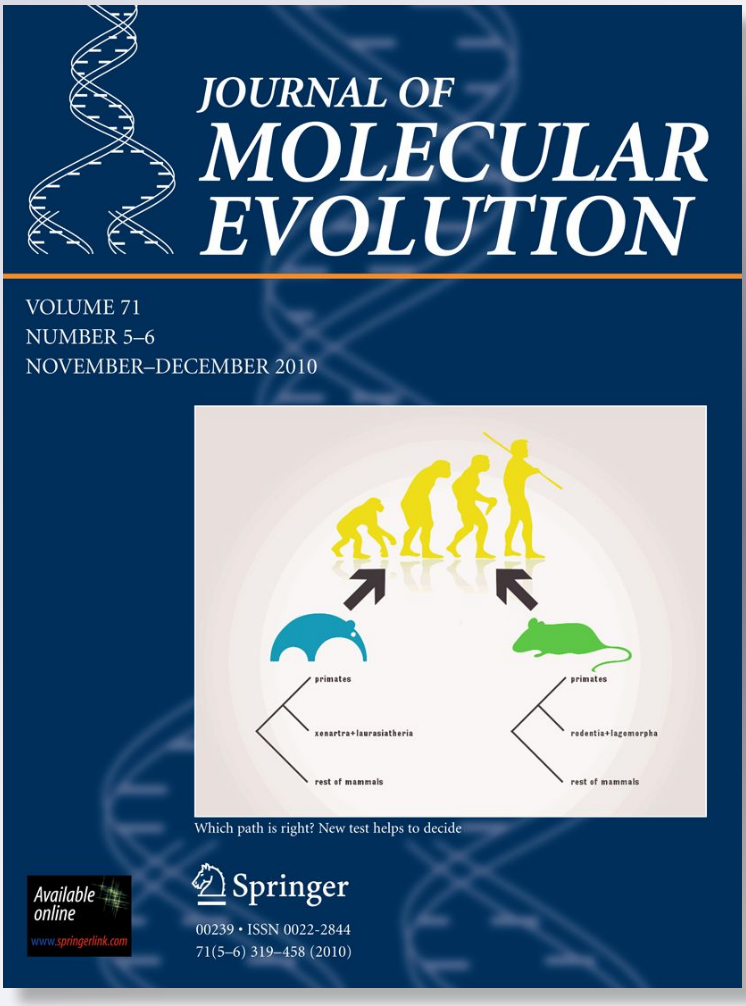
Unassigned Codons, Nonsense Suppression, and Anticodon Modifications in the Evolution of the Genetic Code

Peter T. S. van der Gulik & Wouter D. Hoff

Journal of Molecular Evolution

ISSN 0022-2844
Volume 73
Combined 3-4

J Mol Evol (2011) 73:59-69
DOI 10.1007/s00239-011-9470-3



J Mol Evol (2011) 73:59–69
DOI 10.1007/s00239-011-9470-3

Unassigned Codons, Nonsense Suppression, and Anticodon Modifications in the Evolution of the Genetic Code

Peter T. S. van der Gulik · Wouter D. Hoff

Received: 8 July 2011 / Accepted: 24 October 2011 / Published online: 11 November 2011

Abstract The origin of the genetic code is a central open problem regarding the early evolution of life. Here, we consider two undeveloped but important aspects of possible scenarios for the evolutionary pathway of the translation machinery: the role of unassigned codons in early stages of the code and the incorporation of tRNA anticodon modifications. As the first codons started to encode amino acids, the translation machinery likely was faced with a large number of unassigned codons. Current molecular scenarios for the evolution of the code usually assume the very rapid assignment of all codons before all 20 amino acids became encoded. We show that the phenomenon of nonsense suppression as observed in current organisms allows for a scenario in which many unassigned codons persisted throughout most of the evolutionary development of the code. In addition, we demonstrate that incorporation of anticodon modifications at a late stage is feasible. The wobble rules allow a set of 20 tRNAs fully lacking anticodon modifications to encode all 20 canonical amino acids. These observations have implications for the biochemical plausibility of early stages in the evolution of the genetic code predating tRNA anticodon modifications and allow for effective translation by a relatively small and simple early tRNA set.

Keywords Genetic code · Unassigned codons · Wobble rules · Evolution · RNA modification · tRNA

P. T. S. van der Gulik (✉)
Centrum Wiskunde & Informatica (CWI), P.O. Box 94079,
1090 GB Amsterdam, The Netherlands
e-mail: Peter.van.der.Gulik@cwi.nl

W. D. Hoff
Department of Microbiology and Molecular Genetics,
Oklahoma State University, Stillwater, OK 74078, USA

The origin of the genetic code can be envisioned as starting with a single primordial tRNA, which gave rise to the full complement of tRNAs by a complex series of gene duplication and diversification events. This view of tRNA genes as paralogues pervades thinking about the origin and evolution of the genetic code (Crick 1968; Fitch and Upper 1987; Osawa et al. 1992). While many aspects of tRNA evolution have been considered (cf. Di Giulio 2006; Randau and Söll 2008; Fujishima et al. 2009; Shaul et al. 2010; Rodin et al. 2011), gene duplication and diversification are common themes during the evolutionary development of tRNA sets. Presumably, during this diversification process additional amino acids were incorporated one by one into the developing genetic code. This consideration leads to an important problem facing possible scenarios for the evolution of the code. In very early stages of the development of the standard genetic code (SGC) most codons were unassigned, leading to a situation in which many mutations in an early protein-encoding nucleic acid sequence would result in the introduction of an unassigned codon (Speyer et al. 1963; Sonneborn 1965; Crick 1968).

One can envision two general approaches to this problem of potentially lethal unassigned codons. The first option is that as soon as a small set of amino acids started to be encoded by tRNAs, rapid tRNA gene duplication and mutation of the anticodon resulted in a situation in which all codons were assigned to this initial set of amino acids. An important consequence of this scenario is that subsequent incorporation of novel amino acids into the expanding code requires reassignments of the meaning of codons. A second approach is that the code evolved more slowly, and that for extended periods of evolutionary time indeed many codons were not assigned (Lehman and Jukes 1988; Ikehara 2002; Francis 2011). The introduction of novel amino acids could then proceed without codon reassignment. However, this scenario requires the non-lethality of nonsense mutations during the early evolution of the code. Thus, formulating specific molecular scenarios for the evolution of the genetic code requires a choice: either numerous codon reassignment events or the prolonged existence of nonsense codons. Current thinking strongly favors the first option (e.g., Agris et al. 2007; Higgs 2009; Grosjean et al. 2010).

Here, we examine the strength of the evidence supporting this choice, and use biochemical knowledge regarding nonsense suppression in existing organisms (Beier and Grimm 2001; Kramer and Farabaugh 2007) to support the viability of the second scenario. In addition, we use knowledge on tRNA wobble rules (Crick 1966; Takai and Yokoyama 2003; Agris et al. 2007; Grosjean et al. 2010; Ran and Higgs 2010) and the biochemistry of tRNA anticodon modifications (Muramatsu et al. 1988; Mandal et al. 2010; Ikeuchi et al. 2010) to examine when tRNA anticodon modifications were introduced into the developing genetic code. These considerations lead to a novel scenario for the development of the SGC. All such scenarios are faced with the issue of the temporal order of and interplay between three key developments: (i) the assignment of nonsense codons, (ii) the incorporation of all 20 canonical amino acids into the code, and (iii) the introduction of tRNA anticodon modifications. We present an analysis of relevant available biochemical information that supports a model that contrasts with most published models with respect to the relative order of these three processes. This analysis supports the viability of scenarios involving the persistence of nonsense codons until all 20 amino acids were included in the code, and the incorporation of anticodon modifications at a relatively late stage in the evolution of the code.

Unassigned Codons and Nonsense Suppression

The highly deleterious nature of nonsense codons was vividly described in an influential 1965 paper by Tracey Sonneborn:

> A nonsense mutation resulting in nontranslation of all codons distal to it would as a rule be enormously more detrimental (and therefore more rapidly eliminated) than a sensible (or mis-sensible) mutation which permits translation of the entire message. Hence, neutralizing the detriment of a nonsense mutation by a second mutation or a genic recombination is very much less likely. In short, such nonsense mutations would with high probability have no evolutionary future, and they would by virtue of their detriment be prime targets for elimination by natural selection. On the other hand, mis-sense mutations could sometimes have relatively little detrimental effect and therefore a relatively long persistence and correspondingly greater chance to enter into a lucky genic combination by further mutation or recombination.

This early view on the highly lethal nature of nonsense mutations and the relatively benign character of missense mutations has been solidly incorporated into thinking about the evolution of the genetic code (e.g., Crick 1968; Agris et al. 2007). As a result, the persistence of nonsense codons during most of the evolution of the SGC has not been considered as a viable possibility, while codon reassignments during this process are viewed as realistic and unproblematic. This view has been developed in detail in an important recent paper (Higgs 2009).

While the deleterious effect of nonsense mutations stands unchallenged (Sonneborn 1965; Crick 1968; Agris et al. 2007; Higgs 2009), here we want to reinvestigate its implications for early stages of the genetic code. Specifically, we will examine both the presumed level of lethality of nonsense mutations and the presumed likelihood of codon reassignments in the light of current knowledge of existing organisms.

A significant body of data is available regarding the translational fate of mRNA molecules containing nonsense mutations (Beier and Grimm 2001; Chabelskaya et al. 2004; Doronina and Brown 2006; Lao et al. 2009). These studies have revealed that a significant level of translational readthrough across stop codons occurs. As a result, nonsense mutations even in essential genes often are non-lethal.

Such nonsense suppression can involve mutations in tRNAs as in the amber, ocher, and opal suppressor tRNAs. However, natural nonsense suppression through the reading of stop codons by normal cellular tRNAs, which are called natural suppressors, has also been well documented (Beier and Grimm 2001). In general, a view of translation has emerged in which the meaning of a codon is always a balance between the affinities of several different tRNAs for that codon, and the affinity of release factors for that codon (Kramer and Farabaugh 2007). The current translational machinery in general exhibits a very low error rate. Thus, the amount of full-length protein that is produced in the presence of a stop codon in a coding sequence is significantly reduced, but in a number of cases (e.g., Longstaff et al. 2007; Murina et al. 2010) has been found to allow for viability of the organism.

The degree to which the use of formally unassigned codons diminishes the translational efficiency of an organism will depend on its codon usage. In some organisms, the usage of certain codons can be extremely low (see e.g., Ussery et al. 2004), and inefficient translation of these codons will therefore only affect the synthesis of a small number of proteins. A central factor affecting codon usage is the abundance of the tRNA involved: tRNAs that are rare in the cellular tRNA pool tend to translate codons that are also rare, particularly in highly expressed proteins, presumably to optimize translational efficiency (Akashi 2001). If such rare codons were to become formally unassigned, this event would be expected to result in relatively mild detrimental effects. Indeed, formally unassigned codons are known in current organisms (Oba et al. 1991; Kano et al. 1993), providing a powerful argument against the supposed lethality of unassigned codons due to their introduction into the genome by mutations.

Unassigned Codons, Suppression, and Termination in Primordial Organisms

The experimental work on natural nonsense suppression discussed above has been obtained using contemporary organisms. What to expect in the case of primordial organisms? The first critical consideration is that it appears likely that the fidelity of the early translational system was considerably lower. Thus, the "meaning" of a codon would be determined by its relative affinities for various tRNAs, and would thus be translated as a weighted mixture of various amino acids. Such "statistical proteins" were introduced by Woese (1965), and have also been considered in later work (Sella and Ardell 2006; Higgs 2009). Reduced translational fidelity implies a level of readthrough (and therefore non-lethality) that is higher than that observed in current organisms. The presence of "inaccurate decoding" does not necessarily mean lethality: the acquisition of new evolutionary potentialities as a result of production of "statistical proteins" can even confer growth advantage. This has been experimentally demonstrated using mutants in which the editing function of isoleucinyl-tRNA synthetase was impaired, resulting in the low-level incorporation of non-canonical amino acids like norvaline into the proteome and an increased growth yield (Pezo et al. 2004).

The second critical consideration is that the modern system of release factors provides a rapid and high-fidelity system for recognizing stop codons. The introduction of a dedicated system for the recognition of stop codons during the evolution of the genetic code in general has not received much attention. The most primitive system for handling a stop codon would be that the ribosome stalls when it reaches an unassigned codon and eventually dissociates from the mRNA. In this view, all unassigned codons would have stop codon activity. The actual translation of unassigned codons in such an early translational system would then be a balance between the rate of natural nonsense suppression and spontaneous ribosome dissociation.

Thus, we arrive at a situation in which early translational systems combine a relatively high translational error rate, resulting in the frequent translation of formally unassigned codons, with the absence of an efficient system dedicated to recognizing stop codons. This line of thought thus predicts that formally unassigned codons could be translated either as a stop codon (through spontaneous ribosome dissociation) or as a mixture of amino acid (through nonsense suppression). The relative frequency of these events would be open to optimization through molecular evolution of the components of the early translational system. The essence for the present paper is that "unassigned codons" in effect were to a significant extent not unassigned. The introduction of such codons would thus have likely been somewhat detrimental but not lethal.

Genome size is a third consideration with respect to the proposed process of rapid tRNA gene duplication and mutation to assign all codons to a small set of initial amino acids during an early stage of the evolution of the genetic code. The early genome replication machinery can reasonably be expected to have had limited fidelity. Thus, these early systems would be at considerable risk of facing an error catastrophe in which the chance of deleterious mutations per replication event would overwhelm the rate at which natural selection can purge deleterious mutations (Eigen and Schuster 1977). This effect would result in a strong selection for organisms with very small genomes. Thus, it is not clear if systems in which the development of the genetic code has just started had sufficient genome replication fidelity to allow for a substantial number of different tRNAs. Based on these considerations, we conclude that it is reasonable to consider scenarios for the evolution of the genetic code in which many formally unassigned codons persisted throughout most of the evolutionary development of the code.

In summary, point mutations can introduce formally unassigned codons into the genome of early organisms. Because of the existence of natural nonsense suppression, such mutations will tend to reduce translational efficiency but will often not be lethal. This selective pressure against the use of such unassigned codons will cause these codons to remain rare in early organisms. Thus, the persistence of formally unassigned codons during the evolution of the genetic code is biochemically entirely plausible.

Codon Reassignments are Difficult

The SGC is nearly universal. Most code variants are known from mitochondria (see Sengupta et al. 2007 for an up-to-date treatment of mitochondrial codes and the mechanisms which lead to their emergence), which have an extremely small genome: less than a 100 protein-coding genes. Apart from mitochondria, code variants are extremely rare. Organisms as different as an elephant and an *E. coli* bacterium have exactly the same 64 codon assignments, as stressed in early molecular biology. Apart from mitochondria, only one sense reassignment is known: the 7 serine codons code of certain yeasts (Santos et al. 2011). A handfull of code variants with stop codon reassignments are known, among them are the 4 glutamine codons code of certain ciliates (Hanyu et al. 1986), the 3 cysteine codons code of other ciliates (Meyer et al. 1991), and the 2 tryptophan codons code of *Mycoplasma* bacteria (Yamao et al. 1985). Despite enormous genomics efforts during the last decade, no new non-mitochondrial codon reassignments have emerged.

Several code variants are known to have emerged multiple times, both in the group where they were discovered the first time (e.g., the 4 glutamine codons code in the ciliates, cf. Lozupone et al. 2001) and in other groups (e.g., the 4 glutamine codons code in diplomonads: Keeling and Doolittle 1996; and in certain green algae: Schneider et al. 1989). This shows that certain taxonomic groups (e.g., the ciliates) are prone to reach the rare situation in which codon reassignment can occur. Taken together, this extensive body of work on codon reassignments in current organisms shows that reassignment events are very rare, which implicates that *codon reassignments are very difficult*. This observation contrasts sharply with the ease with which codons are reassigned in origin of the SGC scenarios (e.g., Crick 1968; Higgs 2009).

The functional impact of codon reassignments during the development of the genetic code can be expected to strongly depend on the degree of evolutionary optimization of the proteins in these early systems. On one end of the spectrum, one can envision organisms using statistical proteins with a low level of structure–function optimization. In such a system, the detrimental effects caused by introduction of a substantial number of mutations because of a codon reassignment may be limited. However, it is also possible that the genetic code evolved slowly, and that the functional properties of the proteins in early systems were already quite advanced, with highly optimized amino acid sequences. In that case, most codon reassignments would be expected to have devastating effects on the proteome function.

In recent work, the fitness cost of codon reassignment events was modeled (Higgs 2009). This analysis focused on the presumably rare sites in proteins at which the reassignment will benefit the protein, while the likely damage to protein function caused by the reassignment was not considered. However, a body of recent work regarding the extrapolated amino acid composition of organisms predating the last universal common ancestor (LUCA) has provided support for the presence of a highly optimized proteome (Brooks et al. 2004; Jordan et al. 2005; Fournier and Gogarten 2010).

The analysis of trends in amino acid composition for sets of resurrected ancient proteins offers an interesting approach to explore the proteome of organisms predating the LUCA. A number of independent analyses following different bioinformatics strategies have revealed that amino acids that are often considered to have been added during a late stage of the evolution of the SGC (such as the aromatic amino acids and cysteine) were underrepresented in the LUCA (Brooks et al. 2004; Jordan et al. 2005; Fournier and Gogarten 2010). This result implies that the functions of the proteins in these early systems were already sufficiently evolved to leave detectable traces in the proteins of current organisms. This conclusion suggests that the protein world was already fairly well developed before all 20 amino acids were incorporated. If this inference is correct, then codon reassignment during the evolution of the SGC would have been very difficult.

The above analysis indicates that in current scenarios of the evolution of the SGC, the degree of lethality of nonsense mutations tends to be overestimated, while the difficulties associated with codon reassignments are generally underestimated. We therefore conclude that scenarios in which many unassigned codons persisted throughout most of the evolutionary development of the code should be considered. Such scenarios have the advantage that they do not require codon reassignments. In addition, they allow the developing code to function with a relatively small number of tRNAs, which is attractive in view of the error catastrophe threat in early systems with limited genome replication fidelity.

What properties would be expected for such small tRNA sets during the early stages of the development of the SGC? In general, nonsense suppression relieves the need for the developing translational system to contain tRNAs for the formal assignment of all codons. A second important aspect of the SGC in current organisms is the widespread use of anticodon modifications to achieve the correct assignment of all codons. Did this highly sophisticated system of base modifications develop concomitant with the assignment of codons in the developing code? Or is it biochemically plausible that anticodon modifications were incorporated at a late stage, after the incorporation of all 20 amino acids into the code? In the following, we provide support for the latter possibility, leading to a view in which a small set of tRNAs with unmodified anticodons capable of nonsense suppression allowed the effective functioning of early systems encoding all 20 amino acids. In this scenario, the lack of modifications in the tRNAs specifically regards the three nucleotides in the anticodon. It is entirely possible that other regions of these tRNAs did contain modified bases.

Role of Anticodon Modifications in the SGC

Many tRNA anticodon modifications have been identified. A in the first position of the anticodon is nearly always deaminated to inosine, as already discussed by Crick

(1966). The effect of this is that the tRNA readily recognizes 3 codons instead of 2 (with the complicating factors that the exact effects are different in each codon box (Johansson et al. 2008) and may be taxonomically diverse). U in the first position of the anticodon is nearly always modified, which can occur in various ways. 2-Thiolation results in recognition of both purine-ending codons (e.g., Numata et al. 2006; Phizicky and Hopper 2010). G in the first position of the anticodon can be modified in various different and complex ways, often resulting in increased specificity for the recognition of pyrimidine-ending codons. Modifications in the other positions of the anticodon also occur. A pseudouridine in the second position enlarges the capability of a $\mathrm{tRNA}^{\mathrm{Tyr}}$ to also recognize UAG, which is counteracted by a first position modification (Grosjean et al. 2010). Furthermore, modifications of other residues of the anticodon-loop, and in other parts of the tRNA molecule, can influence the readout properties of the tRNA (see e.g., Beier and Grimm 2001; Johansson et al. 2008). In summary, anticodon modifications in the tRNA molecules of contemporary organisms are widespread, and usually substantially alter the readout properties of the tRNA.

Since anticodon modifications alter the readout properties of tRNAs, the issue of when these tRNA anticodon modifications arose during the development of the SGC is important. Despite the large body of information on the effects of anticodon modifications on the translational properties of tRNA (Takai and Yokoyama 2003; Agris et al. 2007; Johansson et al. 2008; Grosjean et al. 2010), this question has not received much attention in the literature regarding the evolution of the SGC. In the following, we explore the possibility that the machinery to perform anticodon modifications evolved after the 20 amino acids were already incorporated into the developing genetic code.

Wobble Rules for tRNAs with Unmodified Anticodons

When anticodon modifications are taken into account, the wobble rules are complex (see e.g., Agris et al. 2007). However, the wobble behavior of tRNAs with anticodons starting with unmodified G or unmodified C was already described in 1966 (Crick 1966). Regarding the wobble behavior of tRNAs with anticodons starting with unmodified U significant progress has recently been made, as summarized below. Based on this information, we deduce the predicted properties of tRNA sets containing only unmodified anticodons. As discussed below, in this analysis we take the approach that the wobble rules operational during early stages of the evolution of the genetic code were the same as the wobble rules that apply to contemporary organisms.

Wobble Rules and Family Boxes

The boxes of 4 codons in the genetic code table which differ only in the third position and which all encode the same amino acids (e.g., the GCN codons encoding alanine) are referred to as "family boxes". Here, we use the expression "codon box" as a more general concept for collections of 4 codons which only differ in the third position (e.g., the GAN codons are a codon box which is not a family box).

The factor causing the distribution of family boxes in the SGC is a long-standing question in the field (Lagerkvist 1978). Recently, a molecular mechanism was reported explaining this pattern based on hydrogen bonding interactions (Lehmann and Libchaber 2008). When the first two nucleotides of a codon form six hydrogen bonds with the anticodon, the codon box is a family box (codons CCN, CGN, GCN, and GGN). When the first two nucleotides of a codon make only four hydrogen bonds with the anticodon, the codon box is not a family box (codons UUN, UAN, AUN, and AAN). When the first two nucleotides of a codon are able to make five hydrogen bonds with the anticodon, the codon box is a family box only if the middle base of the codon is a pyrimidine (codons UCN, CUN, ACN, and GUN). This is caused by the stabilization of the position of the purine that forms the middle base of the anticodon by a long-range intramolecular hydrogen bond from U_{33} (Lehmann and Libchaber 2008).

For the resulting eight family boxes, the codon–anticodon complex is sufficiently strong to allow the recognition of all three non-cognate nucleotides in the third position of the codon by wobble. Recent experimental results have demonstrated the in vivo importance of this phenomenon in chloroplasts: their ribosomes allow "superwobbling," in which an anticodon with unmodified U in the first position can read all 4 codons in the glycine family box (Rogalski et al. 2008). A recent analysis of the tRNA sets present in bacterial genomes shows that in many bacteria "superwobbling" is widely used (Ran and Higgs 2010).

This information allows the conclusion that a set of 8 tRNAs with the anticodons UGA, UAG, UGG, UCG, UGU, UAC, UGC, and UCC, all starting with unmodified U, suffices to read the 32 codons of the family boxes (Fig. 1).

Wobble Rules and Unmodified-G-Starting Anticodons

The first two codons in a codon box in the SGC always encode the same amino acid. The molecular basis for this pattern is that a single tRNA with an anticodon starting with unmodified G recognizes both Y-ending codons (Crick 1966). The C-ending codon is the cognate codon, and the U-ending codon is recognized by wobble.

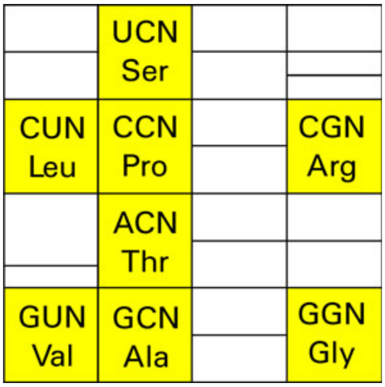


Fig. 1 Coding by tRNAs with anticodons starting with an unmodified U. The codons read by a set of 8 tRNAs with unmodified-U-starting anticodons as based on the wobble rules are indicated. The specific codon sets were selected to reflect the family boxes in the SGC

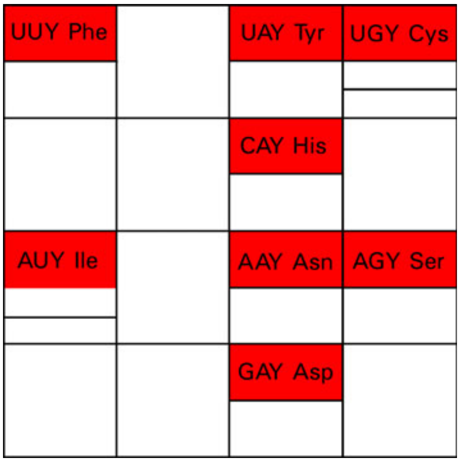


Fig. 2 Coding by tRNAs with anticodons starting with an unmodified G. The codons read by a set of 8 tRNAs with unmodified-G-starting anticodons as based on the wobble rules are indicated. The specific codon sets were selected to reflect the Y-ending codons in the SGC that are not part of family boxes

This pattern implies that a set of 8 tRNAs with the anticodons GAA, GUA, GCA, GUG, GAU, GUU, GCU, and GUC, all starting with unmodified G, suffices to read the 16 Y-ending codons of the codon boxes which are not family boxes (Fig. 2).

Wobble Rules and Unmodified-C-Starting Anticodons

Anticodons starting with unmodified C do not wobble (Crick 1966). Thus, a set of 7 tRNAs with the anticodons CAA, CCA, CUG, CAU, CUU, CCU, and CUC, all starting with unmodified C, suffices to read the seven G-ending sense codons in the codon boxes which are not family boxes (Fig. 3). UAG is a stop codon in the SGC, but might originally have been a universally used pyrrolysine codon (cf. Kavran et al. 2007). For the present purpose, we ignore the UAG codon and focus on the seven G-ending sense codons of the non-family boxes of the SGC.

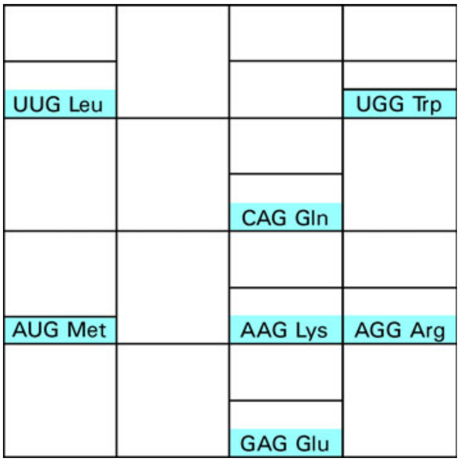


Fig. 3 Coding by tRNAs with anticodons starting with an unmodified C. The codons read by a set of 7 tRNAs with unmodified-C-starting anticodons as based on the wobble rules are indicated. The specific codon sets were selected to reflect the G-ending sense codons of the codon boxes which are not family boxes in the SGC

Considerations Regarding Wobble Rules During the Evolution of the Genetic Code

The degree to which the wobble rules already operated during early stages of the evolution of the genetic code is difficult to ascertain definitively. A specific example is that, based on their work on tRNA sets, Tong and Wong have proposed that the superwobble was a relatively late development that took place in the bacterial domain (Tong and Wong 2004). This would not alter the main conclusions of our manuscript, because the 20 canonical amino acids can be coded, with the canonical assignments, by a small set of codons read by G-starting and C-starting anticodons only. However, the following two arguments provide support for the approach taken here, in which current wobble rules apply to the first stages of the evolution of the SGC. First, the wobble rules are a direct consequence of the physical chemistry of codon–anticodon hydrogen bonding interactions, and thus would be expected to apply as soon as the first codons and anticodons started to interact. Second, two classic regularities in the genetic code are readily interpreted as being direct results of the operation of the wobble rules.

First, the fact that, without exception, both Y-ending codons in a codon box encode the same amino acid is most easily explained as a result of the wobble behavior of unmodified G in the first position of the anticodon. Second the fact that, also without exception, the 32 codons which form the most stable codon–anticodon pairs are organized as family boxes is most easily explained as a result of the superwobble. These regularities are consistent with the view that whenever a single tRNA could read several codons with a reasonable level of efficiency, diversification of the meaning of these codons was blocked. Natural selection favored the appearance of anticodons starting with unmodified U for reading the codons of the 8 family boxes because a minimal number of tRNAs in this way could read a maximal number of codons. The basic structure of the SGC (8 quartets and 8 pairs) can therefore be seen as a reflection of the wobble rules for anticodons starting with unmodified U for the family boxes, and anticodons with unmodified G for the other codon boxes.

These considerations leave ample room for the further development of various aspects of the genetic code, such as those considered by Tong and Wong (2004), since the first organism in which all 20 amino acids were encoded likely was an earlier and more primitive organism than the Last Universal Common Ancestor (LUCA). However, such developments do not affect the main conclusions reached here.

A Set of 20 tRNAs Able to Translate all 20 Amino Acids

From the 23 tRNAs listed above (8 U-starting anticodon tRNAs, 8 G-starting anticodon tRNAs, and 7 C-starting anticodon tRNAs, all with unmodified anticodons), various sets of 20 can be picked such that all 20 canonical amino acids are encoded. This observation leads to the conclusion that no anticodon modification is needed to specifically encode all 20 amino acids. This conclusion is a direct consequence of the wobble rules that has not yet been pointed out in literature, but that is relevant for possible scenarios for the development of the SGC.

Figure 4 compares the coding capabilities of one possible set of 20 tRNAs derived above with that of the SGC. In the above set of 23 tRNAs, Ser, Arg, and Leu are translated by two distinct tRNAs. Here, we describe one specific example of a set of 20 tRNAs encoding all 20 amino acids. A very similar description applies to other 20 tRNA set variants. The main feature of the depicted 20-tRNA code is a striking similarity to the SGC. A few small but systematic deviations are present. First, the three stop codons in the SGC are not assigned in the 20-tRNA code. Second, in the SGC Ile is encoded by three codons, while in the 20-tRNA code this is reduced to two codons. For Lys, Arg, Gln, and Glu, the SGC contains two adjacent codons; in the 20-tRNA code, these are each reduced to a single (G-ending) codon.

The key conclusion is that sets of 20 tRNAs that do not contain anticodon modifications can encode all 20 canonical amino acids in a pattern that is highly similar to that of the SGC. This analysis shows the biochemical feasibility of scenarios for the development of the SGC in which anticodon modifications were introduced only after all 20 canonical amino acids were already incorporated into the developing code. We would like to stress that this finding does not constitute proof for such a relative late development of tRNA anticodon modifications. In addition, it also does not necessarily imply that such a set of 20 tRNAs existed at a specific stage of the evolution of the SGC. For example, for the tRNAs transferring Arg it is entirely possible that two isoacceptors already existed (one reading the codons of the CGN family box, the other reading the AGG codon) before Cys and the aromatic amino acids were added to the amino acid repertoire, and similar considerations apply to Leu and Ser. However, it does demonstrate that this option is biochemically feasible and thus should be considered, since current knowledge does not allow a firm identification of the stage of the development of the SGC at which anticodon modifications were introduced. Similarly, with the above series of tRNAs we do not wish to imply that this sequence of events occurred during the evolution of the SGC. Our conclusion is that small sets of 20–23 tRNAs with unmodified anticodons are capable of encoding all 20 canonical amino acids. In view of the relative simplicity of these tRNA sets and their biochemical plausibility, we propose that scenarios for the evolution of the SGC incorporating such a tRNA set should be considered as a viable possibility.

This view of the evolution of the SGC presents two novel possibilities that (i) nonsense suppression is an important feature of the developing code, and (ii) tRNA anticodon modifications were not introduced until after all 20 amino acids were encoded. In this scenario, eight A-ending codons remained unassigned far longer than generally assumed. It should be noted that this does not mean that these codons

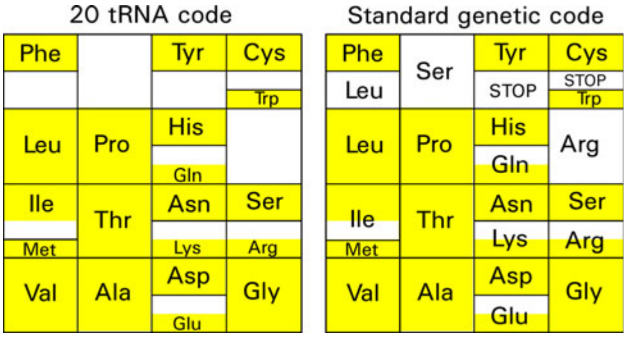


Fig. 4 Comparison of the coding behavior of a set of 20 tRNAs with unmodified anticodons with that of the SGC. In the *left panel* the codons read by a set of 20 tRNAs selected from Figs. 1, 2, and 3 are indicated. This set of 20 is an example in which the UCN (Ser), CGN (Arg), and UUG (Leu) were omitted. To aid visual inspection, all codon sets selected from Figs. 1, 2, and 3 are shaded. Together, this set of 20 tRNAs can translate all 20 canonical amino acids. The *right panel* depicts the SGC with the same pattern of shading

were not used at all in early protein-coding genes, as they could be read by sense suppression. The key attraction of such scenarios for the evolution of the SGC is the relative simplicity of the tRNA set that would allow for the translation of all 20 canonical amino acids. The minimal number of 20-23 tRNAs would be able to perform this translational task in the absence of any machinery for introducing tRNA anticodon modifications. The set of 23 can be reached by a relatively straightforward series of steps involving tRNA gene duplication/anticodon-mutation/mutation in tRNA amino acid charging specificity, and (as discussed further below) can be refined by the subsequent incorporation of tRNA anticodon modifications.

It has been argued that the tRNA set of the archaeon *Methanopyrus kandleri* reflects a relatively early stage of development and resembles that in the LUCA (Tong and Wong 2004; Wong et al. 2007; but see Brochier et al. 2004). In accord with the scenario developed here, the tRNA set of *M. kandleri* resembles the 20 tRNA set depicted in the left panel of Fig. 4. The tRNA set of *M. kandleri* shows a certain "simplicity" (Tong and Wong 2004). In all 8 family boxes of the tRNA set of *M. kandleri* two isoacceptors exist, one in which the anticodon starts with G and another in which the anticodon starts with U. The resemblance with the 20 tRNA set depicted in the left panel of Fig. 4 resides in the fact that these 16 tRNAs could have developed from a primordial set of 8 tRNAs with anticodons starting with unmodified U. In the 5 codon boxes which are not family boxes and which are considered "standard boxes" by Tong and Wong (i.e., the UUN, CAN, AAN, GAN, and AGN codon boxes), the Y-ending codons are read by a tRNA with a G-starting anticodon, and the R-ending codons are read by a tRNA with an U-starting anticodon. This resembles the 20 tRNA set depicted in the left panel of Fig. 4 in the sense that these 10 tRNAs could have developed from a primordial set of 10 tRNAs in which the G-ending codons were read by a tRNA with an anticodon starting with unmodified C, and the A-ending codons were unassigned. The "uniform GU coding" concept of Tong and Wong could in this way be a next step from a more primordial situation in which a more restricted set of codons was read by a set of tRNAs like the one depicted in the left panel of Fig. 4. To make this step, anticodon modification would need to be introduced. An alternative way to look to the tRNA set of *M. kandleri* is to consider the organism as having returned (cf. Brochier et al. 2004) to a simpler set of tRNAs, coming from the more elaborate "uniform GUC coding" (Tong and Wong 2004) predominant in archaea. In that case, *M. kandleri*, like vertebrate mitochondria in a different aspect (superwobbling), used the potential for "simplicity," a potential which was present in the system as a trace of the past. Seen in this light, these simplicities are not entirely new "discoveries," but potentials lurking in the system, because the system had evolved from these simplicities. The resemblance of the tRNA sets of archaea in combination with the proposed resemblance to the LUCA is in excellent agreement with the scenario described here, both when *M. kandleri* is considered as a living fossil, and when *M. kandleri* is seen as a case of return to simpler stage.

Introducing Anticodon Modifications Does Not Require Codon Reassignments

Earlier, we argued that codon reassignments are very rare. Since tRNA modifications alter anticodon readout properties, the introduction of these modifications at a late stage in the development of the genetic code faces the possible problem of highly deleterious changes in the readout of anticodons that are used to encode proteins. In the following, we provide a scenario in which the late introduction of tRNA modifications can proceed without perturbing protein-coding gene sequences.

The introduction of the enzyme that adds a sulfur atom to U-starting anticodons (Numata et al. 2006) also containing U on the second position allows for the appearance of duplicates of the tRNAs with C-starting anticodons for Gln, Lys, and Glu, followed by C-to-U mutations at the first anticodon positions. In this way, the collection of codons specifying, e.g., Lys increases from one (AAG) to two (AAA and AAG). Since in the scenario proposed here these A-ending codons had thus far remained unassigned, no codon reassignments are involved, and no deleterious changes in the existing proteome result from the introduction of the anticodon modification systems. This process is part of a proposed final stage of the process of tRNA repertoire expansion which leads to a situation in which all codons are efficiently and unambiguously encoded. With a pattern of codon assignments as presented in the left panel of Fig. 4 as starting point, anticodon modifications can be introduced without the concomitant introduction of assignment changes of codons used in the protein-coding part of the genome. Similar scenarios can result in the incorporation of the remaining codons in the SGC.

The observation that tRNA anticodon modifications as observed in the SGC can be introduced into the early 20-tRNA set proposed here without deleterious codon reassignments adds to the plausibility of this scenario.

Experimental Evidence for Evolution of Anticodon Modifications After the LUCA: Agmatidine and Lysidine

Tong and Wong (2004) used the analysis of tRNA sets to deduce that the introduction of the inosine modification of

A in the first position of the anticodon was a relatively late evolutionary development. In general, such a relatively late introduction of tRNA anticodon modifications lends support to the scenario presented here. Very recently, detailed biochemical data have become available which imply that modification of the anticodon responsible for decoding the AUA codon occurred after the LUCA.

If the incorporation of tRNA anticodon modifications indeed occurred after all 20 amino acids were incorporated into the developing code, it is possible that this modification system was not yet fully developed in the LUCA. In that case, one would expect differences in the tRNA anticodon modification machinery in the three domains of life. Recent reports on tRNA anticodon modifications in bacteria and archaea indeed provide support for the view that at least some tRNA anticodon modifications were not yet present in the LUCA. Bacteria use the modified nucleoside lysidine to translate AUA as Ile without concomitantly translating AUG as Ile (Muramatsu et al. 1988). Archaea use another modification, agmatidine (Mandal et al. 2010), and another type of modification enzyme (Ikeuchi et al. 2010). This implies that Bacteria and Archaea independently evolved both the modified anticodon nucleoside and the modification enzyme, presumably from a common ancestor in which this anticodon modification was not yet present.

These results indicate that the tRNA anticodon modification machinery is a valuable source of information on the development of the genetic code (Grosjean et al. 2010). The analysis reported here provides a natural framework for understanding this emerging taxonomic diversity in tRNA anticodon modifications: divergent evolution from an earlier translation system lacking these tRNA anticodon modifications. Future studies along these lines should take into account the complicating possibility of inter-domain lateral gene transfer of tRNA anticodon modification enzymes. Inter-domain lateral gene transfer has been documented for the aminoacyl-tRNA synthetases (Woese et al. 2000). This line of research promises to reveal at which stage of the evolution of the SGC the various tRNA anticodon modifications were introduced.

In summary, nonsense suppression can permit the persistence of unassigned codons throughout the evolution of the genetic code, resulting in a small but functional tRNA set, and small sets of tRNAs with unmodified anticodons can efficiently encode all 20 amino acids. These findings allow for a relatively simple early genetic code, specifying *all 20* canonical amino acids, *in the absence of tRNA anticodon modifications*. This proposal appears to be compatible with the main features of influential ideas on the evolution of the SGC (Crick 1968; Wong 2005; Yarus et al. 2005; Di Giulio 2008; Higgs 2009). Future studies on the taxonomic distribution of tRNA anticodon modifications offer a viable avenue to further explore the properties of the genetic code in organisms predating the last common ancestor.

The analysis described here reveals a novel regularity in the genetic code, expanding upon known regularities. In the 1960s it was realized that, without exception, all pairs of Y-ending codons sharing a codon box encode the same amino acid (Crick 1966), and that the middle-U codons are all encoding hydrophobic amino acids, while the middle-C codons are all encoding amino acids of comparable value of polar requirement (Woese et al. 1966). Subsequently, it was pointed out that amino acids encoded by A-starting codons tend to have aspartate as a biosynthetic precursor while amino acids encoded by C-starting codons tend to have glutamate as a biosynthetic precursor (Wong 1975) and that, without exception, all 32 codons which form the most stable codon–anticodon pairs are organized as family boxes (Lagerkvist 1978). Here, we report that no canonical amino acid is encoded by one single A-ending codon only, and that this regularity, in combination with the known wobble behavior of tRNAs with G-starting and C-starting anticodons, has implications for the likely primordial tRNA sets which existed *before* the LUCA.

Acknowledgments The authors gratefully acknowledge the large improvements of the manuscript based on comments of several anonymous referees. PvdG also gratefully acknowledges stimulating discussions with Dave Speijer (University of Amsterdam and Academic Medical Center) and Harry Buhrman (University of Amsterdam and Centrum Wiskunde & Informatica). This study was partly supported by Vici grant 639-023-302 from the Netherlands Organization for Scientific Research (NWO).

References

Agris PF, Vendeix FAP, Graham WD (2007) tRNA's wobble decoding of the genome: 40 years of modification. J Mol Biol 366:1–13

Akashi H (2001) Gene expression and molecular evolution. Curr Opin Gen Dev 11:660–666

Beier H, Grimm M (2001) Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. Nucleic Acids Res 29:4767–4782

Brochier C, Forterre P, Gribaldo S (2004) Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. Genome Biol 5:R17

Brooks DJ, Fresco JR, Singh M (2004) A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. Bioinformatics 20:2251–2257

Chabelskaya S, Kiktev D, Philippe M, Inge-Vechtomov, Zhouravleva G (2004) Nonsense mutations in the essential gene *SUP35* of *Saccharomyces cerevisiae* are non-lethal. Mol Genet Genomics 272:297–307

Crick FHC (1966) Codon-anticodon pairing: the wobble hypothesis. J Mol Biol 19:548–555

Crick FHC (1968) The origin of the genetic code. J Mol Biol 38: 367–379

Di Giulio M (2006) The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the Last Universal Common Ancestor (LUCA). J Theor Biol 240:343–352

Di Giulio M (2008) An extension of the coevolution theory of the origin of the genetic code. Biol Direct 3:37

Doronina VA, Brown JD (2006) When nonsense makes sense and vice versa: noncanonical decoding events at stop codons in eukaryotes. Mol Biol 40:654–663

Eigen M, Schuster P (1977) The hypercycle, a principle of natural self-organization Part A: emergence of the hypercycle. Naturwissenschaften 64:541–565

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 52:759–767

Fournier GP, Gogarten JP (2010) Rooting the ribosomal tree of life. Mol Biol Evol 27:1792–1801

Francis BR (2011) An alternative to the RNA world hypothesis. Trends Evol Biol 3:e2

Fujishima K, Sugahara J, Kikuta K, Hirano R, Sato A, Tomita M, Kanai A (2009) Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea. Proc Natl Acad Sci USA 106:2683–2687

Grosjean H, de Crécy-Lagard V, Marck C (2010) Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Lett 584:252–264

Hanyu N, Kuchino Y, Susumu N, Beier H (1986) Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* tRNAs^{Gln}. EMBO J 5:1307–1311

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

Ikehara K (2002) Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. J Biosci 27:165–186

Ikeuchi Y, Kimura S, Numata T, Nakamura D, Yokogawa T, Ogata T, Wada T, Suzuki T, Suzuki T (2010) Agmatine-conjugated cytidine in a tRNA anticodon is essential for AUA decoding in archaea. Nat Chem Biol 6:277–282

Johansson MJ, Esberg A, Huang B, Bjork GR, Bystrom AS (2008) Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. Mol Cell Biol 28:3301–3312

Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S (2005) A universal trend of amino acid gain and loss in protein evolution. Nature 433:633–638

Kano A, Andachi Y, Ohama T, Osawa S (1993) Unassigned or nonsense codons in *Micrococcus luteus*. J Mol Biol 230:51–56

Kavran JM, Gundllapalli S, O'Donoghue P, Englert M, Söll D, Steitz TA (2007) Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation. Proc Natl Acad Sci USA 104:11268–11273

Keeling PJ, Doolittle WF (1996) A non-canonical genetic code in an early diverging eukaryotic lineage. EMBO J 15:2285–2290

Kramer EB, Farabaugh PJ (2007) The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. RNA 13:87–96

Lagerkvist U (1978) "Two out of three": an alternative method for codon reading. Proc Natl Acad Sci USA 75:1759–1762

Lao NT, Maloney AP, Atkins JF, Kavanagh TA (2009) Versatile dual reporter gene systems for investigating stop codon readthrough in plants. PLoS ONE 4:e7354

Lehman N, Jukes TH (1988) Genetic code development by stop codon takeover. J Theoret Biol 135:203–214

Lehmann J, Libchaber A (2008) Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. RNA 14:1264–1269

Longstaff DG, Blight SK, Zhang L, Green-Church KB, Krzycki JA (2007) In vivo contextual requirements for UAG translation as pyrrolysine. Mol Microbiol 63:229–241

Lozupone CA, Knight RD, Landweber LF (2001) The molecular basis of nuclear genetic code change in ciliates. Curr Biol 11:65–74

Mandal D, Kohrer C, Su D, Russell SP, Krivos K, Castleberry CM, Blum P, Limbach PA, Söll D, RajBhandary UL (2010) Agmatidine, a modified cytidine in the anticodon of archaeal tRNA(Ile), base pairs with adenosine but not with guanosine. Proc Natl Acad Sci USA 107:2872–2877

Meyer F, Schmidt HJ, Plumper E, Hasilik A, Mersmann G, Meyer HE, Engström A, Heckmann K (1991) UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus. Proc Natl Acad Sci USA 88:3758–3761

Muramatsu T, Nishikawa K, Nemoto F, Kuchino Y, Nishimura S, Miyazawa T, Yokoyama S (1988) Codon and amino acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. Nature 336:179–181

Murina OA, Moskalenko SE, Zhouravleva GA (2010) Overexpression of genes encoding tRNA^{Tyr} and tRNA^{Gln} increases the viability of *Saccharomyces cerevisiae* strains with nonsense mutations in the *SUP45* gene. Mol Biol 44:268–276

Numata T, Ikeuchi Y, Fukai S, Suzuki T, Nureki O (2006) Snapshots of tRNA sulphuration via an adenylated intermediate. Nature 442:419–424

Oba T, Andachi Y, Muto A, Osawa S (1991) CGG: An unassigned or nonsense codon in *Mycoplasma capricolum*. Proc Natl Acad Sci USA 88:921–925

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264

Pezo V, Metzgar D, Hendrickson TL, Waas WF, Hazebroucl S, Döring V, Marlière P, Schimmel P, de Crécy-Lagard V (2004) Artificially ambiguous genetic code confers growth yield advantage. Proc Natl Acad Sci USA 101:8593–8597

Phizicky EM, Hopper AK (2010) tRNA biology charges to the front. Genes Dev 24:1832–1860

Ran W, Higgs PG (2010) The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol Biol Evol 27:2129–2140

Randau L, Söll D (2008) Transfer RNA genes in pieces. EMBO Rep 9:623–628

Rodin AS, Szatmáry E, Rodin SN (2011) On origin of genetic code and tRNA before translation. Biol Direct 6:14

Rogalski M, Karcher D, Bock R (2008) Superwobbling facilitates translation with reduced tRNA sets. Nat Struct Biol 15: 192–198

Santos MAS, Gomes AC, Santos MC, Carreto LC, Moura GR (2011) The genetic code of the fungal CTG clade. C R Biol 334: 607–611

Schneider SU, Leible MB, Yang XP (1989) Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of Acetabularia and the occurrence of unusual codon usage. Mol Gen Genet 218:445–452

Sella G, Ardell DH (2006) The coevolution of genes and genetic codes: Crick's Frozen Accident revisited. J Mol Evol 63:297–313

Sengupta S, Yang X, Higgs PG (2007) The mechanisms of codon reassignments in mitochondrial genetic codes. J Mol Evol 64: 662–688

Shaul S, Berel D, Benjamini Y, Graur D (2010) Revisiting the operational code for amino acids: ensemble attributes and their implications. RNA 16:141–153

Sonneborn TM (1965) Degeneracy of the genetic code: extent, nature, and genetic implications. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 377–397

Speyer JF, Lengyel P, Basilio C, Wahba AJ, Gardner RS, Ochoa S (1963) Synthetic polynucleotides and the amino acid code. Cold Spring Harb Symp Quant Biol 28:559–567

Takai K, Yokoyama S (2003) Roles of 5-substituents of tRNA wobble uridines in the recognition of purine-ending codons. Nucleic Acids Res 31:6383–6391

Tong KL, Wong JT (2004) Anticodon and wobble evolution. Gene 333:169–177

Ussery DW, Hallin PF, Lagesen K, Wassenaar TM (2004) Genome Update: tRNAs in sequenced microbial genomes. Microbiology 150:1603–1606

Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552

Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. Proc Natl Acad Sci USA 55:966–974

Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microb Mol Biol Rev 64:202–236

Wong JT (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912

Wong JT (2005) Coevolution theory of the genetic code at age thirty. BioEssays 27:416–425

Wong JT, Chen J, Mat WK, Ng SK, Xue H (2007) Polyphasic evidence delineating the root of life and roots of biological domains. Gene 403:39–52

Yamao F, Muto A, Kawauchi Y, Iwami M, Iwagami S, Azumi Y, Osawa S (1985) UGA is read as tryptophan in *Mycoplasma capricolum*. Proc Natl Acad Sci USA 82:2306–2309

Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: the escaped triplet theory. Ann Rev Biochem 74:179–198