Little evidence the standard genetic code is optimized for resource conservation

Hana Rozhoňová[1,2], Joshua L. Payne[1,2,*]

1. Institute of Integrative Biology, ETH Zurich, Universitätstrasse 16, 8092, Zurich, Switzerland

2. Swiss Institute of Bioinformatics, Quartier UNIL-Sorge, 1015, Lausanne, Switzerland

* Corresponding author: joshua.payne@env.ethz.ch

Abstract: Shenhav & Zeevi (*Science*, 2020, 370:683-687) propose nitrogen and carbon conservation as optimization principles in the standard genetic code that are not confounded by the established optimizations for polar requirement and hydropathy. We show their results for nitrogen conservation are highly sensitive to their choice of null model and their results for carbon conservation are confounded by molecular volume.

Main text:

The missense mutations allowed by the standard genetic code tend to preserve key physicochemical properties of amino acids, particularly polar requirement, hydropathy (*1*), and to a lesser extent, molecular volume (*2*). Quantitative tests of such optimization principles are typically based on comparisons to large numbers of alternative codes (*3*). There are many ways to generate alternative codes (*4, 5*), each of which maintain some aspects of the standard code, while perturbing others. For example, a common approach is to permute the 20 amino acids amongst the codon blocks, which maintains the block structure of the standard code, but changes which codon blocks correspond to which amino acids (*1*).

Shenhav & Zeevi propose resource conservation as an additional optimization in the standard code, which provides "robustness against mutations that increase carbon and nitrogen incorporation into protein sequences" (*6*). They draw this conclusion by comparing the expected random mutation cost (ERMC) of missense mutations allowed by the standard code to the ERMC of missense mutations allowed by alternative codes, measured in carbon or nitrogen atoms. The alternative codes they study are constrained to maintain the number of codons per amino acid (e.g., arginine always has six codons in their alternative codes) and to have the same degeneracy in the third base as the standard code (e.g., the three codons for

isoleucine always have the same first and second base as the codon for methionine in their alternative codes) (*7*).

We repeat Shenhav & Zeevi's analysis of resource conservation for nitrogen and carbon by comparing the ERMC of the standard code to the ERMC of alternative codes generated by amino acid permutation (*1*). We find no statistical support for nitrogen conservation when considering all codon frequencies to be equal and the transition:transversion ratio to be 1:2 ($p = 0.485$, Fig. 1A) or when considering the codon frequencies of the 39 organisms and 11 transition:transversion ratios studied by Shenhav & Zeevi (minimum $p = 0.573$ after Benjamini-Hochberg correction for multiple testing, Fig. 1B).

What explains the qualitative difference between their results and ours? The key is the alternative codes Shenhav & Zeevi study always maintain the number of codons per amino acid, and specifically the number of codons for arginine, which is both the most nitrogen-rich amino acid and one of the three amino acids with the largest number of codons – six. In the standard code, four of the five other amino acids with more than one nitrogen atom (histidine, lysine, glutamine, tryptophan, but not asparagine) can be reached via missense mutation from an arginine codon. Because this is unlikely to be true in alternative codes, maintaining the number of codons per amino acid will almost always result in a higher ERMC for nitrogen. This effect is diminished when generating alternative codes by amino acid permutation. The reason is the number of codons for arginine decreases with high probability and cannot increase, because no amino acid has more than six codons. This is not the case for other amino acids with more than one nitrogen atom, whose number of codons can both increase or decrease. This means that among all missense mutations to amino acids with more than one nitrogen atom, the ones leading to arginine occur relatively less frequently, thus significantly decreasing the mean cost of missense mutations to amino acids with more than one nitrogen atom (1.494 vs. 1.914 nitrogen atoms, $p < 2.2e\text{-}16$, Welch two sample t-test), which renders nitrogen conservation statistically insignificant.

It is well known that the standard code is optimized for polar requirement and hydropathy (*1*). Shenhav & Zeevi used a hierarchical model to argue carbon conservation is not confounded by these established optimizations. While we corroborate their statistical support

for carbon conservation when comparing to alternative codes generated by amino acid permutation ($p = 0.009$ when considering equal codon abundances and transition:transversion ratio 1:2, Fig. 1A; median $p = 0.009$, maximum $p = 0.069$ after Benjamini-Hochberg correction for multiple testing across the codon frequencies of 39 organisms and 11 transition:transversion ratios in Fig. 1B), we reasoned that carbon conservation might be confounded by molecular volume (*8*), because the molecular volume of an amino acid is strongly correlated with its number of carbon atoms (Pearson's correlation 0.906, $p = 3.97e-08$, Fig. 2A) and the changes in amino acids' molecular volume and number of carbon atoms caused by missense mutations are therefore strongly correlated (Pearson's correlation 0.813, $p < 2.2e-16$, Fig 2B). We repeated Shenhav & Zeevi's hierarchical modeling approach, examining the subset of alternative codes generated by amino acid permutation that had a lower ERMC than the standard code for molecular volume, and tested whether this subset is also optimized for carbon. It is not ($p = 0.139$ when considering equal codon abundances and transition:transversion ratio 1:2, Fig. 2C; minimum $p = 0.190$ after Benjamini-Hochberg correction across the codon frequencies from 39 organisms and 11 transition:transversion ratios). It is also not optimized for carbon when using the alternative codes studied by Shenhav & Zeevi ($p = 0.089$ for equal codon abundances and transition:transversion ratio 1:2; minimum $p = 0.428$ after Benjamini-Hochberg correction across the codon frequencies from 39 organisms and 11 transition:transversion ratios). Thus, carbon conservation is confounded by the known conservative nature of the standard code with respect to the molecular volume of amino acids (*2*).

We repeated our analyses using eight additional methods for generating alternative codes, specifically those studied by Caporaso et al. (*4*) and Massey (*5*), as well as a variant of amino acid permutation, in which the number of codons per amino acid is allowed to change by at most two. These methods vary in the aspects of the standard code they maintain and perturb. Considering equal codon abundances and a transition:transversion ratio of 1:2, as well as the codon frequencies of the 39 organisms and 11 transition:transversion ratios studied by Shenhav & Zeevi (430 tests per method), we find that nitrogen conservation is only statistically significant when using one of the two methods that maintain the number of codons per amino acid (the Codon Shuffler from Caporaso et al. (*4*); 430 out of 430 tests, $p < 0.05$, Benjamini-Hochberg correction for multiple testing). For all of the remaining seven methods, nitrogen conservation is statistically insignificant ($p > 0.05$ for 3,010 out of 3,010

tests, Benjamini-Hochberg correction for multiple testing), and across all methods, carbon conservation is almost always confounded by molecular volume ($p > 0.05$ for 3402 out of 3440 tests, Benjamini-Hochberg correction for multiple testing).

We do not dispute Shenhav & Zeevi's finding that resource conservation can manifest in the genetic sequences of organisms adapted to nutrient-limited environments, but we find little evidence it manifests in the standard genetic code. Choosing an appropriate null model for the standard code is challenging, because methods for generating alternative codes are generally agnostic to the evolutionary history of the standard code, which can give the false impression that optimizations are the product of selection rather than a byproduct of the physical processes of gene duplication and mutation (*5*), and because extant non-standard codes, which are derived from the standard code (*9*), violate basic assumptions of the methods used to generate alternative codes. For example, extant non-standard codes do not always maintain the number of codons per amino acid or the block structure of the standard code. Our analyses therefore provide a cautionary tale regarding the choice of appropriate null model for identifying evolutionary design principles (*10-12*) and highlight the frailty of adaptive hypotheses (*13*) for optimizations in the standard genetic code (*5, 14*).

References and Notes:

1. D. Haig, L. D. Hurst, A quantitative measure of error minimization in the genetic code. *J Mol Evol* **33**, 412-417 (1991).
2. D. Haig, L. D. Hurst, A quantitative measure of error minimization in the genetic code. *J Mol Evol* **49**, 708 (1999).
3. S. J. Freeland, T. Wu, N. Keulmann, The case for an error minimizing standard genetic code. *Orig Life Evol Biosph* **33**, 457-477 (2003).
4. J. G. Caporaso, M. Yarus, R. Knight, Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J Mol Evol* **61**, 597-607 (2005).
5. S. E. Massey, A neutral origin for error minimization in the genetic code. *J Mol Evol* **67**, 510-516 (2008).
6. L. Shenhav, D. Zeevi, Resource conservation manifests in the genetic code. *Science* **370**, 683-687 (2020).
7. C. Alff-Steinberger, The genetic code and error transmission. *Proc Natl Acad Sci U S A* **64**, 584-591 (1969).
8. R. Grantham, Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-864 (1974).
9. R. D. Knight, S. J. Freeland, L. F. Landweber, Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet* **2**, 49-58 (2001).
10. Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, L. Stone, Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science* **305**, 1107; author reply 1107 (2004).
11. M. Lynch, The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet* **8**, 803-813 (2007).
12. R. W. Lusk, M. B. Eisen, Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLoS Genet* **6**, e1000829 (2010).
13. M. Lynch, The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A* **104 Suppl 1**, 8597-8604 (2007).
14. A. Stoltzfus, L. Y. Yampolsky, Amino acid exchangeability and the adaptive code hypothesis. *J Mol Evol* **65**, 456-462 (2007).

Acknowledgments: We thank Sinisa Bratulic and David M. McCandlish for discussions and feedback on this manuscript.

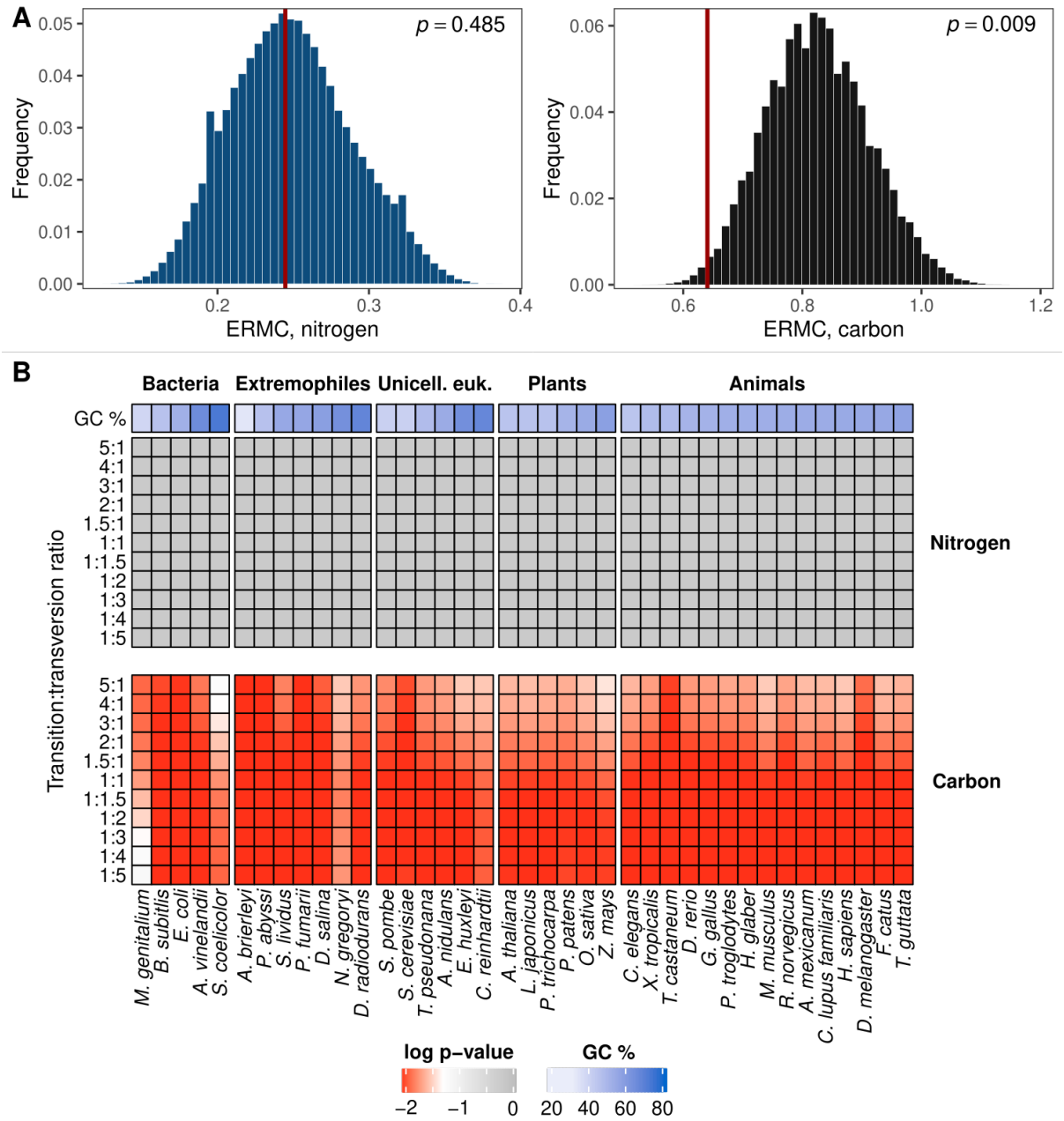

Fig. 1. Nitrogen conservation is sensitive to choice of null model.

1A - Histograms of the ERMC in 1 million alternative codes generated by amino acid permutation (*1*) for nitrogen (blue) and carbon (black). The vertical red line denotes the ERMC of the standard code.

1B - *P*-values of the ERMC of the standard code compared to 1 million randomized codes for nitrogen (top) and carbon (bottom), using codon abundances from 39 different organisms and 11 different transition:transversion ratios (*6*). Shades of red correspond to statistically significant *p*-values ($p<0.05$; darker = more significant) and shades of gray to statistically insignificant *p*-values ($p>0.05$; darker = less significant). The *p*-values were adjusted using Benjamini-Hochberg correction for multiple testing. Organisms in each group are ordered based on the GC content of their coding sequences. Unicell. euk. = unicellular eukaryotes.

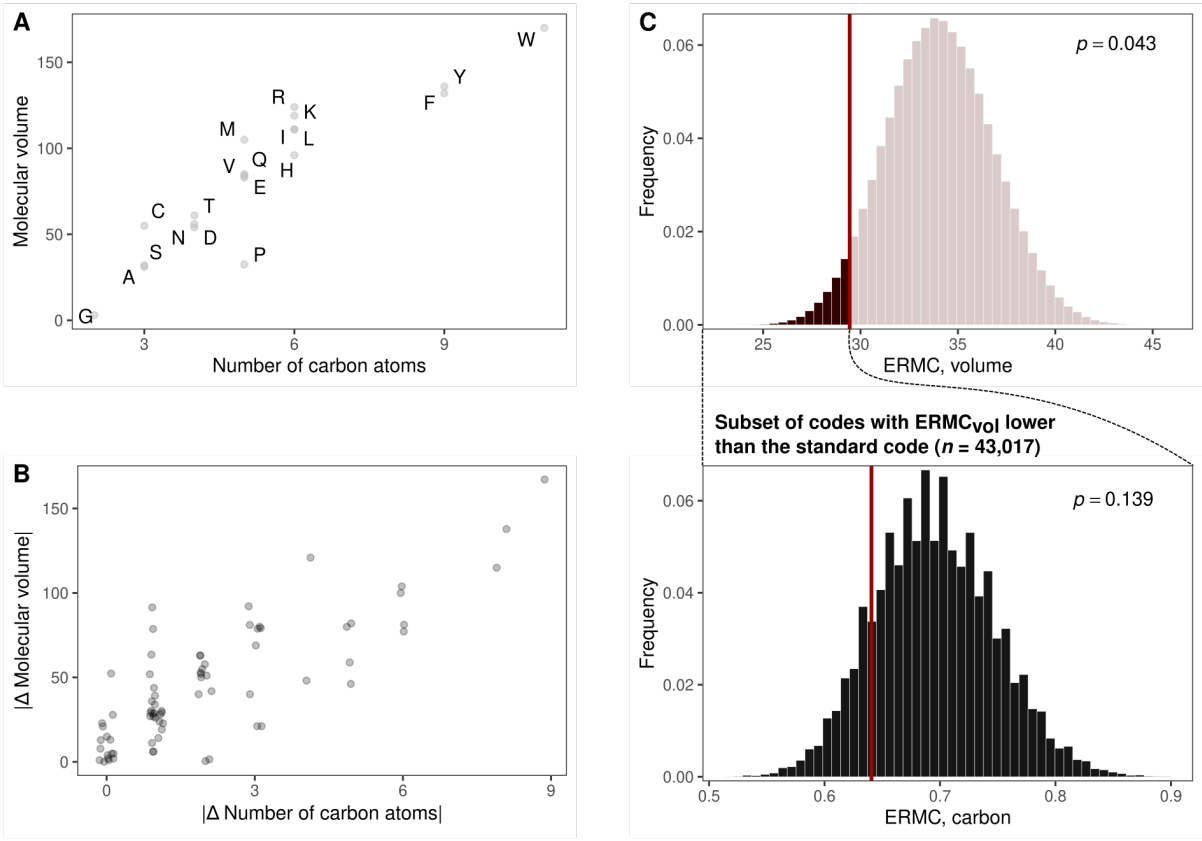

Fig. 2. Carbon conservation is confounded by amino acid volume.

2A - Scatter plot of the number of carbon atoms and the molecular volume of the twenty proteinogenic amino acids.

2B - Scatter plot of the absolute change in the number of carbon atoms and the absolute change in molecular volume for the 75 amino acid pairs that are connected by a missense mutation in the standard code. Jitter applied in the x-axis for visualization.

2C - Histograms of ERMC for amino acid volume in 1 million alternative codes generated by amino acid permutation (top) and for carbon conservation in the subset of 43,017 codes that perform better than the standard code in molecular volume conservation (bottom).