Evolution Letters, 2024, **XX**(XX), 1–13
https://doi.org/10.1093/evlett/qrae061
Letter

Does metabolic rate influence genome-wide amino acid composition in the course of animal evolution?

Wei Wang[1,] and De-Xing Zhang[1,2,]

[1]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China
[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China

Corresponding authors: Group of Molecular Ecology and Evolution, Institute of Zoology, Chinese Academy of Sciences, 1 Beichen West Road, Chaoyang District, Beijing 100101, China. Email: wangwei@ioz.ac.cn; Group of Molecular Ecology and Evolution, Institute of Zoology, Chinese Academy of Sciences, 1 Beichen West Road, Chaoyang District, Beijing 100101, China. Email: dxzhang@ioz.ac.cn

Abstract

Natural selection is believed to shape amino acid usage of the proteome by minimizing the energy cost of protein biosynthesis. Although this hypothesis explains well the amino acid frequency ($AA_{frequency}$) difference among the 20 common amino acids within a given genome (species), whether it is applicable to cross-species difference remains to be inspected. Here, we proposed and tested a "metabolic rate hypothesis," which suggests that metabolic rate impacts genome-wide $AA_{frequency}$, considering that the energy allocated to protein biosynthesis is under selection pressure due to metabolic rate constraint. We performed integrated phylogenetic comparative analyses on proteomic sequence and metabolic rate data of 166 species covering 130 eumetazoan orders. We showed that resting metabolic rate (RMR) was significantly linked to $AA_{frequency}$ variation across animal lineages, with a contribution comparable to or greater than genomic traits such as GC content and codon usage bias. Consistent with the metabolic rate hypothesis, low-energy-cost amino acids are observed to be more likely at higher frequency in animal species with high (residual) metabolic rate. Correlated evolution of RMR and $AA_{frequency}$ was further inferred being driven by adaptation. The relationship between RMR and $AA_{frequency}$ varied greatly among amino acids, most likely reflecting a trade-off among various interacting factors. Overall, there exists no "one-size-fits-all" predictor for $AA_{frequency}$, and integrated investigation of multilevel traits is indispensable for a fuller understanding of $AA_{frequency}$ variation and evolution in animal.

Keywords: amino acid abundance, proteome evolution, metabolic theory, adaptation, interspecific comparison, phylogenetic model

Lay Summary

Natural selection has previously been proposed to underlie compositional differences of amino acids within a given animal species: to minimize the energy cost of protein biosynthesis, selection should result in an increased abundance of energy-cheap amino acids and a decreased abundance of energy-expensive ones in the proteome. However, amino acid composition varies greatly among species, yet the underlying evolutionary processes are largely unclear. Here, we propose that interspecific variations in metabolic rate exert an effect on amino acid composition due to natural selection, which is referred to as the "metabolic rate hypothesis." The present study intended to test this hypothesis by examining the relationship between molecular-level amino acid frequency and organism-level metabolic rate (as a proxy for energy cost) in an evolutionary framework. Phylogeny-based analysis revealed significant connection of these two traits, indicating they have evolved most likely in concert, and metabolic rate is an important confining factor to the compositional variation of amino acid in animal. Consistent with the metabolic rate hypothesis, low-energy-cost amino acids are observed to be more likely at higher frequency in animal species with high residual metabolic rate.

Introduction

Proteins are among the most fundamental molecules of life and evolved under natural selection. Signatures of selection on proteins have been identified at various levels, such as the conservation of specific amino acid residue(s), secondary and tertiary structures, as well as compositional patterns of amino acids (i.e., the relative abundance or frequency of amino acids making up the entire proteome) (Akashi & Gojobori, 2002; Chen et al., 2022; de Jong et al., 2023; Liu et al., 2008; Suzuki & Gojobori, 1999; Yang et al., 2000). The last characteristic bears some interesting peculiarities: while for a given species there exist up to dozens of folds of difference in the relative abundance among amino acids (e.g., from 0.39% for Cys to 14.5% for Leu; Krick et al., 2014), the compositional patterns vary greatly among species (Chen & Nielsen, 2022; Knight et al., 2001; Krick et al., 2014; Moura et al., 2013). This suggests that complex interwinding factors exist, shaping the amino acid compositional patterns during evolution.

Received March 8, 2023; revisions received October 18, 2024; accepted October 26, 2024

Two major hypotheses have been proposed (Krick et al., 2014) for explaining the observed compositional patterns within a given species (genome) in particular. The "genetic code hypothesis" suggests that amino acid residues distribute according to the permutation of genetic code, which is largely a random-drift explanation, predicting a positive relationship between amino acid frequency (hereafter $AA_{frequency}$) and the permutation of genetic code (Dyer, 1971; King & Jukes, 1969). The "synthesis cost hypothesis" suggests that bias in amino acid usage arises from cost-minimization of protein biosynthesis, which is a natural selection explanation, predicting a universal negative relationship between $AA_{frequency}$ and the energy (ATP) cost of amino acid synthesis (Akashi & Gojobori, 2002; Seligmann, 2003; Swire, 2007). Compared with genetic code permutation, synthesis cost appears to better explain the variance of $AA_{frequency}$ within a given proteome, since synthesis cost shows higher correlation with $AA_{frequency}$ for species from all three domains of life (Krick et al., 2014). However, the current "synthesis cost hypothesis" faced difficulties in explaining the great interspecific variation of amino acids, because the energy cost for a given amino acid is generally considered constant across species due to the highly conserved biochemical nature of metabolic pathways (Krick et al., 2014; Smith & Morowitz, 2004; Swire, 2007; Zhang et al., 2018).

We noticed that the physiological trait metabolic rate, which is related to energy cost, also manifests great interspecific variation. Metabolic rate, representing the rate at which organisms obtain, allocate, and expend energy, is the fundamental measure of the energy cost of life functions and biological activities (Brandl et al., 2023; Brown et al., 2004; Burger et al., 2019). Metabolic rate is well-known to differ substantially among species (Brown et al., 2004; White & Kearney, 2013). Among the many factors contributing to the metabolic rate variations among animals, body mass and body temperature are the two best-known and most remarkable ones, which together can explain more than 90% of the metabolic rate variations (Brown et al., 2004; Gillooly et al., 2001; White & Kearney, 2013). Nevertheless, after accounting for body mass and temperature effects, several-fold differences are still observed in the residual metabolic rate (i.e., body mass and temperature-independent metabolic rate) among animals (e.g., Giacometti et al., 2022; Gillooly et al., 2001; Hayssen & Lacy, 1985; Lighton et al., 2001). It is recognized that residual metabolic rate is associated with ecological and evolutionary processes (Carter et al., 2023; Giacometti et al., 2022; McNab, 2015), thus being a good trait for inspecting evolutionary adaptation. So far, it remains unknown whether metabolic rate, the residual metabolic rate in particular, is linked with $AA_{frequency}$ in animals.

Unlike plants, animals are heterotrophs that cannot produce their own food and must obtain energy from other organisms to support various physiological processes for survival, development, and reproduction. Since energy resources are generally neither cost-free nor unlimited for animals (McCue, 2010), the energy budget that can be devoted to specific metabolic pathways (including protein biosynthesis) is basically under strong selection pressure (Akashi & Gojobori, 2002; Heizer et al., 2006; Seligmann, 2003). Notably, a considerable amount of energy is allocated to protein biosynthesis, e.g., about 20% of total oxygen consumption (a proxy for metabolic rate) is used by protein synthesis in mammals (Rolfe & Brown, 1997), and there exists great variation among amino acids in the energy expended to their biosynthesis (Craig & Weber, 1998; Heizer et al., 2006); protein biosynthesis itself thus can be optimized by natural selection (Akashi & Gojobori, 2002; Heizer et al., 2006) under the constraint of the overall metabolic rate adopted during evolution. Therefore, energy-cheap amino acids should have been preferred in the process of protein biosynthesis by natural selection during animal evolution (Krick et al., 2014; Seligmann, 2003; Swire, 2007). We can then conjecture that the energy-cheaper amino acids would be more frequently employed in proteins and thus present at higher frequency in high-energy-expenditure species. Namely, where possible, the biosynthetic cost of proteins should have been minimized in order to maximize the total reproductive output of an animal by safeguarding the essential high-energy-expenditure function(s) (such as flying in insects). Interspecific variation in $AA_{frequency}$ among animals is thus expected as a result of diversifying evolution in energy budget and metabolic rate in different lineages. For convenience, we refer hereafter this conjecture as to the "metabolic rate hypothesis" on interspecific compositional variation in amino acids, while acknowledging that a metabolic rate hypothesis has been proposed for explaining variation in the rates of nucleotide substitution among evolutionary lineages (Martin & Palumbi, 1993).

In the present report, we intend to use an integrated phylogenetic comparative analysis on proteomic and metabolic rate data to test the "metabolic rate hypothesis" by examining whether there exists any interconnection between metabolic rate (as a proxy for organism-level energetic cost) and the $AA_{frequency}$ across animal lineages. Note that the "metabolic rate hypothesis" assumed that metabolic rate affects the amino acid composition rather than *vice versa*. Many well-known observations support this argument. For example, amino acid composition itself is influenced by numerous factors (e.g., codon usage bias, functional constraint, genome compositional bias, environment) (Baeza et al., 2021; Berthelot et al., 2019; Cutter et al., 2006; Tekaia et al., 2002), and thus shows great variation at almost all levels: within a genome, within an individual (between genomes), within and between species (Knight et al., 2001; Krick et al., 2014; Moura et al., 2013). This thus leaves great room for the optimization of energetic costs in protein synthesis by natural selection (Akashi & Gojobori, 2002; Heizer et al., 2006), on which the metabolic rate of an animal adopted during evolution would impose serious constraints. Furthermore, molecular mechanisms connected to $AA_{frequency}$ are indirectly affected by metabolic rate; for instance, metabolic rate can influence mutation rate by affecting DNA damage (Gillooly et al., 2007; Martin & Palumbi, 1993), and mutation process is known to be a major determinant of amino acid composition (Akashi & Gojobori, 2002; King & Jukes, 1969). Similarly, small cell and small genome size could be favored physiologically (Hessen et al., 2013; Szarski, 1983), while genome size is correlated with the use of amino acids (Du et al., 2018; Seligmann, 2003). By dissecting the metabolic rate into three components (i.e., the body mass component, body temperature component, residual component) and using proteomic sequence data from 166 animal species representing 130 eumetazoan ("true animal") orders, we examined the "metabolic rate hypothesis" focusing on the following two questions: (1) Does the metabolic rate (an organism-level phenotype) influence the $AA_{frequency}$ (a molecular-level phenotype)? (2) What are the evolutionary drivers underlying the relationship between metabolic rate and $AA_{frequency}$?

Methods

Compilation of the amino acid data: Proteome sequences, $AA_{frequency}$, and energy costs of amino acid biosynthesis

Proteomic sequence data were chosen with the aim to cover more animal groups at the order level and to reduce data imbalance among taxa, e.g., more proteomes from mammals and

less from other vertebrates. First, all proteomes available for Eumetazoa ("true animal") were retrieved from NCBI RefSeq database. Second, one species (proteome) was selected within each order, according to the following data-selection criteria: species with metabolic rate data available were selected first, then species with chromosome-level genome, and finally species whose genome has higher contig N50. This procedure yielded an initial dataset containing 130 species. Third, because there exists serious taxonomic sampling bias in the initial dataset (about two thirds of the 130 species are vertebrates), 36 invertebrate species (covering 35 families) and their corresponding proteomes were supplemented into the dataset, following a similar data-selection criteria as described above. This yielded a proteomic sequence dataset that included 84 vertebrate and 82 invertebrate species covering 130 orders in Eumetazoa (see Supplementary Table S1).

$AA_{frequency}$ was calculated as the percentage of the number of each of the 20 common amino acids divided by the count of all amino acids in the non-redundant proteome after redundancy removal in proteomic sequences using CD-HIT (Li & Godzik, 2006) following the similar procedures in previous studies (Akashi & Gojobori, 2002; Heizer et al., 2006). $AA_{frequency}$ was estimated separately for each species.

The energy cost of amino acid biosynthesis was measured as the ATP molecules per amino acid molecule needed during amino acid biosynthesis. These synthesis cost values were obtained from previous studies with correction for decay rate and assuming similar energy choices of essential and non-essential amino acids (Krick et al., 2014; Swire, 2007; Zhang et al., 2018).

Compilation of the resting metabolic rate data

Resting metabolic rate (RMR) and the RMR-related data were compiled for each of the above 166 animal species. The RMR, body mass (*M*), and body temperature (*T*) were collected by literature search (Supplementary Table S1). When multiple RMR units were available for the same species, the raw RMR data expressed in carbon dioxide production (VCO_2) or oxygen consumption (VO_2) was preferred (e.g., in μl CO_2 per hour). Generally, RMR data measured with animals under stressed conditions were not considered. The respiratory quotient (equal to VCO_2/VO_2) values were also compiled if available. All RMR data were ultimately transformed to microwatts (μW) based on respiratory quotient values (Lighton, 2008). Body mass and body temperature were transformed to gram (g) of wet body mass and Kelvin (K), respectively. Dry body mass was transformed to wet body mass assuming a ratio of 3 (unless otherwise specified by the primary literature) between the latter and the former according to previous study (Makarieva et al., 2008).

Phylogenetic tree and phylogenetic signal

For the studied animals, an initial phylogenetic tree was constructed, with both tree topology and branch length obtained from the TimeTree database (www.timetree.org). To deal with several unresolved branches, the initial tree was manually updated according to phylogenies reported by Blaxter (2009), Delsuc et al. (2018), and Arribas et al. (2020). For instance, tunicate species *Styela clava* was currently not included in the TimeTree database, and thus this species was manually added as sister lineage to the vase tunicate *Ciona intestinalis* (which was included in TimeTree database), while their divergence time was estimated as 388.5 million years ago (Mya) (Delsuc et al., 2018). The updated time tree was utilized for all subsequent analyses unless otherwise clarified.

The phylogenetic signal of studied traits (e.g., RMR) was evaluated by Pagel's λ (Pagel, 1999), which was estimated using the maximum-likelihood method implemented in the package *phytools* (Revell, 2012) in R environment (version 4.1.2). Compared with other indices of phylogenetic signal, Pagel's λ appeared to perform better especially for discriminating between complex models of traits evolution (Münkemüller et al., 2012). This index was introduced as a phylogeny transformation parameter, which measures the phylogenetic dependence of trait data by a value varying between 0 and 1. Generally, λ closer to 1 indicates strong phylogenetic signal and thus close relatives are more likely to have similar traits than distant relatives; λ closer to 0 indicates low phylogenetic signal and close relative are not more similar. Statistical significance of λ was evaluated by a likelihood-ratio test ($p < 0.05$) by comparing the observed λ with the null hypothesis (no phylogenetic signal).

Genomic traits

Several genomic traits (characteristics) have been proposed to affect the amino acid composition (e.g., Akashi & Gojobori, 2002; Cutter et al., 2006; Moura et al., 2013; Seligmann, 2003; Sueoka, 1961). Six potential traits were closely inspected here, viz: the genomic GC content (GC-genome), GC content of genomic protein-coding sequences (GC-CDS), genome size, protein expression level, protein evolutionary rate, and the genetic code permutation. The last three traits will be described in detail in the following paragraphs.

Protein expression level: CAI-based codon usage bias and experiment-based proteome abundance

It is known that amino acid usage can be influenced by differences in protein/gene expression levels (Akashi & Gojobori, 2002). Herein two types of expression data were used. The first one is codon usage bias, which can be used to predict protein expressivity given their association (Carbone et al., 2003; Sharp & Li, 1987). To quantitatively measure the extent of codon usage bias, the widely used Codon Adaptation Index (CAI) was calculated (Carbone et al., 2003; Sharp & Li, 1987). CAI values were first estimated for each individual gene. The gene CAI value ranges from 0 to 1, with low values suggesting biased codon usage and high values similar codon usage. As the present study focused mainly on interspecific variations, species-level CAI values were then estimated by calculating the average of CAIs of all genes in a species (genome). Species with low CAI imply biased protein expression and species with high CAI unbiased protein expression (Botzman & Margalit, 2011; Sharma et al., 2023). In this respect, variations in species-level CAI signal differences in gene/protein expression across species.

The second type of expression data was the experiment-based proteome abundance, which was obtained from the PaxDB database (Huang et al., 2023). We chose the abundance data from the whole organism unless it was unavailable (see Supplementary Table S2). Because such data were only available in 15 species (Supplementary Table S2), we did not include them in subsequent dimension reduction analysis (see below) due to the small sample size. To evaluate the effect of experiment-based proteome abundance on RMR-$AA_{frequency}$ relationship, we divided the proteomes into two subdatasets, namely, the high-expression proteins and the low-expression proteins, and then tested whether they yielded different results under the analysis. To identify the low- and high-expression proteins for a given species, we pooled all quantified proteins together and computed the 1st, 2nd, and

3rd quartiles of the overall distribution. Proteins with abundance below the 1st quartile and above the 3rd quartile were determined as low-expression and high-expression proteins, respectively. The high-expression proteins showed abundance about 10–10,000-fold greater (depending on the species) than the low-expression ones. Limitation in sample size did not allow us to further contrast the two subdatasets.

Protein evolutionary rate (Evol Rate)

The rate of protein evolution was estimated in absolute term (i.e., amino acid substitutions per site per million years). First, single-copy orthologous (SCO) genes across studied species were inferred using the BUSCO metazoan_odb10 database (Simão et al., 2015). Protein sequences of the identified SCO genes were then aligned by MUSCLE (Edgar, 2004), and the yielded alignments were trimmed by BMGE to select phylogenetic informative regions (Criscuolo & Gribaldo, 2010). Second, for each SCO gene, the gene tree was reconstructed using maximum-likelihood method with IQ-TREE (Nguyen et al., 2015) based on the trimmed alignments. IQ-TREE analysis was performed with the aforementioned time tree topology acting as a backbone constraint. Evolutionary model of protein sequence evolution was determined by ModelFinder (Kalyaanamoorthy et al., 2017). Third, for each SCO gene, its protein evolutionary rate was estimated as terminal lengths (expressed in amino acid substitutions per site) in gene tree divided by their lengths (expressed in million years) in time tree. Finally, the species-level protein evolutionary rates were obtained by calculating the average of evolutionary rates of all SCO genes identified in a species (36–56 SCO genes per species, with the mean being 52 SCO genes per species).

Genetic code permutation as suggested by the "genetic code hypothesis"

The genetic code effect was quantified as the expected $AA_{frequency}$ due to random permutations of genetic codon bases. By multiplying the frequency of bases (A, G, C, U) to obtain codon frequency, the expected $AA_{frequency}$ was equal to the sum of frequencies of amino acid codons (Dyer, 1971; King & Jukes, 1969). This was done separately for each of the studied species.

Modeling the relationship between $AA_{frequency}$ and RMR

Prior to the inference of the relationship between RMR and $AA_{frequency}$, RMR was dissected into three parts to clarify its sources (see *Introduction* section). Additionally, due to the obvious interactions among the six genomic traits mentioned above (e.g., between GC-genome and GC-CDS), principal component analysis (PCA) was conducted to reduce the data dimensions of genomic traits.

Disentangling the components of metabolic rate

The present study dissected the RMR into three parts: the body mass component, the body temperature component, and the component independent of body mass and temperature (i.e., residual component; see *Introduction* section). To estimate these components, a mathematical model was fitted by modifying the fundamental equation of the metabolic theory of ecology (MTE) (Brown et al., 2004; Gillooly et al., 2001):

$$\ln R = a + b \ln M + c\left(\frac{1}{kT}\right) \quad (1)$$

where R is RMR; a is a constant; b is the mass-scaling exponent (in the unmodified MTE equation the b is fixed to 0.75); M is body mass (in g); c is the negative value of activation energy (in eV); k is Boltzmann constant (0.0000862 eV K^{-1}); and T is body temperature (in K). To make it more readable, the above equation was written as:

$$\ln R = R_{mass} + R_{temperature} + R_{residual} \quad (2)$$

where R_{mass} is the body mass component of RMR and equal to $b\ln M$; $R_{temperature}$ is the temperature component of RMR and equal to $c(1/kT)$; $R_{residual}$ is equal to $\ln(RM^{-b}e^{-c/kT})$ that represents the RMR component independent of body mass and temperature (Brown et al., 2004; Gillooly et al., 2001). Parameters a, b, and c were inferred from the phylogenetic generalized least squares (PGLS) method (Grafen, 1989) in the *nlme* package to control for phylogenetic non-independence across species (see paragraphs below).

Dimension reduction of genomic traits by PCA

For the genomic traits investigated (GC-genome, GC-CDS, genome size, CAI, protein evolutionary rate, and genetic code permutation), PCA was employed to reduce their complexity before subsequent analyses. Because the genetic code permutations varied not only across species but also among amino acids, PCA was run for each of the 20 common amino acids separately. PC1~3 were selected for subsequent analyses. These PCs together can explain most (>86%) of variances of the studied genomic traits (see *Results* section). PCA was performed via the *stats* and *FactoMineR* packages in R.

Modeling the relationship between AAfrequency and metabolic rate

Relationships were modeled between the $AA_{frequency}$ and the three RMR components (R_{mass}, $R_{temperature}$, $R_{residual}$) and three principal components (PC1~3 of six genomic traits, see above): "$\ln(AA_{frequency}) \sim R_{mass} + R_{temperature} + R_{residual} + PC1 + PC2 + PC3$." Because PC1~3 varied among different amino acids, it was modeled for each of the 20 amino acids, respectively. We fitted the model using both ordinary least squares (OLS) and PGLS methods. Compared with OLS, PGLS can convert phylogeny to a variance–covariance matrix. The phylogenetic correlation structure (variance–covariance matrix) was constructed with five evolutionary models using corBrownian, corPagel, corGrafen, corBlomberg, and corMartins in the *ape* package (Paradis & Schliep, 2019). Fitting performance of OLS and different PGLS models was evaluated by the corrected value of Akaike information criterion (AIC_c): lower AIC_c indicating better fit and difference in AIC_c (ΔAIC_c) > 4 indicating strong support (Burnham & Anderson, 2002).

The normality and homogeneity assumptions of OLS and PGLS regressions were checked by investigating the distribution of quantile against quantile and the trend of residuals against fitted values, respectively (Garamszegi, 2014; Quinn & Keough, 2002). Assumption of lack of collinearity was tested by the Variance Inflation Factor (VIF) value, with VIF <10 indicating no collinearity among variables (Quinn & Keough, 2002). The independence assumption is often violated in trait data across species due to their shared evolutionary history (phylogeny), and PGLS method was implemented to deal with such a phylogenetic non-independence problem (Garamszegi, 2014). Since some PGLS regressions violated the normality/homogeneity assumptions, non-parametric bootstrap approach (with 1,000 replicates) was performed via the *car* package, and 95% CIs of estimates were calculated by the bias corrected and accelerated method (or percentile method if the bias-corrected and accelerated method failed).

Regression slopes of predictors R_{mass}, $R_{temperature}$, and $R_{residual}$ were used to describe how the $AA_{frequency}$ responded to the body mass

component of RMR, body temperature component of RMR, and RMR component independent of body mass and temperature. For example, significantly positive (or negative) slopes of R_{mass} and $R_{temperature}$ indicated positive (or negative) relationships between the $AA_{frequency}$ with body mass and body temperature, respectively. Statistical significance was determined when the slope 95% CIs calculated by the bootstrap approach (see above) did not overlap with zero ($p < 0.05$).

Inferring the contribution of RMR to $AA_{frequency}$

To estimate how much variance of the $AA_{frequency}$ was explained by RMR, we calculated the likelihood-based square value of the correlation coefficient (hereafter r^2_{lik}) via package *rr2* (Ives, 2019). In contrast to r^2 (square of Pearson's *r*) and other modifications of r^2, r^2_{lik} is particularly appropriate when the question is to identify the relative contribution of a specific factor within a given model (Ives, 2019). While r^2 was estimated directly under OLS model, r^2_{lik} was estimated by comparing a full PGLS model with a reduced PGLS model that removes a specific variable. For instance, r^2_{lik} for RMR was inferred from comparing the full PGLS model "$\ln(AA_{frequency}) \sim R_{mass} + R_{temperature} + R_{residual}$ + Phylogeny" and the reduced model "$\ln(AA_{frequency})$ ~ Phylogeny." Similarly, r^2_{lik} for genomic trait, e.g., CAI, was estimated from comparing the full model "$\ln(AA_{frequency})$ ~ CAI + Phylogeny" and the reduced model "$\ln(AA_{frequency})$ ~ Phylogeny." In these models, phylogeny was included as the covariance in the residual variation of the PGLS fitting (Ives, 2019; Wang et al., 2022). Comparisons of r^2_{lik} values for RMR and the six genomic traits were done via the Wilcoxon signed-rank test (two tailed) in the *stats* package. Statistical significance was adopted when the Wilcoxon test *p*-value was below 0.05 (i.e., $p < 0.05$).

Great heterogeneities exist among the studied taxa in the six genomic traits aforementioned. This will certainly exert an influence on the relative contribution of RMR to $AA_{frequency}$. Therefore, it is essential to filter the data to reduce the confounding internal variation so that the general relative contribution of RMR can be estimated. Preliminary analyses showed that GC content of genomic protein-coding sequences (GC-CDS) bears the greatest confounding influence among the studied species; hence, we factored out the species with highly biased GC-CDS (i.e., the GC-CDS-biased species, see below) and analyzed the r^2_{lik} for RMR based on the remaining species. To identify species with biased GC-CDS, we pooled all species together and calculated the 1st, 2nd, and 3rd quartiles of the overall distribution. Species with GC-CDS below 1st quartile and above 3rd quartile was identified as low-GC-CDS and high-GC-CDS species, respectively. Both low- and high-GC-CDS species were considered as GC-CDS-biased species and excluded.

Testing correlated evolution and evolutionary adaptation

Theoretically, there are at least two kinds of evolutionary mechanisms in causing the relationship between two traits (Blomberg et al., 2003; Prinzing et al., 2001; Wang et al., 2022). The first one is correlated evolution due to evolutionary adaptation. The second one is phylogenetic constraint, in which two traits have evolved independently of each other, and their seeming "relationship" is a reflection that both traits are constrained by their shared phylogenetic history (i.e., legacy from their common ancestors) (Blomberg et al., 2003; Wang et al., 2022). Hence, only when the frequency of a given amino acid was significantly related to at least one of the three RMR components under PGLS rather than OLS approach ($p < 0.05$), can correlated evolution be considered responsible for the RMR-$AA_{frequency}$ relationship.

Adaptive and non-adaptive random-drift processes of trait evolution were described by the Ornstein–Uhlenbeck (OU) and Brownian motion (BM) model, respectively (Felsenstein, 1985; Hansen, 1997; Ingram & Mahler, 2013). The BM model assumes that trait changes are random events and trait may evolve indefinitely to any value, while the OU model assumes trait changes toward either single adaptive optimum (single-regime OU, OU1) or multiple adaptive optima (multiple-regime OU, OUM). Therefore, adaptive evolution can be monitored by OU process in which different lineages evolved to either same or different optima (Hansen, 1997; Ingram & Mahler, 2013). Three evolutionary models (adaptive OU1, adaptive OUM, and non-adaptive BM model) were fitted separately on RMR components and $AA_{frequency}$ data, using package *geiger* (Pennell et al., 2014) and the SURFACE method implemented in package *surface* (Ingram & Mahler, 2013). Model comparison was evaluated by ΔAIC_c (see above). Selection was considered as the potential driver of the correlated evolution of RMR and $AA_{frequency}$ only when all of the following conditions were satisfied: (1) For a given amino acid, its frequency should be correlated significantly ($p < 0.05$) with at least one of the three RMR components under PGLS approach; (2) For a given amino acid, the evolution of its frequency should follow OU rather than BM model; (3) For a given amino acid, the evolution of its relevant RMR component(s) should follow OU rather than BM model.

Calculating *Z*-score values for each variable for heatmap visualization

Z-score values were used just for the heatmap visualization to present the interspecific variations in variables studied herein ($AA_{frequency}$, RMR components, genomic traits). With $AA_{frequency}$ as an example, its *Z*-score was calculated by taking the difference between a species and the overall mean of all species and then dividing the difference by standard deviation (resulting in each amino acid having zero mean and standard deviation of one).

Results

Sources of interspecific variations in RMR and dimension reduction of genomic traits

As expected, RMR positively covaried with both body mass and body temperature (Figure 1A and B; note that in Figure 1B, the body temperature *T* was contained in the denominator in the $1/kT$ term of *x*-axis), with more than 95% of the interspecific RMR variance being explained by the two factors (not shown). Body mass manifested a higher correlation to RMR than body temperature ($r^2 = 0.956$, $p < 2.2e{-}16$ versus $r^2 = 0.561$, $p < 2.2e{-}16$; Figure 1A and B). The mass-scaling exponent (*b*) was estimated to be 0.805 and activation energy (*–c*) 0.583 eV. These two parameters were used for the calculation of the three RMR components R_{mass}, $R_{temperature}$, and $R_{residual}$ (see *Methods* section).

PCA results (Figure 1C and D) showed that (1) more than 86% of the variance in the six genomic traits can be explained by the first three principal components (i.e., PC1~3), and (2) PC1 was highly correlated with the following three genomic traits: GC-CDS, GC-genome, and CAI ($|r| \geq 0.84$). Both observations held true regardless of the amino acid concerned (note that we performed PCA separately for each of the 20 common amino acids, see *Methods* section). PC2 and PC3 were mainly correlated with protein evolutionary rate and genome size, and also with genetic code permutation for several amino acids (Asp, Cys, Gln, Glu, Ser, Thr, and Val; note that this genomic trait was highly correlated to PC1 in the remaining amino acids; $|r| \geq 0.78$). PC1~3, which was independent of each other, can thus largely represent the studied

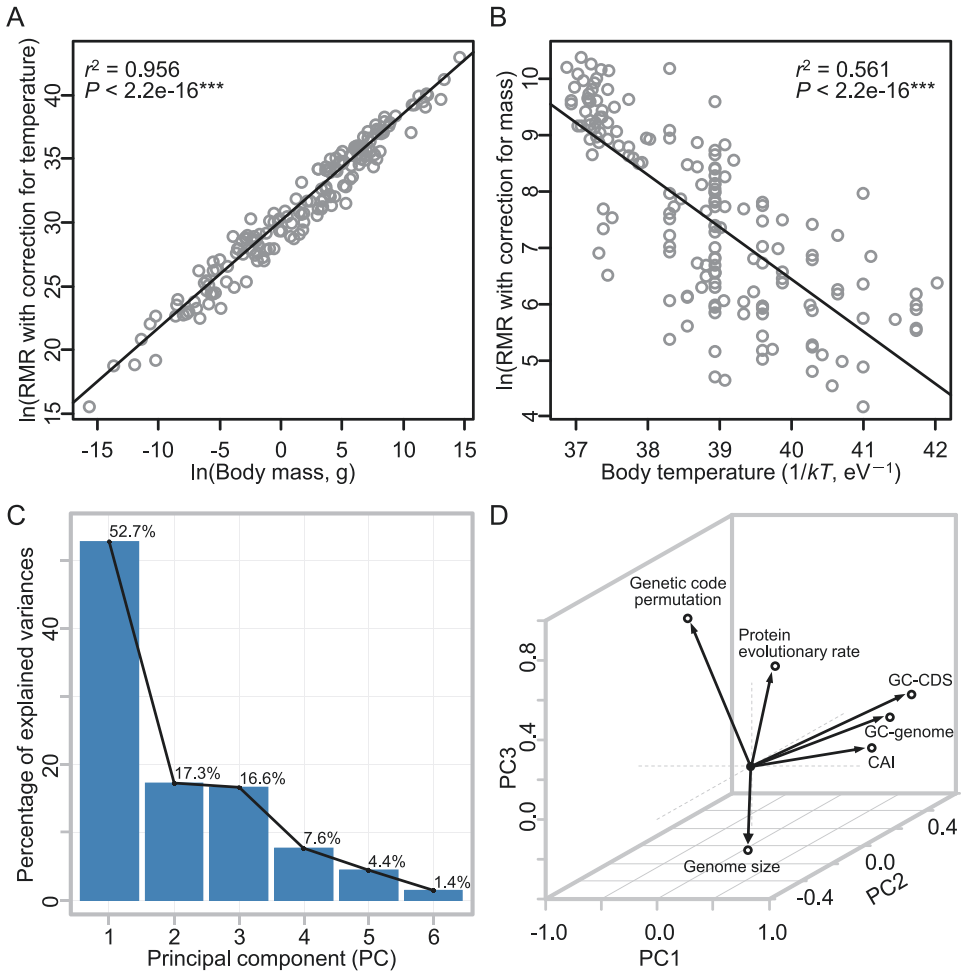

Figure 1. Evaluation of sources of variations in resting metabolic rate (RMR) and dimension reduction of six genomic traits. (A) Correlation of RMR with body mass. The ln(RMR) is a measure of RMR corrected for the effect of body temperature with controlling for phylogenetic non-independence across species. The r^2 is square value of Pearson's correlation coefficient. *** indicates $p < 0.001$. (B) Correlation of RMR with body temperature. The ln(RMR) is a measure of RMR corrected for the effect of body mass with controlling for phylogenetic non-independence across species. The parameter k is Boltzmann constant (0.0000862 eV K^{-1}), and T is body temperature (see *Equation* 1 in main text). (C) Principal component analysis shows variances of the six genomic traits that are explained by principal components (PC1~6). Shown is an example with all data pooled together. The six genomic traits are the codon adaptation index (CAI), genomic GC content (GC-genome), GC content of genomic protein-coding sequences (GC-CDS), protein evolutionary rate, amino acid frequency expected from random permutations of genetic codon bases (Genetic code permutation), and the genome size. (D) The correlation coefficient between six genomic traits with PC1~3 was extracted by principal component analysis. Shown is an example with all data pooled together. The x-, y-, and z-axis values indicate the correlation coefficient.

genomic traits, and was then included as confounding factors in the model describing the relationship between $AA_{frequency}$ and RMR components (see *Methods* section).

Phylogenetic patterns of $AA_{frequency}$ and RMR across major animal lineages

The animal species studied here covered the major lineages of extant Eumetazoa ("true animal"), including vertebrates such as fishes, frogs, birds, and mammals, as well as invertebrates such as crustaceans, insects, arachnids, and mollusks (Figure 2). As shown in Figure 2, $AA_{frequency}$ and RMRs appeared to be more conserved in vertebrates than in invertebrates (Figure 2), and we acknowledged that the invertebrate lineages studied here covered a much greater phylogenetic depth.

With all species as a whole, significant strong phylogenetic signals (Pagel's λ close to one) were detected for both $AA_{frequency}$ (λ = 0.95–1.00, $p < 2.1e{-}29$) and RMR components (λ = 0.84–0.91, $p < 3.1e{-}25$). This suggested that compared with distantly related animals, closely related ones had more similar RMR and $AA_{frequency}$ in their proteomes, regardless of the RMR components and amino acid types concerned (Figure 2). The maximum difference in $AA_{frequency}$ ranged from 1.25 (Leu) to 2.66 folds (Asn) in pairwise species comparisons.

Evolutionary relationship between RMR and $AA_{frequency}$ across animals

In fitting the relationship between the $AA_{frequency}$ with RMR components (with PC1~3 of six genomic traits as confounding factors, see above), PGLS outperformed OLS approach for each of the 20 common amino acids ($\Delta AIC_c > 33$). Results showed that there was no general relationship: the regression slopes under PGLS approach varied and could be positive, negative, or non-significant among different amino acids and different RMR components (R_{mass}, $R_{temperature}$, $R_{residual}$) (Figure 3; Supplementary Table S3). For instance, the frequency of amino acid Pro had positive slope with R_{mass} (Figure 3A) and $R_{residual}$ (Figure 3C), yet non-significant

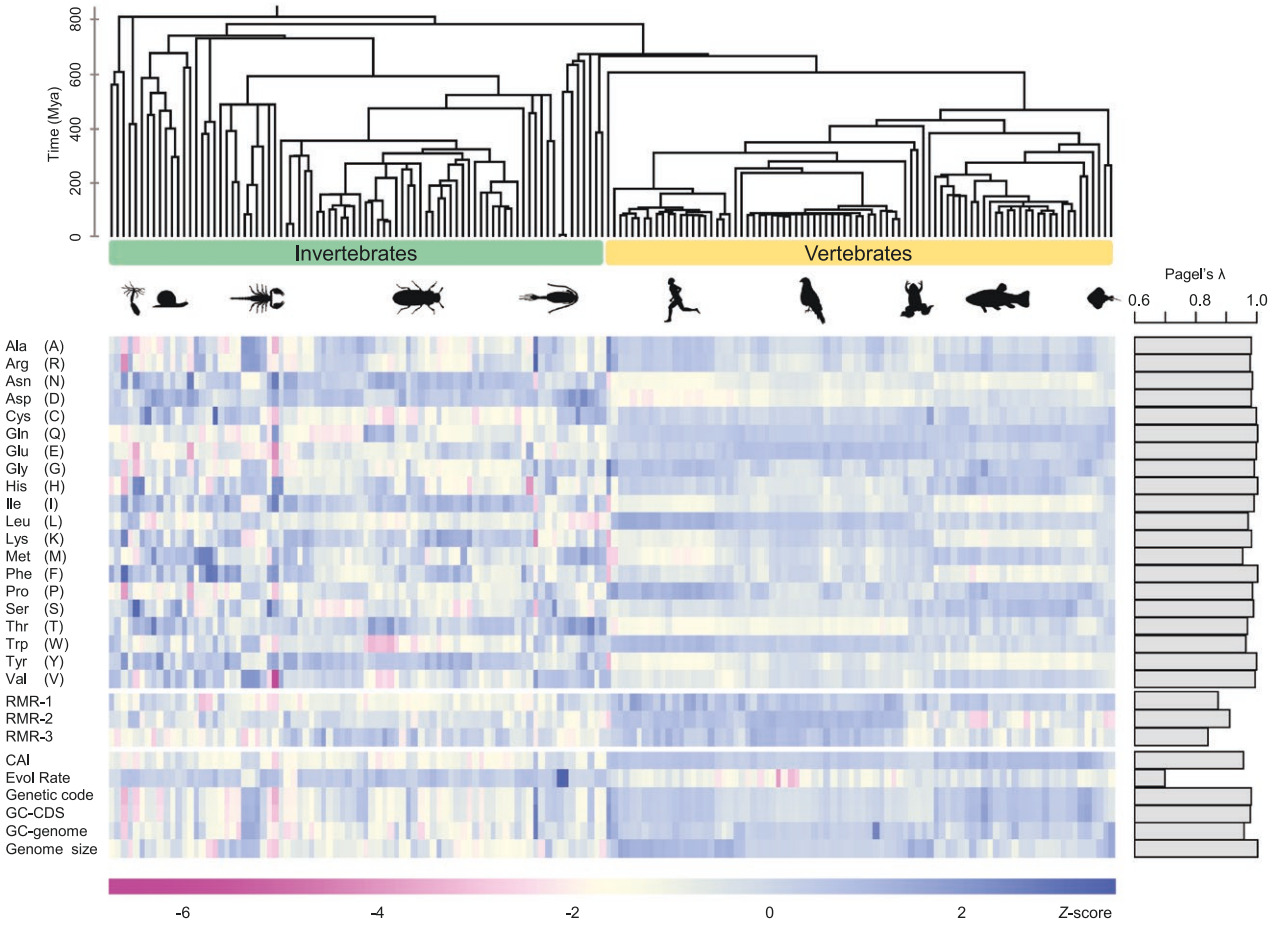

Figure 2. Interspecific variations and phylogenetic patterns of amino acid frequencies, resting metabolic rate (RMR), and six genomic traits among 166 animal species. Upper panel: evolutionary tree with branch lengths equal to time. Mya: million years ago. Lower panel: heatmap showing variations in amino acid frequencies, RMR, and six genomic traits expressed as *Z*-score values (see *Methods* section). One-letter amino acid abbreviations are shown in parentheses. RMR-1, RMR-2, and RMR-3 represent the body mass component of RMR (R_{mass}), the body temperature component of RMR ($R_{temperature}$), and the residual RMR ($R_{residual}$; the RMR component independent of body mass and temperature), respectively. The six genomic traits are codon adaptation index (CAI), protein evolutionary rate (Evol Rate), genetic code permutation as suggested by the "genetic code hypothesis" (Genetic code), GC content of genomic protein-coding sequences (GC-CDS), genomic GC content (GC-genome), and genome size. Right panel: barplot showing the phylogenetic signal (Pagel's **λ**) for each trait. The **λ** generally ranges between 0 and 1, with **λ** close to 1 indicating strong phylogenetic signal (phylogenetic dependence of the trait value).

(95% CIs overlapped with zero) slope with $R_{temperature}$ (Figure 3B), whereas that of the amino acid Ile had significantly negative slope with each of the three RMR components (Figure 3). Overall, under the PGLS approach, significant relationship between $AA_{frequency}$ with RMR components (95% CIs did not overlap with zero) existed in eight amino acids including Ala, Arg, Gly, Ile, Lys, Met, Phe, and Pro (Figure 3).

To further explore why the interspecific relationship varied and was not consistent, we analyzed the interrelationship between amino acid synthesis cost and each of the three RMR components (displayed as regression slopes in Figure 3). The results were shown in Figure 4. It indicated that although there was no significant correlation between the synthesis cost and either R_{mass} ($r = -0.16$, $p = 0.50$) (Figure 4A) or $R_{temperature}$ slope ($r = 0.18$, $p = 0.44$) (Figure 4B), there did exist a significant negative correlation between synthesis cost and slope of $R_{residual}$ ($r = -0.55$, $p = 0.01$) (Figure 4C). This leads to a pattern that low-energy-cost amino acids were more likely to be at higher frequency in animal species with high residual metabolic rate (see *Discussion* section). Hence, the inconsistent $R_{residual}$-$AA_{frequency}$ relationship was at least partially due to difference in amino acid synthesis cost.

We also tested whether protein abundance (experiment-based expression level) influenced the above observations by examining comparatively low- and high-expression proteins (see *Methods* section). Our results indicated that in most cases the slopes (85%) were similar (95% CIs overlapped) in both subdatasets, with only 15% of slopes differing significantly (95% CIs did not overlap). Likewise, a similar correlation (95% CIs overlapped) existed between amino acid synthesis cost and the slopes in both sub-datasets (Supplementary Figure S1). These observations thus indicated that protein abundance did not appear to significantly shape the observed RMR-$AA_{frequency}$ relationship, while we acknowledged that limitation in sample size due to the lack of enough expression data in animals did not allow us to conduct further test.

Quantitative contribution of RMR to $AA_{frequency}$ across animals

In terms of r^2_{lik} values, the studied predictors (RMR and six genomic traits) together explained up to 81.2% of variations in $AA_{frequency}$ across animals. To mitigate the confounding influence of genomic traits on the relative contribution of RMR to $AA_{frequency}$, species

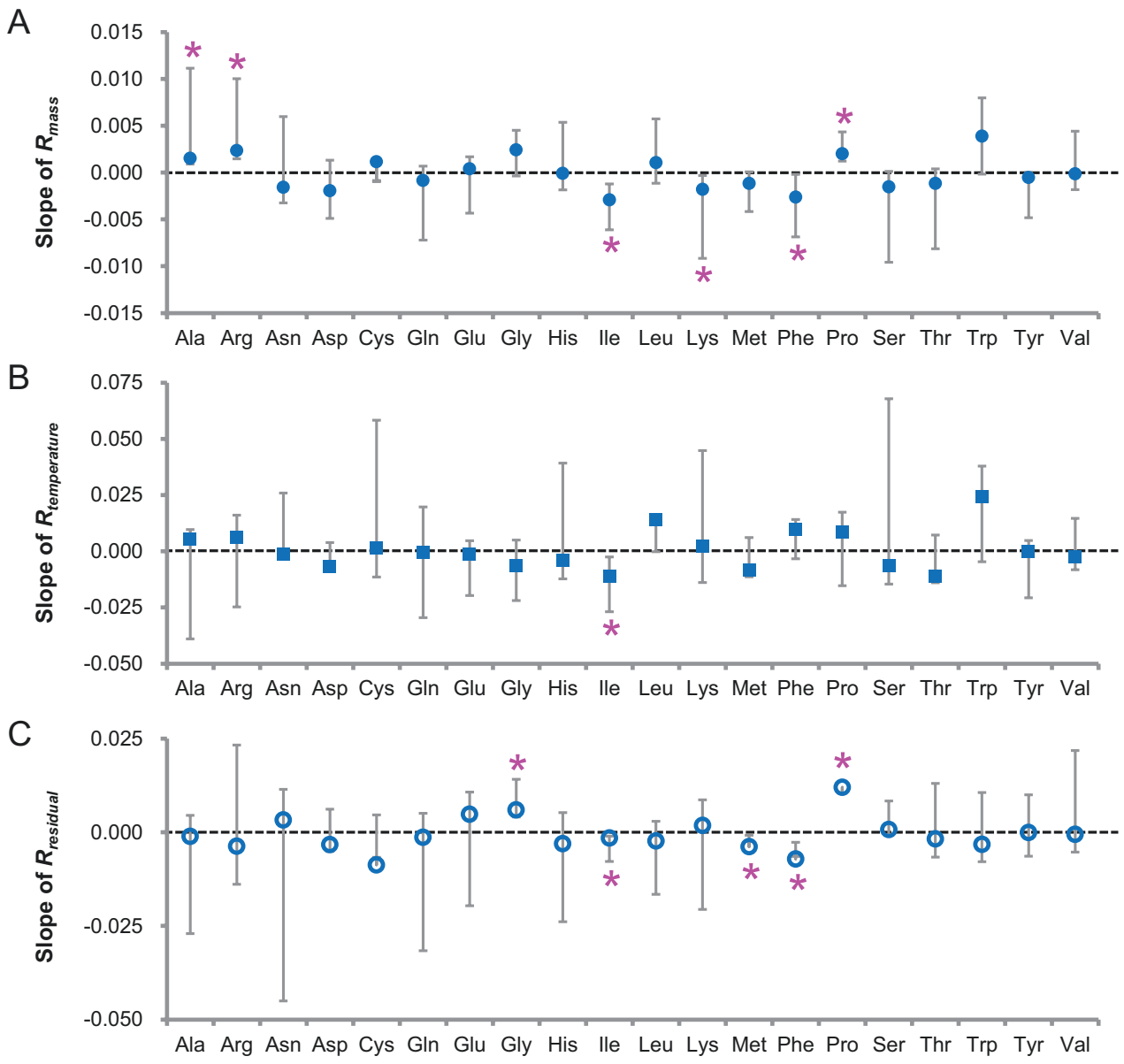

Figure 3. Results of the linear regressions between the frequencies of 20 common amino acids (in log unit) and the resting metabolic rate (RMR) using the phylogenetic generalized least squares (PGLS) approach. (A) Slopes of PGLS regressions between the body mass component of RMR (R_{mass}) and amino acid frequency ($AA_{frequency}$) of the 20 common amino acids. The error bar indicates the 95% CIs calculated by bootstrap approach. (B) Slopes of PGLS regressions between the body temperature component of RMR ($R_{temperature}$) and $AA_{frequency}$ of the 20 common amino acids. (C) Slopes of PGLS regressions between the residual RMR ($R_{residual}$; see *Methods* section) and $AA_{frequency}$ of the 20 common amino acids. * indicates significant slope, for which the 95% CIs did not overlap with zero. Estimates of the regression intercepts and slopes of other predictors are displayed in Supplementary Table S3.

with biased GC-CDS were excluded (see *Methods* section), yielding a dataset with 82 species (52 vertebrates and 30 invertebrates). On average, RMR made a relative contribution of 27.7% (r^2_{lik} for RMR = 2.0%–69.2%) to $AA_{frequency}$ (Figure 5). This level of contribution was significantly greater (Wilcoxon test $p < 0.05$) than those of GC-genome (mean r^2_{lik} = 11.9%), genome size (mean 2.9%), and protein evolutionary rate (mean 2.8%), and comparable to ($p = 0.48$–0.88) the remaining three genomic traits including GC-CDS, genetic code permutation, and CAI (mean 27.2%–33.2%) (Figure 5).

Evolution of RMR and $AA_{frequency}$ was adaptive

To examine whether the RMR-$AA_{frequency}$ relationship was driven by selection or random drift and to infer the evolutionary patterns, three evolutionary models (BM, OU1, and OUM) were compared. The BM model describes a random process, whereas OU model describes the adaptive process with either single (OU1) or multiple (OUM) adaptive optima caused by selection pressure. For each of the 20 amino acids and the three RMR components, OUM model was always the best fit among the three models ($\Delta AIC_c > 57$) (Table 1). This indicated that these variables had evolved following adaptive rather than random-drift process, and there existed multiple rather than a single adaptive optima across different evolutionary lineages. Selection was further suggested to underline the significant $AA_{frequency}$-RMR relationships observed for eight amino acids (Ala, Arg, Gly, Ile, Lys, Met, Phe, Pro; Figure 3).

Discussion

Connection between metabolic rate and $AA_{frequency}$ across animal lineages

Our results indicate that RMR played an important role in the variation of $AA_{frequency}$ across animal lineages, with a contribution

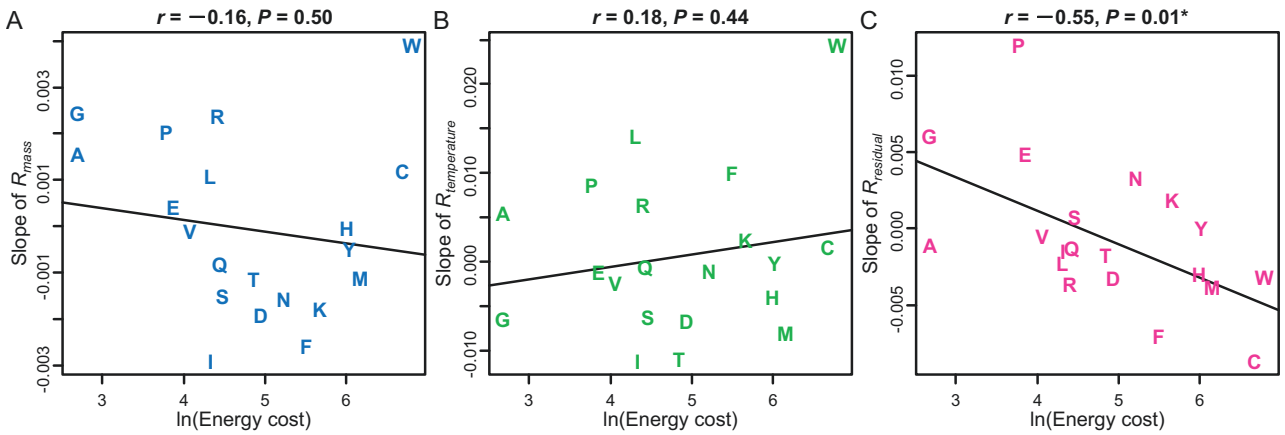

Figure 4. Effects of the energy cost of amino acid biosynthesis on regression slope between resting metabolic rate (RMR) and amino acid frequency. (A) The effect of biosynthetic energy cost on the regression slope between amino acid frequency and body mass component of RMR (i.e., the slope of R_{mass}). Values of regression slopes are obtained from Figure 3 and Supplementary Table S3. *r*: Pearson's correlation coefficient. The 20 common amino acids are indicated as one-letter amino acid abbreviations, i.e., A: Ala; C: Cys; D: Asp; E: Glu; F: Phe; G: Gly; H: His; I: Ile; K: Lys; L: Leu; M: Met; N: Asn; P: Pro; Q: Gln; R: Arg; S: Ser; T: Thr; V: Val; W: Trp; Y: Tyr. (B) The effect of biosynthetic energy cost on the slope of regression between amino acid frequency and body temperature component of RMR (the slope of $R_{temperature}$). (C) The effect of biosynthetic energy cost on the slope of regression between amino acid frequency and residual RMR (the slope of $R_{residual}$; see *Methods* section for details about $R_{residual}$). * indicates statistical significance ($p < 0.05$).

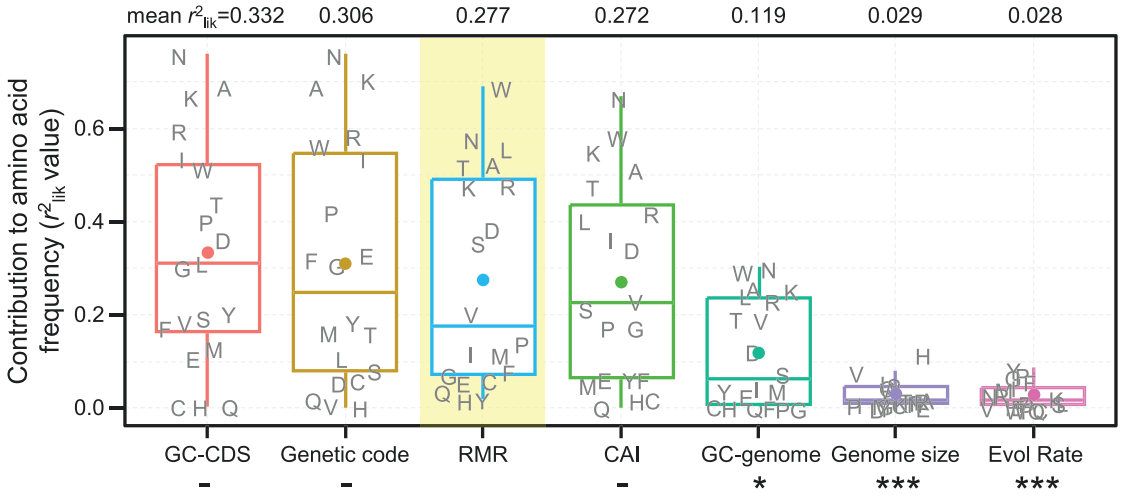

Figure 5. Contributions of predictors to variations in the frequency of 20 common amino acids among 82 species with similar GC content. Contributions are quantified by the likelihood-based square values of the correlation coefficient (r^2_{lik}). Predictors shown here include the resting metabolic rate (RMR highlighted in the graph) as well as six genomic traits: GC content of genomic protein-coding sequences (GC-CDS), genetic code permutation as suggested by the "genetic code hypothesis" (Genetic code), codon adaptation index (CAI), genomic GC content (GC-genome), genome size, and protein evolutionary rate (Evol Rate). The r^2_{lik} values are inferred from animal species with exclusion of those showing biased GC-CDS content. See *Methods* section for details on how the GC-CDS-biased species are identified. Predictors are ordered according to their mean r^2_{lik} values, which were shown as colored dots in the boxplot (with their values also shown on the top of the boxplot). * indicates Wilcoxon test $p < 0.05$ in comparison between RMR with one of the genomic traits; *** indicates $p < 0.001$; - indicates $p > 0.05$. The 20 common amino acids are indicated as one-letter amino acid abbreviations, i.e., A: Ala; C: Cys; D: Asp; E: Glu; F: Phe; G: Gly; H: His; I: Ile; K: Lys; L: Leu; M: Met; N: Asn; P: Pro; Q: Gln; R: Arg; S: Ser; T: Thr; V: Val; W: Trp; Y: Tyr.

comparable to or even much greater than the genomic traits investigated here. We showed that RMR explains on average 27.7% of the variance among animal lineages where the confounding effects of nucleotide compositional biases were controlled (Figure 5). Our comparative analysis of the experiment-based proteome abundance data indicated that protein abundance did not appear to significantly shape the observed RMR-$AA_{frequency}$ relationship, although this preliminary conclusion deserves further testing when more expression data become available.

Nevertheless, the RMR-$AA_{frequency}$ relationship varies greatly in different amino acids, even with an account of the differences from several genomic traits. For example, high RMR appears to exert a positive influence on the amino acid Pro (hence at high frequency) but a negative influence on Ile (hence at low frequency) in the proteomes. That is, no general significant relationship can be detected, as shown by the regression slopes under PGLS approach, which varied among amino acids and RMR components and can be positive, negative, or non-significant (Figure 3). Such a varying influencing pattern of RMR was in parallel with the effect of GC content observed by Moura et al. (2013). They reported that higher level of GC content resulted in increased $AA_{frequency}$ for several amino acids such as Gly and Val, but decreased $AA_{frequency}$ for some other ones such as Ile and Lys (Moura et al., 2013). Taken together, these observations signal

Table 1. Model-fitting results showing that the evolution of amino acid frequency and resting metabolic rate (RMR) follows adaptive evolution.

Amino acid and RMR	BM model		OU1 model		OUM model	
	AIC_c	ΔAIC_c	AIC_c	ΔAIC_c	AIC_c	ΔAIC_c
Ala	−326.15	98.71	−324.07	100.78	**−424.86**	0
Arg	−444.06	144.41	−441.99	146.48	**−588.47**	0
Asn	−311.25	134.24	−309.17	136.31	**−445.48**	0
Asp	−735.93	128.93	−733.86	131.00	**−864.86**	0
Cys	−616.26	156.05	−614.18	158.13	**−772.31**	0
Gln	−646.43	212.75	−644.36	214.82	**−859.18**	0
Glu	−711.55	156.66	−709.48	158.74	**−868.22**	0
Gly	−459.46	134.24	−457.39	136.32	**−593.70**	0
His	−701.33	116.91	−699.25	118.98	**−818.24**	0
Ile	−345.38	107.28	−343.30	109.35	**−452.66**	0
Leu	−767.66	143.01	−765.59	145.08	**−910.67**	0
Lys	−416.07	163.65	−414.00	165.73	**−579.72**	0
Met	−609.24	153.57	−609.74	153.06	**−762.81**	0
Phe	−552.66	87.09	−550.58	89.17	**−639.75**	0
Pro	−451.04	100.18	−448.97	102.26	**−551.22**	0
Ser	−696.79	134.39	−694.71	136.47	**−831.18**	0
Thr	−650.89	96.07	−649.12	97.85	**−746.96**	0
Trp	−558.83	203.09	−566.94	194.98	**−761.92**	0
Tyr	−551.51	103.51	−549.44	105.59	**−655.03**	0
Val	−650.79	214.68	−648.72	216.76	**−865.47**	0
R_{mass}	896.99	156.25	877.54	136.81	**740.74**	0
$R_{temperature}$	303.50	231.47	274.55	202.53	**72.02**	0
$R_{residual}$	414.97	72.53	400.07	57.63	**342.44**	0

Note. The *R* represents the resting metabolic rate (RMR). R_{mass} is the body mass component of RMR; $R_{temperature}$ is the body temperature component of RMR; $R_{residual}$ is the RMR component independent of body mass and temperature. BM = Brownian motion model; OUM = multiple-regime Ornstein–Uhlenbeck (OU) model; OU1 = single-regime OU model. OUM and OU1 are adaptive models, whereas BM is non-adaptive model. The lower the AIC_c value, the better the model. The ΔAIC_c value is calculated relative to the lowest AIC_c (in bold); $\Delta AIC_c > 4$ indicates good support. OUM is thus determined as the best model with strong support.

that interspecific $AA_{frequency}$ variations are shaped by multiple interacting factors during evolution, and the observed outcomes reflect the trade-off among these factors in a specific evolutionary context conforming to the physiology and ecology of the organisms under consideration (see below).

The mechanisms for the aforementioned evolutionary trade-off are likely functioning hierarchically. Namely, while RMR can significantly affect $AA_{frequency}$ via the optimization of energy allocation to protein biosynthesis, other factors acting at some other levels could dim this effect in the course of evolution. The synthesis cost of amino acid is one of them (see the next section; Figure 4C), and another crucial factor could be the requirement to maximize the protein sequence entropy to enable protein function diversity (Krick et al., 2014). The trade-off between optimization in energy utilization and diversification in protein sequence/function may help to reconcile the varying RMR-$AA_{frequency}$ relationships observed here. Taking the amino acid Leu as an example, decreased Leu content in numerous proteins is favored by many organisms that have adapted to the low-temperature environment (Berthelot et al., 2019). Meanwhile, the low environmental temperature would increase the metabolic rate ($R_{residual}$) of many animals (Addo-Bediako et al., 2002; Kovac et al., 2022; White et al., 2012). Therefore, low environmental temperature became connected with low Leu frequency and high $R_{residual}$ in these situations, thereby resulting in a negative relationship between these two traits. Such a negative relationship due to low-temperature adaptation masked the positive effect of RMR, hence ultimately leading to the observed non-significant slope between Leu frequency and $R_{residual}$ in Figure 3C. Our observations thus provide evidence for the supposition that effects of minimizing protein synthesis cost at molecular level usually vary with different ecological and metabolic strategies at the whole-organism level (Seligmann, 2003).

The above analyses lead to a general conclusion that there does not exist a universally applicable "one-size-fits-all" predictor for the frequencies of the 20 common amino acids (Figure 5; Knight et al., 2001; Figure 1B therein; Moura et al., 2013, Table 3 therein), and integrated investigations on interacting traits across different levels of biological hierarchy of organization are indispensable, so that the evolution of $AA_{frequency}$ can be both qualitatively and quantitatively elucidated.

Relationship between RMR and $AA_{frequency}$ was the evolutionary consequence of adaptation

Our data demonstrated that the $R_{residual}$-$AA_{frequency}$ relationship (the regression slope) is significantly and negatively associated with the energy cost of amino acid biosynthesis (Figure 4C). This means that low-energy-cost amino acids are more likely to display a positive response (i.e., slope > 0) to $R_{residual}$ than high-energy-cost amino acids. The logic is as follows. Species with high $R_{residual}$ will have more energy allocated to amino acid biosynthesis and thus may show high $AA_{frequency}$. However, given the cost variations in amino acid biosynthesis, there exist two possible strategies on the energy allocation: (1) more energy allocated to high-cost amino acids, or alternatively (2) more energy allocated to low-cost amino acids. Because energy resources are generally limited for animals, energy allocation to protein biosynthesis has to be optimized (Akashi & Gojobori, 2002; Heizer et al., 2006; Seligmann, 2003) given the adaptive nature of metabolic rate. Therefore, if natural selection exerts influence on $AA_{frequency}$ via the optimization of energy utilization, the second rather than the first strategy is expected, since more low-cost amino acids can be biosynthesized than high-cost ones with a given amount of energy flux. Consequently, less energetically costly amino acids are favored and thus have higher frequency in proteomes of high-$R_{residual}$ species. This is exactly what we have observed in the analysis

(Figure 4C), which is consistent with the pattern proposed by the "synthesis cost hypothesis" (see *Introduction* section). Our finding has hence provided evidence to earlier observation that the energy-cheap amino acids should have been preferred due to selection pressure (Krick et al., 2014; Seligmann, 2003; Swire, 2007) and importantly, extended it to among-species scenarios. Therefore, natural selection serves as a general explanation for both intraspecific and interspecific variations in $AA_{frequency}$.

Our analyses of the evolutionary model fitting further support that the observed RMR-$AA_{frequency}$ relationship is the outcome of correlated evolution driven by adaptation. While in comparative analyses across species, two traits can be seemingly related because they are a legacy from the ancestors (i.e., due to shared phylogenetic history) rather than truly correlated as a result of correlated evolution (Blomberg et al., 2003; Prinzing et al., 2001; Wang et al., 2022), the PGLS approach adopted here took into account of the influence of phylogenetic relatedness. Our results showed that correlated evolution is responsible for the correlations between RMR and the frequencies of eight amino acids (Ala, Arg, Gly, Ile, Lys, Met, Phe, and Pro), as shown by their significant slopes under PGLS (Figure 3). Our evolutionary model-fitting analyses furthermore showed that the adaptive OUM model fitted the data best than other models (including the random-drift model), regardless of which RMR component or amino acid is concerned (Table 1). These results indicate that evolutionary adaptation is the driving force of the observed correlated evolution. This is not unexpected given that both metabolic rate and $AA_{frequency}$ have been documented to manifest adaptive response to variables such as environmental temperature and resources (Addo-Bediako et al., 2002; Arnqvist et al., 2022; McNab, 2015; Moura et al., 2013; Tekaia et al., 2002).

Conclusions

The results of our exploratory investigation on the relationship between the organism-level RMR and molecular-level $AA_{frequency}$ were in line with the "metabolic rate hypothesis" of $AA_{frequency}$ evolution in animals. We revealed that RMR makes a non-negligible contribution to the variance of $AA_{frequency}$ (mean 27.7%, 2.0%–69.2%) across animal lineages, with an effect not less than any of the genomic traits studied when the nucleotide compositional bias was controlled. We found that low-energy-cost amino acids are more likely to be at higher frequency in animal species with high residual metabolic rate. We further showed that the relationship between RMR and $AA_{frequency}$ varies greatly among amino acids, and such a variation is most likely an outcome of trade-offs among various interacting factors (e.g., metabolic optimum and protein sequence diversity) in an evolutionary context conforming to the physiology/ecology of the lineages under consideration. We also demonstrated that the significant RMR-$AA_{frequency}$ correlations observed were driven by adaptation rather than random drift. Clearly, given that there is unlikely any universally applicable "one-size-fits-all" predictor, integrated investigations on various interacting factors across multiple levels of biological hierarchy of organization are vital for a fuller understanding of the mechanisms of interspecific $AA_{frequency}$ variations in animals. In particular, more data of metabolic rate and proteome expression from phylogenetically diverse animal species will greatly promote a better understanding of the RMR-$AA_{frequency}$ relationship and thus allow to further test the "metabolic rate hypothesis."

Supplementary material

Supplementary material is available online at *Evolution Letters*.

Data and code availability

Data details are provided in Supplementary Table S1.

Author contributions

W.W. and D.-X.Z. designed the study. D.-X.Z. supervised the study. W.W. compiled the data and performed the analysis. W.W. drafted the manuscript. W.W. and D.-X.Z revised the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (grant no. 31672271) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDB13030200).

Conflict of interest. The authors declare no conflict of interest.

Acknowledgments

The authors thank the researchers for submitting, managing, and sharing the genomic sequence data in NCBI. We are grateful to the editors and anonymous reviewers for their critical and constructive comments and suggestions.

References

Addo-Bediako, A., Chown, S. L., & Gaston, K. J. (2002). Metabolic cold adaptation in insects: A large-scale perspective. *Functional Ecology*, *16*(3), 332–338.

Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(6), 3695–3700. https://doi.org/10.1073/pnas.062526999

Arnqvist, G., Rönn, J., Watson, C., Goenaga, J., & Immonen, E. (2022). Concerted evolution of metabolic rate, economics of mating, ecology, and pace of life across seed beetles. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(33), e2205564119. https://doi.org/10.1073/pnas.2205564119

Arribas, P., Andújar, C., Moraza, M. L., Linard, B., Emerson, B. C., & Vogler, A. P. (2020). Mitochondrial metagenomics reveals the ancient origin and phylodiversity of soil mites and provides a phylogeny of the Acari. *Molecular Biology and Evolution*, *37*(3), 683–694. https://doi.org/10.1093/molbev/msz255

Baeza, M., Zúñiga, S., Peragallo, V., Barahona, S., Alcaino, J., & Cifuentes, V. (2021). Identification of stress-related genes and a comparative analysis of the amino acid compositions of translated coding sequences based on draft genome sequences of Antarctic yeasts. *Frontiers in Microbiology*, *12*, 623171. https://doi.org/10.3389/fmicb.2021.623171

Berthelot, C., Clarke, J., Desvignes, T., William Detrich H III, Flicek, P., Peck, L. S., Peters, M., Postlethwait, J. H., & Clark, M. S. (2019). Adaptation of proteins to the cold in Antarctic fish: A role for methionine? *Genome Biology and Evolution*, *11*(1), 220–231.

Blaxter M. (2009). Nematodes (Nematoda). In S. B. Hedges & S. Kumar (Eds.), *The timetree of life* (pp. 247–250). Oxford University Press.

Blomberg, S. P., Garland, T. Jr, & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, *57*(4), 717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x

Botzman, M., & Margalit, H. (2011). Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biology*, *12*(10), R109. https://doi.org/10.1186/gb-2011-12-10-r109

Brandl, S. J., Lefcheck, J. S., Bates, A. E., Rasher, D. B., & Norin, T. (2023). Can metabolic traits explain animal community assembly and functioning? *Biological Reviews of the Cambridge Philosophical Society*, *98*(1), 1–18. https://doi.org/10.1111/brv.12892

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., & West, G. B. (2004). Toward a metabolic theory of ecology. *Ecology*, *85*(7), 1771–1789. https://doi.org/10.1890/03-9000

Burger, J. R., Hou, C., & Brown, J. H. (2019). Toward a metabolic theory of life history. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(52), 26653–26661. https://doi.org/10.1073/pnas.1907702116

Burnham KP, Anderson DR. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer-Verlag.

Carbone, A., Zinovyev, A., & Képès, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, *19*(16), 2005–2015. https://doi.org/10.1093/bioinformatics/btg272

Carter, M. J., Cortes, P. A., & Rezende, E. L. (2023). Temperature variability and metabolic adaptation in terrestrial and aquatic ectotherms. *Journal of Thermal Biology*, *115*, 103565. https://doi.org/10.1016/j.jtherbio.2023.103565

Chen, Q., Yang, H., Feng, X., Chen, Q., Shi, S., Wu, C. I., & He, Z. (2022). Two decades of suspect evidence for adaptive molecular evolution-negative selection confounding positive-selection signals. *National Science Review*, *9*(5), nwab217. https://doi.org/10.1093/nsr/nwab217

Chen, Y., & Nielsen, J. (2022). Yeast has evolved to minimize protein resource cost for synthesizing amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2114622119. https://doi.org/10.1073/pnas.2114622119

Craig, C. L., & Weber, R. S. (1998). Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by *Escherichia coli*. *Molecular Biology and Evolution*, *15*(6), 774–776. https://doi.org/10.1093/oxfordjournals.molbev.a025981

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, *10*, 210. https://doi.org/10.1186/1471-2148-10-210

Cutter, A. D., Wasmuth, J. D., & Blaxter, M. L. (2006). The evolution of biased codon and amino acid usage in nematode genomes. *Molecular Biology and Evolution*, *23*(12), 2303–2315. https://doi.org/10.1093/molbev/msl097

de Jong, M. J., van Oosterhout, C., Hoelzel, A. R., & Janke, A. (2023). Moderating the neutralist-selectionist debate: Exactly which propositions are we debating, and which arguments are valid? *Biological Review*, *99*(1), 23–55. https://doi.org/10.1111/brv.13010

Delsuc, F., Philippe, H., Tsagkogeorga, G., Simion, P., Tilak, M. K., Turon, X., López-Legentil, S., Piette, J., Lemaire, P., & Douzery, E. J. P. (2018). A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biology*, *16*(1), 39. https://doi.org/10.1186/s12915-018-0499-2

Du, M. Z., Zhang, C., Wang, H., Liu, S., Wei, W., & Guo, F. B. (2018). The GC content as a main factor shaping the amino acid usage during bacterial evolution process. *Frontiers in Microbiology*, *9*, 2948. https://doi.org/10.3389/fmicb.2018.02948

Dyer, K. F. (1971). The quiet revolution: A new synthesis of biological knowledge. *Journal of Biological Education*, *5*(1), 15–24. https://doi.org/10.1080/00219266.1971.9653663

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, *125*(1), 1–15. https://doi.org/10.1086/284325

Garamszegi LZ. (2014). *Modern phylogenetic comparative methods and their application in evolutionary biology*. Springer-Verlag.

Giacometti, D., Bars-Closel, M., Kohlsdorf, T., de Carvalho, J. E., & Cury de Barros, F. (2022). Environmental temperature predicts resting metabolic rates in tropidurinae lizards. *Journal of Experimental Zoology. Part A, Ecological and Integrative Physiology*, *337*(9-10), 1039–1052. https://doi.org/10.1002/jez.2656

Gillooly, J. F., Brown, J. H., West, G. B., Savage, V. M., & Charnov, E. L. (2001). Effects of size and temperature on metabolic rate. *Science*, *293*(5538), 2248–2251. https://doi.org/10.1126/science.1061967

Gillooly, J. F., McCoy, M. W., & Allen, A. P. (2007). Effects of metabolic rate on protein evolution. *Biology Letters*, *3*(6), 655–659. https://doi.org/10.1098/rsbl.2007.0403

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *326*(1233), 119–157. https://doi.org/10.1098/rstb.1989.0106

Hansen, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, *51*(5), 1341–1351. https://doi.org/10.1111/j.1558-5646.1997.tb01457.x

Hayssen, V., & Lacy, R. C. (1985). Basal metabolic rates in mammals: Taxonomic differences in the allometry of BMR and body mass. *Comparative Biochemistry and Physiology A Comparative Physiology*, *81*(4), 741–754. https://doi.org/10.1016/0300-9629(85)90904-1

Heizer, E. M. Jr, Raiford, D. W., Raymer, M. L., Doom, T. E., Miller, R. V., & Krane, D. E. (2006). Amino acid cost and codon-usage biases in 6 prokaryotic genomes: A whole-genome analysis. *Molecular Biology and Evolution*, *23*(9), 1670–1680. https://doi.org/10.1093/molbev/msl029

Hessen, D. O., Daufresne, M., & Leinaas, H. P. (2013). Temperature-size relations from the cellular-genomic perspective. *Biological Reviews of the Cambridge Philosophical Society*, *88*(2), 476–489. https://doi.org/10.1111/brv.12006

Huang, Q., Szklarczyk, D., Wang, M., Simonovic, M., & von Mering, C. (2023). PaxDb 5.0: Curated protein quantification data suggests adaptive proteome changes in yeasts. *Molecular and Cellular Proteomics*, *22*(10), 100640. https://doi.org/10.1016/j.mcpro.2023.100640

Ingram, T., & Mahler, D. L. (2013). SURFACE: Detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods in Ecology and Evolution*, *4*(5), 416–425. https://doi.org/10.1111/2041-210x.12034

Ives, A. R. (2019). R^2s for correlated data: Phylogenetic models, LMMs, and GLMMs. *Systematic Biology*, *68*(2), 234–251. https://doi.org/10.1093/sysbio/syy060

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

King, J. L., & Jukes, T. H. (1969). Non-Darwinian evolution. *Science*, *164*(3881), 788–798. https://doi.org/10.1126/science.164.3881.788

Knight, R. D., Freeland, S. J., & Landweber, L. F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, *2*(4), RESEARCH0010. https://doi.org/10.1186/gb-2001-2-4-research0010

Kovac, H., Käfer, H., Petrocelli, I., & Stabentheiner, A. (2022). The respiratory metabolism of overwintering paper wasps gynes (*Polistes dominula* and *Polistes gallicus*). *Physiological Entomology*, *47*(1), 62–71.

Krick, T., Verstraete, N., Alonso, L. G., Shub, D. A., Ferreiro, D. U., Shub, M., & Sánchez, I. E. (2014). Amino acid metabolism conflicts with protein diversity. *Molecular Biology and Evolution*, *31*(11), 2905–2912. https://doi.org/10.1093/molbev/msu228

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lighton JRB. (2008). *Measuring metabolic rates: A manual for scientists*. Oxford University Press.

Lighton, J. R. B., Brownell, P. H., Joos, B., & Turner, R. J. (2001). Low metabolic rate in scorpions: Implications for population biomass and cannibalism. *Journal of Experimental Biology*, *204*(Pt 3), 607–613. https://doi.org/10.1242/jeb.204.3.607

Liu, J., Zhang, Y., Lei, X., & Zhang, Z. (2008). Natural selection of protein structural and functional properties: A single nucleotide polymorphism perspective. *Genome Biology*, *9*(4), R69. https://doi.org/10.1186/gb-2008-9-4-r69

Makarieva, A. M., Gorshkov, V. G., Li, B. L., Chown, S. L., Reich, P. B., & Gavrilov, V. M. (2008). Mean mass-specific metabolic rates are strikingly similar across life's major domains: Evidence for life's metabolic optimum. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(44), 16994–16999. https://doi.org/10.1073/pnas.0802148105

Martin, A. P., & Palumbi, S. R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(9), 4087–4091. https://doi.org/10.1073/pnas.90.9.4087

McCue, M. D. (2010). Starvation physiology: Reviewing the different strategies animals use to survive a common challenge. *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, *156*(1), 1–18. https://doi.org/10.1016/j.cbpa.2010.01.002

McNab, B. K. (2015). Behavioral and ecological factors account for variation in the mass-independent energy expenditures of endotherms. *Journal of Comparative Physiology B, Biochemical, Systemic, and Environmental Physiology*, *185*(1), 1–13. https://doi.org/10.1007/s00360-014-0850-z

Moura, A., Savageau, M. A., & Alves, R. (2013). Relative amino acid composition signatures of organisms and environments. *PLoS One*, *8*(10), e77319. https://doi.org/10.1371/journal.pone.0077319

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., & Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, *3*(4), 743–756. https://doi.org/10.1111/j.2041-210x.2012.00196.x

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Pagel, M. D. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884. https://doi.org/10.1038/44766

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., Alfaro, M. E., & Harmon, L. J. (2014). geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*(15), 2216–2218. https://doi.org/10.1093/bioinformatics/btu181

Prinzing, A., Durka, W., Klotz, S., & Brandl, R. (2001). The niche of higher plants: Evidence for phylogenetic conservatism. *Proceedings Biological Sciences*, *268*(1483), 2383–2389. https://doi.org/10.1098/rspb.2001.1801

Quinn, G. P., & Keough, M. J. (2002). *Experimental designs and data analysis for biologists*. Cambridge University Press.

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217–223. https://doi.org/10.1111/j.2041-210x.2011.00169.x

Rolfe, D. F., & Brown, G. C. (1997). Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiological Reviews*, *77*(3), 731–758. https://doi.org/10.1152/physrev.1997.77.3.731

Seligmann, H. (2003). Cost-minimization of amino acid usage. *Journal of Molecular Evolution*, *56*(2), 151–161. https://doi.org/10.1007/s00239-002-2388-z

Sharma, A., Gupta, S., & Paul, K. (2023). Codon usage behavior distinguishes pathogenic *Clostridium* species from the non-pathogenic species. *Gene*, *873*, 147394. https://doi.org/10.1016/j.gene.2023.147394

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index – A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. https://doi.org/10.1093/nar/15.3.1281

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smith, E., & Morowitz, H. J. (2004). Universality in intermediary metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(36), 13168–13173.

Sueoka, N. (1961). Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harbor Symposia on Quantitative Biology*, *26*, 35–43. https://doi.org/10.1101/sqb.1961.026.01.009

Suzuki, Y., & Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, *16*(10), 1315–1328. https://doi.org/10.1093/oxfordjournals.molbev.a026042

Swire, J. (2007). Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *Journal of Molecular Evolution*, *64*(5), 558–571. https://doi.org/10.1007/s00239-006-0206-8

Szarski, H. (1983). Cell size and the concept of wasteful and frugal evolutionary strategies. *Journal of Theoretical Biology*, *105*(2), 201–209. https://doi.org/10.1016/s0022-5193(83)80002-2

Tekaia, F., Yeramian, E., & Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. *Gene*, *297*(1-2), 51–60. https://doi.org/10.1016/s0378-1119(02)00871-5

Wang, G., Ives, A. R., Zhu, H., Tan, Y., Chen, S. C., Yang, J., & Wang, B. (2022). Phylogenetic conservatism explains why plants are more likely to produce fleshy fruits in the tropics. *Ecology*, *103*(1), e03555. https://doi.org/10.1002/ecy.3555

White, C. R., Alton, L. A., & Frappell, P. B. (2012). Metabolic cold adaptation in fishes occurs at the level of whole animal, mitochondria and enzyme. *Proceedings Biological Sciences*, *279*(1734), 1740–1747. https://doi.org/10.1098/rspb.2011.2060

White, C. R., & Kearney, M. R. (2013). Determinants of inter-specific variation in basal metabolic rate. *Journal of Comparative Physiology B, Biochemical, Systemic, and Environmental Physiology*, *183*(1), 1–26. https://doi.org/10.1007/s00360-012-0676-5

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, *155*(1), 431–449. https://doi.org/10.1093/genetics/155.1.431

Zhang, H., Wang, Y., Li, J., Chen, H., He, X., Zhang, H., Liang, H., & Lu, J. (2018). Biosynthetic energy cost for amino acids decreases in cancer evolution. *Nature Communications*, *9*(1), 4124. https://doi.org/10.1038/s41467-018-06461-1