Yeast as a model organism for studying the evolution of non-standard genetic codes

Raquel M. Silva, Isabel Miranda, Gabriela Moura and Manuel A. S. Santos

Abstract

During the last 30 years, a number of alterations to the standard genetic code have been uncovered both in prokaryotes and eukaryotic nuclear and mitochondrial genomes. But, the study of the evolutionary pathways and molecular mechanisms of codon identity redefinition has been largely ignored due to the assumption that non-standard genetic codes can only evolve through neutral evolutionary mechanisms and that they have no functional significance. The recent discovery of a genetic code change in the genus *Candida* that evolved through an ambiguous messenger RNA decoding mechanism is bringing to an abrupt end that naive assumption by showing in a rather dramatic way that genetic code changes have profound physiological and evolutionary consequences for the species that redefine codon identity. In this paper, the recent data on the evolution of the *Candida* genetic code are reviewed and an experimental framework based on forced evolution, molecular genetics and comparative and functional genomics methodologies is put forward for the study of non-standard genetic codes and genetic code ambiguity in general. Additionally, the importance of using *Saccharomyces cerevisiae* as a model organism to elucidate the evolutionary pathway of the *Candida* and other genetic code changes is emphasised.

INTRODUCTION

The discovery of several genetic code changes in bacteria and eukaryotic nuclear and mitochondrial systems (Figure 1) prompted the development of two theories to explain their evolution — namely, the ‘Codon Capture’ and the ‘Ambiguous Intermediate’ theories. The ‘Codon Capture’ theory proposes a neutral mechanism for the evolution of genetic code changes and was developed from the observation that biased GC content has a strong impact on codon usage and, in extreme cases, can drive codons to extinction (Figure 2A,B). This theory is supported by the disappearance of the CGG codon in *Mycoplasma capricolum* (25 per cent of genome is GC) and the AGA and AUA codons in *Micrococcus luteus* (75 per cent of genome is GC) (Figure 2B). The theory postulates that, once a codon disappears from the genome, it can be reintroduced by genetic drift. In this case, a misreading transfer RNA (tRNA) from a non-cognate codon family captures the reintroduced codon, thus changing its identity. At this stage, the existence of a misreading tRNA is critical for identity redefinition, otherwise the codon remains unassigned (ie does not code for any amino acid). The usage of the reintroduced codon having a new identity — due to its capture by a non-cognate misreading tRNA — can then increase over time. This neutral theory excludes any function for genetic code changes and postulates, in a rather elegant and robust manner, that genetic code changes result from biased genome GC content arising from mutations in DNA polymerases or DNA repair systems (Figure 2A). The discovery of additional genetic code changes involving both sense and non-sense codons has questioned the ‘Codon Capture’ theory in that it became apparent that certain codons change identity without vanishing from the
This is supported by the observation that GC content along the genome is not evenly distributed and, consequently, GC pressure at the third codon position — the one that would allow codon disappearance without negative impact on protein sequence — is not the same in all parts of the genome. Additionally, some codons that change identity do so in a manner that is unrelated to genome GC content. So, how do codons change their identity without disappearing from the genome? The most probable hypothesis is that mutant tRNAs with double identity (recognised by more than one aminoacyl–tRNA synthetase) or with expanded decoding properties (altered anticodons or unusual tertiary structures) provide a route for driving genetic code changes through an ambiguous codon decoding mechanism. The existence of natural non-sense suppressor tRNAs — natural or mutant tRNAs that decode one of the three termination codons UAA, UAG or UGA — combined with the observation that stop codons are often reassigned both in eukaryotic cytoplasmic and mitochondrial translation systems and also in several bacteria (Figure 1) provide strong support for the hypothesis that codon identity redefinition can be mediated through mechanisms involving ambiguous codon decoding (Figure 3A,B).

The above observations prompted Yarus and Schultz to propose the ‘Ambiguous Intermediate’ theory for...
the evolution of non-standard genetic codes.9–11 This is a non-neutral theory that postulates that ambiguous codon decoding provides a critical initial step for gradual codon identity change (Figure 3A,B). In this scenario, wild-type or mutant misreading tRNAs have the potential to decode non-cognate codons and, in extreme cases, take over their decoding from their cognate tRNAs; however, complete codon identity redefinition can only be accomplished if the wild-type cognate tRNA decoders disappear from the genome by negative selection.9,10 Intriguingly, this rather puzzling mechanism that destabilises the proteome and reduces fitness has been given strong support by the discovery that certain extant Candida species redefined the identity of the CUG codon from leucine to serine through an ambiguous decoding mechanism.19–22 Remarkably, in C. zeylanoides, the novel Ser–tRNA (Ser–tRNA<sub>CAG</sub>) that mediates this genetic code change is still ambiguous because it is charged with both leucine (3–5 per cent) and serine.

---

**Figure 2**: Codon unassignment and capture through biased genome GC pressure. (A) Strong GC pressure arising from mutations in DNA polymerases and/or DNA repair systems alters codon usage significantly. In extreme cases, biased GC pressure may lead to codon disappearance from the genome. (B) The disappearance of the arginine AGA and the isoleucine AUA codons from the genome of Microccocus luteus (74 per cent GC) and the arginine CGG codon from the genome of Mycoplasma capricolum (25 per cent GC) provides strong support for this mechanism.7 A similar mechanism is apparently in action in the AT-rich mitochondrial genomes which display a rather high genetic code diversity. In these cases, degeneracy of the genetic code works as a buffer permitting codon disappearance without changing protein sequence. Once a codon has disappeared from the genome, it can be reintroduced by genetic drift, allowing for redefinition of the identity of the newly introduced codon if a tRNA from a non-cognate codon family misreads it. The existence of a misreading tRNA (wild-type or mutant) is critical to ensure decoding of the reintroduced codon — if the codon is not decoded, it will stall the ribosome during mRNA translation, blocking protein synthesis. The misreading tRNA can then acquire a cognate anticodon for the ‘new’ codon which will allow for its efficient decoding during mRNA translation and full capture of the reintroduced codon6–8.
(95–97 per cent), thus providing direct evidence for the hypothesis that CUG identity redefinition evolved through an ambiguous decoding mechanism. The functional role of this codon decoding ambiguity still present in some Candida species is not yet understood.

**Figura 3:** Mechanismos de redefinición de identidad de codón a través de ambigüedad del código genético. (A) Redefinición de identidad de codones de sentido. La identidad de los codones de sentido puede redefinirse gradualmente a través de un mecanismo molecular que implica la lectura desambiguizada durante un estado intermedio de redefinición de identidad de codón.9–11 Esta ambigüedad puede surgir a través de la competencia entre una tRNA adecuada y una tRNA no adecuada o a través de la carga incorrecta de una tRNA adecuada por un aminoacil–tRNA sintetasa de una familia de tRNA isoacceptoras diferentes. La ambigüedad del codón es particularmente relevante porque la carga de la tRNA en vivo depende del equilibrio correcto entre aminoacil–tRNA sintetasa y tRNAs.18 La identidad del codón es completamente redefinida con el desaparecimiento de la tRNA que decodificó el codón siguiendo las reglas del código genético estándar. La ilustración muestra la redefinición de la identidad del leucina CUG codón a serina en el género Candida. (B) Redefinición de identidad de codones no-senso. La identidad de los codones no-senso también puede redefinirse gradualmente a través de la lectura desambiguizada involucrando un estado intermedio caracterizado por la competencia entre tRNAs suppressor no-senso y el complejo de factores de liberación, que reconoce los codones parados UAA, UAG o UGA y termina la elongación de la cadena polipeptídica durante la traducción de mRNA. La ilustración muestra la redefinición de la identidad del UAA codón de parada a glutamina que ocurre en varias especies de ciliados, no tetrahymena thermophila. Para el cambio de identidad completo del UAA codón de parada, el anticodón (5'-UUU-3') del tRNA Gln muta creando un anticodón (5'-UUG-3') que es reconocio para el UAA codón. Adicionalmente, eRF1 pierde su capacidad para reconocer el UAA codón de parada (su capacidad para reconocer los otros codones parados no se altera), lo que garantiza una codificación eficiente del UAA codón de parada como glutamina. La misma secuencia de eventos es válida para alterar la identidad de los otros dos codones no-senso, que ocurren en mitochondrias, bacterias y ciliados5,16

**COMPARATIVE AND EVOLUTIONARY GENOMICS**
An important characteristic of genetic code changes is that they evolve from the standard genetic code and do not represent ancestral lineages of alternative genetic codes that existed in the RNA.
Therefore, timing the evolutionary pathway of genetic code alterations is of fundamental importance in understanding better the mechanisms and forces driving change. This assumes particular relevance in the context of genetic code changes that evolve through ambiguous codon decoding as these evolve gradually and introduce significant proteome destabilisation whose relevance can only be appreciated on an evolutionary time-scale.

By using ribosomal RNA (rRNA) and tRNA molecular phylogeny approaches, Massey and colleagues have shed the very first light into the time-scale of the Candida CUG identity redefinition from leucine to serine. These authors have shown that high-level serine—CUG decoding is approximately 171±27 million years (My) old; however, the ser—tRNACAG that decodes it as serine originated at least 272±25 My ago. These molecular phylogeny studies showed that the probability of the change in CUG identity and the ser—tRNACAG appearing at the same time — within 10 My of each other — is extremely low (p < 0.0006). Interestingly, the same molecular phylogeny methodology indicated that the genus Candida and Saccharomyces separated from each other 178±19 My, implying that the CUG codon was highly ambiguous in the ancestor of these yeasts for approximately 100 My (Figure 4).

During the early stages of CUG identity redefinition, decoding ambiguity arose via competition between the leu—tRNACAG and the new ser—tRNACAG for the decoding of the CUG codon at the ribosome A site during mRNA translation. The consequences of this codon decoding ambiguity on an evolutionary time-scale raise two fundamental biological questions:

**Figure 4:** The evolutionary pathway of CUG reassignment in Candida. In Candida, the identity of the leucine—CUG codon has been redefined to serine by a novel serine—tRNA (ser—tRNACAG) that acquired a 5′-CAG-3′ anticodon allowing it to decode the CUG codon as serine. Molecular phylogeny studies using both rRNA and tRNA sequences showed that the ser—tRNACAG appeared approximately 272 My ago while the genera Candida and Saccharomyces separated from each other 170 My. This reassignment pathway suggests that the Saccharomyces ancestor was ambiguous for at least 100 My and that the genus Schizosaccharomyces was not affected by this genetic code change. The ser—tRNACAG was maintained in the lineage that originated the genus Candida but was lost in the lineage leading to the genus Saccharomyces. Complete CUG reassignment required the disappearance of the cognate leu—tRNA.
• Does decoding ambiguity affect the usage of the ambiguous codons?
• What are the physiological and evolutionary consequences of ‘living’ with an unstable proteome on an evolutionary time-scale?

Considering that codon identity redefinition has the potential to affect the entire set of genes of a genome and destabilise all or almost all proteins encoded by that genome, these questions can only be tackled using comparative (Figure 5) and functional genomics approaches (Figures 6–8). For this, the powerful genomics and proteomics tools available to yeast and the availability of the complete genome sequences of several yeast species — namely, S. cerevisiae, C. albicans and Schizosaccharomyces pombe — are of paramount importance.

The impact of ambiguous decoding on

<table>
<thead>
<tr>
<th>Orthologous gene set</th>
<th>Serine codons</th>
<th>Leucine codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. albicans</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. glabrata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. pombe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5:** Comparative genomics permits tracing of the fate of codons undergoing identity change. Evolution of genetic code alterations through ambiguous codon decoding or biased GC pressure (Codon Capture) has a strong impact on the usage of codons that undergo identity change and also on codons belonging to its codon family. This effect was unveiled using comparative genomics methodologies. For this, a set of orthologous genes from C. albicans, S. cerevisiae and S. pombe were aligned and the level of conservation of the CUG codon in the three genomes was determined. The data showed that ambiguous CUG decoding drove most of the ‘old’ CUG codons from the genome of the Candida ancestor (A,B). This is clearly supported by the observation that CUG codons present in C. albicans genes are represented in the other genomes by serine and not leucine codons (A). The reciprocal alignment showed that CUG codons present in S. cerevisiae genes are represented by other leucine — but not serine — codons (mainly the leu—UUG codon) in the C. albicans homologous genes (B). Therefore, the 17,000 CUG codons existent in the genome of C. albicans are unrelated to the CUG codons existent in S. cerevisiae and S. pombe genomes and evolved during the period of redefinition of the identity of the CUG codon, ie during the last 272 My²⁵
the usage of the CUG codon in *C. albicans* was elucidated by aligning a set of orthologous genes from *C. albicans*, *S. cerevisiae* and *S. pombe* and identifying in the alignment the positions occupied by CUG codons (Figure 5). This study showed that CUG double identity (leucine and serine) forced the disappearance of approximately 30,000 CUG codons existent in the *Candida* ancestor; that is, like biased genome GC pressure, codon decoding ambiguity is able to drive codons from the genome. 

Interestingly, a CUG scan of the *C. albicans* genome identified 17,000 CUG codons. This apparent contradiction between CUG disappearance in the ancestor and the existence of 17,000 CUG codons in the genome of *C. albicans* was resolved by reciprocal scoring of CUG codons in the alignment of the orthologous gene set (Figure 5A,B). The refined sequence alignment showed that only 2 per cent of all CUG codons present in the *C. albicans* genome are represented in the other genomes by one of the six leucine codons. The remaining 98 per cent of CUG codons appear in
positions represented by serine codons or codons coding for conserved amino acids of serine (Figure 5A,B); that is, the 30,000 and 26,000 CUG codons present in the *S. cerevisiae* and *S. pombe* genomes, respectively, are not related to the 17,000 CUG codons present in the *C. albicans* genome, indicating that the latter codons, which are decoded as serine and not leucine, evolved during the last 272 My from codons coding for serine rather than leucine. This implies that CUG identity redefinition was mediated through a dynamic molecular mechanism that involved gradual disappearance of ‘old’ ambiguous leu–CUG codons (coding for leucine plus serine) and simultaneous emergence of ‘new’ ambiguous ser–CUG codons arising via mutation of codons that coded for serine or conserved amino acids of serine.25

This novel pathway of codon identity redefinition exposes in a rather dramatic manner an unsuspected flexibility in the genetic code and shows that *C. albicans*, and most likely all other organisms, are well equipped to tolerate the negative impact of proteome disruption caused by codon identity redefinition. Additionally, it implies that the *C. albicans* proteome has been unstable during the last 272±25 My because both the disappearing (old) and emerging (new) codons remained ambiguous since the appearance of the ser–tRNA$_{CAG}$ to the present day.23

Interestingly, the appearance of the novel Ser–tRNA$_{CAG}$ was a major evolutionary force shaping the usage of all six leucine

**Figure 7:** Functional genomics methodologies are powerful tools for studying the impact of codon identity redefinition and genetic code ambiguity on gene expression. Genetic code ambiguity in general and codon identity redefinition in particular are global phenomena whose impact can only be tackled using global gene expression methodologies — namely, proteomics, DNA microarrays and metabolomics. Large-scale synthesis of mutant proteins results in accumulation of misfolded and aggregated proteins and protein degradation. Cells respond to this by overshoohing expression of molecular chaperones and the protein degradation machinery, ie the proteasome in eukaryotic cells. Accumulation of misfolded proteins, induced by ambiguous mRNA decoding, mimics stress conditions, namely heat shock,23 triggering the stress response. The availability of these DNA microarrays and annotated two-dimensional maps for *S. cerevisiae* makes it an ideal model organism for studying both the effect of codon identity redefinition and genetic code ambiguity in general.
codons, i.e. UUA, UUG, CUA, CUG, CUC and CUU, thus showing that redefinition of the identity of a single codon affects the usage of all codons belonging to that codon family.\textsuperscript{25}

FUNCTIONAL GENOMICS OF NON-STANDARD GENETIC CODES

Other important biological questions arising from the evolution of non-standard genetic codes through ambiguous codon decoding relate to the impact of proteome destabilisation on physiology, genetic variability and adaptation (Figure 6). The most relevant questions are:

- How do organisms cope with large-scale synthesis of mutant/aberrant proteins?
- What is the destiny of these proteins?
- What is the relevance of proteome destabilisation on physiology, gene expression, genome stability and, more importantly, the evolution of the species that define codon identity?

Reconstruction of the Candida genetic code change in S. cerevisiae is proving to be an excellent experimental approach for

Figure 8: S. cerevisiae as a model to elucidate the physiological and evolutionary significance of genetic code alterations. The evolutionary significance of non-standard genetic codes can be elucidated by studying the impact of codon identity redefinition at physiological and ecological levels. The reconstruction of the evolutionary pathway of Candida CUG identity redefinition in S. cerevisiae is an excellent experimental model for tackling the high complexity associated with the global effects of ambiguous mRNA decoding. By studying ambiguous yeast cell lines using forced evolution methodologies and phenotype arrays and analysing their metabolome and karyotype, one is bound to provide new insights into the global effects of codon decoding ambiguity and identity redefinition on genome stability and physiology and, ultimately, in understanding at the molecular level the surprising selective advantage displayed by ambiguous cell lines in adaptation to environmental stress conditions and ultimately new ecological niches\textsuperscript{15,28}.
tackling these questions because of the availability of powerful genetics and genomics tools for *S. cerevisiae*. Additionally, the CUG codon is used at low frequency in both yeasts — there are 30,000 and 17,000 CUG codons in *S. cerevisiae* and *C. albicans*, respectively — and the *C. albicans* ser–tRNA<sub>CAG</sub> is correctly processed and charged in *S. cerevisiae*. Such reconstruction, involving transformation of *S. cerevisiae* with the *C. albicans* ser–tRNA<sub>CAG</sub>, creates ambiguous cell lines — the CUG codon is decoded as both leucine by the endogenous leu–tRNA<sub>UAG</sub> and serine by the transformed *C. albicans* ser–tRNA<sub>CAG</sub> — that can be used to study the impact of codon identity redefinition and ambiguity on physiology and adaptation using global genomics approaches, namely phenotypic arrays, DNA microarrays and proteomics (Figures 7 and 8).

Preliminary data validate this experimental approach. For example, phenotypic characterisation of ambiguous *S. cerevisiae* cells showed that, despite having a slower growth rate, they are tolerant to heavy metals, drugs, ethanol, oxidants and sodium chloride. In other words, despite the growth disadvantage observed in rich media, ambiguous cells undergoing codon identity redefinition have a selective advantage under specific environmental conditions — namely, extreme environmental stress conditions. This surprising result suggests that accumulation of aberrant/unfolded proteins synthesised through ambiguous mRNA decoding triggers expression of stress proteins — namely, the molecular chaperones Hsp104 and Hsp70 — that protect ambiguous cells on exposure to severe stress.

Since ambiguous mRNA decoding destabilises a significant part of the proteome, which results in permanent overexpression of stress proteins (proteins normally expressed under stress conditions or in stationary phase), this permanent alteration in gene expression results in pleiotropic phenotypic effects. Therefore, the redefinition of the CUG identity over 272 My is likely to have had a major impact on the evolution of the genus *Candida*. The same principle should apply to all organisms that use non-standard genetic codes. Ongoing studies in the authors’ laboratory aim at elucidating these questions.

**CONCLUSIONS**

Despite the discovery of the first genetic code change more than 30 years ago, it is only now that we are beginning to understand the functional and evolutionary significance of non-standard genetic codes. Molecular phylogeny data clearly show that non-standard genetic codes evolve from the standard genetic code, excluding the hypothesis that they represent molecular relics of alternative codes that may have existed in the RNA world. The full evolutionary implications of changing the genetic code are far from being understood and one is urged to develop methodologies, such as those outlined in this paper, to tackle their study experimentally.

The full diversity of the genetic code is also far from being understood; however, the diversity already uncovered suggests that it is likely that new alterations to the standard code will be discovered in the future. Thus, this has practical implications for genome annotation, gene homology studies and heterologous gene expression. For example, non-standard genetic codes introduce significant noise in protein sequence alignments and, more importantly, genes from organisms with non-standard genetic codes may not express functional proteins in host organisms with the standard genetic code. This is well illustrated by the expression of certain *C. albicans* genes in *S. cerevisiae* and also by the converse expression of heterologous reporter genes in *C. albicans*, namely the commonly used green fluorescent protein (GFP), which is non-functional in the latter.

Finally, reconstruction of the *Candida* genetic code change in its close relative...
S. cerevisiae is proving to be a very powerful tool in dissecting the molecular mechanisms and elucidating the functional implications of this genetic code change for the evolution of the genus Candida. This, and expansion of the genetic code through genetic engineering approaches, clearly shows that the genetic code is flexible and that organisms are well equipped to deal with genetic code ambiguity.\textsuperscript{31–34} More importantly, these pioneering experiments define an experimental framework for shedding new light on, and hopefully elucidating, how and why certain organisms evolve alternative genetic codes.

References

25. Massey, S. E., Moura, G., Beltrão, P. et al. (2003), ‘Comparative evolutionary genomics unveils the molecular mechanisms of


