FEBS Letters 584 (2010) 334–341

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Development of the genetic code: Insights from a fungal codon reassignment

Gabriela R. Moura, João A. Paredes, Manuel A.S. Santos *

Department of Biology and CESAM, University of Aveiro, 3810-193 Aveiro, Portugal

ARTICLE INFO

Article history:
Received 29 October 2009
Revised 17 November 2009
Accepted 18 November 2009
Available online 24 November 2009

Edited by Michael Ibba

Keywords:
Genetic code
Protein synthesis
tRNA
Mistranslation
Codon reassignment
Codon usage
Evolution
Candida

ABSTRACT

The high conservation of the genetic code and its fundamental role in genome decoding suggest that its evolution is highly restricted or even frozen. However, various prokaryotic and eukaryotic genetic code alterations, several alternative tRNA-dependent amino acid biosynthesis pathways, regulation of tRNA decoding by diverse nucleoside modifications and recent in vivo incorporation of non-natural amino acids into prokaryotic and eukaryotic proteins, show that the code evolves and is surprisingly flexible. The cellular mechanisms and the proteome buffering capacity that support such evolutionary processes remain unclear. Here we explore the hypothesis that codon misreading and reassignment played fundamental roles in the development of the genetic code and we show how a fungal codon reassignment is enlightening its evolution.
© 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

1. Introduction

Life is based on the extraordinary capacity of cells to translate the nucleic acids information of their genomes into the amino acids information of their proteomes. The genetic code determines how gene words (codons) are translated into protein words (amino acids), highlighting the fundamental role of 20 aminoacyl-tRNA synthetases (aaRSs) in genome decoding [1]. Each aaRS binds and activates a specific amino acid and transfers it to a cognate tRNA, producing aminoacylated tRNAs (aa-tRNAs) [2,3]. The latter read mRNA codons translating the nucleic acids alphabet into the amino acids alphabet through specific ribosome dependent decoding rules [4]. The genetic code is therefore established by specific attachment of amino acids onto tRNA adaptor molecules by aaRSs and by direct reading of mRNA codons by aa-tRNA anticodons in the ribosome. This suggests that reconstruction of the evolutionary pathways that established the genetic code requires deep structural, biochemical, functional and evolutionary knowledge of aaRSs, tRNAs, mRNAs and of the ribosome. To date, many crystal structures of these molecules have been obtained, and detailed biochemical and biophysical characterization of the tRNA aminoacylation and decoding reactions [2,5–7], as well as large scale phylogenetic analysis of the various components of the genetic code have been carried out [8]. Despite these extraordinary advances, the evolution of the genetic code remains an open biological question.

The Frozen Accident Theory proposed by Crick in 1968 postulates that the code is immutable because any alteration to it would be lethal or highly detrimental to life [9]. However, a number of genetic code alterations discovered over the last 40 years indicate that the code has intrinsic flexibility and can evolve (reviewed in [10,11]). We discuss below how these genetic code alterations are enlightening the evolution of the genetic code and we raise the hypothesis that codon reassignment processes played an important role in the code development. The origin of the genetic code, i.e., the origin of tRNAs, aaRSs, the ribosome and the mechanisms of incorporation of the first 10 prebiotic amino acids into the code, which mediated the transition of life from the RNA to the protein worlds, are beyond the scope of this review and will not be addressed. We mention briefly the main theories that have been proposed to explain the origin of the genetic code in order to provide an integrated view of the code evolution.

2. Origin and early evolution of the genetic code

There are three main theories to explain the origin and structure of the genetic code, namely: (i) the Stereochemical Theory, (ii) the Adaptive Theory and (iii) the Coevolution Theory (reviewed in [12]). The Stereochemical Theory posits that codon and amino acid assignments were determined by physicochemical affinities

* Corresponding author. Fax: +351 234426408.
E-mail address: msantos@ua.pt (M.A.S. Santos).

0014-5793/$36.00 © 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2009.11.066

between amino acids and nucleic acids [13,14]. This theory is supported by experimental data arising from selection-amplification of small RNAs (SELEX) which show that at least 8 of the 20 natural amino acids select RNA sequences enriched in cognate codon or anticodon binding motifs [15,16]. Indeed, RNA aptamers selected in the presence of Trp contained Trp CCA anticodons while small RNAs selected in the presence of Ile were enriched in Ile UAU anticodons [17–19], however the statistical significance and the strength of the associations between RNA aptamers and amino acids has been questioned and the Stereochemical Theory requires further validation [20].

The Adaptive Theory postulates that the evolution of the genetic code is mainly driven by the selective forces that minimize the effects of protein synthesis errors, being them from mutational origin or from mRNA misreading [21,22]. The observation that amino acids with similar chemical properties are assigned to similar codons plus statistical and computational evidence for a strong bias towards error minimization pressure in the code provide important support for this theory [12,23,24].

The Coevolution Theory postulates that the structure of the genetic code reflects directly the evolution of amino acid biosynthetic pathways [25]. This theory assumes that the number of amino acids that existed in the prebiotic earth was small (10 or so) and that the other amino acids of the genetic code were derived from the prebiotic ones through biosynthetic processes. The theory is supported by the identification of precursor-product pairs of amino acids and by the discovery of tRNA-dependent biosynthesis of Gln, Asn, Cys and Sec in various prokaryotes and eukaryotes (see below) [26].

The evolutionary scenarios described above, in particular the one proposed by the Coevolution Theory, suggest the existence of three critical moments (steps) in the development of the genetic code (Fig. 1A). An initial step (Phase-1) characterized by the incorporation of the prebiotic amino acids Gly, Ala, Ser, Asp, Glu, Val, Leu, Ile, Pro and Thr. An intermediate step (Phase-2) involving the incorporation of 7 additional amino acids derived from the prebiotic ones through biosynthetic means, namely Phe, Tyr, Arg, His, Trp, Lys and Met. And, a final step (Phase-3) where the five amino acids whose synthesis is tRNA-dependent or is mediated through non-canonical biosynthetic pathways, namely Asn, Gln, Cys, selenocysteine (Sec) and pyrrolysine (Pyl), were incorporated into the genetic code [12,25,26].

We discuss below the mechanistic and structural implications of this stratified evolution of the genetic code under the assumption of the following postulates for the Phase-1 of the code development:

1. The triplet nucleotide nature of codons and the translational machinery were largely established during the incorporation of the first 10 prebiotic amino acids into the genetic code.
2. The basic structure of the tRNA molecule and the codon–anticodon decoding principles were defined.
3. An essential proteome was synthesized with the 10 prebiotic amino acids.

The simultaneous existence of only 10 prebiotic amino acids and 64 codons suggests that some codons were initially unassigned (did not code for any amino acid) or that the 10 prebiotic amino acids were assigned to more than one codon family box (Fig. 1B), as is the case for Leu, Ser and Arg, in extant organisms (Fig. 1D). Indeed, Leu is still encoded by the CUN (N = any nucleotide) codon family plus the UUA/G codons of the UUN codon family box (Fig. 1C and D). The other two codons of the UUN codon family box (UUU/C codons) encode Phe, which was incorporated late into the genetic code [26]. Therefore, during Phase-1 of the code development Leu must have been assigned to both the CUN and the UUN

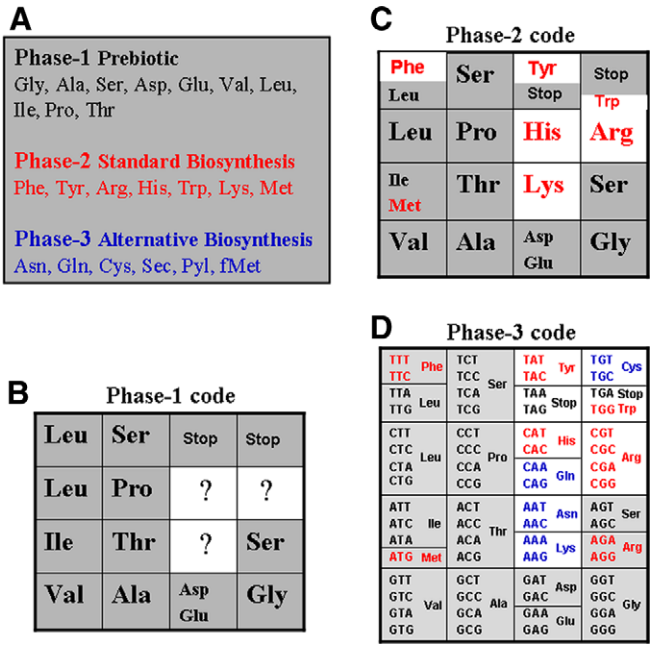

Fig. 1. Scheme outlining the putative evolution of the genetic code. The tables highlight the gradual incorporation of amino acids into the genetic code, according to the Coevolution Theory. (A) The development of the genetic code into three phases follows the evolution of amino acids biosynthetic processes and highlights the requirement of codon reassignments to accommodate new amino acids into the code, beyond the 10 prebiotic ones. (B and C) The tables show the incorporation of the Phase-2 and 3 amino acids into the primordial genetic code. The distribution of codons in the Phase-1 and 2 tables follow that indicated in the Phase-3 table. (D) Genetic code of most extant organisms. The boxes in white colour highlight codon boxes where incorporation of the indicated amino acids involved capture and reassignment of codons. The UUC/U codons were reassigned from Leu to Phe, the AUG codon was reassigned from Ile to Met, the UAU/C codons were reassigned from stop to Tyr and the UGG codon was reassigned from stop to Trp. The full set of codon reassignments was completed with the incorporation of Phase-3 amino acids into the code.

codon family boxes (Fig. 1B). Phe addition to the code required a new (mutant) $tRNA^{Phe}$ to capture the UUU/C codons from Leu. Complete reassignment of these codons to Phe required the loss of the ancestral $tRNA^{Leu}$ that decoded them (Fig. 1C). The same principle of codon capture followed by reassignment can be applied to the incorporation of the other Phase-2 amino acids (Fig. 1C and D). An alternative explanation would be that UUU/UUC, as well as the other codons of split codon families, were initially unassigned and that their late assignment to new amino acids escaped reassignment from one amino acid to another. However, tRNAs with U at the wobble position are able to decode the four codons of codon family boxes and it is likely that these rather than more sophisticated tRNAs bearing nucleoside modifications or expanded sets of tRNA isoacceptors were originally used to decode the 61 sense codons of the genetic code. Furthermore, the pairs of codons of split codon family boxes end with a purine or a pyrimidine and consequently cannot be unassigned simultaneously by genome G + C pressure alone. Therefore, it is unlikely that codon unassignment played a relevant role in the early amino acid assignments.

The Phase-3 amino acids (Asn, Gln, Cys, Sec, Pyl, fMet) are particularly interesting because their alternative biosynthesis suggests that they were incorporated rather late into the genetic code [27,28]. In various bacterial and archaeal species, Asn is still synthesized on a $tRNA^{Asn}$ which is charged with Asp by a non-discriminating AspRS, generating a mischarged Asp-$tRNA^{Asn}$ [29]. A similar mechanism is used in archaea, in most bacteria and in chloroplasts for the synthesis of Gln. In this case, a $tRNA^{Gln}$ is charged

with Glu by a non-discriminating GluRS generating a mischarged Glu-tRNA^{Gln} [30]. Downstream reactions catalysed by the amido transferases Asp-AdT and Glu-AdT convert Asp and Glu into Asn and Gln, respectively [31]. A slightly different mechanism is used in methanogenic archaea to synthesize cysteine (Cys). In this case, a tRNA^{Cys} is initially charged with *O*-phosphoserine (Sep) by the enzyme O-phosphoseryl-tRNA synthase (SepRS), generating the mischarged Sep-tRNA^{Cys}, a Sep-tRNA:Cys-tRNA synthase (SepCysS) then transforms Sep-tRNA^{Cys} into Cys-tRNA^{Cys} [32]. On the other hand, selenocysteine is synthesized on a tRNA$^{[Ser]Sec}$ which is first aminoacylated with serine by a SerRS. In bacteria, a Sec synthetase (SecS) converts Ser into Sec [33], while in archaea and in eukaryotes the seryl moiety is *O*-phosphorylated by the *O*-phosphoseryl-tRNA kinase (PSTK) and the phosphate group is then converted into selenocysteinyl-tRNA^{Sec} (Sec-tRNA^{Sec}) by the Sep-tRNA:Sec-tRNA synthase [34–36]. Sec is inserted at specific UGA codons which are recoded by a mRNA cis-element named SECIS [33,37]. Finally, pyrrolysine is synthesized in *Methanosarcinaceae* using a metabolic pathway involving the PylB, PylC and PylD genes and D-ornithine as precursor. Pyl is charged directly on the tRNA^{Pyl} by a PylRS and is then incorporated into the genetic code in response to specific UAG codons [38,39].

The late incorporation of Asn, Gln, Cys, Sec and Pyl into the genetic code and their assignment to codons that belong to split codon families (Fig. 1D) suggest that their codons were reassigned during Phase-3 of the code development. If so, incorporation of Gln into the genetic code should have required the reassignment of the CAA/G codons which were originally assigned to His (Phase-2 amino acid) (Fig. 1C). Similarly, Asn incorporation into the code should have required capture of the AAU/C codons from Lys (Phase-2 amino acid). Since, codon reassignments have the potential to disrupt the proteome, cause lethality or decrease fitness, as postulated by the Frozen Accident theory, this raises the puzzling question of how did the genetic code expand beyond the 10 prebiotic amino acids? We describe below how extant genetic code reassignments in the mitochondria of vertebrates, invertebrates, fungi and in bacteria and also in the nuclear genome of unicellular eukaryotes are helping to clarify this question.

3. Genetic code alterations

To date, a number of alterations to the standard genetic code have been discovered in various organisms. Stop codons have been reassigned to Trp, Glu, Gln, Cys and Tyr and have also been used to expand the genetic code to Sec and Pyl (reviewed in [10]). The AUA codon has been reassigned from Ile to Met, the AGA/G (Arg) codons have been reassigned to Ser, Gly or to Stop and the AAA (Lys) codon has been reassigned to Asn (Fig. 2). In the mitochondria of several yeast species, including *Saccharomyces cerevisiae*, the CUN (Leu) codon family has been reassigned to Thr. In bacteria, the UGA stop codon has been reassigned to Trp in *Mycoplasma* spp. and *Spiroplasma* spp. [10] and is ambiguously decoded as Stop and Trp in *Bacillus subtilis* [40]. The A + T rich AUA (Ile) and AGA (Arg) codons are unassigned in the G + C rich genome of *Micrococcus* spp. (75% GC), and the CGG (Arg) codon is unassigned in the A + T rich genome of *Mycoplasma* spp. (25% GC) [41,42]. Different species of ciliates also reassigned stop codons. For example, the UGA stop has been reassigned to cysteine in *Euplotes* spp. and is decoded as cysteine or selenocysteine in *Euplotes crassus* by two different UGA decoders, namely the tRNA$_{UCA}^{Sec}$ and the tRNA$_{UCA}^{Cys}$ [43]. The latter tRNA inserts cysteine in the thioredoxin reductase (eTR1) and (eTR2) proteins in response to 6 of the 7 UGAs present in their genes. The other UGA is recoded by a SECIS element present in the eTR1 and eTR2 mRNAs and is decoded by the tRNA$_{UCA}^{Sec}$ which inserts Sec in the respective proteins [43]. In other ciliates, including the model *Tetrahymena thermophila*, in the green algae *Acetabularia* spp. and *Batophora cesthedi* the UAA and UAG stop codons have been reassigned to Gln (reviewed in [10]).

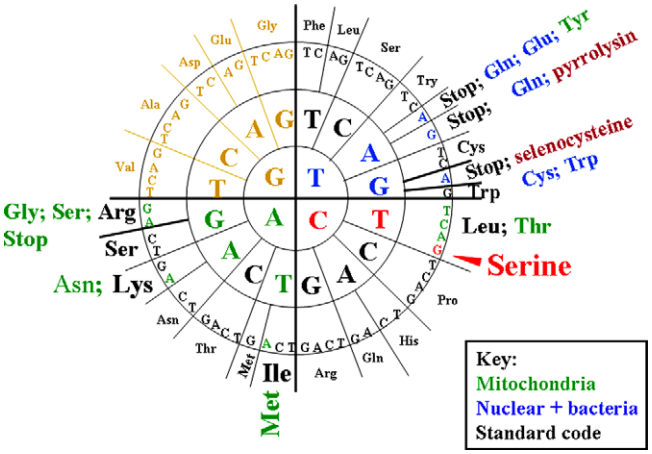

Fig. 2. Diagram showing the genetic code alterations discovered so far. To date 19 genetic code alterations have been discovered in mitochondria (green colour), in bacteria and several unicellular eukaryotes (blue colour). The bacterial and eukaryotic genetic code alterations are a subset of the mitochondrial ones. The diagram indicates that, with exception of the leucine CUN codon family, codon reassignments involve codons of the ANN or UNN types, which suggests that the strength of the first codon–anticodon base pair is important for the evolution of genetic code alterations. The UGA and UAG stop codons are also involved in the expansion of the genetic code to selenocysteine and pyrrolysine, respectively.

The evolution of genetic code alterations can be explained by the Codon Capture and the Ambiguous Intermediate theories [44,45]. The Codon Capture theory postulates that G + C pressure

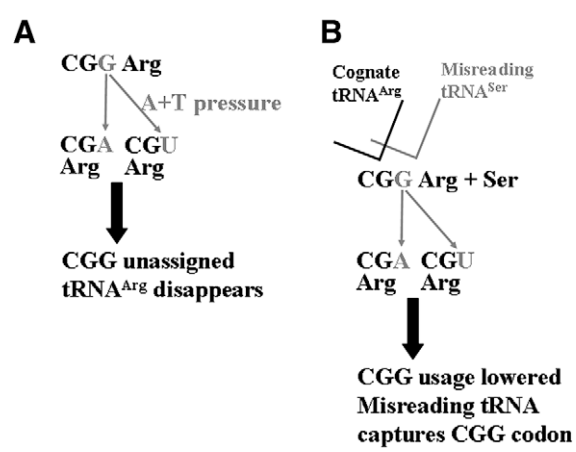

Fig. 3. The molecular mechanisms of codon reassignment. Two main mechanisms explain the unassignment and reassignment of codons. (A) Genome replication biases (A + T or G + C pressure) alter codon usage and may lead to total disappearance of codons from the genome. Codon unassignment leads to disappearance of the respective decoding tRNA due to lack of selective pressure to maintain its gene in the genome. If genome replication biases change overtime the unassigned codons may be gradually reintroduced in genes from mutation of other codons. Sudden reintroduction of missing codons may disrupt mRNA translation due to lack of a cognate tRNA to decode them, however non-cognate tRNAs may misread the newly reintroduced codons and capture them. In such cases the codons can be reassigned to the amino acid family of the non-cognate tRNA(s). This process is neutral since there is no alteration in the proteins primary structure. The diagram highlights the unassignment of the Arg CGG codon in *Mycoplasma* spp. whose genome is A + T rich (75% AT). (B) Alternatively, mutant or wild-type tRNAs may misread codons and introduce codon ambiguity. This decreases fitness and should be eliminated by natural selection. However, in certain cases, it may create advantageous phenotypes that select ambiguity overtime. The negative impact of the misreading phenotype is minimized through gradual reduction of the codon usage.

plays a major role in the evolution of genetic code alterations via its biased effects on codon usage [44,46] (Fig. 3A). The theory posits that codons can disappear from genomes due to strong G + C or A + T replication pressure, and is supported by the unassignment of the AGA, AUA codons in *Microccocus* spp. (75% GC) and the CGG codon in *Mycoplasma* spp. (25% GC). The theory also postulates that such unassigned codons promote reassignment if they reappear in the genome, due to alteration in the DNA replication bias. Their reassignment is mediated by non-cognate tRNAs that misread them [47]. However, if such misreading tRNAs do not exist, the re-emerged codons block mRNA decoding and can be toxic [48], but the theory does not provide a mechanism to circumvent such toxicity. Also, this theory cannot explain reassignment of codons in the absence of DNA replication biases or in cases where the usage of the reassigned codon is favoured by such bias. Examples of such exceptions are the reassignment of the UGA stop codon to Trp, the UAA from Stop to Tyr, the UAU from Ile to Met, the AAA from Lys to Asn and the AGA from Arg to Ser, Gly or Stop in A + T rich mitochondria (Fig. 2). The reassignment of the entire Leu CUN codon family to Thr in fungal mitochondria or the reassignment of the Leu CUG codon to Ser in some fungal species also escape the Codon Capture theory [10,11]. These codon reassignments are better explained by the Ambiguous Intermediate theory, which postulates that misreading tRNAs can capture codons from their cognate tRNAs through a selection-driven process involving gradual increase of misreading efficiency and subsequent disappearance of cognate tRNAs (Fig. 3B) [45,49]. In this process, the codons become ambiguous and their reassignment introduces significant proteome disruption. The theory does not explain how codon ambiguity is selected, but it is strongly supported by CUG reassignment from Leu to Ser in fungi (see below).

The prevalence of genetic code alterations in mitochondria highlights yet another important feature of the evolution of the genetic code, namely that proteome size imposes strong negative pressure on codon reassignment. This has been demonstrated by a large scale comparative genomics study showing a negative correlation between the number of genetic code alterations and the number of genes encoded by mitochondrial genomes [50]. This principle is nicely illustrated in human mitochondria where only 13 of the 900 or so proteins of its proteome are encoded by its genome [51]. Since nuclear encoded proteins are synthesized in the cytoplasm using the standard genetic code and are transported into the mitochondria using a signal peptide translocation system their synthesis escapes the disruption caused by mitochondrial codon reassignments. This is in line with the Genome Minimization hypothesis which postulates that replication speed imposes a strong negative pressure on the mitochondrial genome, leading to selection of small size genomes [52]. In other words, changes in codon usage that relax the pressure to maintain certain tRNA and release factor genes lead to their disappearance and favour codon unassignments and/or reassignments [49]. Therefore, the data available indicate that low codon usage, codon unassignment, genome G + C pressure, genome minimization, small proteome size and tRNA disappearance, are critical players in the evolution of the genetic code. Interestingly, plant mitochondria escape somehow the effects of these evolutionary forces and maintain the standard genetic code.

4. A fungal genetic code alteration

In several species of the genus *Candida* and *Debaryomyces*, the so-called CTG clade, Leu CUG codons are decoded as Ser by a novel seryl-$tRNA_{CAG}$ ($tRNA^{Ser}_{CAG}$) (Fig. 4A) [11,53–57]. Since these species have sophisticated genomes encoding thousands of genes (7000 genes on average) [56] and Leu and Ser are chemically distinct – Leu is hydrophobic and is located in the hydrophobic core of proteins while Ser is polar and is located on the surface of proteins in direct contact with the solvent – the reassignment of CUG codons should have caused maximal protein structural disruption [58]. Moreover, the ancestor of the CTG clade species reassigned between 26 000 and 30 000 CUG codons which existed in approximately 50% of its 7000 or so genes at an average frequency of 1–6 CUGs per gene, indicating that more than half of the proteins of the fungal ancestor had their structure affected [56,59].

The reassignment of those CUG codons was initiated 275 ± 25 million years (My) ago by a mutant Ser tRNA that acquired a 5′-CAG-3′ Leu anticodon ($tRNA^{Ser}_{CAG}$) [59] (Fig. 4A). During the early stages of CUG reassignment, the $tRNA^{Ser}_{CAG}$ competed with a

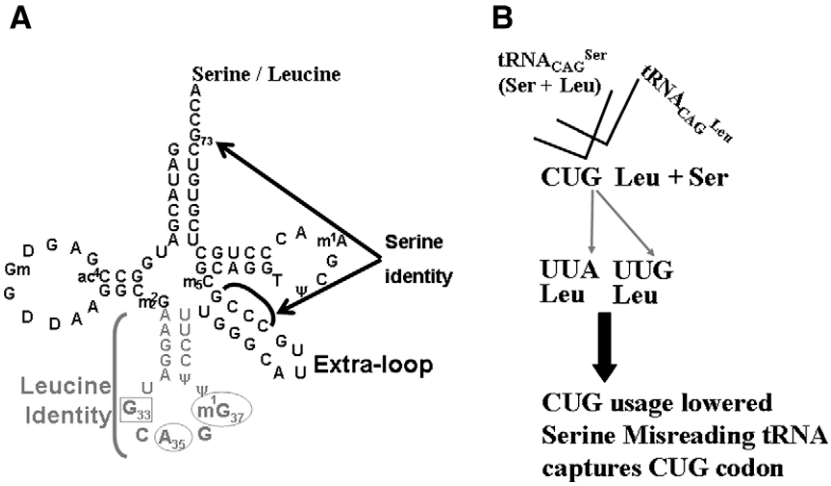

Fig. 4. The evolution of CUG reassignment in the fungal CTG clade. (A) Secondary structure of the *Candida albicans* $tRNA^{Ser}_{CAG}$ which decodes Leu CUG codons as both Ser and Leu. This unique tRNA has identity elements for both the seryl-tRNA synthetase (SerRS) and the leucyl-tRNA synthetase (LeuRS) and is charged with both Ser (major; 97–99%) and Leu (1–3%). Leu mischarging can increase up to 28% without visible effects on growth rate. The SerRS recognizes the $(GC)_3$ helix of the extra-loop and the discriminator base of the tRNA (G_{73}), indicated by the arrow. The LeuRS recognizes A_{35} and m^1G_{37} in the anticodon-loop (circled). The G_{33} distorts the anticodon-arm of the tRNA and lowers the efficiency of leucylation of the tRNA. (B) Misreading of CUG codons due to mischarging of the $tRNA^{Ser}_{CAG}$ and/or competition between this $tRNA^{Ser}_{CAG}$ and the cognate $tRNA^{Leu}_{CAG}$ forced a mutational movement of CUGs to other Leu codons, mainly to UUA and UUG codons. This had a very strong negative impact on CUG usage. The reduction of CUG usage minimized the deleterious effects of CUG ambiguity in the fungal ancestor proteome. For reasons not yet fully understood, the $tRNA^{Leu}_{CAG}$ disappeared from the genome and the $tRNA^{Ser}_{CAG}$ was selected, leading to total CUG reassignment from Leu to Ser.

$tRNA^{Leu}_{CAG}$ for CUG codons and consequently both Leu and Ser were incorporated at CUG positions creating CUG ambiguity (Fig. 4B). For yet unclear reasons the $tRNA^{Leu}_{CAG}$ disappeared and the mutant serine $tRNA^{Ser}_{CAG}$ was selected leading to complete CUG reassignment [60]. This supports the Ambiguous Intermediate theory and suggests that complete codon reassignments can be mediated by low level tRNA misreading followed by codon capture and final elimination of competitor tRNAs. It raises several fundamental questions. For example, how did the CTG clade ancestor cope with CUG misreading? what were the short and long term cellular consequences of CUG misreading? what kind of selective advantages emerged from CUG misreading to allow for selection of the mutant $tRNA^{Ser}_{CAG}$ and elimination of the cognate $tRNA^{Leu}_{CAG}$? and, what was the impact of CUG misreading on the evolution of the CTG clade? Could CUG misreading have driven evolution of the CTG clade?

In order to answer the above questions Ser misincorporation at CUG codons was reconstructed in *S. cerevisiae*. The data were most enlightening. Firstly, CUG mistranslation increased ploidy (up to 4N) and created aneuploid cell lines. It also blocked mating and sexual reproduction and created polyploid asexual cell lineages [61]. Interestingly, recent comparative genomics analysis showed that key components of the mating and meiosis pathways are missing in species of the *Candida* clade. The human pathogen *Candia albicans* uses a parasexual mating system that does not involve nuclear fusion or meiosis, other species apparently do not mate, while a subgroup of species use homothallic and others use heterothallic sexual cycles [56]. Such mating diversity is related to high genetic diversity of the mating locus (MTL) which controls mating in fungi. There is obviously insufficient experimental evidence to unequivocally demonstrate that CUG reassignment is responsible for such mating locus alterations in the CTG clade. However, the data provide convincing evidence for a pivotal role of CUG reassignment in the emergence of a genetic barrier that prevented mating. It also created diploid or aneuploid ancestral lineages that gradually diverged away from each other and from the other fungi. Such genetic barrier could have worked as a speciation mechanism for the CTG clade species [61].

Beyond the alterations mentioned above, CUG mistranslation also altered the expression of molecular chaperones, proteasome activity and carbohydrate metabolism. Cell wall structural proteins were up-regulated while protein synthesis and amino acid metabolism were down regulated. Interestingly, the genome and gene expression alterations induced by Ser misincorporation at CUG positions in *S. cerevisiae* also produced important phenotypic alterations. Colony morphology, cell shape and size were highly heterogeneous, and the resistance to several stress agents, namely nutrient starvation, cadmium and hydrogen peroxide increased, suggesting that the deleterious effects caused by codon mistranslation can be overcome under certain environmental conditions [61].

5. The effects of CUG ambiguity in *C. albicans*

Further insight into the evolution of CUG reassignment in the fungal CTG clade was obtained using the human pathogen *C. albicans* as a model system. The first important discovery was that the $tRNA^{Ser}_{CAG}$ gene evolved from a $tRNA^{Ser}_{CGA}$ gene via insertion of a single adenosine in the second position of the anticodon (A_{35}) of this serine tRNA isoacceptor [62]. This created a hybrid tRNA containing the "body" of a serine tRNA and the anticodon of a leucine tRNA (5′-CAG-3′ anticodon) (Fig. 4A). The identity of this mutant tRNA was maintained because eukaryotic SerRSs do not contact the anticodon of $tRNA^{Ser}$ and, consequently, the above mutation had no effect on the serylation of the tRNA (see above). However, the adenosine insertion (A_{35}) had important implications for the decoding properties of the $tRNA^{Ser}_{CAG}$. Firstly, the leucine 5′-CAG-3′ anticodon of the mutant $tRNA^{Ser}_{CAG}$ was not adapted to the serine tRNA anticodon-arm and this lowered the decoding efficiency of the $tRNA^{Ser}_{CAG}$ [57]. This in turn reduced the toxicity caused by CUG mistranslation because the wild-type $tRNA^{Leu}$ could compete more efficiently with the $tRNA^{Ser}_{CAG}$ for CUG codons. That is, serine misincorporation at CUG positions had a small impact on viability. As the reassignment of the CUG codons progressed, the anticodon-arm of the $tRNA^{Ser}_{CAG}$ was reshaped to increase decoding efficiency. This left unique structural and functional fingerprints in the $tRNA^{Ser}_{CAG}$ of *Candida* species which were unveiled by comparative genomics analysis of the CTG clade $tRNA^{Ser}_{CAG}$ genes [11]. Furthermore, the A_{35} of the 5′-CAG-3′anticodon and the G_{37} of the anticodon-loop (3′ to the anticodon) are directly recognized by the leucyl-tRNA synthetase (LeuRS) (Fig. 4A) [63]. The G_{37} is characteristic of $tRNA^{Leu}$ and is required for LeuRS recognition and also for high decoding efficiency. Therefore, A_{35} and G_{37} allowed for charging of the $tRNA^{Ser}_{CAG}$ with Leu by its cognate LeuRS and did not interfere with serylation of the $tRNA^{Ser}_{CAG}$ [63]. This dual identity of the $tRNA^{Ser}_{CAG}$ maintained CUG ambiguity to the present day.

Another important consequence of CUG misreading was a drastic reduction in the usage of this codon in the ancestor of the CTG clade species [56,59]. Comparative genomics of orthologous genes of CTG and non-CTG clade species demonstrated that the original CUG codons present in the genome of the ancestor of the CTG clade species were eliminated (Fig. 4B) and that the CUGs present in the genomes of extant CTG clade species evolved recently from Ser rather than Leu codons [56,59]. The biological implications of this finding are profound. Firstly, they confirm that codon decoding fidelity is a major selective force in the evolution of codon usage, which is compatible with the Adaptive Theory [21,22]. Secondly, it demonstrates that codons can disappear from genomes due to tRNA misreading rather than biased genome G + C pressure. Thirdly, it shows unequivocally that mutant tRNAs with novel decoding properties can capture codons from unrelated codon families (different amino acids) and that codons can be rapidly reintroduced in the genome from mutation of other codons, even in cases where 2 or 3 simultaneous mutations are required (Fig. 4B).

6. Genetic code ambiguity as a phenotypic diversity generator

The study of CUG reassignment in *C. albicans* has also shown that CUG misreading is an important phenotypic diversity generator (Fig. 5) [64,65]. This is in line with the phenotypes observed in *S. cerevisiae* (see above), however the phenotypic variation observed in *C. albicans* was far more extensive and relevant to adaptation than that observed in *S. cerevisiae*. In the former, CUG misreading activated morphogenetic pathways that led to expression of an array of highly variable cell and colony morphological phenotypes [64,65] (Fig. 5). *C. albicans* colony phenotypes were somewhat unstable and ranged from high frequency white-opaque switching to various types of rough surfaces with visible spontaneous aerial hypha formation. Cell variability was high, in particular size and shape were altered and pseudohyphal and hyphal formation occurred at very high rate. Beyond this, CUG misreading increased secretion of extracellular hydrolases, namely lipases and proteases, and had a strong effect on cell adhesion on solid surfaces and flocculation in liquid media [65]. It is not yet clear how CUG misreading induces expression of such a vast variety of phenotypes, however, gene expression alterations, together with increased ploidy and large chromosomal rearrangements were observed, and it is likely that such phenotypic variability results from the cumulative effects of those alterations. More importantly, CUG misreading generates phenotypes of high adaptive potential

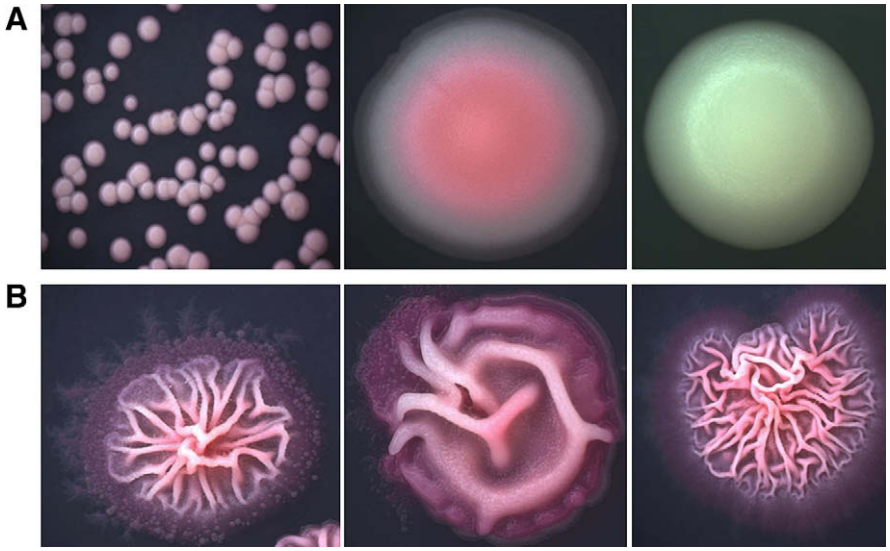

Fig. 5. High level leucine misincorporation at CUG codons generates phenotypic diversity. Artificial increase in the level of Leu misincorporation at CUG codons in *C. albicans* up to 28% did not reduce growth rate, rather it generated extensive morphological variation and hypha formation suggesting that CUG ambiguity is a major generator of phenotypic diversity. (A) Colonies of control *C. albicans* cells and (B) colonies of *C. albicans* expressing a heterologous $tRNA^{Leu}_{CAG}$ that elevated Leu misincorporation at CUG codons.

which may have allowed for selection of the genetic code alteration (Fig. 6).

7. Conclusions

Understanding the evolution of the genetic code is fundamental to elucidate the origin of life. Over the last 60 years remarkable progress has been made on the structural and functional analysis of the various components of the genetic code, namely tRNAs, aaRSs, the ribosome and the overall mRNA translation process. However, the development of the genetic code remains unclear. The hypothesis that codon reassignment played an important role during the early evolution of the code is supported by the gradual incorporation of amino acids into it and by its expansion from 20 to 22 amino acids. The existence of genetic code alterations in extant organisms which evolved from the standard code explains how codons can be reassigned.

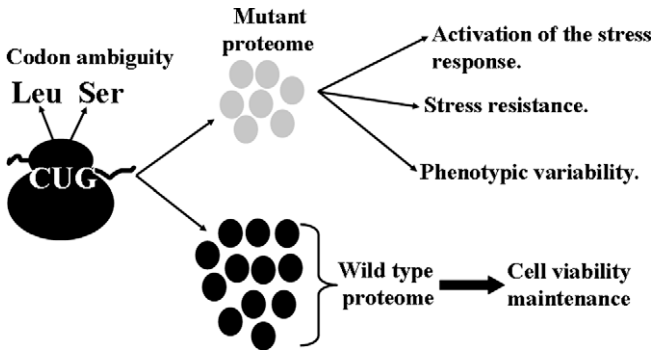

Fig. 6. Selective advantages created by codon misreading. The evolution of codon reassignments through codon ambiguity, as postulated by the Ambiguous Intermediate Theory (see main text), creates the difficulty of disrupting the proteome and reducing fitness. However, this may be overcome if new proteins with new advantageous functions are created through codon misreading. Reconstruction of the fungal CTG clade genetic code alteration in *S. cerevisiae* showed that putative advantages brought about by codon misreading may also be indirect due to activation of the stress response, up-regulation of molecular chaperones and many other stress genes and also due to stress cross protection. These alterations pre-adapted cells to stress allowing them to survive in toxic environments [61]. Those studies also showed that codon mistranslation may generate advantageous phenotypes upon which natural selection can operate. The diagram highlights various consequences of the ambiguous synthesis of mutant proteomes. Part of the aberrant proteome is degraded but the other part is rescued by chaperone systems and remains active.

Considering that most codon reassignments cannot be explained by neutral mechanisms, it is likely that codon ambiguity was a prevalent mechanism in the early evolution of the genetic code, raising the question of how did the code evolve under negative selection. The phenotypic diversity generated in *S. cerevisiae* and *C. albicans* by codon misreading shows that codon ambiguity may have positive outcomes. Also, the incorporation of selenocysteine and pyrrolysine into the genetic code shows how incorporation of new amino acids can expand the proteome and create new functional classes of proteins which bring about selective advantages. Therefore, codon misreading is not necessarily disadvantageous. Since its effects can be largely overcome by proteome novelty and also by indirect advantages such as phenotypic diversity and adaptation to new environmental conditions.

Extrapolation of the cellular consequences of codon misreading and reassignment in living organisms, such as fungi, to the early evolution of the genetic code requires a degree of caution. Clearly the differences between modern cells and the primordial forms of life that existed before the last universal common ancestor (LUCA) are paramount. The variety of extant genetic code alterations also suggest that the forces and mechanisms that mediate the evolution of genetic code alterations are complex and diverse, thus preventing generalizations. However, genome minimization and the role of small proteome size in mitochondrial codon reassignments, suggest that the small size of the pre-LUCA proteomes may have facilitated codon reassignments during Phase-2 and Phase-3 of the code development. In this scenario, the proteome novelty arising from incorporation of new amino acids into the genetic code should have been the major driving force for the expansion of the genetic code up to 22 amino acids, as is amply demonstrated from the advantages generated from the introduction of selenocysteine and pyrrolysine into the genetic code.

Acknowledgements

The Santos group is supported by the Portuguese Foundation for Science and Technology (FCT) through Projects PTDC/BIA-BMC/

64745/2006 and 72251/2006 and also by the European FP7 project MEPHITIS.

References

[1] Woese, C.R., Olsen, G.J., Ibba, M. and Soll, D. (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. 64, 202–236.

[2] Giege, R. (2008) Toward a more complete view of tRNA biology. Nat. Struct. Mol. Biol. 15, 1007–1014.

[3] Giege, R. (2006) The early history of tRNA recognition by aminoacyl-tRNA synthetases. J. Biosci. 31, 477–488.

[4] Marck, C. and Grosjean, H. (2002) TRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA 8, 1189–1232.

[5] Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289, 905–920.

[6] Nissen, P., Hansen, J., Ban, N., Moore, P.B. and Steitz, T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. Science 289, 920–930.

[7] Schmeing, T.M. and Ramakrishnan, V. (2009) What recent ribosome structures have revealed about the mechanism of translation. Nature 461, 1234–1242.

[8] De Pouplana, L.R. and Schimmel, P. (2001) Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends Biochem. Sci. 26, 591–596.

[9] Crick, F.H. (1968) The origin of the genetic code. J. Mol. Biol. 38, 367–379.

[10] Knight, R.D., Freeland, S.J. and Landweber, L.F. (2001) Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. 2, 49–58.

[11] Miranda, I., Silva, R. and Santos, M.A. (2006) Evolution of the genetic code in yeasts. Yeast 23, 203–213.

[12] Koonin, E.V. and Novozhilov, A.S. (2009) Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61, 99–111.

[13] Pelc, S.R. and Welton, M.G. (1966) Stereochemical relationship between coding triplets and amino-acids. Nature 209, 868–870.

[14] Welton, M.G. and Pelc, S.R. (1966) Specificity of the stereochemical relationship between ribonucleic acid-triplets and amino-acids. Nature 209, 870–872.

[15] Yarus, M., Caporaso, J.G. and Knight, R. (2005) The origins of the genetic code: the escaped triplet theory. Annu. Rev. Biochem. 74, 179–198.

[16] Yarus, M., Widmann, J.J. and Knight, R. (2009) RNA-amino acid binding: a stereochemical era for the genetic code. J. Mol. Evol., doi:10.1007/s00239-009-9270-1.

[17] Majerfeld, I. and Yarus, M. (2005) A diminutive and specific RNA binding site for L-tryptophan. Nucleic Acids Res. 33, 5482–5493.

[18] Majerfeld, I. and Yarus, M. (1998) Isoleucine: RNA sites with associated coding sequences. RNA 4, 471–478.

[19] Legiewicz, M. and Yarus, M. (2005) A more complex isoleucine aptamer with a cognate triplet. J. Biol. Chem. 280, 19815–19822.

[20] Ellington, A.D., Khrapov, M. and Shaw, C.A. (2000) The scene of a frozen accident. RNA 6, 485–498.

[21] Epstein, C.J. (1966) Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. Nature 210, 25–28.

[22] Woese, C.R. (1965) On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54, 1546–1552.

[23] Freeland, S.J. and Hurst, L.D. (1998) The genetic code is one in a million. J. Mol. Evol. 47, 238–248.

[24] Novozhilov, A.S., Wolf, Y.I. and Koonin, E.V. (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.

[25] Wong, J.T.F. (1975) A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA 72, 1909–1912.

[26] Wong, J.T. (2005) Coevolution theory of the genetic code at age thirty. Bioessays 27, 416–425.

[27] Ambrogelly, A., Palioura, S. and Soll, D. (2007) Natural expansion of the genetic code. Nat. Chem. Biol. 3, 29–35.

[28] Sheppard, K., Yuan, J., Hohn, M.J., Jester, B., Devine, K.M. and Soll, D. (2008) From one amino acid to another: tRNA-dependent amino acid biosynthesis. Nucleic Acids Res. 36, 1813–1825.

[29] Curnow, A.W., Ibba, M. and Soll, D. (1996) TRNA-dependent asparagine formation. Nature 382, 589–590.

[30] Wilcox, M. (1969) Gamma-glutamyl phosphate attached to glutamine-specific tRNA A precursor of glutaminyl-tRNA in *Bacillus subtilis*. Eur. J. Biochem. 11, 405–412.

[31] Jahn, D., Chen, M.W. and Soll, D. (1991) Purification and functional characterization of glutamate-1-semialdehyde aminotransferase from *Chlamydomonas reinhardtii*. J. Biol. Chem. 266, 161–167.

[32] Borup, B. and Ferry, J.G. (2000) Cysteine biosynthesis in the Archaea: *Methanosarcina thermophila* utilizes O-acetylserine sulfhydrylase. FEMS Microbiol. Lett. 189, 205–210.

[33] Forchhammer, K., Boesmiller, K. and Bock, A. (1991) The function of selenocysteine synthase and SELB in the synthesis and incorporation of selenocysteine. Biochimie 73, 1481–1486.

[34] Carlson, B.A., Xu, X.M., Kryukov, G.V., Rao, M., Berry, M.J., Gladyshev, V.N. and Hatfield, D.L. (2004) Identification and characterization of phosphoseryl-tRNA[Ser]Sec kinase. Proc. Natl. Acad. Sci. USA 101, 12848–12853.

[35] Yuan, J., Palioura, S., Salazar, J.C., Su, D., O'Donoghue, P., Hohn, M.J., Cardoso, A.M., Whitman, W.B. and Soll, D. (2006) RNA-dependent conversion of phosphoserine forms selenocysteine in eukaryotes and archaea. Proc. Natl. Acad. Sci. USA 103, 18923–18927.

[36] Xu, X.M., Carlson, B.A., Mix, H., Zhang, Y., Saira, K., Glass, R.S., Berry, M.J., Gladyshev, V.N. and Hatfield, D.L. (2007) Biosynthesis of selenocysteine on its tRNA in eukaryotes. PLoS Biol. 5, e4.

[37] Bock, A. (2001) Molecular biology. Invading the genetic code. Science 292, 453–454.

[38] Namy, O., Zhou, Y., Gundllapalli, S., Polycarpo, C.R., Denise, A., Rousset, J.P., Soll, D. and Ambrogelly, A. (2007) Adding pyrrolysine to the *Escherichia coli* genetic code. FEBS Lett. 581, 5282–5288.

[39] Hao, B., Gong, W., Ferguson, T.K., James, C.M., Krzycki, J.A. and Chan, M.K. (2002) A new UAG-encoded residue in the structure of a methanogen methyltransferase. Science 296, 1462–1466.

[40] Karow, M.L., Rogers, E.J., Lovett, P.S. and Piggot, P.J. (1998) Suppression of TGA mutations in the *Bacillus subtilis spoIIR* gene by *prfB* mutations. J. Bacteriol. 180, 4166–4170.

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

[42] Kano, A., Ohama, T., Abe, R. and Osawa, S. (1993) Unassigned or nonsense codons in *Micrococcus luteus*. J. Mol. Biol. 230, 51–56.

[43] Turanov, A.A., Lobanov, A.V., Fomenko, D.E., Morrison, H.G., Sogin, M.L., Klobutcher, L.A., Hatfield, D.L. and Gladyshev, V.N. (2009) Genetic code supports targeted insertion of two amino acids by one codon. Science 323, 259–261.

[44] Osawa, S. (1995) Evolution of the Genetic Code, Oxford University Press, New York.

[45] Schultz, D.W. and Yarus, M. (1996) On the malleability in the genetic-code. J. Mol. Evol. 42, 597–601.

[46] Osawa, S., Jukes, T.H., Watanabe, K. and Muto, A. (1992) Recent evidence for evolution of the genetic code. Microbiol. Rev. 5, 229–264.

[47] Osawa, S. and Jukes, T.H. (1995) On codon reassignment. J. Mol. Evol. 41, 247–249.

[48] Kowal, A.K. and Oliver, J.S. (1997) Exploiting unassigned codons in *Micrococcus luteus* for tRNA-based amino acid mutagenesis. Nucleic Acids Res. 25, 4685–4689.

[49] Knight, R.D., Landweber, L.F. and Yarus, M. (2001) How mitochondria redefine the code. J. Mol. Evol. 53, 299–313.

[50] Massey, S.E. and Garey, J.R. (2007) A comparative genomics analysis of codon reassignments reveals a link with mitochondrial proteome size and a mechanism of genetic code change via suppressor tRNAs. J. Mol. Evol. 64, 399–410.

[51] Elstner, M., Andreoli, C., Ahting, U., Tetko, I., Klopstock, T., Meitinger, T. and Prokisch, H. (2008) MitoP2: an integrative tool for the analysis of the mitochondrial proteome. Mol. Biotechnol. 40, 306–315.

[52] Andersson, S.G. and Kurland, C.G. (1998) Reductive evolution of resident genomes. Trends Microbiol. 6, 263–268.

[53] Santos, M.A.S. and Tuite, M.F. (1995) The CUG codon is decoded *in vivo* as serine and not leucine in *Candida albicans*. Nucleic Acids Res. 23, 1481–1486.

[54] Ohama, T., Suzuki, T., Mori, M., Osawa, S., Ueda, T., Watanabe, K. and Nakase, T. (1993) Non-universal decoding of the leucine codon CUG in several *Candida* species. Nucleic Acids Res. 21, 4039–4045.

[55] Yokogawa, T., Suzuki, T., Ueda, T., Mori, M., Ohama, T., Kuchino, Y., Yoshinari, S., Motoki, I., Nishikawa, K., Osawa, S. and Watanabe, K. (1992) Serine tRNA complementary to the nonuniversal serine codon CUG in *Candida cylindracea*: evolutionary implications. Proc. Natl. Acad. Sci. USA 89, 7408–7411.

[56] Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A., Sakthikumar, S., Munro, C.A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J.L., Agrafioti, I., Arnaud, M.B., Bates, S., Brown, A.J., Brunke, S., Costanzo, M.C., Fitzpatrick, D.A., de Groot, P.W., Harris, D., Hoyer, L.L., Hube, B., Klis, F.M., Kodira, C., Lennard, N., Logue, M.E., Martin, R., Neiman, A.M., Nikolaou, E., Quail, M.A., Quinn, J., Santos, M.C., Schmitzberger, F.F., Sherlock, G., Shah, P., Silverstein, K.A., Skrzypek, M.S., Soll, D., Staggs, R., Stansfield, I., Stumpf, M.P., Sudbery, P.E., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N.A., Lorenz, M.C., Birren, B.W., Kellis, M. and Cuomo, C.A. (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature 459, 657–662.

[57] Santos, M.A.S., Keith, G. and Tuite, M.F. (1993) Non-standard translational events in *Candida albicans* mediated by an unusual seryl-tRNA with a 5′-CAG-3′ (leucine) anticodon. EMBO J. 12, 607–616.

[58] Tuite, M.F. and Santos, M.A. (1996) Codon reassignment in *Candida* species: an evolutionary conundrum. Biochimie 78, 993–999.

[59] Massey, S.E., Moura, G., Beltrao, P., Almeida, R., Garey, J.R., Tuite, M.F. and Santos, M.A. (2003) Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in *Candida* spp.. Genome Res. 13, 544–557.

[60] Santos, M.A.S., Moura, G., Massey, S.E. and Tuite, M.F. (2004) Driving change: the evolution of alternative genetic codes. Trends Genet. 20, 95–102.

[61] Silva, R.M., Paredes, J.A., Moura, G.R., Manadas, B., Lima-Costa, T., Rocha, R., Miranda, I., Gomes, A.C., Koerkamp, M.J., Perrot, M., Holstege, F.C., Boucherie, H. and Santos, M.A. (2007) Critical roles for a genetic code alteration in the evolution of the genus *Candida*. EMBO J. 26, 4555–4565.

[62] Suzuki, T., Ueda, T., Yokogawa, T., Nishikawa, K. and Watanabe, K. (1994) Characterization of serine and leucine tRNAs in an asporogenic yeast *Candida*

cylindracea and evolutionary implications of genes for tRNA(Ser)CAG responsible for translation of a non-universal genetic code. Nucleic Acids Res. 22, 115–123.

[63] Suzuki, T., Ueda, T. and Watanabe, K. (1997) The 'polysemous' codon–a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. EMBO J. 16, 1122–1134.

[64] Gomes, A.C., Miranda, I., Silva, R.M., Moura, G.R., Thomas, B., Akoulitchev, A. and Santos, M.A. (2007) A genetic code alteration generates a proteome of high diversity in the human pathogen *Candida albicans*. Genome Biol. 8, R206.

[65] Miranda, I., Rocha, R., Santos, M.C., Mateus, D.D., Moura, G., Carreto, L. and Santos, M.A.S. (2007) A genetic code alteration is a phenotype diversity generator in the human pathogen *Candida albicans*. PLoS ONE 2, e996.