ANNUAL REVIEWS

Annual Review of Genetics

Origin and Evolution of the Universal Genetic Code

Eugene V. Koonin[1] and Artem S. Novozhilov[2]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA; email: koonin@ncbi.nlm.nih.gov

[2]Department of Mathematics, North Dakota State University, Fargo, North Dakota 58108, USA

Annu. Rev. Genet. 2017. 51:45–62

First published as a Review in Advance on August 30, 2017

The *Annual Review of Genetics* is online at genet.annualreviews.org

https://doi.org/10.1146/annurev-genet-120116-024713

Keywords

standard genetic code, universal genetic code, codons, anticodons, aminoacyl–tRNA synthetases, evolution of translation, error minimization, stereochemical theory, coevolution theory, frozen accident

Abstract

The standard genetic code (SGC) is virtually universal among extant life forms. Although many deviations from the universal code exist, particularly in organelles and prokaryotes with small genomes, they are limited in scope and obviously secondary. The universality of the code likely results from the combination of a frozen accident, i.e., the deleterious effect of codon reassignment in the SGC, and the inhibitory effect of changes in the code on horizontal gene transfer. The structure of the SGC is nonrandom and ensures high robustness of the code to mutational and translational errors. However, this error minimization is most likely a by-product of the primordial code expansion driven by the diversification of the repertoire of protein amino acids, rather than a direct result of selection. Phylogenetic analysis of translation system components, in particular aminoacyl–tRNA synthetases, shows that, at a stage of evolution when the translation system had already attained high fidelity, the correspondence between amino acids and cognate codons was determined by recognition of amino acids by RNA molecules, i.e., proto-tRNAs. We propose an experimentally testable scenario for the evolution of the code that combines recognition of amino acids by unique sites on proto-tRNAs (distinct from the anticodons), expansion of the code via proto-tRNA duplication, and frozen accident.

INTRODUCTION

The genetic code, which defines the rules of translation from the 4-letter nucleic acid alphabet to the 20-letter alphabet of proteins, has a special place in all of biology (16, 57, 58, 84). Indeed, the code is arguably the single central informational invariant of all life forms. Despite many variations that continue to emerge through the study of protein coding in diverse life forms, the basic structure of the code and the majority of the codon assignments are genuinely universal (44, 72).

As soon as the codon table was established in 1965, it became apparent that the code encompasses distinct patterns begging for explanations (14, 84). The 64 codons are neatly organized in sets of 4 or 2, with the third base of a codon typically being synonymous (i.e., changes in this position do not lead to amino acid replacement) (**Figure 1**). Furthermore, the assignment of codons to amino acids across the code table is clearly nonrandom: Related amino acids typically occupy contiguous areas in the table. The second position of a codon is the most important specificity determinant, and three of the four columns of the codon table encode related, chemically similar amino acids. For example, all codons with a U in the second position correspond to hydrophobic amino acids. Even a simple qualitative examination shows that the code is robust to mutational or translational errors. Substitutions and translation errors in synonymous positions (typically, the third position in a codon) have no effect on the protein (although this does not necessarily imply such substitutions are selectively neutral), whereas substitutions in the first position most often lead to incorporation of an amino acid similar to the correct one, thus decreasing the damage.

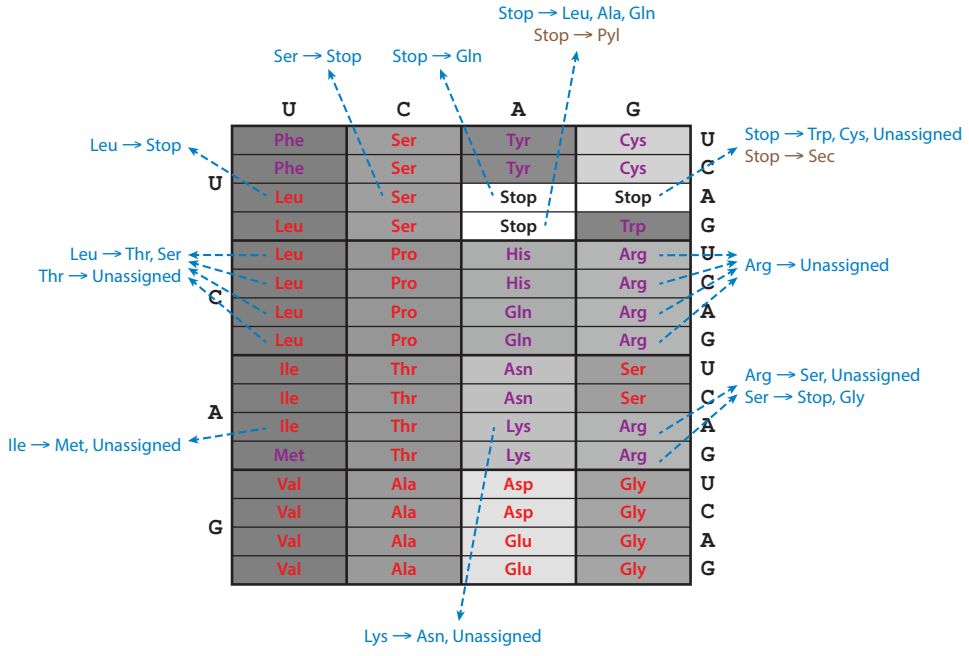

Figure 1

The standard genetic code (SGC), early and late amino acids, and code changes in extant organisms. The cells in the code table are shaded according to the Polar Requirement scale values of the respective amino acids to emphasize clustering of similar amino acids in the SGC. Shown here are the early amino acids (*red*), late amino acids (*purple*), and deviations in the code detected in extant organisms (*blue and brown*) (72). The first, second, and third codon positions correspond to the letters along the left side, the top, and the right side of the table, respectively. Abbreviations: Pyl, pyrrolysine; Sec, selenocysteine.

In a pioneering 1991 study, Haig & Hurst (29) developed a quantitative metric to measure the robustness of codes to error, often referred to as error minimization. Refining that metric further, Freeland & Hurst (25) have shown that the code is one in a million, i.e., that it is more robust than approximately a million randomly chosen codes. Subsequent efforts employing much more sophisticated models have revealed even greater robustness of the code (27). Nevertheless, the standard genetic code (SGC) is not optimal: Given the overall hyperastronomical number of possible codes ($>10^{84}$), billions of variants are more robust than the actual universal code (44).

SGC: standard genetic code

aaRS: aminoacyl–tRNA synthetase

In the translation system that is universally conserved in all extant cellular life forms (40), the accurate decoding of an mRNA sequence, i.e., the incorporation of the cognate amino acid in response to a given codon, is ensured by a complex, indirect mechanism. This involves an aminoacyl–tRNA synthetase (aaRS) recognizing a unique tRNA and charging it with the cognate amino acid. Thus, translation of the code does not involve direct recognition of the codons (or anticodons) by amino acids, which brings up a burning question: Why are the codon assignments what they are? In other words, why is it the case that, for instance, glycine is encoded by GGN codons rather than, say, CCN codons (the latter of which encode proline in the SGC)? The principal debate revolves around four competing perspectives on the code's origin and evolution (18, 39, 44). The stereochemical theory posits that the codon table actually reflects affinity between amino acids and codons (or anticodons) that is thought to have played a role in primordial translation before being replaced by the extant indirect mechanism (98). The coevolution theory postulates that the structure of the code reflects the evolution of the pathways for amino acid biosynthesis (90). The error minimization theory holds that the code has evolved to its current form from a possibly random ancestral state under the pressure of selection and primarily for maximum robustness to errors introduced during replication, transcription, or translation (7). Finally, the frozen accident view (which hardly qualifies as a theory) is that codons were initially assigned by sheer chance, and once these assignments were fixed, further major change of the code became lethal; therefore, only a few minor codon reassignments occurred in some groups of organisms (15). Clearly, these scenarios of code evolution do not have to be mutually exclusive, further complicating the problem (44).

The above is only one of the many "why" questions that are repeatedly asked about the features of the code, albeit perhaps the most fundamental one. Here are some others:

- Why is the code triplet?
- Why are there 20 amino acids encoded by the code table—no more and no less (notwithstanding a few additional amino acids encoded by certain groups of organisms)?
- Why is the code highly robust to error?
- Why, although highly robust, is the code still far from being globally optimal?
- Why is the code universal? And when the code does deviate, why does this occur in some organisms but not others?

Ultimately, all of these questions are about the origin and evolution of the code. If we succeed in developing a defendable scenario for the origin and subsequent evolution of the code, we should be able to answer most if not all of the "why" questions. The trouble is that the problem appears extremely and unusually hard. Evolution of the code is intimately linked to the origin and evolution of the translation apparatus itself, and this is one of the most fundamental and hardest problems in all of biology (84, 88). The study of the evolution of translation involves an intrinsic and formidable paradox: A high-fidelity translation system requires a number of functional proteins, but the maintenance of such proteins is impossible without a reliable translation system. Naturally, a variety of scenarios have been developed, but all appear to remain within the domain of speculation. Thus in many studies, exploration of the code, as a small but apparently sustainable research field on its own, largely abstracts itself from the problem of the origin of the translation

system—it more or less takes translation for granted. Although such analyses can be productive for understanding the salient features of the code, there remains a distinct possibility that the code evolution problem can be solved ultimately only in conjunction with the evolution of the translation mechanisms themselves.

Theoretical and experimental research into the origin and evolution of the code spans more than five decades, during which an extensive body of literature has accumulated and has been covered by review articles at different stages in the development of the field (18, 36, 44, 69, 72, 78). In 2009, we attempted to critically assess and synthesize the main lines of evidence on the nature and evolution of the code (44). This review article is a logical extension of the one from 2009, and therefore we focus mostly on the recent advances, both theoretical and experimental, and take stock of the current state of the code enigma.

THE STANDARD CODE, ITS VARIANTS, AND ITS RECENT EVOLUTION

The SGC table maps the 20 canonical amino acids and the three translation stop signals to the 64 codons. It is most convenient to examine the table by considering codon quartets in which the first two bases are identical whereas the third position is occupied by each of the four bases; as discussed below, this structure seems to reflect the course of the code evolution. Exactly half of the quartets comprise 8 codon sets encoding a single amino acid each, with a synonymous, uninformative third position; five quartets form 10 sets of 2 synonymous codons each; of the remaining three quartets, two include the 3 stop codons along with an amino acid assigned to 2 codons or 1 codon, and the third quartet contains the single codon for Met that also serves as the translation start signal, along with 3 Ile codons (**Figure 1**).

The code is nearly universal in all extant life forms but shows limited evolvability manifested in changes to the standard codon assignment in many groups of organisms, particularly in organelles with tiny genomes and in some parasitic and endosymbiotic bacteria with highly reduced genomes (47, 68, 72). More than 20 nonstandard codes have been described, and new variants continue to emerge with the progress of genomic and metagenomic sequencing. Modifications to the code belong to three major categories: (*a*) reassignment of codons within the canonical set of 21, including the stop signal; (*b*) loss (unassignment) of codons, and (*c*) incorporation of new amino acids. Stop codons are strongly overrepresented among the code modifications. Of the 23 nonstandard codes surveyed by Sengupta & Higgs (72), there are 8 cases of stop codons being reassigned or acquired, 8 cases of codon loss, and 10 reassignments of a codon from one amino acid to another (**Figure 1**). Given that there are only three stop codons, the same changes to the code occurred in parallel in different groups of organisms. By far the most common one is the reassignment of the UGA stop codon to Trp, which has been identified in many mitochondria as well as some bacteria and ciliates. Recruitment of the UAA/G stop codons to encode Gln also has been detected in diverse organisms. Strikingly, two deviant codes that apparently lack any dedicated stop codons have been recently discovered in two species of ciliates (74) and in a group of trypanosomatids (100). These variants combine the recruitment of stop codons to encode both Trp and Gln, which previously have been shown to occur separately in different deviant codes. Apparently, these groups of organisms evolved still poorly characterized mechanisms for dual, context-dependent decoding of the (former) stop codons.

There are two well-characterized noncanonical amino acids that so far have been shown to be co-opted into the code: selenocysteine, which is represented in varying sets of proteins in diverse organisms from all three domains of life (3, 55), and pyrrolysine, currently detected only in some archaea. The mechanisms that lead to the incorporation of these two amino acids are completely

different: Pyrrolysine is accommodated via the reassignment of a stop codon in an arrangement analogous to code changes within the canonical amino acid set, whereas selenocysteine is incorporated only when a stop codon directs recoding in the presence of a distinct, regulatory sequence element (49, 99, 102).

The evolutionary mechanisms that lead to codon reassignment and emergence of deviant codes are not thoroughly understood, but clearly they must involve changes in tRNA specificity or to the evolution of new specificities in the case of stop codon recruitment. Sengupta, Higgs, and coworkers (71–73) have captured these mechanisms within a gain and loss framework, where gain refers to acquisition of a new tRNA specificity, often following duplication of a tRNA gene, and loss refers to the elimination of a tRNA specificity, typically via deletion. There is no clear evidence that any modern code variants are associated with adaptation. Most likely, they emerge via neutral evolutionary processes, namely genetic drift and mutational pressure that drives small genomes toward low GC content.

Although the subject is not discussed in any detail in this review, it has to be mentioned that recently developed methods of synthetic biology have allowed substantial artificial alteration of the code in bacteria (11, 48, 56, 91). The fitness of the bacteria with altered codes has not been studied in detail, but their viability itself seems to support the view that the fitness of different codes might not differ dramatically. Therefore, the exceptional evolutionary stability of the SGC could be caused by high fitness barriers separating it from other codes, or in other words, the low fitness of intermediates (72).

ANCIENT EVOLUTION OF THE CODE: EARLY AND LATE AMINO ACIDS

A major tenet that the current primordial evolution scenarios hold in common is that the earliest proteins did not start with the modern set of 20 amino acids. Indeed, the canonical amino acids differ substantially in their chemical complexity and stability. Ten amino acids are consistently produced in prebiotic chemistry experiments and also have been identified in meteorites in the following order of relative abundance: Gly, Ala, Asp, Glu, Val, Ser, Ile, Leu, Pro, Thr (10, 13, 61, 101). Calculations have shown that the ranks of amino acids in this list strongly correlate with the free energy of their syntheses: the cheapest—thermodynamically—are on top of the list (33). A great variety of criteria have been applied in attempts to determine the timing of amino acid incorporation into the code, and consensus approaches have converged on a very similar ranking (81, 82). Notably, a completely independent approach—analysis of the fluxes of amino acids in the recent evolution of proteins in diverse organisms—has shown that the concentrations of the putative early amino acids in the above list are mostly decreasing, whereas those of the remaining, late amino acids are typically increasing (38). This convergence of widely different methods of inference suggests that the above 10 amino acids can be confidently considered old; i.e., they were represented already in the first proteins.

WHY IS THE CODE (NEARLY) UNIVERSAL AND WHEN CAN IT DEVIATE?

At first glance, the universality of the code seems to stem from the (retrospectively) obvious frozen accident argument that was articulated by Crick (15) in the seminal 1968 paper on code evolution. The frozen accident perspective holds that, once the codon table is defined, reassignment of even a single codon in a sufficiently large genome will exert a prohibitive deleterious effect. This argument does not necessarily require that the original choice of codon assignment is literally and strictly random. Any number of factors could contribute to the initial codon assignments, but once the choice

HGT: horizontal gene transfer

is made, these assignments are frozen; i.e., only rare and minor changes may be allowed. All discovered codon reassignments in extant organisms are indeed quite limited in scope (**Figure 1**). Using the language of fitness landscapes, the frozen accident perspective implies that there can be many different codes occupying fitness peaks, but these are separated by deep valleys of low fitness (60).

However, the early stages of cellular and especially precellular evolution could and actually must have been a different matter. This early evolution necessarily involved competition between ensembles of virus-like genetic elements and selection at the level of such collectives (43, 75, 76). It hardly can be imagined that translation evolved within a single such ensemble, and if it emerged on multiple occasions in different ensembles, there is every reason to postulate that there were numerous, different codes initially. Why would it be the case that a single code survived? In other words, why was there only one frozen accident (that is, if the actual codon assignments are indeed accidental)?

In a breakthrough paper, Vetsigian et al. (83) came up with a strikingly simple but powerful idea: There was one frozen accident because extensive horizontal gene transfer (HGT) was an essential part of early evolution, without which the transition to the cellular level of complexity simply could not have occurred. Obviously, even a small change in the code has a prohibitive effect on HGT. Vetsigian et al. turned to mathematical modeling to investigate the proposition that the requirement for HGT results in the survival of a single, universal code during evolution. Starting with a random ensemble of codes, the simulated evolution experiments lead to code diversification in the absence of HGT but to survival of only a few code variants when HGT is allowed. In their original analysis of the link between the code evolution and HGT, Vetsigian et al. explored a deterministic model in an infinite population approximation. A more realistic recent model that took into account stochastic effects of the finite population size produced a single, universal code, with a structure resembling the SGC, within a certain range of HGT rates (1, 70).

It is important to note that any substantial reassignment of codons has a dual effect on the respective microbial population. On the one hand, HGT is abrogated, but on the other hand, the population becomes resistant to parasitic genetic elements, such as viruses and plasmids. Hence, there is a trade-off between the deleterious and beneficial effects of HGT elimination. In the long run, the benefits appear to outrun the drawbacks.

Recent theoretical, comparative genomic and experimental research on microbial evolution clearly indicates that HGT is the principal factor in microbial adaptation and innovation (46, 65, 66, 80). Moreover, under a wide range of parameters of the evolutionary process, clonal populations are subject to the imminently fatal effect of Muller's ratchet in the absence of a sufficient HGT rate (77). It has been shown that, for most microbial groups, the critical HGT rate required to avoid Muller's ratchet is greater than the threshold of persistence of parasitic genetic elements (35). Thus in general, curtailment of HGT resulting from code modification leads to extinction of the respective group of microbes. However, it appears likely that protection against genetic parasites outweighs the benefits of HGT under certain conditions in the shorter term, and this allows limited changes in the code to be fixed in evolution. In other cases, codon reassignments might be fixed by random drift in isolated microbial populations, such as those of intracellular parasites that do not experience much HGT and either are headed toward eventual extinction or, like endosymbiotic organelles, are maintained by specific selective forces. Such organisms often have small genomes, minimizing the cost of codon reassignment. Thus, there seem to be two key factors that effectively lock the code into its universal configuration: The direct cost of codon reassignment resulting from maladapted protein formation prohibits major changes, whereas the deleterious effect of HGT curtailment severely limits even local codon reassignment.

RECENT PROGRESS ON THE THREE PRINCIPAL SCENARIOS OF CODE ORIGIN AND EVOLUTION

The structure of the SGC—that is, the mapping of 64 codons to 20 amino acids and the stop signal—is manifestly regular with respect to multiple criteria (44, 84). This nonrandomness of the code seems to require an explanation. The three most advanced and coherent concepts that strive to explain regularities in the code are the stereochemical, coevolution, and error minimization theories (18, 39, 44). We apply the term theory hereinafter for the sake of brevity and following the tradition, although none of the scenarios of code evolution seems to meet the criteria of a theory in the strict sense. Arguably, the frozen accident perspective also has to be taken into account in conjunction with any of the above concepts.

Polar Requirement (PR): a scale of amino acid physicochemical properties based on their solubility in pyridine

Stereochemical Theory

Gamow (26) proposed the stereochemical theory together with the very first formulation of the coding problem. However, the initial attempts for a direct experimental demonstration of interaction between amino acids and the cognate codons or anticodons, primarily by Woese and coworkers (85 and references therein), were generally unconvincing, resulting in a long lull in the pursuit of the stereochemical account of the code. Nevertheless, this early work was productive methodologically and resulted in the development of the Polar Requirement (PR) scale based on the solubility of amino acids in pyridine or its derivatives (pyrimidine analogs) (54). Amino acids differ widely on the PR scale, which has made it a popular metric for assessing the potential interactions between pyrimidine bases and amino acids as well as the robustness of the code (see discussion below).

Progress in aptamer technology brought about a renaissance of the stereochemical theory. A key observation was that short RNA sequences (aptamers) selected from random mixtures by amino acid binding showed, at least for some amino acids, significant enrichment by cognate triplets. These were codons in some cases and anticodons in others (92, 93, 96). Additional experimental evidence was recently presented by Yarus et al. (98), who analyzed the sequences of 337 independent binding sites for eight amino acids (Arg, Gln, His, Ile, Leu, Phe, Trp, Tyr) containing 18,551 nucleotides to test for interactions between amino acids and cognate coding triplets. Statistically significant affinity for cognate triplets has been observed for five of the eight amino acids (excluding Gln, Leu, and Tyr—the latter narrowly missing significance). These results were interpreted as evidence of the existence of an early stereochemical era during the code evolution. Moreover, it has been concluded that the majority (approximately 75%) of the modern amino acids were co-opted into the code at this stage (94).

In our previous review of the code evolution, we presented several arguments to the effect that the statistical evidence from the aptamer experiments did not provide for direct conclusions about the stereochemical origin of the code (44). Notwithstanding the additional data and counterargument (98), our reasoning still appears to hold. Without going into much detail and leaving aside potential problems with statistical significance (which is less than overwhelming for all amino acids other than Arg), it is important to note that all the significant findings on aptamer–amino acid binding (98) involve amino acids that are believed to be late additions to the genetic code (see the section titled Ancient Evolution of the Code: Early and Late Amino Acids). Of the five amino acids that showed significant cognate aptamer binding, only one, Ile, belongs to the early group. From the chemical standpoint, this is not at all surprising because the late-addition amino acids have large side chains that can interact with oligonucleotides. All the evidence from aptamer experiments that appears compatible with the existence of physicochemical affinity between amino acids and cognate nucleotide triplets involves complex amino acids. This trend does not bode well for the idea that such affinity stands behind the origin of the code.

A complementary take on the stereochemical scenario would involve evidence from the biology of extant life forms. Such evidence, however, is scarce and circumstantial. It has been shown that codons for Arg, which has the strongest statistical support from the aptamer experiments, also confer binding specificity for Arg in the self-splicing group I introns (97). A much larger, if less direct, body of evidence has been derived from the analysis of the spatial adjacency of amino acid residues in ribosomal proteins to rRNA nucleotide triplets in the structures of bacterial and archaeal ribosomes (37). Enrichment for anticodons and codons was observed with moderate statistical significance for 11 and 8 amino acids, respectively. Simulations with randomized codes have shown that the association with cognate triplets was considerably more significant for the 11 anticodon-associated amino acids than for the 8 codon-associated ones. Given that the ribosome is the most ancient, universal molecular structure in all cells, Johnson & Wang (37) concluded that amino acid–anticodon interaction was central to the origin of the code and remains at the heart of the modern translation machinery.

More recently, Zagrovic and coworkers (31, 34, 62–64) have published a series of extensive analyses on the apparent complementarity between the amino acid sequences of proteins and the nucleotide sequences of cognate mRNAs. These whole proteome analyses identified strong negative correlations between the mean PR scale values of proteins and the content of pyrimidines in the respective mRNAs, indicating that proteins with high affinity to pyrimidine analogs are encoded by pyrimidine-rich mRNAs. Similar conclusions were reached from an independent, more general approach whereby data on nucleotide–amino acid interactions in diverse ribonucleoprotein and deoxyribonucleoprotein complexes were compared with the content of the respective bases in the mRNAs. Extension of this approach to codons resulted in a more complicated picture. For the early amino acids, strong correlations were observed between the G and C contents of the codons and the binding preferences, whereas for the late amino acids, these correlations were not significant. Conversely, the A content of the codons for late amino acids showed a negative correlation with the A-binding propensity. Zagrovic and coworkers (62) interpret these findings as support for the old idea of Woese on the importance of direct interactions between amino acids and the cognate triplets. However, it remains questionable whether such weak specificity of amino acid interaction with RNA—observed for peptides rather than free amino acids—could play a central role in the evolution of the code.

Coevolution (Metabolic) Theory

The coevolution theory, first proposed by Wong (89) in 1975, holds that the genetic code is shaped by the precursor–product relationships between amino acids (see also 19, 20, 90). Thus, under this scenario, the code evolved from an ancestral version that included only simple amino acids produced abiogenically and then expanded to incorporate the more complex amino acids in parallel with the evolution of their respective biosynthetic pathways (i.e., there was code–pathway coevolution). The importance of biosynthetic pathways for the code evolution is almost self-evident because amino acids could not be incorporated into the code unless they were available. Under this coevolution theory, the code evolved by subdivision: In the ancestral code, large blocks of codons encoded the same amino acid but were split to encode two amino acids upon the evolution of the respective metabolic pathways. The specific pattern of codon reassignment is determined by the precursor–product relationships between amino acids, whereby a product takes over some of the codons in the block that initially encoded the precursor. The evolution of the code is thought to be driven by positive selection for diversification of protein functionality enabled by the incorporation of new amino acids.

Implications of the coevolution theory are not limited to the self-evident importance of biosynthetic pathways for the inclusion of the late amino acids into the code. On the basis of the precursor–product relationships between amino acids in these pathways, the theory makes specific and readily falsifiable predictions about the subdivision of the primordial blocks of codon assignments (as any genuine theory, of course, should do). As we show in the next section, these specific inferences do not seem to fare particularly well.

Error Minimization Theory

The grouping of similar amino acids within the same column of the code table (**Figure 1**) immediately indicates that the code is robust to mutational and translational errors, i.e., that it is organized in a way that reduces the deleterious effect of such errors. Beyond this qualitative assessment, how unusual is the robustness of the SGC compared with random codes? Answering this question in a meaningful way is a nontrivial task that requires a well-defined cost function to assign to each version of the code a numerical value quantifying its ability to tolerate errors. In order to quantify the relative robustness of each code, these values are then compared with those obtained for randomized code versions. The cost function depends on the metric of amino acid relatedness that quantifies the fitness effect of replacing one amino acid in a protein with each other amino acid. The metric that is most often employed in the study of code robustness is the PR scale, but additional properties of amino acids—e.g., the hydropathy index or molecular volume—also have been used either as alternatives or in combination (17, 21, 32, 72). Under alternative approaches, metrics that reflect the measured fitness effect of amino acid substitutions in evolving proteins were used instead of physicochemical metrics (9). The method of generating the sample of random codes also matters because searching the entire space of $>10^{84}$ potential codes or even any substantial part of that space is impractical. In an early formal analysis of the code's robustness, Freeland & Hurst (25) showed that the code is one in a million, i.e., that the chance that a random code is as robust as the SGC is approximately 10^{-6}. More sophisticated variations of this analysis using different cost functions and randomization schemes have in subsequent studies pointed to even greater levels of optimization in the SGC depending on the parameters of analysis (27, 30, 51–53, 59, 60, 67, 79, 103). The unanimous consensus from these studies is that the SGC is well adapted to mitigate the effect of errors but also that there is a huge number of codes (even if it is a small fraction of all of them) that are even more robust. Furthermore, it appears that the SGC is not a local peak on the code fitness landscape because certain local rearrangements can increase the level of error minimization; a quantitative estimate has shown that the SGC is positioned roughly halfway from an average random code to the summit of the corresponding local peak (60).

The typical conclusion of studies demonstrating the high robustness of the SGC is that the code evolved under selection for error minimization. However, notwithstanding its intuitive appeal, this conclusion is not necessarily justified. Code optimization analyses traditionally focus on rearrangements of the standard code table, which allows a formal estimation of robustness but not of the evolutionary processes that lead to it. A substantially different approach involves reconstruction of the routes of code expansion that might produce error minimization as a selectively neutral by-product of evolution driven by other factors (51–53).

The idea of neutral evolution of error minimization goes back to the seminal 1968 paper of Crick (15), in which he wrote that "similar amino acids would tend to have similar codons" (p. 375) in the course of postulated code expansion toward the frozen accident version. In a recent series of publications, Massey (51–53) provided a comprehensive simulation model that gives strong statistical support to the neutral emergence of code robustness. Three specific scenarios of code expansion were explored: the 2-1-3 model of Massey (50), which is similar to expansion schemes

independently proposed by Higgs (32) and Francis (24); the ambiguity reduction model (23); and a scheme that follows the precursor–product expansion code as specified by the coevolution theory (89, 90). The 2-1-3 model and similar schemes, inspired by the properties of the SGC (see the section titled The Standard Code, Its Variants, and Its Recent Evolution), postulate that, in the ancestral code, only the second base of the codon was informative, and this code expanded by assigning specificity to the first, and then—in some codon series—the third bases. The ambiguity reduction model postulates that in the ancestral code, codon series ambiguously encoded groups of amino acids, such that its subsequent evolution involved a gradual increase in the specificity of codon–amino acid mapping. A code's fitness cost, based on several measures of amino acid similarity, was applied, and random code evolution was simulated under each of the three expansion schemes. The simulations have shown that both the 2-1-3 model and the ambiguity reduction model readily yield codes with error minimization levels exceeding that of the SGC. In contrast, the coevolution model, although producing some level of error minimization, was found to be inferior to the other scenarios; this was also demonstrated in an earlier analysis by Higgs (32). These findings imply that, although substantial error minimization is clearly a property of the SGC, this feature likely evolved as a by-product of code expansion rather than by direct selection for code robustness. Moreover, the evolutionary moves that are typically involved in the computational search for highly robust codes, namely swaps of codon assignments by 4-codon or 2-codon blocks, are hardly biologically realistic. Within a broader biological context, it is worth noting in this regard that extreme error minimization is not necessarily a beneficial feature of the code because it severely constrains the exploration of the sequence space in the course of evolution (22).

Given the consensus on the set of ancient amino acids and the unavailability of the late amino acids prior to the advent of complex biosynthetic pathways (see the section titled Ancient Evolution of the Code: Early and Late Amino Acids), code expansion appears to be an essential part and parcel of the early stages of code evolution. Postulating that the third letter of the codons was uninformative in the primordial code and reducing the code table to the codons for the 10 early amino acids resulted in the reconstruction of the ancient doublet code, albeit with several codon series remaining ambiguous (**Figure 2**) (59). Notably, this reconstructed primordial code shows an exceptional level of error minimization and is much closer to the global optimum than the SGC. This observation implies that at the early stage of the code evolution, when the fidelity of primitive translation might have been low, selection for error minimization could have been a more important factor in evolution of the code than it was at the later stages. Again, however, it cannot be ruled out that the high level of error minimization in the doublet code is a simple consequence of the expansion of the most primitive code, e.g., under the 2-1-3 and the 4-column scenarios.

As discussed above, most studies on the error minimization property of the code rely on more or less arbitrary cost functions that provide a numerical measure of the code's ability to

	U	C	A	G
U	Leu/???	Ser	???	???
C	Leu	Pro	???	???
A	Ile	Thr	???	Ser/???
G	Val	Ala	Asp	Gly

Figure 2

Reconstruction of the ancient doublet code. A 2-letter code consisting of 16 supercodons is shown with assignments inferred from the list of early amino acids. Question marks show uncertain codon assignments. The first and second codon positions correspond to the letters along the left side and the top of the table, respectively. The cell shading is the same as in **Figure 1**.

withstand mutations and translation errors. More biologically realistic alternatives to the cost function have been explored. In particular, in vitro evolution experiments, in which the fitness cost of mutations was assessed by their effect on the properties of proteins, have shown that the structure of the code measurably constrains evolution. As a result, the highest fitness mutants evolved along consecutive single substitution trajectories were far inferior to many variants selected from exhaustive mutagenesis experiments (22). The same study showed that the code minimizes the fitness cost of mutations in the sense that the mean cost of an amino acid replacement caused by a single nucleotide substitution is substantially lower than the cost of replacements associated with two or three nucleotide substitutions. Moreover, and much less intuitively, the fraction of adaptive amino acid replacements is greater among those that are accessible via a single substitution than among all replacements. The two properties, minimization of the deleterious effect of mutations and enhancement of the adaptive effect, are probably two sides of the same coin—namely, code organization that strongly favors replacement of amino acids with similar ones resulting from single nucleotide substitutions. Thus, robustness of the code to errors appears to be intrinsically coupled with enhancement of protein evolvability as previously suggested by some theoretical analyses (24, 32). These findings are compatible with selection for robustness as a driving force in the evolution of the code, but they cannot rule out emergence of the code as a by-product under the other evolutionary scenarios. Regardless, and despite their limited scope, experimental studies of code features appear crucially important for validating and expanding the conclusions of previous theoretical analyses.

Last universal common ancestor (LUCA): the hypothetical ancestor of all extant cellular life forms

COEVOLUTION OF THE CODE AND THE TRANSLATION SYSTEM: THE PROTEIN EVOLUTION PARADOX AND THE EXTINCT PRIMORDIAL STEREOCHEMICAL CODE

As already pointed out, the origin and evolution of the translation system is a forbiddingly difficult problem; therefore, many studies treat it formally as a separate issue and approach it almost like a mathematical puzzle (84). Ultimately, however, it appears almost certain that evolution of the code can be understood only in the context of the evolution of translation. An attempt to systematically address the problem of the code's evolution in the context of the molecular biology of the translation system—particularly the interactions among tRNAs, mRNAs, and the ribosome—has been recently undertaken by Grosjean & Westhof (28). The amino acids are divided into three groups that differ by the free energy of the codon–anticodon interaction that obviously correlates with the GC content (**Figure 1**). These groups are (*a*) strong, or having a mean free energy of −3.1 kcal/mol, which includes Gly, Ala, Pro, and Arg; (*b*) weak, or having a mean free energy of −1.0 kcal/mol, which includes Asn, Ile, Leu, Met, Phe, and Tyr; and (*c*) intermediate, or having a mean free energy of −2.2 kcal/mol, which includes the remaining nine amino acids. Obviously, three of the four strong amino acids (with the exception of Arg) are among the simplest and most abundant in the early set. Grosjean & Westhof argue that the code evolved from the earliest, GC-rich stage with four amino acids only (probably including a precursor molecule in place of Arg). It then incorporated the intermediate and then the weak amino acids. Accurate decoding of the codons for the latter group requires an extended anticodon with a modified base in the tRNA molecule (2), which could be an additional argument for the late arrival of these amino acids in the code.

Although the translation system is universally conserved among extant cellular life forms, many protein components of the translation apparatus are paralogs that, furthermore, belong to large protein families. Therefore, phylogenetic analysis of the respective protein families opens a window into the phase of evolution preceding the last universal common ancestor (LUCA). The

results of such analyses bring up a conundrum that can be termed the protein evolution paradox. The aaRSs, which ensure the accurate matching between amino acids and cognate codons in the modern translation system, belong to two classes of paralogs with 10 specificities in each class (86, 87). As demonstrated particularly convincingly for the Class I aaRSs, which contain Rossmann fold catalytic domains, the diversification of the aaRSs is a late stage in the evolution of the Rossmann fold protein superfamily (5, 6). By the time the common ancestor of the aaRSs gave rise to the 10 specificities through a series of duplications, extensive evolution of the Rossmann fold superfamily had already produced a substantial diversity of other enzymatic and nucleotide-binding domains. Similar conclusions can be reached for the evolution of Class II aaRSs, which belong to the biotin synthase superfamily, although the evolution of these proteins has not been examined to the same level of detail (4, 8). Clearly, for this early protein evolution to occur, high-fidelity translation was essential, and given that the different aaRS specificities had not yet evolved, the inevitable conclusion was that the specificity of amino acid–codon correspondence was determined by RNA molecules (41).

The conclusion that mRNA decoding in the early translation system was performed by RNA molecules that were, conceivably, the evolutionary precursors of modern tRNAs implies a stereochemical model of code origin and evolution distinct from traditional models of this type (**Figure 3**) (42). Under this model, specific RNA–amino acid interactions would not involve the anticodon (let alone the codon), so it therefore could be chosen arbitrarily and fixed through frozen accident. Instead, following the reasoning presented previously (88), the amino acids are postulated to have been recognized by unique pockets in the tertiary structure of the proto-tRNAs. The grouping of codons for related amino acids underlying error minimization naturally follows from code expansion through duplication of the proto-tRNAs; the molecules resulting from such duplications obviously would be structurally similar and accordingly would bind similar amino acids (**Figure 3**). This part of the scenario is a version of the model of code expansion that could be driven by the benefits of diversifying the repertoire of protein amino acids and would yield robustness as a by-product.

Once the specificity determinants migrated from the proto-tRNAs to the aaRSs, the amino acid–binding pockets in the tRNAs deteriorated such that modern tRNAs show no detectable affinity to the cognate amino acids. Attempts to decipher that primordial stereochemical code by comparative analysis of modern translation system components are likely to be futile. The aptamer experiments do not appear to be up to the task, either, because more complex RNA molecules are required. It appears that the best hope for cracking the ancient code lies in experiments on in vitro evolution of aminoacylating ribozymes, which themselves seem to recapitulate a key aspect of the primordial translation system (12, 45, 95).

CONCLUDING REMARKS

The problems of the nature, origin, and evolution of the genetic code appear to be unique in combining extreme outward simplicity with excruciating difficulty. The code literature reviewed here spans more than 50 years, from 1965 to 2017, but the analyses of Woese and Crick from the 1960s remain remarkably relevant. Certainly, there have been many developments with regard to the quantification of code properties, but the fundamental framework of evolutionary ideas laid out in the classic papers has not been overhauled. Unfortunately, this is due less to the successes of the early research than to the limited and questionable progress achieved over the next half century. The prescience of Woese and Crick cannot be questioned, but nevertheless, they were far from an adequate understanding of the code's evolution. Notwithstanding the complete transformation of biology that occurred over these decades, we do not seem to be much closer to the solution.

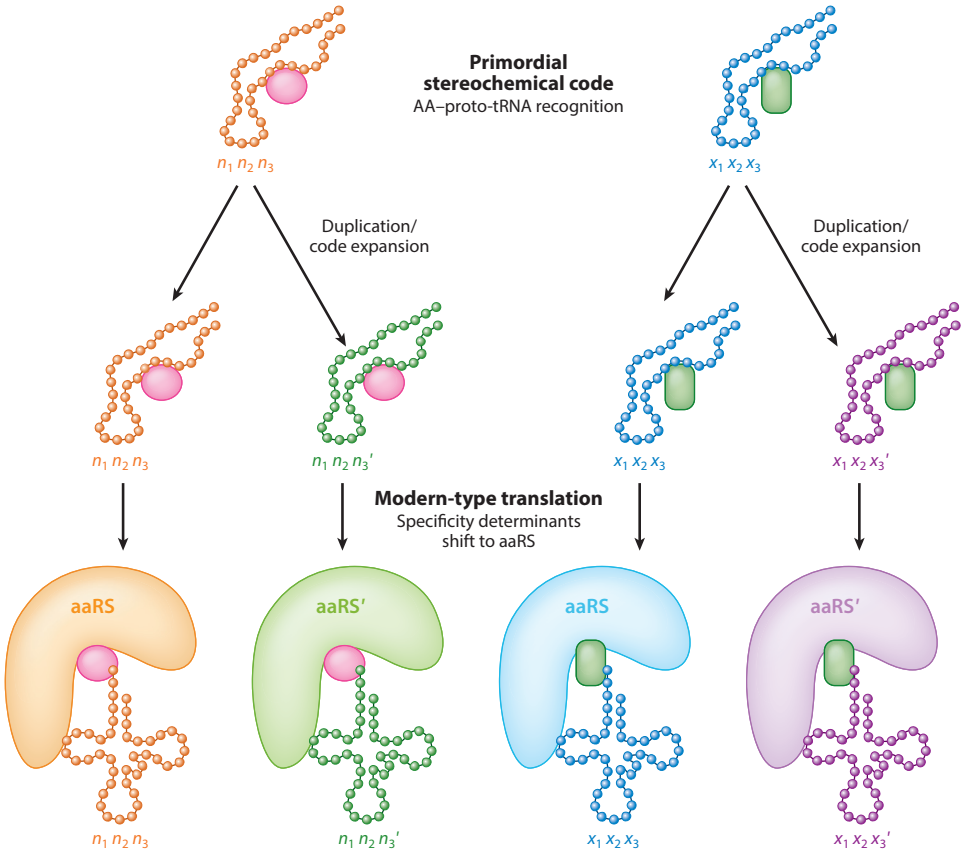

Figure 3

Evolution of the putative primordial stereochemical code to the standard code by expansion and frozen accident. Related amino acids (AAs) and aminoacyl–tRNA synthetases (aaRSs) are shown in similar colors. In the code expansion phase, the anticodons are shown changing in the third position (corresponding to the first position of the codon), resulting in recruitment of related amino acids.

Recalling the list of "why" questions we asked in the introduction, we find that it is hardly possible to answer any of them definitively.

In a difficult situation like this, it seems useful to take stock of what we actually can and cannot be confident about regarding the evolution of the code. So let us make a short list of well-established properties of the code and aspects of its evolution:

- The code is effectively universal: Departures from code universality in extant organisms are minor and of secondary origin.
- The code is nonrandomly organized and is highly robust to errors, although it is far from being globally optimal.
- Evolution of the code involved expansion from a limited set of primordial amino acids toward the modern canonical set.

This seems to be the extent of what we can confidently claim about the code's evolution. The three main theories of code evolution seem to have largely exhausted their potential for constructive analysis, at least in their traditional form. The stereochemical theory, focusing on

codons and anticodons, fails to provide clear solutions. The coevolution and error minimization theories each capture important aspects of the code evolution, but they do not seem to provide major insights beyond the solid conclusion about code expansion most likely involving duplication of proto-tRNAs. Such code expansion would have been contingent on the evolution of amino acid biosynthesis pathways—although it would not necessarily reflect the metabolic connections precisely—and it would yield error minimization as a by-product.

Unlike the table of codon assignments itself, the origin and evolution of the code hardly can be cracked like a puzzle. Arguably, these processes can be reconstructed only by a comprehensive analysis of the early stages in the evolution of translation. This is a daunting task that, at a superficial glance, might seem virtually hopeless given the universal conservation of the core translation system components in all extant cellular life forms. Nevertheless, extensive duplication within the translation system opens a window into the deep pre-LUCA past. In particular, phylogenetic analysis of the aaRSs indicates that, at a stage of evolution when the translation system had already attained high fidelity, the correspondence between amino acids and cognate codons was determined by recognition of the former by RNA molecules, i.e., by proto-tRNAs. Hence we propose an experimentally testable scenario for code evolution that combines a new flavor of the stereochemical hypothesis—in which amino acids are recognized via unique sites in the 3D structure of proto-tRNAs rather than by anticodons—with the notions of code expansion and frozen accident (**Figure 3**).

In our view, theoretical study of the genetic code as a cryptographic problem has largely run its course. The best hope for further progress in understanding the origin and evolution of the code seems to lie with the technically challenging but conceptually clear experimentation aiming at recapitulation of the inferred steps in the translation system evolution.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

E.V.K. is supported by intramural funds of the US Department for Health and Human Services to the National Library of Medicine.

LITERATURE CITED

1. Aggarwal N, Bandhu AV, Sengupta S. 2016. Finite population analysis of the effect of horizontal gene transfer on the origin of an universal and optimal genetic code. *Phys. Biol.* 13:036007
2. Allner O, Nilsson L. 2011. Nucleotide modifications and tRNA anticodon–mRNA codon interactions on the ribosome. *RNA* 17:2177–88
3. Ambrogelly A, Palioura S, Soll D. 2007. Natural expansion of the genetic code. *Nat. Chem. Biol.* 3:29–35
4. Anantharaman V, Koonin EV, Aravind L. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* 30:1427–64
5. Aravind L, Anantharaman V, Koonin EV. 2002. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. *Proteins* 48:1–14
6. Aravind L, Mazumder R, Vasudevan S, Koonin EV. 2002. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* 12:392–99

7. Archetti M. 2004. Selection on codon usage for error minimization at the protein level. *J. Mol. Evol.* 59:400–15
8. Artymiuk PJ, Rice DW, Poirrette AR, Willet P. 1994. A tale of two synthetases. *Nat. Struct. Biol.* 1:758–60
9. Buhrman H, van der Gulik PT, Kelk SM, Koolen WM, Stougie L. 2011. Some mathematical refinements concerning error minimization in the genetic code. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8:1358–72
10. Burton AS, Stern JC, Elsila JE, Glavin DP, Dworkin JP. 2012. Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chem. Soc. Rev.* 41:5459–72
11. Chin JW. 2014. Expanding and reprogramming the genetic code of cells and animals. *Annu. Rev. Biochem.* 83:379–408
12. Chumachenko NV, Novikov Y, Yarus M. 2009. Rapid and simple ribozymic aminoacylation using three conserved nucleotides. *J. Am. Chem. Soc.* 131:5257–63
13. Cleaves HJ II. 2010. The origin of the biologically coded amino acids. *J. Theor. Biol.* 263:490–98
14. Crick FH. 1966. The genetic code—yesterday, today, and tomorrow. *Cold Spring Harb. Symp. Quant. Biol.* 31:3–9
15. Crick FH. 1968. The origin of the genetic code. *J. Mol. Biol.* 38:367–79
16. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. 1961. General nature of the genetic code for proteins. *Nature* 192:1227–32
17. de Oliveira LL, de Oliveira PS, Tinos R. 2015. A multiobjective approach to the genetic code adaptability problem. *BMC Bioinform.* 16:52
18. Di Giulio M. 2005. The origin of the genetic code: theories and their relationships, a review. *Biosystems* 80:175–84
19. Di Giulio M. 2008. An extension of the coevolution theory of the origin of the genetic code. *Biol. Direct* 3:37
20. Di Giulio M. 2016. The lack of foundation in the mechanism on which are based the physico-chemical theories for the origin of the genetic code is counterposed to the credible and natural mechanism suggested by the coevolution theory. *J. Theor. Biol.* 399:134–40
21. Di Giulio M, Medugno M. 2001. The level and landscape of optimization in the origin of the genetic code. *J. Mol. Evol.* 52:372–82
22. Firnberg E, Ostermeier M. 2013. The genetic code constrains yet facilitates Darwinian evolution. *Nucleic Acids Res.* 41:7420–28
23. Fitch WM, Upper K. 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* 52:759–67
24. Francis BR. 2013. Evolution of the genetic code by incorporation of amino acids that improved or changed protein function. *J. Mol. Evol.* 77:134–58
25. Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. *J. Mol. Evol.* 47:238–48
26. Gamow G. 1954. Possible relation between deoxyribonucleic acid and protein structures. *Nature* 173:318
27. Goodarzi H, Nejad HA, Torabi N. 2004. On the optimality of the genetic code, with the consideration of termination codons. *Biosystems* 77:163–73
28. Grosjean H, Westhof E. 2016. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res.* 44:8020–40
29. Haig D, Hurst LD. 1991. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* 33:412–17
30. Haig D, Hurst LD. 1999. Erratum for "A quantitative measure of error minimization in the genetic code." *J. Mol. Evol.* 49:708
31. Hajnic M, Osorio JI, Zagrovic B. 2014. Computational analysis of amino acids and their sidechain analogs in crowded solutions of RNA nucleobases with implications for the mRNA–protein complementarity hypothesis. *Nucleic Acids Res.* 42:12984–94
32. Higgs PG. 2009. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol. Direct* 4:16
33. Higgs PG, Pudritz RE. 2009. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* 9:483–90

34. Hlevnjak M, Polyansky AA, Zagrovic B. 2012. Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. *Nucleic Acids Res.* 40:8874–82
35. Iranzo J, Puigbò P, Lobkovsky AE, Wolf YI, Koonin EV. 2016. Inevitability of genetic parasites. *Genome Biol. Evol.* 8:2856–69
36. Jestin JL, Kempf A. 2009. Optimization models and the structure of the genetic code. *J. Mol. Evol.* 69:452–57
37. Johnson DB, Wang L. 2010. Imprints of the genetic code in the ribosome. *PNAS* 107:8298–303
38. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, et al. 2005. A universal trend of amino acid gain and loss in protein evolution. *Nature* 433:633–38
39. Knight RD, Freeland SJ, Landweber LF. 1999. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci.* 24:241–47
40. Koonin EV. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* 1:127–36
41. Koonin EV. 2011. *The Logic of Chance: The Nature and Origin of Biological Evolution*. Upper Saddle River, NJ: FT Press
42. Koonin EV. 2017. Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. *Life* 7:E22
43. Koonin EV, Martin W. 2005. On the origin of genomes and cells within inorganic compartments. *Trends Genet.* 21:647–54
44. Koonin EV, Novozhilov AS. 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61:99–111
45. Kumar RK, Yarus M. 2001. RNA-catalyzed amino acid activation. *Biochemistry* 40:6998–7004
46. Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLOS Biol.* 3:e130
47. Ling J, O'Donoghue P, Soll D. 2015. Genetic code flexibility in microorganisms: novel mechanisms and impact on physiology. *Nat. Rev. Microbiol.* 13:707–21
48. Liu CC, Schultz PG. 2010. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* 79:413–44
49. Lobanov AV, Turanov AA, Hatfield DL, Gladyshev VN. 2010. Dual functions of codons in the genetic code. *Crit. Rev. Biochem. Mol. Biol.* 45:257–65
50. Massey SE. 2006. A sequential "2-1-3" model of genetic code evolution that explains codon constraints. *J. Mol. Evol.* 62:809–10
51. Massey SE. 2008. A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* 67:510–16
52. Massey SE. 2015. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. *Life* 5:1301–32
53. Massey SE. 2016. The neutral emergence of error minimized genetic codes superior to the standard genetic code. *J. Theor. Biol.* 408:237–42
54. Mathew DC, Luthey-Schulten Z. 2008. On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* 66:519–28
55. Mukai T, Englert M, Tripp HJ, Miller C, Ivanova NN, et al. 2016. Facile recoding of selenocysteine in nature. *Angew. Chem. Int. Ed. Engl.* 55:5337–41
56. Neumann H, Wang K, Davis L, Garcia-Alai M, Chin JW. 2010. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* 464:441–44
57. Nirenberg MW. 2004. Historical review: deciphering the genetic code—a personal account. *Trends Biochem. Sci.* 29:46–54
58. Nirenberg MW, Matthaei JH. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *PNAS* 47:1588–602
59. Novozhilov AS, Koonin EV. 2009. Exceptional error minimization in putative primordial genetic codes. *Biol. Direct* 4:44
60. Novozhilov AS, Wolf YI, Koonin EV. 2007. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol. Direct* 2:24
61. Pizzarello S. 2006. The chemistry of life's origin: a carbonaceous meteorite perspective. *Acc. Chem. Res.* 39:231–37

62. Polyansky AA, Hlevnjak M, Zagrovic B. 2013. Analogue encoding of physicochemical properties of proteins in their cognate messenger RNAs. *Nat. Commun.* 4:2784
63. Polyansky AA, Hlevnjak M, Zagrovic B. 2013. Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code. *RNA Biol.* 10:1248–54
64. Polyansky AA, Zagrovic B. 2013. Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res.* 41:8434–43
65. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* 12:66
66. Puigbò P, Wolf YI, Koonin EV. 2010. The tree and net components of prokaryote evolution. *Genome Biol. Evol.* 2:745–56
67. Salinas DG, Gallardo MO, Osorio MI. 2016. Local conditions for global stability in the space of codons of the genetic code. *Biosystems* 150:73–77
68. Santos MA, Moura G, Massey SE, Tuite MF. 2004. Driving change: the evolution of alternative genetic codes. *Trends Genet.* 20:95–102
69. Sella G, Ardell DH. 2006. The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J. Mol. Evol.* 63:297–313
70. Sengupta S, Aggarwal N, Bandhu AV. 2014. Two perspectives on the origin of the standard genetic code. *Orig. Life Evol. Biosph.* 44:287–91
71. Sengupta S, Higgs PG. 2005. A unified model of codon reassignment in alternative genetic codes. *Genetics* 170:831–40
72. Sengupta S, Higgs PG. 2015. Pathways of genetic code evolution in ancient and modern organisms. *J. Mol. Evol.* 80:229–43
73. Sengupta S, Yang X, Higgs PG. 2007. The mechanisms of codon reassignments in mitochondrial genetic codes. *J. Mol. Evol.* 64:662–88
74. Swart EC, Serra V, Petroni G, Nowacki M. 2016. Genetic codes with no dedicated stop codon: context-dependent translation termination. *Cell* 166:691–702
75. Szathmary E, Demeter L. 1987. Group selection of early replicators and the origin of life. *J. Theor. Biol.* 128:463–86
76. Szathmary E, Maynard Smith J. 1997. From replicators to reproducers: the first major transitions leading to life. *J. Theor. Biol.* 187:555–71
77. Takeuchi N, Kaneko K, Koonin EV. 2014. Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. *G3* 4:325–39
78. Tlusty T. 2010. A colorful origin for the genetic code: information theory, statistical mechanics and the emergence of molecular codes. *Phys. Life Rev.* 7:362–76
79. Torabi N, Goodarzi H, Shateri Najafabadi H. 2007. The case for an error minimizing set of coding amino acids. *J. Theor. Biol.* 244:737–44
80. Treangen TJ, Rocha EP. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLOS Genet.* 7:e1001284
81. Trifonov EN. 2000. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* 261:139–51
82. Trifonov EN. 2004. The triplet code from first principles. *J. Biomol. Struct. Dyn.* 22:1–11
83. Vetsigian K, Woese C, Goldenfeld N. 2006. Collective evolution and the genetic code. *PNAS* 103:10696–701
84. Woese CR. 1967. *The Genetic Code*. New York: Harper & Row
85. Woese CR. 1968. The fundamental nature of the genetic code: prebiotic interactions between polynucleotides and polyamino acids or their derivatives. *PNAS* 59:110–17
86. Woese CR, Olsen GJ, Ibba M, Soll D. 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* 64:202–36
87. Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* 9:689–710

88. Wolf YI, Koonin EV. 2007. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol. Direct* 2:14
89. Wong JT-F. 1975. A co-evolution theory of the genetic code. *PNAS* 72:1909–12
90. Wong JT-F, Ng S-K, Mat W-K, Hu T, Xue H. 2016. Coevolution theory of the genetic code at age forty: pathway to translation and synthetic life. *Life* 6:12
91. Xie J, Schultz PG. 2006. A chemical toolkit for proteins—an expanded genetic code. *Nat. Rev. Mol. Cell Biol.* 7:775–82
92. Yarus M. 1998. Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J. Mol. Evol.* 47:109–17
93. Yarus M. 2000. RNA-ligand chemistry: a testable source for the genetic code. *RNA* 6:475–84
94. Yarus M. 2010. *Life from an RNA World: The Ancestor Within*. Cambridge, MA: Harvard Univ. Press
95. Yarus M. 2011. The meaning of a minuscule ribozyme. *Philos. Trans. R. Soc. B* 366:2902–9
96. Yarus M, Caporaso JG, Knight R. 2005. Origins of the genetic code: the escaped triplet theory. *Annu. Rev. Biochem.* 74:179–98
97. Yarus M, Christian EL. 1989. Genetic code origins. *Nature* 342:349–50
98. Yarus M, Widmann JJ, Knight R. 2009. RNA–amino acid binding: a stereochemical era for the genetic code. *J. Mol. Evol.* 69:406–29
99. Yuan J, O'Donoghue P, Ambrogelly A, Gundllapalli S, Sherrer RL, et al. 2010. Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett.* 584:342–49
100. Zahonova K, Kostygov AY, Sevcikova T, Yurchenko V, Elias M. 2016. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr. Biol.* 26:2364–69
101. Zaia DA, Zaia CT, De Santana H. 2008. Which amino acids should be used in prebiotic chemistry studies? *Orig. Life Evol. Biosph.* 38:469–88
102. Zhang Y, Baranov PV, Atkins JF, Gladyshev VN. 2005. Pyrrolysine and selenocysteine use dissimilar decoding strategies. *J. Biol. Chem.* 280:20740–51
103. Zhu W, Freeland S. 2006. The standard genetic code enhances adaptive evolution of proteins. *J. Theor. Biol.* 239:63–70

Annual Review of Genetics

Volume 51, 2017

Contents

Errata

An online log of corrections to *Annual Review of Genetics* articles may be found at http://www.annualreviews.org/errata/genet