Rates of Protein Evolution in the Self-Fertilizing Plant Arabidopsis thaliana than in the Out-Crossers Arabidopsis lyrata and Arabidopsis halleri** *Genome Biology and Evolution* **10**:895–900

Plotkin JB, Dushoff J, Desai MM, Fraser HB. (2006) **Codon Usage and Selection on Proteins** *Journal of Molecular Evolution* **63**:635–653

Plotkin JB, Kudla G. (2011) **Synonymous but not the same: the causes and consequences of codon bias** *Nature Reviews Genetics* **12**:32–42

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, Loire E, Lourenco JM, Nabholz B, Roux C, Tsagkogeorga G, Weber AAT, Weinert LA, Belkhir K, Bierne N, Glémin S, Galtier N. (2014) **Comparative population genomics in animals uncovers the determinants of genetic diversity** *Nature* **515**:261–263

Romiguier J, Roux C. (2017) **Analytical Biases Associated with GC-Content in Molecular Evolution** *Frontiers in Genetics* **8**

Schad E, Tompa P, Hegyi H. (2011) **The relationship between proteome size, structural disorder and organism complexity** *Genome Biology*

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. (2005) **Variation in the strength of selected codon usage bias among bacteria** *Nucleic Acids Research* **33**:1141–1153

Sharp PM, Emery LR, Zeng K. (2010) **Forces that influence the evolution of codon bias** *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**:1203–1212

Sharp PM, Li W-H. (1987) **The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications** *Nucleic Acids Research* **15**:1281–1295

Smith NGC, Eyre-Walker A. (2001) **Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C-Rich Genes in Humans** *Molecular Biology and Evolution* **18**:982–986

Subramanian S. (2008) **Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes** *Genetics* **178**:2429–2432

Sun X, Yang Q, Xia X. (2012) **An Improved Implementation of Effective Number of Codons (Nc)** *Molecular Biology and Evolution* **30**:191–196

Urrutia AO, Hurst LD. (2001) **Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection** *Genetics* **159**

Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. (2009) **Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity** *Cell* **138**:198–208

Vicario S, Moriyama EN, Powell JR. (2007) **Codon usage in twelve species of Drosophila** *BMC Evolutionary Biology* **7**

Wangen JR, Green R. (2020) **Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides** *eLife* **9**

Weber CC, Nabholz B, Romiguier J, Ellegren H. (2014) **Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection** *Genome Biology* **15**

Wright F. (1990) **The 'effective number of codons' used in a gene** *Gene* **87**:23–29

Xue B, Dunker AK, Uversky VN. (2012) **Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life** *Journal of Biomolecular Structure and Dynamics* **30**:137–149

Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, Yu J. (2012) **Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance** *BMC Bioinformatics*

Article and author information

Catherine A. Weibel
Department of Mathematics, University of Arizona, Tucson, Arizona 85721, USA, Department of Physics, University of Arizona, Tucson, Arizona 85721, USA, Department of Applied Physics, Stanford University, California, USA
ORCID iD: 0000-0003-1837-5209

Andrew L. Wheeler
Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, Arizona 85721, USA
ORCID iD: 0000-0002-5347-5419

Jennifer E. James
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA, Department of Ecology and Genetics, Evolutionary Biology Center, Uppsala University, Sweden
ORCID iD: 0000-0003-0518-6783

Sara M. Willis
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA, University Information Technology Services, University of Arizona, Tucson, Arizona 85721, USA
ORCID iD: 0000-0002-1605-6426

Joanna Masel
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA
For correspondence: masel@arizona.edu
ORCID iD: 0000-0002-7398-2127

Copyright

Editors

Reviewing Editor
C Brandon Ogbunugafor
Yale University, United States of America

Senior Editor
Molly Przeworski
Columbia University, United States of America

Reviewer #1 (Public Review):

In this manuscript, the authors propose a new codon adaptation metric, Codon Adaptation Index of Species (CAIS), which they present as an easily obtainable proxy for effective population size. To permit between-species comparisons, they control for both amino acid frequencies and genomic GC content, which distinguishes their approach from existing ones. Having confirmed that CAIS negatively correlates with vertebrate body mass, as would be expected if small-bodied species with larger effective populations experience more efficient selection on codon usage, they then examine the relationship between CAIS and intrinsic structural disorder in proteins.

The idea of a robust species-level measure of codon adaptation is interesting. If CAIS is indeed a reliable proxy for the effectiveness of selection, it could be useful to analyze species without reliable life history- or mutation rate data (which will apply to many of the genomes becoming available in the near future).

A key question is whether CAIS, in fact, measures adaptation at the codon level. Unfortunately, CAIS is only validated indirectly by confirming a negative correlation with body mass. As a result, the observations about structural disorder are difficult to evaluate.

A potential problem is that differences in GC between species are not independent of life history. Effective population size can drive compositional differences due to the effects of GC-biased gene conversion (gBGC). As noted by Galtier et al. (2018), genomic GC correlates negatively with body mass in mammals and birds. It would therefore be important to examine how gBGC might affect CAIS, and to what extent it could explain the relationship between CAIS and body mass.

Suppose that gBGC drives an increase in GC that is most pronounced at 3rd codon positions in high-recombination regions in small-bodied species. In this case, could observed codon usage depart more strongly from expectations calculated from overall genomic GC in small vertebrates compared to large ones? The authors also report that correcting for local intergenic GC was unsuccessful, based on the lack of a significant negative relationship with body mass (Figure 3D). In principle, this could also be consistent with local GC providing a relatively more appropriate baseline in regions with high recombination rates. Considering these scenarios would clarify what exactly CAIS is capturing.

Given claims about "exquisitely adapted species", the case for using CAIS as a measure of codon adaptation would also be stronger if a relationship with gene expression could be demonstrated. RSCU is expected to be higher in highly expressed genes. Is there any evidence that the equivalent GC-controlled measure behaves similarly?

The manuscript is overall easy to follow, though some additional context may be helpful for the general reader. A more detailed discussion of how this work compares to the approach taken by Galtier et al. (2018), which accounted for GC content and gBGC when examining codon preferences, would be appropriate, for example. In addition, it would have been useful to mention past work that has attempted to explicitly quantify selection on codon usage.

Reviewer #2 (Public Review):

Summary:

The goal of the authors in this study is to develop a more reliable approach for quantifying codon usage such that it is more comparable across species. Specifically, the authors wish to

estimate the degree of adaptive codon usage, which is potentially a general proxy for the strength of selection at the molecular level. To this end, the authors created the Codon Adaptation Index for Species (CAIS) that controls for differences in amino acid usage and GC% across species. Using their new metric, the authors find a previously unobserved negative correlation between the overall adaptiveness of codon usage and body size across 118 vertebrates. As body size is negatively correlated with effective population size and thus the general strength of natural selection, the negative correlation between CAIS and body size is expected. The authors argue this was previously unobserved due to failures of other popular metrics such as Codon Adaptation Index (CAI) and the Effective Number of Codons (ENC) to adequately control for differences in amino acid usage and GC content across species. Most surprisingly, the authors also find a positive relationship between CAIS and the overall "disorderedness" of a species protein domains. As some of these results are unexpected, which is acknowledged by the authors, I think it would be particularly beneficial to work with some simulated datasets. I think CAIS has the potential to be a valuable tool for those interested in comparing codon adaptation across species in certain situations. However, I have certain theoretical concerns about CAIS as a direct proxy for the efficiency of selection sN_e when the mutation bias changes across species.

Strengths:

(1) I appreciate that the authors recognize the potential issues of comparing CAI when amino acid usage varies and correct for this in CAIS. I think this is sometimes an under-appreciated point in the codon usage literature, as CAI is a relative measure of codon usage bias (i.e. only considers synonyms). However, the strength of natural selection on codon usage can potentially vary across amino acids, such that comparing mean CAI between protein regions with different amino acid biases may result in spurious signals of statistical significance (see Cope et al. Biochemica et Biophysica Acta - Biomembranes 2018 for a clear example of this).

(2) The authors present numerous analysis using both ENC and mean CAI as a comparison to CAIS, helping given a sense of how CAIS corrects for some of the issues with these other metrics. I also enjoyed that they examined the previously unobserved relationship between codon usage bias and body size, which has bugged me ever since I saw Kessler and Dean 2014. The result comparing protein disorder to CAIS was particularly interesting and unexpected.

(3) The CAIS metric presented here is generally applicable to any species that has an annotated genome with protein-coding sequences.

Weaknesses:

(1) The main weakness of this work is that it lacks simulated data to confirm that it works as expected. This would be particularly useful for assessing the relationship between CAIS and the overall effect of protein structure disorder, which the authors acknowledge is an unexpected result. I think simulations could also allow the authors to assess how their metric performs in situations where mutation bias and natural selection act in the same direction vs. opposite directions. Additionally, although I appreciate their comparisons to ENC and mean CAI, the lack of comparison to other popular codon metrics for calculating the overall adaptiveness of a genome (e.g. dos Reis et al.'s S statistic, which is a function of tRNA Adaptation Index (tAI) and ENC) may be more appropriate. Even if results are similar to S,

CAIS has a noted advantage that it doesn't require identifying tRNA gene copy numbers or abundances, which I think are generally less readily available than genomic GC% and protein-coding sequences.

The authors mention the selection-mutation-drift equilibrium model, which underlies the basic ideas of this work (e.g. higher N_e results in stronger selection on codon usage), but a more in-depth framing of CAIS in terms of this model is not given. I think this could be valuable, particularly in addressing the question "are we really estimating what we think we're estimating?"

Let's take a closer look at the formulation for RSCUS. From here on out, subscripts will only be used to denote the codon and it will be assumed that we are only considering the case of $r = genome$ for some species s.

$$RSCUS_i = \frac{O_i}{E_i}$$

I think what the authors are attempting to do is "divide out" the effects of mutation bias (as given by E_i, such that only the effects of natural selection remain, i.e. deviations from the expected frequency based on mutation bias alone represent adaptive codon usage. Consider Gilchrist et al. MBE 2015, which says that the expected frequency of codon i at selection-mutation-drift equilibrium in gene g for an amino acid with N_a synonymous codons is

$$E_{i,g} = \frac{e^{-\Delta M_i - \Delta\eta_i\phi_g}}{\sum_{j=1}^{N_a} e^{-\Delta M_j - \Delta\eta_j\phi_g}}$$

where ΔM is the mutation bias, $\Delta\eta$ is the strength of selection scaled by the strength of drift, and ϕ_g is the gene expression level of gene (g). In this case, <a href="https://imgur.com/F2ikUov">ΔM and $\Delta\eta$ reflect the strength and direction of mutation bias and natural selection relative to a reference codon, for which $\Delta M_ref, \Delta\eta_ref = 0$. Assuming the selection-mutation-drift equilibrium model is generally adequate to model the true codon usage patterns in a genome (as I do and I think the authors

do, too), the $E_{i,g}$ could be considered the expected observed frequency codon i in gene g $E[O_{i,g}]$.

Let's re-write the $E_i = \frac{p_i}{\sum_{j=1}^{N_a} p_j}$ in the form of Gilchrist et al., such that it is a function of mutation bias ΔM. For simplicity, we will consider just the two-codon case and assume the amino acid sequence is fixed. Assuming GC% is at equilibrium, the term g_r and $1 - g_r$ can be written as

$$g_r = \frac{\mu_{AT\to GC}}{\mu_{AT\to GC}+\mu_{GC\to AT}} \quad 1 - g_r = \frac{\mu_{GC\to AT}}{\mu_{AT\to GC}+\mu_{GC\to AT}}$$

where $\mu_{x\to y}$ is the mutation rate from nucleotides x to y. As described in Gilchrist et al. MBE 2015 and Shah and Gilchrist PNAS 2011, the mutation bias $\Delta M_{NNA,NNG} = log(\frac{\mu_{AT\to GC}}{\mu_{GC\to AT}})$. This can be expressed in terms of the equilibrium GC content by recognizing that

$$\frac{g_r}{1-g_r} = \frac{\mu_{AT\to GC}}{\mu_{GC\to AT}} \implies \frac{g_r}{1-g_r} = e^{\Delta M}$$

As we are assuming the amino acid sequence is fixed, the probability of observing a synonymous codon i at an amino acid becomes just a Bernoulli process.

$$p_i = g_r^x (1 - g_r)^{(1-x)}$$

If we do this, then

$$E_{NNA} = \frac{p_{NNA}}{p_{NNA}+p_{NNG}} = \frac{1-g_r}{g_r+(1-g_r)} = \frac{1}{\frac{g_r}{1-g_r}+1} = \frac{1}{e^{\Delta M}+1} = \frac{e^{-\Delta M}}{1+e^{-\Delta M}}$$

Recall that in the Gilchrist et al. framework, the reference codon has $\Delta M_{NNG,NNG} = 0 \implies e^{-\Delta M_{NNG,NNG}} = 1$. Thus, we have recovered the Gilchrist et al.

model from the formulation of E_i under the assumption that natural selection has no impact on codon usage and codon NNG is the pre-defined reference codon. To see this, plug in 0 for $\Delta\eta$ in equation (1).

We can then calculate the expected RSCUS using equation (1) (using notation $E[O_i]$ and equation (6) for the two codon case. For simplicity assume, we are only considering a gene of average expression (defined as $\phi_g = 1$. Assume in this case that NNG is the reference codon $(\Delta M_{NNG}, \Delta\eta_{NNG} = 0)$.

$$E\left[RSCUS_{NNA}\right] = \frac{E[O_{NNA}]}{E_{NNA}} = \frac{e^{-\Delta\eta_{NNA}}\left(e^{-\Delta M_{NNA}} + e^{-\Delta M_{NNG}}\right)}{e^{-\Delta M_{NNA}-\Delta\eta_{NNA}} + e^{-\Delta M_{NNG}-\Delta\eta_{NNG}}} = \frac{e^{-\Delta M_{NNA}-\Delta\eta_{NNA}} + e^{-\Delta M_{NNG}-\Delta\eta_{NNA}}}{e^{-\Delta M_{NNA}-\Delta\eta_{NNA}} + e^{-\Delta M_{NNG}-\Delta\eta_{NNG}}} =$$

$$\frac{e^{-\Delta M_{NNA}-\Delta\eta_{NNA}} + e^{-\Delta\eta_{NNA}}}{e^{-\Delta M_{NNA}-\Delta\eta_{NNA}} + 1}$$

This shows that the expected value of RSCUS for a two-codon amino acid is expected to increase as the strength of selection $\Delta\eta$ increases, which is desired. Note that $\Delta\eta$ in Gilchrist et al. is formulated in terms of selection *against* a codon relative to the reference, such that a negative value represents that a codon is favored relative to the reference. If $\Delta\eta = 0$ (i.e. selection does not favor either codon), then $E[RSCUS] = 1$. Also note that the expected RSCUS does not remain independent of the mutation bias. This means that even if $_sN_e$ (i.e. the strength of natural selection) does not change between species, changes to the strength and direction of mutation bias across species could impact RSCUS. Assuming my math is right, I think one needs to be cautious when interpreting CAIS as representative of the differences in the efficiency of selection across species except under very particular circumstances. One such case could be when it is known that mutation bias varies little across the species of interest. Looking at the species used in this manuscript, most of them have a GC content ranging around 0.41, so I suspect their results are okay.

Although I have not done so, I am sure this could be extended to the 4 and 6 codon amino acids.

Another minor weakness of this work is that although the method is generally applicable to any species with an annotated genome and the code is publicly available, the code itself contains hard-coded values for GC% and amino acid frequencies across the 118 vertebrates. The lack of a more flexible tool may make it difficult for less computationally-experienced researchers to take advantage of this method.

Author Response

The primary concern of Reviewer 1 is that Ne might affect gBGC and hence GC, and this might act as a confounding effect. The reviewer suggests that we should investigate how gBGC (with GC presumably as its proxy) might affect CAIS, and to what extent any relationship here could explain the relationship between CAIS and body mass. We believe that we have already dealt with this both in Supplementary Figure S5A (where we regret having inserted the wrong figure panel, a mistake we will correct), and its PIC-corrected counterpart in S5B. These two panels show (or will show) that CAIS is not correlated with GC. Note that we expect our genomic-GC-based codon usage expectations to reflect unchecked gBGC in an average genomic region, independently of whether that species has high or low Ne. Our working model is that mutation biases, including but not limited to the strength of gBGC, vary among species, and that they rather than selection determine each species' genome-wide %GC. By correcting for genome-wide %GC, our CAIS thus corrects for mutation bias, in order to isolate the effects of selection.

Reviewer 1 also suggests that we examine the relationship between gene expression and GC corrected RSCU, as we would expect codon adaptation to be stronger in more highly expressed genes, as was previously shown in the non-GC corrected CAI metric (Sharp et al 1987). Correlations with gene expression are outside the scope of the current work, which is focused on producing a single value of codon adaptation per species. It is indeed possible that our general approach could be useful in future work investigating differences among genes.

One key difference between our work and that of Galtier et al. 2018 is that our approach does not rely on identifying specific codon preferences per species. Our approach thus remains appropriate even for scenarios e.g. where different cell types, different environmental conditions, and/or different genes have different codon preferences (Gingold et al. 2014 https://doi.org/10.1016/j.cell.2014.08.011). At a high level, our results are in broad agreement with those of Galtier et al., 2018, who found that gBGC affected all animal species, regardless of Ne, and who like us, found that the degree of selection on codon usage depended on Ne. Through use of a more sensitive methodology, we believe we have expanded our ability to detect codon adaptation into animals of somewhat higher Ne than in previous work.

We thank Reviewer 2 for explicitly laying out the math that was implicit in our Figures 1 and 2. In our revisions, we will more clearly acknowledge that the per-site codon adaptation bias depicted in Figure 1 has limited sensitivity to s*Ne. We believe our approach worked despite this because the phenomenon is driven by what is shown in Figure 2. I.e., where Ne makes a difference is by determining the proteome-wide fraction of codons subject to significant codon adaptation, rather than by determining the strength of codon adaptation at any particular site or gene.

Simulated datasets would be great, but we think it a nice addition rather than must-have, in particular because we are skeptical about whether our understanding of all relevant processes is good enough such that simulations would add much to our more heuristic argument along the lines of Figure 2. E.g. we believe the complications documented by Gingold et al. 2014 cited above are pertinent, but incorporating them into simulations would require a complex set of assumptions.

In response to the final comment of reviewer 2, the reason that we hard-coded genome-wide %GC values is that we took them from the previous study of James et al. (2023) https://doi.org/10.1093/molbev/msad073. As summarized in the manuscript, genome-wide %GC was a byproduct of a scan conducted in that work, of all six reading frames across genic and intergenic sequences available from NCBI with access dates between May and July 2019. The code used in the current work to calculate the intergenic %GC, as well as that used to calculate amino acid frequencies, is located at https://github.com/MaselLab/Codon-

Adaptation-Index-of-Species. We agree that more user-friendly tools would be useful, but producing robust tools falls outside the scope of the current manuscript.