Causes and Effects of N-Terminal Codon Bias in Bacterial Genes

Daniel B. Goodman *et al.*

Science **342**, 475 (2013);
DOI: 10.1126/science.1241934

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

some secondary path, such as oxidative decarboxylation of DHG (*30*).

Our proposed mechanism begins with SAM bound at the N-terminal cluster and tyrosine bound at the C-terminal cluster (Fig. 3, step I), with reductive cleavage of SAM generating 5′-dA• (Fig. 3, step II). This radical is quenched by the abstraction of a solvent-exchangeable H atom (see mass spectrometric data, fig. S5), consistent with the proposal (*20*) that 5′-dA• abstracts the phenolic H of free tyrosine. The resulting Tyr• radical is ligated to the C-terminal 4Fe-4S cluster (Fig. 3, step III) and is not currently observed in the RFQ EPR experiments. The specific coordination geometry of the Tyr• bound to the C-terminal 4Fe-4S cluster may play a direct role in directing the subsequent C_α-C_β bond cleavage along the heterolytic pathway, concomitantly forming the observed 4OB• and DHG ligated to the 4Fe-4S cluster (Fig. 3, step IV). Free DHG is unstable and rapidly hydrolyzes to produce glyoxylate and ammonia, observed as by-products in the ThiH reaction in the absence of the other required thiazole precusors (*19*). In the case of HydG, scission of the 4Fe-4S–bound DHG occurs to yield Fe-bound CO and CN^-, concomitant with electron and proton transfer to 4OB• to form the *p*-cresol product of the HydG reaction (Fig. 3, step V). Given the facile interconversion between 4Fe-4S and 3Fe-4S forms observed in proteins such as aconitase (*22*, *23*, *31*) and ferredoxins (*32*), this unique CO- and CN^--loaded Fe may then be inserted into the assembly of the 2Fe subunit of the H cluster. We are currently using a variety of spectroscopy techniques to reveal further details of this fascinating metallocofactor assembly process.

References and Notes

1. P. M. Vignais, B. Billoud, *Chem. Rev.* **107**, 4206–4272 (2007).
2. K. A. Vincent, A. Parkin, F. A. Armstrong, *Chem. Rev.* **107**, 4366–4413 (2007).
3. J. W. Peters, W. N. Lanzilotta, B. J. Lemon, L. C. Seefeldt, *Science* **282**, 1853–1858 (1998).
4. A. S. Pandey, T. V. Harris, L. J. Giles, J. W. Peters, R. K. Szilagyi, *J. Am. Chem. Soc.* **130**, 4533–4540 (2008).
5. G. Berggren *et al.*, *Nature* **499**, 66–69 (2013).
6. J. Esselborn *et al.*, *Nat. Chem. Biol.* **9**, 607–609 (2013).
7. D. W. Mulder *et al.*, *Structure* **19**, 1038–1052 (2011).
8. P. A. Frey, A. D. Hegeman, F. J. Ruzicka, *Crit. Rev. Biochem. Mol. Biol.* **43**, 63–88 (2008).
9. J. L. Vey, C. L. Drennan, *Chem. Rev.* **111**, 2487–2506 (2011).
10. http://sfld.rbvi.ucsf.edu/django/superfamily/29/.
11. E. Pilet *et al.*, *FEBS Lett.* **583**, 506–511 (2009).
12. R. C. Driesener *et al.*, *Angew. Chem. Int. Ed.* **49**, 1687–1690 (2010).
13. E. M. Shepard *et al.*, *J. Am. Chem. Soc.* **132**, 9247–9249 (2010).
14. J. M. Kuchenreuther, S. J. George, C. S. Grady-Smith, S. P. Cramer, J. R. Swartz, *PLOS ONE* **6**, e20346 (2011).
15. J. K. Rubach, X. Brazzolotto, J. Gaillard, M. Fontecave, *FEBS Lett.* **579**, 5055–5060 (2005).
16. C. Tron *et al.*, *Eur. J. Inorg. Chem.* **2011**, 1121–1127 (2011).
17. Y. Nicolet, J. C. Fontecilla-Camps, *J. Biol. Chem.* **287**, 13532–13540 (2012).
18. M. Kriek *et al.*, *J. Biol. Chem.* **282**, 17413–17423 (2007).
19. M. Kriek, F. Martins, M. R. Challand, A. Croft, P. L. Roach, *Angew. Chem. Int. Ed.* **46**, 9223–9226 (2007).
20. Y. Nicolet, L. Martin, C. Tron, J. C. Fontecilla-Camps, *FEBS Lett.* **584**, 4197–4202 (2010).
21. J. M. Kuchenreuther, R. D. Britt, J. R. Swartz, *PLOS ONE* **7**, e45850 (2012).
22. H. Beinert, M. C. Kennedy, C. D. Stout, *Chem. Rev.* **96**, 2335–2374 (1996).
23. M. C. Kennedy *et al.*, *J. Biol. Chem.* **259**, 14463–14471 (1984).
24. A. J. M. Richards, A. J. Thomson, R. H. Holm, K. S. Hagen, *Spectrochim. Acta Part A Molec. Biomolec. Spectrosc.* **46**, 987–993 (1990).
25. C. J. Fugate *et al.*, *J. Am. Chem. Soc.* **134**, 9042–9045 (2012).
26. B. A. Barry, G. T. Babcock, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7099–7103 (1987).
27. M. L. Gilchrist Jr., J. A. Ball, D. W. Randall, R. D. Britt, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9545–9549 (1995).
28. R. J. Hulsebosch *et al.*, *J. Am. Chem. Soc.* **119**, 8685–8694 (1997).
29. F. Dole, B. A. Diner, C. W. Hoganson, G. T. Babcock, R. D. Britt, *J. Am. Chem. Soc.* **119**, 11540–11541 (1997).
30. R. Michaels, L. V. Hankes, W. A. Corpe, *Arch. Biochem. Biophys.* **111**, 121–125 (1965).
31. T. A. Kent *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1096–1100 (1982).
32. S. J. George, F. A. Armstrong, E. C. Hatchikian, A. J. Thomson, *Biochem. J.* **264**, 275–284 (1989).

Acknowledgments: This work was funded by the Division of Chemical Sciences, Geosciences, and Biosciences (R.D.B. award no. DE-FG02-11ER16282) and the Division of Material Sciences and Engineering (J.R.S. award no. DE-FG02-09ER46632) of the Office of Basic Energy Sciences of the U.S. Department of Energy. We thank D. Suess for useful discussions on the proposed HydG mechanism.

Supplementary Materials

www.sciencemag.org/content/342/6157/472/suppl/DC1
Materials and Methods
Figs. S1 to S5
Table S1
References (*33*–*48*)

13 June 2013; accepted 20 September 2013
10.1126/science.1241859

Causes and Effects of N-Terminal Codon Bias in Bacterial Genes

Daniel B. Goodman,[1,2,3] George M. Church,[1,2]* Sriram Kosuri[1]*

Most amino acids are encoded by multiple codons, and codon choice has strong effects on protein expression. Rare codons are enriched at the N terminus of genes in most organisms, although the causes and effects of this bias are unclear. Here, we measure expression from >14,000 synthetic reporters in *Escherichia coli* and show that using N-terminal rare codons instead of common ones increases expression by ~14-fold (median 4-fold). We quantify how individual N-terminal codons affect expression and show that these effects shape the sequence of natural genes. Finally, we demonstrate that reduced RNA structure and not codon rarity itself is responsible for expression increases. Our observations resolve controversies over the roles of N-terminal codon bias and suggest a straightforward method for optimizing heterologous gene expression in bacteria.

Codon usage is biased in natural genes and can strongly affect heterologous expression (*1*). Many organisms are enriched for poorly adapted codons at the N terminus of genes (*2*–*5*). Several studies suggest that these codons slow ribosomal elongation during initiation and lead to increased translational efficiency (*2*, *4*, *6*). Most organisms also display reduced mRNA secondary structure at the N terminus (*7*), and studies using synthetic codon gene variants have resulted in conflicting theories on which mechanisms are causal for expression changes (*7*, *8*). Information about the causes and effects of codon bias has been restricted to relations inferred from natural sequences using genome-wide correlation (*2*, *3*, *5*, *9*, *10*), conservation among species (*4*), or relatively small libraries of synthetic genes with synonymous codon changes (*3*, *8*, *11*–*15*). Here, we separate and quantify the factors controlling expression at the N terminus of genes in *Escherichia coli* by building and measuring expression from a large synthetic library of defined sequences.

We used array-based oligonucleotide libraries (*16*) to generate 14,234 combinations of promoters, ribosome binding sites (RBSs), and 11 N-terminal codons in front of super-folder green fluorescent protein (sfGFP) on a plasmid that constitutively coexpresses mCherry (fig. S1) (*17*–*19*). The sequences for the N-terminal peptides correspond to the first 11 amino acids (including the initiating methionine) of 137 endogenous *E. coli* essential genes (*20*) that utilize the entire codon repertoire (fig. S2). We expressed these sfGFP fusions from two promoters and three RBSs of varying strengths (*19*). We also included the natural RBS for each endogenous gene. For each combination of promoter, RBS, and peptide sequence, we designed a set of 13 codon variants to represent a wide range of codon usages and secondary structure free energies across the translation initiation region. We studied the interactions between the 5′ untranslated region (UTR) and N-terminal codon usage because initiation is thought to be the rate-limiting step for translation (*1*), this region has been previously implicated in determining most expression variation (*8*), N-terminal codons are more highly conserved (*21*), and rare codons are enriched at the N terminus of natural genes and especially those that are highly expressed (*2*).

[1]Wyss Institute for Biologically Inspired Engineering, 3 Blackfan Circle, Boston, MA 02115, USA. [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. [3]Harvard-MIT Health Sciences and Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

*Corresponding author. E-mail: sri.kosuri@wyss.harvard.edu (S.K.); gchurch@genetics.med.harvard.edu (G.M.C.)

We measured DNA, RNA, and protein levels from the entire library using a multiplex assay (Fig. 1C and figs. S3 and S4) (*19*). DNA and RNA levels were determined using DNA sequencing (DNASeq) and RNASeq. Protein levels were determined by FlowSeq; 7327 (51.5%) constructs were within the quantitative range of our assay [coefficient of determination (R^2) = 0.955, $P < 2 \times 10^{-16}$] (fig. S5). We normalized the expression measurements across each 13-member codon variant set as fold change from log-average to control for changes in promoters, RBSs, and peptide sequence (fig. S6).

Changing synonymous codon usage in the 11–amino acid N-terminal peptide resulted in a mean 60-fold increase in protein abundance from the weakest to strongest codon variant even though >96% of the gene remained unchanged. For over 160 codon variant sets (25% of sets within range), the difference was >100-fold. For each codon variant set, we included sequences encoding the most common or rare synonymous codon in *E. coli* for every amino acid. The rare codon constructs displayed a mean 14-fold (median 4-fold)

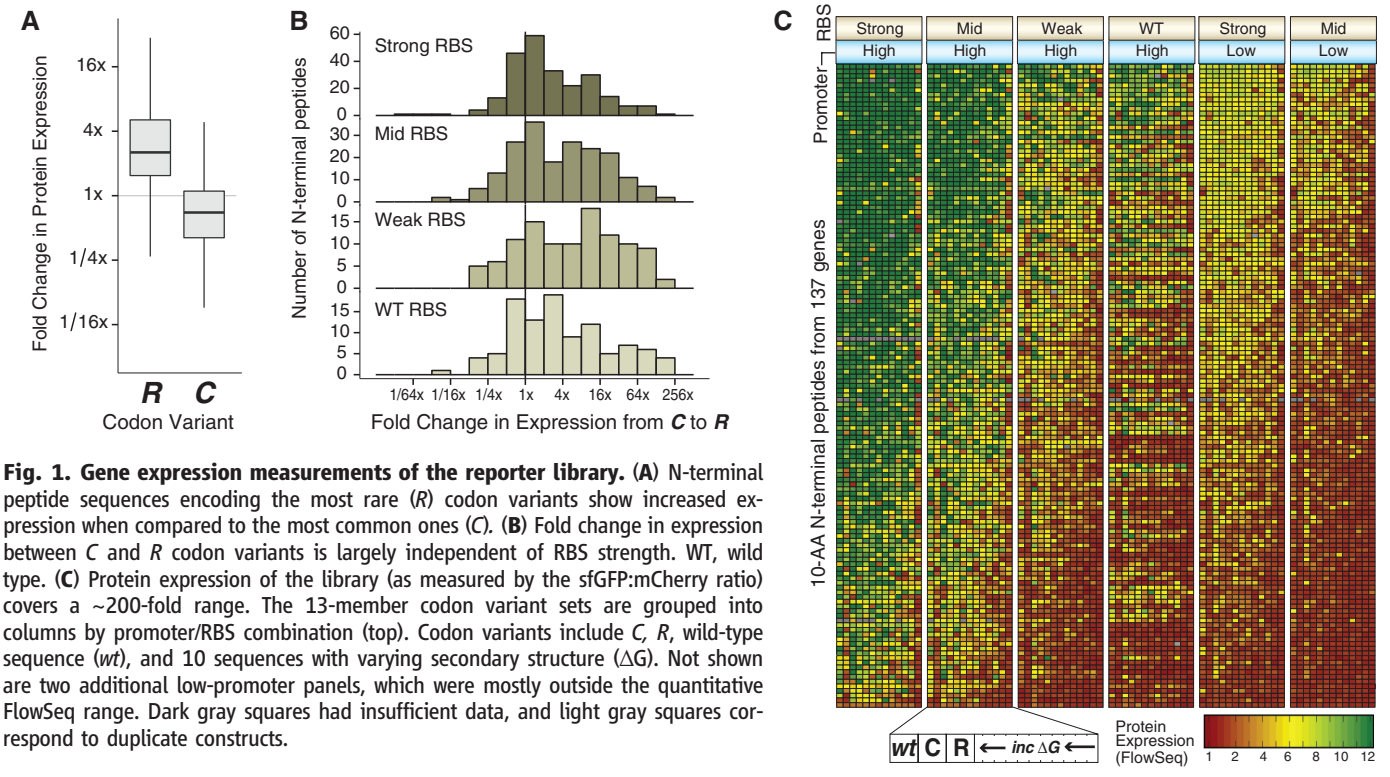

Fig. 1. Gene expression measurements of the reporter library. (**A**) N-terminal peptide sequences encoding the most rare (*R*) codon variants show increased expression when compared to the most common ones (*C*). (**B**) Fold change in expression between *C* and *R* codon variants is largely independent of RBS strength. WT, wild type. (**C**) Protein expression of the library (as measured by the sfGFP:mCherry ratio) covers a ~200-fold range. The 13-member codon variant sets are grouped into columns by promoter/RBS combination (top). Codon variants include *C*, *R*, wild-type sequence (*wt*), and 10 sequences with varying secondary structure (ΔG). Not shown are two additional low-promoter panels, which were mostly outside the quantitative FlowSeq range. Dark gray squares had insufficient data, and light gray squares correspond to duplicate constructs.

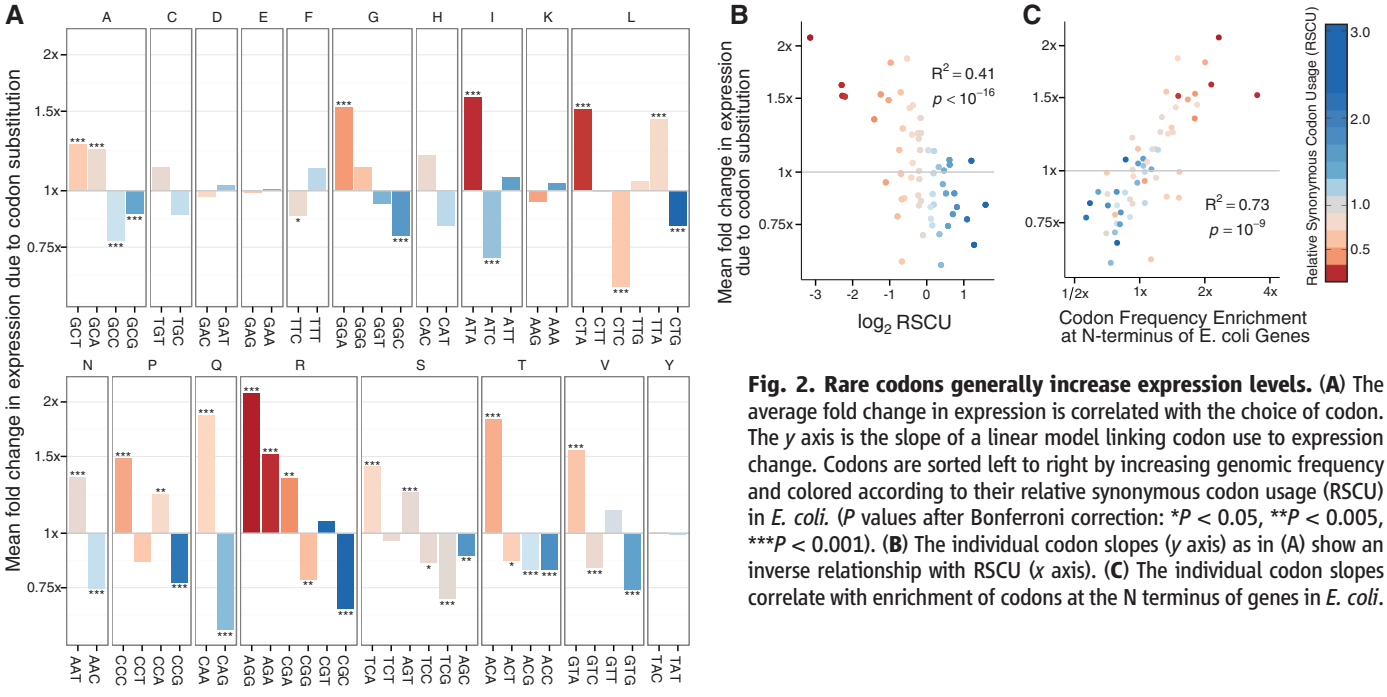

Fig. 2. Rare codons generally increase expression levels. (**A**) The average fold change in expression is correlated with the choice of codon. The *y* axis is the slope of a linear model linking codon use to expression change. Codons are sorted left to right by increasing genomic frequency and colored according to their relative synonymous codon usage (RSCU) in *E. coli*. (*P* values after Bonferroni correction: **P* < 0.05, ***P* < 0.005, ****P* < 0.001). (**B**) The individual codon slopes (*y* axis) as in (A) show an inverse relationship with RSCU (*x* axis). (**C**) The individual codon slopes correlate with enrichment of codons at the N terminus of genes in *E. coli*.

increase in protein abundance compared with common codon constructs (Fig. 1A) ($P < 2 \times 10^{-16}$, two-tailed t test) even though common codons are generally thought to increase protein expression and fitness (*1*, *9*, *22*, *23*).

To understand why rare codons cause increased expression, we first examined several codon usage metrics, but they could only explain <5% of expression differences (fig. S7A). New metrics that take into account both transfer RNA (tRNA) availability and usage [normalized translational efficiency (nTE)] show stronger N-terminal enrichment (*4*). We calculated nTE scores for *E. coli* and found that nTE scores that were similar to the tRNA adaptation index (tAI) ($R^2 = 0.847$, $P < 2 \times 10^{-16}$) did not correlate well with N-terminal codon enrichment in the *E. coli* genome ($R^2 = 0.107$, $P = 0.00654$), and did not significantly correlate with codons that increased protein expression in our data set ($R^2 = 0.024$, $P = 0.124$). Others have proposed that slow ribosome progression at the N terminus due to rare codons increases translational efficiency (*2*, *13*, *14*). This "codon ramp" hypothesis should apply primarily in the context of strong translation, but we found that using rare codons at the N terminus increases expression regardless of translation strength (Fig. 1B). Finally, ribosome occupancy profiling in *E. coli* has shown that tRNA abundance does not correlate to translation rate but that specific rare codons can create internal Shine-Dalgarno–like motifs that can alter translational efficiency (*6*). We looked for an association between the presence of internal Shine-Dalgarno–like motifs and changes in expression, and found it to be weak but statistically significant ($R^2 = 0.002$, $P < 1.3 \times 10^{-5}$).

We built a simple linear regression model correlating the use of each individual synonymous codon with expression changes (Fig. 2A and fig. S8). For most amino acids, we found a link between the rarity of the codon and increased expression (Fig. 2B). There is a strong correlation between codons that affected expression and their relative N-terminal enrichment in *E. coli* ($R^2 = 0.73$, $P < 2.3 \times 10^{-9}$) (Fig. 2C). Using relative translation efficiency instead of relative expression produced similar results (fig. S9).

Decreased GC-content correlated with increased protein expression ($R^2 = 0.12$, $P < 2 \times 10^{-16}$) (Fig. 3A). Rare codons in *E. coli* are frequently A/T-rich at the third position, and codons ending in A/T more frequently correlate with increased expression than synonymous codons ending in G/C. (fig. S10). This association suggested a link to mRNA transcript secondary structure (*8*), and so we computationally predicted RNA structure over the first 120 bases of each transcript using NUPACK (*24*). We found that increased secondary structure was correlated with decreased expression, which explained more variation than any other variable we measured ($R^2 = 0.34$, $P < 2 \times 10^{-16}$) (Fig. 3B). We made a similar linear regression model relating individual codon substitution

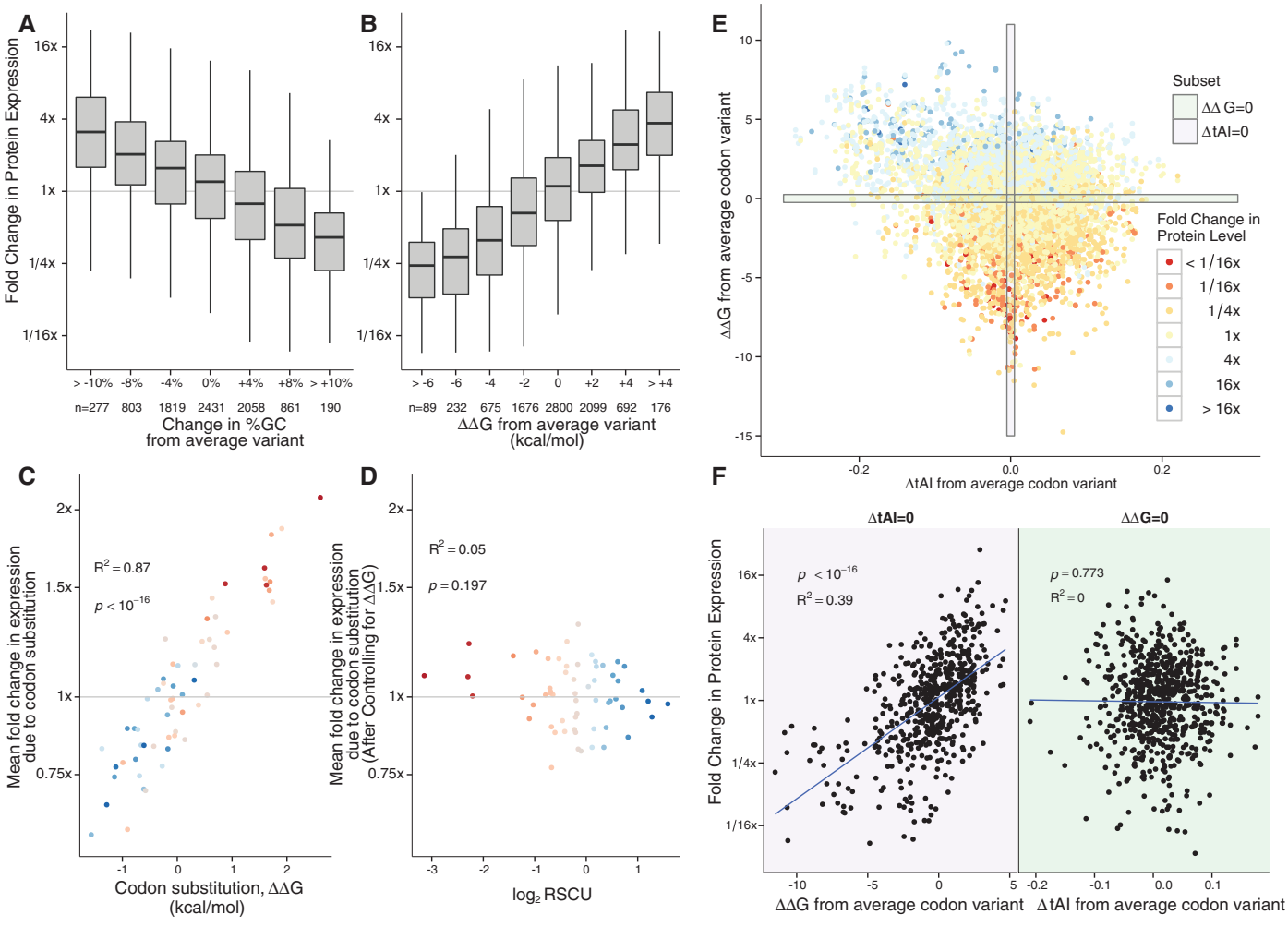

Fig. 3. Rare codons alter expression by reducing mRNA secondary structure. (**A**) Expression changes are correlated with relative changes in %GC content. Each boxplot includes ±2% of centered value. (**B**) Expression increases correlate to relative increases in free energy of folding at the front of the transcript (ΔΔG). Each boxplot includes ±2 kcal/mol of centered value. (**C**) Individual codon slopes (same as Fig. 2A y axis) correlate with the ΔΔG per individual codon substitution. (**D**) After controlling for ΔΔG with a multiple linear regression, there is no longer any relation between individual codon slopes and RSCU (compare with Fig. 2B). (**E**) The ΔΔG versus change in tAI is plotted for all constructs within the quantitative range. Constructs are colored by their relative fold change in expression from the average codon variant within the set. (**F**) Subsets of constructs corresponding to the shaded boxes in (E). (Left) Points with constant codon adaptation and varied secondary structure, (right) points with constant secondary structure and varied codon adaptation.

to change in secondary structure free energy, rather than expression levels, and found a strong correlation between codons that decreased secondary structure and those that increased protein expression ($R^2 = 0.87$, $P < 2 \times 10^{-16}$) (Fig. 3C). In addition, codon adaptation metrics at the N terminus correlate as well to change in secondary structure free energy as they do to change in protein expression (fig. S7B).

We used multiple regression to control for the secondary structure changes between codon variants and found that no relation remained between N-terminal codon adaptation and increased expression ($R^2 = 0.05$, $P = 0.197$) (Fig. 3D). Additionally, constructs with constant tAI still show a correlation between expression and secondary structure, but constructs with constant secondary structure have no correlation between tAI and expression. (Fig. 3, E and F). Finally, if secondary structure is the dominant factor, we would expect a disproportionate enrichment of A over T due to G-U wobble pairing. Indeed, nucleotide triplets with A at the wobble position were more consistently correlated with expression in our data set and with enrichment at the N terminus of *E. coli* genes (fig. S11).

Kudla *et al*. show that local RNA structure in the region between –4 and +38 of translation start is most correlated with expression change (*8*). Our data indicate that the region centered on +10 is most correlated with expression changes (Fig. 4 and figs. S12 to S14), which closely matches in vitro translation studies (*25*). This region remained the most correlated for the subset of constructs with no change in total free energy of folding across the N-terminal region (figs. S15 and S16). Although secondary structure is known to affect the RBS (*26*), when only codon usage is altered, RNA structure after the start codon, and not at the RBS, is the major contributor to expression differences. A multiple linear regression model that combines promoter and RBS choice, as well as N-terminal secondary structure and GC content, still explains only 54% of variation in expression levels. Amino acid composition effects on sfGFP folding and inadequacies in computational RNA structure prediction could be partially responsible. However, there are likely additional effects left to uncover, and the extent to which codon usage beyond the N-terminal region alters gene expression remains unresolved (*8*, *14*).

The N terminus of genes in almost all bacteria displays reduced secondary structure, but enrichment of poorly adapted N-terminal codons is only found in bacteria with GC content of at least 50% (*3*). Recent work further shows that AT-rich codons as opposed to rare codons themselves are preferentially selected, which implicates secondary structure as the driving force for N-terminal codon selection in most bacteria (*5*). Despite mechanistic differences in translation between prokaryotes and eukaryotes, both single- and multicell eukaryotes also have reduced N-terminal secondary structure (*7*). For synthetic GFP templates in yeast, secondary structure is more correlated with expression changes than codon adaptation metrics (*10*). Here, we do not examine other factors that might shape natural sequence, such as codon pair bias (*1*, *27*), cotranslational folding (*4*, *12*, *28*), or growth conditions (*11*, *15*). Natural genomic sequence is often not suited to distinguish between conflicting hypotheses of how sequence affects function; multiplexed assays of large synthetic DNA libraries provide a powerful method to examine such hypotheses in a controlled manner.

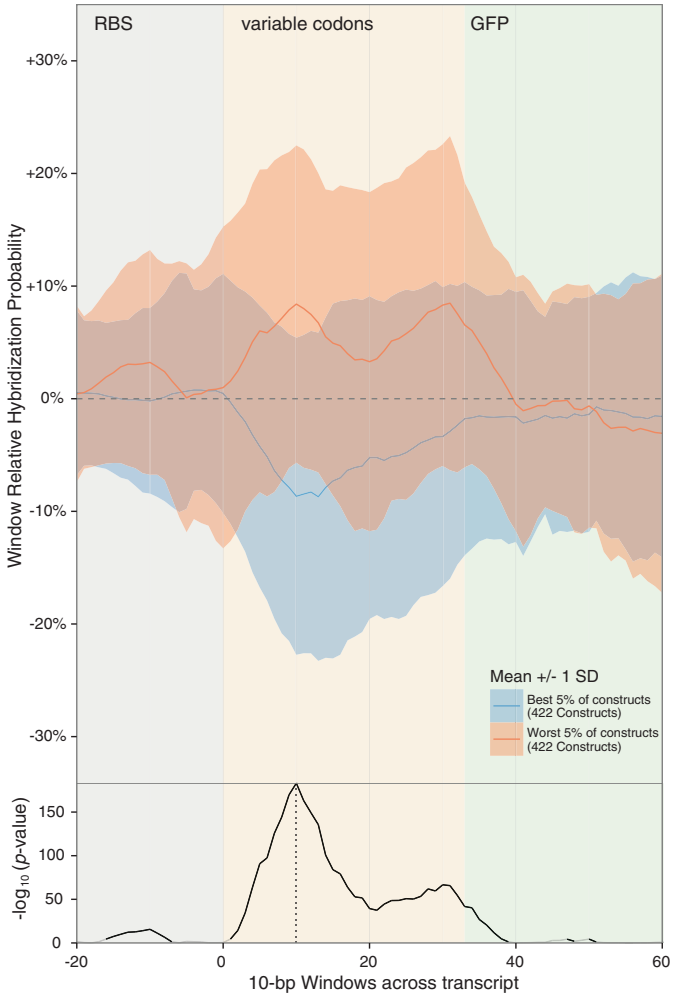

Fig. 4. mRNA structure downstream of start codon is most correlated with reduced expression. Relative hybridization probabilities averaged in 10-nuclotide windows are plotted against their correlation with expression change as a function of position (–20 to +60 from ATG). (**Top**) The best and worst 5% of constructs—as ranked by relative expression within a codon variant set—are grouped and plotted as blue and red ribbons, respectively. The ribbon tops and bottoms are one standard deviation from the mean, which is shown as a solid line. (**Bottom**) The *P* value for linear regressions, correlating hybridization probabilities within each window to expression fold change in all constructs.

References and Notes

1. J. B. Plotkin, G. Kudla, *Nat. Rev. Genet.* **12**, 32–42 (2011).
2. T. Tuller *et al*., *Cell* **141**, 344–354 (2010).
3. M. Allert, J. C. Cox, H. W. Hellinga, *J. Mol. Biol.* **402**, 905–918 (2010).
4. S. Pechmann, J. Frydman, *Nat. Struct. Mol. Biol.* **20**, 237–243 (2013).
5. K. Bentele, P. Saffert, R. Rauscher, Z. Ignatova, N. Blüthgen, *Mol. Syst. Biol.* **9**, 675 (2013).
6. G.-W. Li, E. Oh, J. S. Weissman, *Nature* **484**, 538–541 (2012).
7. W. Gu, T. Zhou, C. O. Wilke, *PLOS Comput. Biol.* **6**, e1000664 (2010).
8. G. Kudla, A. W. Murray, D. Tollervey, J. B. Plotkin, *Science* **324**, 255–258 (2009).
9. M. dos Reis, R. Savva, L. Wernisch, *Nucleic Acids Res.* **32**, 5036–5044 (2004).
10. P. Shah, Y. Ding, M. Niemczyk, G. Kudla, J. B. Plotkin, *Cell* **153**, 1589–1601 (2013).
11. M. Welch *et al*., *PLOS ONE* **4**, e7002 (2009).
12. M. Zhou *et al*., *Nature* **495**, 111–115 (2013).
13. S. Navon, Y. Pilpel, *Genome Biol.* **12**, R12 (2011).
14. T. Tuller, Y. Y. Waldman, M. Kupiec, E. Ruppin, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3645–3650 (2010).
15. A. R. Subramaniam, T. Pan, P. Cluzel, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2419–2424 (2013).
16. E. M. LeProust *et al*., *Nucleic Acids Res.* **38**, 2522–2540 (2010).
17. J.-D.Pédelacq , S. E. P. Cabantous, T. Tran, T. C. Terwilliger, G. S. Waldo, *Nat. Biotechnol.* **24**, 79–88 (2006).
18. N. C. Shaner *et al*., *Nat. Biotechnol.* **22**, 1567–1572 (2004).
19. S. Kosuri *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14024–14029 (2013).
20. Y. Yamazaki, H. Niki, J.-I. Kato, *Methods Mol. Biol.* **416**, 385–389 (2008).
21. D. L. Hartl, E. N. Moriyama, S. A. Sawyer, *Genetics* **138**, 227–234 (1994).
22. M. Gouy, C. Gautier, *Nucleic Acids Res.* **10**, 7055–7074 (1982).
23. P. M. Sharp, W. H. Li, *Nucleic Acids Res.* **15**, 1281–1295 (1987).
24. J. N. Zadeh *et al*., *J. Comput. Chem.* **32**, 170–173 (2011).
25. D. Voges, M. Watzele, C. Nemetz, S. Wizemann, B. Buchberger, *Biochem. Biophys. Res. Commun.* **318**, 601–614 (2004).
26. M. H. de Smit, J. van Duin, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 7668–7672 (1990).
27. J. R. Coleman *et al*., *Science* **320**, 1784–1787 (2008).
28. A. A. Komar, *Trends Biochem. Sci.* **34**, 16–24 (2009).

Acknowledgments: We thank J. C. Way, E. R. Daugharthy, and R. T. Sauer for comments. The research was supported by the U.S. Department of Energy (DE-FG02-02ER63445 to G.M.C.), NSF SynBERC (SA5283-11210 to G.M.C.), Office of Naval Research (N000141010144 to G.M.C. and S.K.), Agilent

Technologies, Wyss Institute, and an NSF Graduate Research Fellowship to D.B.G. Data can be accessed on the National Center for Biotechnology Information, NIH, Sequence Read Archive (SRA) (SRP029609). pGERC reporter can be obtained from AddGene (#47441). Accession numbers: The Project accession at the SRA is SRP029609. The sample accession is SRS477429. There are three experiments, one for DNA, one for RNA, one for FlowSeq: RNA, SRX346948; DNA, SRX346944; and FlowSeq, SRX346268.

Supplementary Materials

www.sciencemag.org/content/342/6157/475/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S16
Table S1
References (*29*–*35*)

14 June 2013; accepted 13 September 2013
Published online 26 September 2013;
10.1126/science.1241934

2000 Years of Parallel Societies in Stone Age Central Europe

Ruth Bollongino,[1]* Olaf Nehlich,[2,3] Michael P. Richards,[2,3,4] Jörg Orschiedt,[5] Mark G. Thomas,[6] Christian Sell,[1] Zuzana Fajkošová,[1] Adam Powell,[1] Joachim Burger[1]

Debate on the ancestry of Europeans centers on the interplay between Mesolithic foragers and Neolithic farmers. Foragers are generally believed to have disappeared shortly after the arrival of agriculture. To investigate the relation between foragers and farmers, we examined Mesolithic and Neolithic samples from the Blätterhöhle site. Mesolithic mitochondrial DNA sequences were typical of European foragers, whereas the Neolithic sample included additional lineages that are associated with early farmers. However, isotope analyses separate the Neolithic sample into two groups: one with an agriculturalist diet and one with a forager and freshwater fish diet, the latter carrying mitochondrial DNA sequences typical of Mesolithic hunter-gatherers. This indicates that the descendants of Mesolithic people maintained a foraging lifestyle in Central Europe for more than 2000 years after the arrival of farming societies.

The Mesolithic-Neolithic transition marks a shift from a foraging to an agricultural way of life. It first appeared around 8500 BC in present-day southeastern Anatolia and Syria. About 3000 years later, this subsistence strategy reached Central Europe through the expansion of the Neolithic Linear Pottery culture (LBK). Whether the first European farmers descended from hunter-gatherers or migrated in from the Near East has been debated extensively in the archaeological literature. Over the last decade, a number of palaeogenetic studies have contributed substantially to current understanding of the Mesolithic-Neolithic transition in Europe [(*1*) and references therein]. Taken together, these findings strongly support a demic diffusion of early farmers into Central Europe, most likely originating in the southeast of the continent (*2*). Little is known about how long hunter-gatherers persisted in Central Europe, as there are no unambiguous signs of their presence in the archaeological record after the Early Neolithic. In this study, we present both ancient DNA and isotopic data, which, when combined, provide persuasive evidence for the prolonged coexistence of genetically distinct hunter-gatherer and farming groups over the course of the Neolithic in Central Europe.

Ancient DNA and sulfur, nitrogen, and carbon isotope ratios were analyzed from bones and teeth of 29 individuals from a burial cave site that contained around 450 remains from both Mesolithic hunter-gatherers and Neolithic individuals. The Blätterhöhle site is situated in Hagen, Germany (*3*) (Fig. 1), and because of its long and narrow geological structure, it is very likely that the human remains were deposited deliberately. Because the layers inside the cave have been disturbed by bioturbation, all samples used in this study were ^{14}C dated by accelerator mass spectrometry. The ^{14}C dates reveal two occupation phases ranging from 9210 to 8340 calibrated BCE (cal BC) (Mesolithic) and from 3986 to 2918 cal BC (Neolithic), respectively (*4*) (Table 1, table S1, and fig. S1).

We applied both a polymerase chain reaction, with subsequent Sanger sequencing, and a capture next-generation sequencing approach to establish partial or complete mitochondrial genomes. Out of 29 samples, 25 yielded reproducible mitochondrial hypervariable region I (HVRI) sequences (*4*) (Table 1 and tables S4 and S6). Complete mitochondrial genomes with coverage from 3.6× up to 39.8× were obtained for one Mesolithic and five Neolithic samples (*4*) (tables S4 and S6).

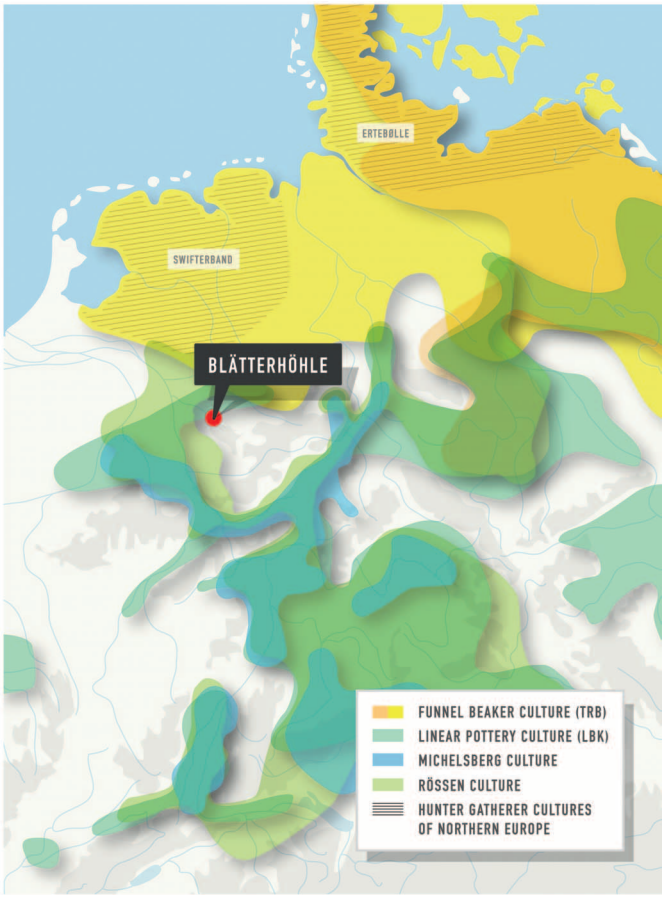

Fig. 1. Geographic location of the Blätterhöhle cave site with schematic representation of the distribution of relevant archaeological cultures in Central and Northern Europe (*27*).

[1]Palaeogenetics Group, Institute of Anthropology, Johannes Gutenberg University, 55099 Mainz, Germany. [2]Department of Anthropology, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada. [3]Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. [4]Department of Archaeology, University of Durham, Durham, DH1 3LE, UK. [5]Institute of Prehistoric Archaeology, Free University of Berlin, 14195 Berlin, Germany. [6]Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK.

*Corresponding author. E-mail: bollongi@uni-mainz.de