A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes

Stefanie Mühlhausen,1 Peggy Findeisen,1 Uwe Plessmann,2 Henning Urlaub,2,3 and Martin Kollmar1

1Group Systems Biology of Motor Proteins, Department of NMR-Based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, 37077 Göttingen, Germany; 2Bioanalytical Mass Spectrometry, Max-Planck-Institute for Biophysical Chemistry, 37077 Göttingen, Germany; 3Bioanalytics Group, Department of Clinical Chemistry, University Medical Center Göttingen, 37075 Göttingen, Germany

The genetic code is the cellular translation table for the conversion of nucleotide sequences into amino acid sequences. Changes to the meaning of sense codons would introduce errors into almost every translated message and are expected to be highly detrimental. However, reassignment of single or multiple codons in mitochondria and nuclear genomes, although extremely rare, demonstrates that the code can evolve. Several models for the mechanism of alteration of nuclear genetic codes have been proposed (including "codon capture," "genome streamlining," and "ambiguous intermediate" theories), but with little resolution. Here, we report a novel sense codon reassignment in Pachysolen tannophilus, a yeast related to the Pichiaceae. By generating proteomics data and using tRNA sequence comparisons, we show that Pachysolen translates CUG codons as alanine and not as the more usual leucine. The Pachysolen tRNACAG is an anticodon-mutated tRNA^A^ containing all major alanine tRNA recognition sites. The polyphyly of the CUG-decoding tRNAs in yeasts is best explained by a tRNA loss driven codon reassignment mechanism. Loss of the CUG-tRNA in the ancient yeast is followed by gradual decrease of respective codons and subsequent codon capture by tRNAs whose anticodon is not part of the aminoacyl-tRNA synthetase recognition region. Our hypothesis applies to all nuclear genetic code alterations and provides several testable predictions. We anticipate more codon reassignments to be uncovered in existing and upcoming genome projects.

[Supplemental material is available for this article.]

The genetic code determines the translation of nucleotide sequences into amino acid sequences. It is commonly assumed that, as any change in the code altering the meaning of a codon would introduce errors into almost every translated message, such codon reassignments would be highly detrimental or lethal (Knight et al. 2001a). Therefore, regardless of whether it is optimal or not (Freeland and Hurst 1998; Freeland et al. 2000), the canonical genetic code was long thought to be immutable and was termed a "frozen accident" of history (Crick 1968). However, reassignment of single or multiple codons in mitochondria (Knight et al. 2001b) and nuclear genomes (Lozupone et al. 2001; Miranda et al. 2006) demonstrates that the code can evolve (Knight et al. 2001a; Koonin and Novozhilov 2009; Moura et al. 2010). Several codons have been reassigned in independent lineages. Most nuclear code alterations reported so far are stop codon and CUG-codon reassignments.

Three theories have been proposed to explain reassigments in the genetic code. The codon capture hypothesis states that first a codon and, subsequently, its then meaningless cognate tRNA must disappear from the coding genome before a tRNA with a mutated anticodon appears, changing the meaning of the codon (Osawa and Jukes 1989; Osawa et al. 1992). Genome GC or AT pressure (for reasons often unclear) is thought to cause codon disappearance. In contrast, the ambiguous intermediate hypothesis postulates that either mutant tRNAs, which are charged by more than one aminoaacyl-tRNA synthetase, or misreading tRNAs drive genetic code changes (Schultz and Yarus 1994, 1996). The ambiguous codon decoding leads to a gradual codon identity change that is completed upon loss of the wild-type cognate tRNA. The alternative CUG encoding as serine instead of leucine in Candida and Debaryomyces species (the so-called alternative yeast code [AYCU]) has been strongly promoted as an example for the ambiguous intermediate theory. It is supposed that CUG-codon decoding is ambiguous in many extant Candida species (Tuite and Santos 1996; Suzuki et al. 1997), that the CUG-codon decoding can—at least in part—be converted (Santos et al. 1996; Bezerra et al. 2013), and that the origin of the tRNA^Ser^ has been estimated to precede the separation of the Candida and Saccharomyces genera by ~100 Myr (Massey et al. 2003). However, the ambiguous decoding of the CUG triplet in extant "CTG clade" species is caused by slightly inaccurate charging of the tRNA^Ser^, and not by competing tRNAs.

The genome streamlining hypothesis notes that codon changes are driven by selection to minimize the translation machinery (Andersson and Kurland 1995). This best explains the many codon reassignments and losses in mitochondria. In Saccharomycetales mitochondria, for example, 10 to 25 sense codons are unused, and the CUG codons are usually translated as threonine by a tRNA^Thr^.

© 2016 Mühlhausen et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Corresponding author: mako@nmr.mpibpc.mpg.de

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.200931.115.
which has evolved from a tRNA\textsubscript{Ala} ancestor (Su et al. 2011). In the Enneothecium subbranch, the CUG codons are decoded by alanine, but this modified code did not originate by capture of the CUG codons by an anticondor-mutated tRNA\textsubscript{Ala} but by switching the acceptor stem identity determinants of the Saccharomycetacea tRNA\textsubscript{Thr} from threonine to alanine (Ling et al. 2014).

From the analysis of the conservation of amino acid types and CUG-codon positions in motor and cytoskeletal proteins, we recently showed that these proteins allow us to unambiguously assign the code employed by any given species of yeast (Mühlhausen and Kollmar 2014). Plotting the assigned code onto the yeast phylogeny demonstrates that the AYC\textsubscript{U} to appear to be polythetic in origin, with the “CTG clade” species and Pachysolen tannophilus grouping in different branches. Pachysolen is especially noteworthy with regards to its possible genetic code. Sequence conservation-based analysis indicates that Pachysolen does not encode leucine by CUG and that its CUG-encoded residues are also not present at conserved serine positions. In addition, Pachysolen shares only a few CUG-codon positions with yeasts using the standard genetic code and no CUG-codon positions with “CTG clade” species. This prompted us to determine the identity of the Pachysolen CUG encoding by molecular phylogenetic and proteome analyses.

**Results**

**A new nuclear genetic code in the yeast P. tannophilus**

We determined the tRNA\textsubscript{CAG}\textsubscript{5} in 60 sequenced yeast species (Supplemental Table S1) and aligned them against known tRNA\textsubscript{CAG}\textsubscript{5} and tRNA\textsubscript{Ser}\textsubscript{CAG}\textsubscript{5}. While tRNA\textsubscript{Leu}\textsubscript{CAG}\textsubscript{5} and tRNA\textsubscript{Ser}\textsubscript{CAG}\textsubscript{5} could clearly be classified, the identified Pachysolen tRNA\textsubscript{CAG} sequence was dissimilar to both (Fig. 1A). Comparison to all Candida albicans cytoplasmic tRNAs suggested a close relationship to alanine tRNAs. The alanine identity of the Pachysolen tRNA\textsubscript{CAG} was verified by molecular phylogenetic analyses based on extensive sequence and taxonomic sampling (Fig. 1B; Supplemental Figs. S1–S4; Supplemental Table S1). The sequence identity between the Pachysolen tRNA\textsubscript{CAG} and all identified yeast GCN-decoding tRNAs (752 Saccharomyces and Kollmar 2014). Plotting the assigned code onto the yeast phylogeny demonstrates that the AYC\textsubscript{U} to appear to be polythetic in origin, with the “CTG clade” species and Pachysolen tannophilus grouping in different branches. Pachysolen is especially noteworthy with regards to its possible genetic code. Sequence conservation-based analysis indicates that Pachysolen does not encode leucine by CUG and that its CUG-encoded residues are also not present at conserved serine positions. In addition, Pachysolen shares only a few CUG-codon positions with yeasts using the standard genetic code and no CUG-codon positions with “CTG clade” species. This prompted us to determine the identity of the Pachysolen CUG encoding by molecular phylogenetic and proteome analyses.

CUG codons can unambiguously translate as alanine

To verify the translation of the CUG codons to alanine, we analyzed a cytoplasmic extract of laboratory-grown Pachysolen by high-resolution tandem mass spectrometry (LC-MS/MS), generating approximately 460,000 high-quality mass spectra (Supplemental Fig. S5). Spectra processing resulted in 27,126 nonredundant peptide matches with a median mass measurement error of about 240 parts per billion (Fig. 2A; Supplemental Fig. S6). We identified 5% (2817) of the 5288 predicted proteins with median protein sequence coverage of ~20%. The median numbers of peptides and corresponding peptide spectrum matches (PSMs) identified per protein are six and nine, respectively (Supplemental Fig. S6).

CUG-codon translation affects 4210 (80%) of Pachysolen protein coding genes (Supplemental Fig. S5). Of the 16,824 CUG-codon positions in Pachysolen protein coding genes, 1433 (8.5%) are covered by nonredundant PSMs (3835 PSMs in total, 2.9-fold average coverage) (Fig. 2B–E). Of these unique CUG positions, 907 are covered by PSMs containing sequences with CUG codons fully supported by b- and/or y-type fragment ions. Almost all of these (97.2%) contained the CUG codons translated as alanine (Fig. 2F; Supplemental Fig. S7).

The remaining rare incidences can be classified into two groups. For eight of the fully supported CUG-codon positions, PSMs with ambiguously translated CUGs (alanine plus another amino acid) were found (Fig. 2E). For an additional 18 CUG-codon positions, PSMs were found that translate CUG as an amino acid other than alanine (Fig. 2E). Both of the above minority cases might be due to differences between our and the sequenced Pachysolen strain (Liu et al. 2012), might be due to transcription and translation errors, or might represent spurious mischarging of the Pachysolen tRNA\textsubscript{Ala}\textsubscript{CAG}. For comparison, we analyzed the unambiguously decoded AUG codon and found similar numbers of differences (Supplemental Fig. S8). Accordingly, the CUG codon is as unambiguous as the unambiguous, related codon AUG. Substantial mischarging of Candida tRNA\textsubscript{Ser}\textsubscript{CAG} by leucines has been shown in vitro and in vivo (Suzuki et al. 1997), but other potential mischarging have never been analyzed.

History of the CUG-decoding tRNA

To reconstruct the history and origin of all yeast tRNA\textsubscript{CAG}, we performed in-depth phylogenetic analyses of all UCN-decoding tRNAs (serine), GCN-decoding tRNAs (alanine), and CUN-decoding tRNAs (leucine) (Supplemental Figs. S9–S11). These analyses support our previous assumption (Mühlhausen and Kollmar
Figure 1. The *Pachysolen* CUG-tRNA is an Ala-tRNA. (A) Secondary structures of the *Pachysolen* tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Ala}}, the *Candida* tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Ser}}\textsubscript{\textit{CAG}}\textsubscript{\textit{Ser}}, and the *Yarrowia* tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Leu}}\textsubscript{\textit{CAG}}\textsubscript{\textit{Leu}}. The consensus *Saccharomycetales* tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Ala}} (based on 752 sequences), tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Ser}}\textsubscript{\textit{CAG}} (748 sequences), and tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Leu}}\textsubscript{\textit{CAG}} (300 sequences) are shown for comparison. The discriminator base “N73” is denoted by the corresponding nucleotide number of the respective tRNA gene. It is obvious that the *Pachysolen* tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Leu}}\textsubscript{\textit{CAG}} and the *Candida* tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Ser}}\textsubscript{\textit{CAG}} share more consensus elements with the alanine and serine tRNAs, respectively, than with the leucine tRNAs. tRNA identity determinants are indicated by red boxes and dashed lines. (B) Unrooted phylogenetic network of 172 tRNA sequences generated using the neighbor-net method as implemented in SplitsTree v4.1.3.1. Methionine, isoleucine, arginine, and threonine tRNAs were included as outgroup. tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Leu}}\textsubscript{\textit{CAG}}, tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Ser}}\textsubscript{\textit{CAG}}, and the *Pachysolen* tRNA\textsubscript{\textit{CAG}}\textsubscript{\textit{Leu}}\textsubscript{\textit{CAG}} are highlighted in dark green, blue, and red, respectively.
2014) that all “CTG clade” species’ tRNA_{CAG} are serine-tRNAs (Supplemental Fig. S9) and that all Saccharomycetales, Phaffiomycetales, and “Pichiaceae” species’ tRNA_{CAG} are leucine-tRNAs (Fig. 3; Supplemental Fig. S10). Monophyly of the tRNA_{CAG} indicates a common origin in the ancestor of the “CTG clade.”

The UCN-decoding tRNAs split into two major subbranches: a group of UCG-decoding tRNAs to which the tRNA_{CAG} belong, and a group of UCU-decoding tRNAs. This supports the prior notion (Massey et al. 2003) that the ancient tRNA_{CAG} originated from an UCG-tRNA by insertion of an A into the anticodon rather than from an UCU-tRNA by insertion of a C directly before the anticodon.

In contrast to the monophyletic tRNA_{CAG}, the tRNA_{CAG} are polyphyletic, and many yeasts contain multiple tRNA_{CAG} derived from gene duplication of cognate and isoacceptor tRNAs (Fig. 3). For example, Yarrowia lipolytica contains 13 tRNA_{CAG}, of which two were most probably derived from either an ancestral tRNA_{CAG} or tRNA_{Leu} and 11 were derived by gene duplication and mutation from a tRNA_{CAG}. The Phaffiomycetales, Kluveromycetes, Lachancea khyveri, and Eremothecium tRNA_{CAG} were derived from an ancestral tRNA_{Leu} of the Lachancea thermotolerans and Lachancea waltii tRNA_{CAG}. The Pachysolen tRNA_{CAG}, most probably originated by duplication of a tRNA_{CAG} (Supplemental Fig. S11), followed by mutation to the isoacceptor tRNA_{CAG} and, finally, insertion of an A into the anticodon similar to the origin of the ancient tRNA_{CAG}.
The CUN family box tRNAs are split into two major groups: a group of tRNALeu\textsubscript{UAG} and a group of tRNALeu\textsubscript{AAG}. The tRNACAGs are only present in Saccharomycetaceae and have been derived from an ancestral tRNALeu\textsubscript{AAG} (Fig. 3). The tRNALeu\textsubscript{UAG} form two subgroups, which most probably originated after the split of Y. lipolytica. One of the subgroups is restricted to Yarrowia, Pachysolen, and the "Pichiaceae"; the other is common to all yeasts and contains the Saccharomyces tRNALeu\textsubscript{UAG}. The latter is unique because it is the only Saccharomyces cerevisiae tRNA with an unmodified uridine in the wobble position of the anticodon triplet (Randerath et al. 1979; Johansson and Byström 2005). Unmodified U34s have otherwise only been found in mitochondria, chloroplasts, and Mycoplasma species. The Saccharomyces tRNALeu\textsubscript{UAG} is also unique as it is able to translate all six leucine codons (Weissenbach et al. 1977).

All sequenced yeasts have distinct NNA-decoding and NNG-decoding tRNAs for all respective two-codon families and most four-codon families. Modifications at U34 in NNA-decoding tRNAs enable these to also read G-ending codons, and accordingly, many NNA-decoding tRNAs are competing with NNG-decoding tRNAs when reading NNG codons (Johansson et al. 2008). However, it is also known that some U34-modified tRNAs, such as 5′-methoxy carbonylmethyl-2-thiouridine (mcm5s2)–modified Gln- and Glu-decoding tRNAs in most if not all prokaryotic and eukaryotic species, are not able to read G-ending codons in vivo (Johansson et al. 2008; Rezgui et al. 2013).

Figure 3. The history of CUG-codon reassignment in yeasts. Decoding and origin of the tRNACAGs were plotted onto a time-calibrated yeast phylogeny adapted from Mühlhausen and Kollmar (2014). Colored lines denote the different tRNACAG types, with green representing tRNACAG\textsubscript{CUC}, blue denoting the "CTG clade" tRNACAG\textsubscript{CUG}, red marking the Pachysolen tRNACAG\textsubscript{CAG}, and black denoting the absence of a tRNACAG in the respective branch. The schemes with the four circles represent the reconstructed CUN tRNA family at each branch, with filled and empty circles denoting the presence and absence, respectively, of the CUN family box tRNAs. Arrows indicate tRNA gene duplications followed by anticodon mutations, except for the tRNACUC\textsubscript{UAG} duplication in the ancestor of the SP–PC branches that resulted in orthologous tRNACUC\textsubscript{UAG} subtypes. Branches with identical decoding schemes have been collapsed. The percentage of CUG codons per species/taxon were derived from Mühlhausen and Kollmar (2014). "Pichiaceae," as shown here, is not a commonly agreed taxon. Because all species of the respective branch in the presented tree have at least one synonymous name starting with "Pichia," we termed the entire branch "Pichiaceae" for simplicity. The scheme at the bottom shows how the four CUN codons can be decoded by several combinations of isoacceptor tRNAs using standard and wobble base-pairing.
tRNA\textsubscript{Leu\textsubscript{CAG}} also contained a modified U34. Given the unambiguous decoding of the CUG codons by the tRNA\textsubscript{Leu\textsubscript{CAG}} in Pachysolen, the Pachysolen tRNA\textsubscript{Leu\textsubscript{CAG}} may be presumed to contain a U34 modification, such as mcm\textsuperscript{5}\textsubscript{U} in Glu-decoding tRNAs or pseudouridine in Ile-decoding tRNA\textsubscript{Ala\textsubscript{CAG}} to prevent competitive decoding of the CUG codons.

**Discussion**

How did such diversity in tRNA origin and CUG-codon decoding (leucine vs. serine vs. alanine) evolve? While several testable predictions of each of the codon reassignment hypotheses have been summarized (Knight et al. 2001a), these predictions, however, did not include decoding of CUG codons by an Ala-tRNA\textsubscript{CAG}. Does the presence of the Pachysolen tRNA\textsubscript{Ala\textsubscript{CAG}} still fit into the existing models?

**The yeast CUG-codon reassignments do not accord with the codon capture theory**

According to the codon capture theory, CUG codons need to have disappeared before their reassignment at the split of the “CTG clade” (Fig. 4; Supplemental Fig. S12). The time frame for codon disappearance is determined by the split of the Saccharomyces/P. halophila clade branches (hereafter called SP and PC branches, respectively). Codon disappearance must have occurred either in the very short time frame between the split of the SP and PC branches and the divergence of the “CTG clade” (Supplemental Fig. S12) or before the SP-PC split (Fig. 4), which would then necessarily include reappearance of the tRNA\textsubscript{Leu\textsubscript{CAG}} and CUG codons at original positions in the SP branch. Subsequently, the Ser-, Ala-, and Leu-tRNA\textsubscript{CAG}S could have captured the still unassigned CUG codon in the “CTG clade,” Pachysolen, and the “Pichiaceae” branches independently from each other. However, disappearance of an entire codon from a genome by neutral mutations is extremely unlikely to happen in such a short time. It is similarly unlikely that AT/GC bias, the main force driving codon reassignment according to the codon capture theory, caused only one codon to disappear.

**The yeast CUG-codon reassignments do not accord with the ambiguous intermediate theory**

The ambiguous intermediate theory assumes the simultaneous assignment of a codon to two tRNAs. In case of the yeasts, it was proposed that both the cognate tRNA\textsubscript{Ser\textsubscript{CAG}} and the new tRNA\textsubscript{CAG} were present before the split of the SP and PC branches (Massey et al. 2003). In order to fit the tRNA\textsubscript{Ala\textsubscript{CAG}} into this scenario, one has to assume either the presence of the tRNA\textsubscript{CAG} at the same time or two successive ambiguous intermediate states. The presence of three tRNA\textsubscript{CAG}S at the same time would give rise to an even more ambiguous decoding and seems to be highly unlikely. More likely seems the scenario including two successive ambiguous intermediates (Fig. 4; Supplemental Fig. S12). In this scenario, a time span with another ambiguous CUG-codon decoding (tRNA\textsubscript{Leu\textsubscript{CAG}} competing with tRNA\textsubscript{Leu\textsubscript{CAG}}S) would have followed the split of the “CTG clade” or would have started in the Pachysolen branch. If the ambiguous intermediate theory were true, tRNA\textsubscript{Ser\textsubscript{CAG}}\textsubscript{5} and tRNA\textsubscript{Leu\textsubscript{CAG}}\textsubscript{5} in extant species should have been derived from the same ancestral Ser- and Leu-tRNAs. While this is true for the tRNA\textsubscript{Ser\textsubscript{CAG}}\textsubscript{5} of the “CTG clade,” the tRNA\textsubscript{Leu\textsubscript{CAG}}\textsubscript{5} are polyphyletic and clearly have different origins (Fig. 3; Supplemental Figs. S9, S10). Accordingly, the ambiguous intermediate theory would at least require—in case of the more likely separate ambiguous intermediate events—the independent loss of the tRNA\textsubscript{Leu\textsubscript{CAG}}\textsubscript{5} in the SP and the Pachysolen/“Pichiaceae” branches (PP branches) and the independent loss of the tRNA\textsubscript{Leu\textsubscript{CAG}}\textsubscript{5} in multiple branches, including the branches with altered decoding. The polyphyly of the tRNA\textsubscript{Leu\textsubscript{CAG}}\textsubscript{5} is not explained in either of the scenarios.

The scenario of two successive CUG-codon reassignments is further weakened by the frequent nature of the CUG codon. Although phylogenetic mapping of variant codes has shown that the same codons have independently been reassigned both in nuclear genomes and in mitochondrial genomes, those reassignments only affected rare codons (Knight et al. 2001a). However, the CUG codon is not a rare codon, and it seems extremely unlikely that the same frequently used sense codon became ambiguous in two subbranches of the same taxon within a very short time in independent events, as it would be required for the two CUG-codon reassignments in yeasts according to the ambiguous intermediate theory. A preference for further ambiguous intermediate events because of CUG-codon usage reduction is similarly unlikely. If this preference would exist, many more CUG-codon reassignment events would be expected in all branches of the Saccharomyces.

In the other scenario assuming the simultaneous ambiguous CUG decoding by three tRNA\textsubscript{CAG}S, one would expect to observe CUG-codon positions conserved between species decoding CUG as leucine, serine, and alanine. A comparison of cytoskeletal and motor protein sequences from 60 yeast species, however, showed that CUG codons from “CTG-clade” species were not found even at moderately (≥50%) conserved leucine positions with CUG codons and that CUG codons from yeasts using standard codon usage were not found at moderately conserved serine positions with CUG codons (Mühlhausen and Kollmar 2014). *Pachysolen* did not share any CUG-codon positions with the “CTG-clade” species but did share some CUG-codon positions with CUG codons from yeasts using standard codon usage (Mühlhausen and Kollmar 2014). However, these shared positions were at alignment positions with low sequence conservation. The few shared CUG codons at nonconserved sequence positions probably do not represent original CUG codons but more likely resulted from random reassignments/mutations. Thus, *Pachysolen* and “CTG-clade” species have independently reassigned the CUG codons: *Pachysolen* mainly at conserved alanine positions and “CTG-clade” species mainly at conserved serine positions. These findings contradict a scenario of simultaneous ambiguous CUG decoding by three tRNA\textsubscript{CAG}S.

**The tRNA loss driven codon reassignment mechanism presents a parsimonious explanation**

The observed polyphyly of the tRNA\textsubscript{CAG}S is best described by a tRNA loss driven codon reassignment process as follows (Fig. 4): The ancestor of the SP and PC clades lost its tRNA\textsubscript{Leu\textsubscript{CAG}} by gene loss or mutation. This loss was accompanied or followed by the appearance of the ancestral U34-modified tRNA\textsubscript{Leu\textsubscript{CAG}} by gene duplication of a “normal” tRNA\textsubscript{Leu\textsubscript{CAG}}. This new tRNA\textsubscript{Leu\textsubscript{CAG}}S evolved the characteristic U34 encoding tRNA\textsubscript{Leu\textsubscript{CAG}}S after its appearance or later in the SP branch. The loss of the tRNA\textsubscript{Leu\textsubscript{CAG}}S might not have caused considerable viability issues because CUG codons could still be decoded as leucine by, although probably inefficiently, wobble base-pairing (Crick 1966) involving the ancestral U34-modified tRNA\textsubscript{Leu\textsubscript{CAG}}. This process was presumably supported by the doubling in tRNA copy number.

Subsequent to the loss of the original tRNA, reduction in translational fidelity (Gromadski et al. 2006) might have caused
**Figure 4.** The mechanism of CUG-codon reassignment. The scheme contrasts the presence of tRNA$^{Sac}_{CAG}$ and evolution of CUG codons according to the tRNA loss driven codon reassignment hypothesis with assumptions based on the codon capture and ambiguous intermediate theories. Only the most probable sequence of events is shown for each hypothesis. Alternative but less likely scenarios are presented in Supplemental Figure S12. Mixed models, for example, ambiguous decoding of the CUG codons in the ancestor of the SP and CP clades followed by loss of the tRNA$^{Sac}_{CAG}$ in the ancient Pachysolen and capture by the tRNA$^{Sac}_{CAG}$, seem extremely unlikely and are not shown. The codon capture theory distinguishes from the tRNA loss driven codon reassignment hypothesis mainly by the order of events (tRNA loss after or before the reduction of CUG codons, respectively), the degree of CUG-codon loss (loss of all codons vs. a reduction of CUG-codon usage, respectively), and the cause of reduction of CUG codons (AT pressure vs. tRNA loss and decreased decoding fidelity, respectively). In contrast to the ambiguous intermediate hypothesis, the tRNA loss driven codon reassignment theory does not assume any ambiguous decoding of the CUG codons. In addition, capture of the CUG codon by different tRNAs is an elemental characteristic of the tRNA loss driven codon reassignment theory and does not need to be split into independent events.

The number of CUG codons to gradually decrease. Compared with the most ancient yeast species *Yarrowia*, all analyzed yeasts have considerably decreased numbers of CUG codons (Fig. 3; Mühlhausen and Kollmar 2014). Even the highly GC-rich genomes of *Ogataea parapolymorpha* (Ravin et al. 2013) and *Kataashiella capsulata* (Morales et al. 2013) have fewer CUG codons, suggesting a general strong reduction of CUG-codon usage after the split of *Yarrowia*.

Many subbranches of the Saccharomycetales independently lost their tRNA$^{Sac}_{CAG}$, which was not accompanied by further
CUG-codon losses (Fig. 3). These subbranches demonstrate that the CUG codons can efficiently be translated by noncognate tRNAs, most probably the U34-unmodified tRNA\textsubscript{Leu}\text{AUG}. The tRNA\textsubscript{AUG}\text{ independent CUG decoding might have developed into full functionality during the time from the split of \textit{Yarrowia} to the divergence of the SP branch, and until achievement of full functionality, CUG-codon usage considerably decreased. CUG-codon reduction most probably happened by transitions and transversions to other leucine codons before the divergence of the SP branch. Within protein coding regions, codon changes within the CTN family box, and also between CTG and TTG, are extremely frequent. Indeed, even closely related \textit{Saccharomyces} species have few conserved leucine codons (Mühlhausen and Kollmar 2014).

Subsequent to the loss of the ancient tRNA\textsubscript{Leu}\text{LACG} and the reduction of CUG-codon usage, the unassigned CUG codon became free to be captured by other tRNAs with mutated anticodons. Capturing by isoacceptor tRNAs is most straightforward and happened in the ancestor of the SP branch and within the “Pichiaceae” species independently from each other as shown by the polyphyly of the tRNA\textsubscript{Leu}\text{LACG} (Fig. 3; Supplemental Fig. S8). In addition to isoacceptor tRNAs, only tRNAs whose anticodon is not part of the identity determinants of their respective aminoacyl-tRNA synthetases (aaRSs) could also capture the free CUG codon. In the “CTG clade,” a mutated Ser-tRNA captured the CUG codon. This event was triggered or supported by loss of the tRNA\textsubscript{Leu}\text{LACG}, which significantly reduced the possibility of further ambiguous CUG decoding. In an independent event, \textit{Pachysolen} acquired the tRNA\textsubscript{AlaLACG} by duplication and subsequent mutation of a GCU-decoding tRNA.

**Characteristics of the capturing tRNAs**

In principle, it is highly unlikely that sense codons are captured by mutant tRNAs charged with noncognate amino acids because of the high recognition accuracy of the respective tRNAs by the aaRSs (Crick 1968; Saks et al. 1994). Aminoacyl-tRNA synthetases usually recognize their cytoplasmic cognate tRNAs in at least two different regions: the most prominent being the discriminator nucleotide “N73” and the acceptor stem, and the other consisting of the anticodon and neighboring nucleotides (Saks et al. 1994; Giegé et al. 1998; Giegé and Eriani 2015). There are four principal scenarios to account for how codons could be captured by noncognate tRNAs and aaRS still be maintained. All four scenarios have been observed in nature (Fig. 5): (1) tRNAs could be mutated in the anticodon retaining recognition by the original aaRSs; (2) tRNAs could be mutated at other discriminator bases disrupting tRNA recognition by the cognate aaRS and enabling acylation by other aaRSs; (3) mutations in the respective aaRSs might relax anticodon discrimination, and this might happen either without or in combination with tRNA anticodon mutations; and (4) new orthogonal tRNA/aaRS pairs might evolve.

The first scenario includes the capture of the CUG codons in “CTG clade” yeasts and \textit{Pachysolen}. The recognition sites of seryl- and alanyl-aaRSs do not include the respective anticodons, providing an explanation as to why it is that only these tRNAs (in addition to other leucine tRNAs) could capture the free CUG codon (Fig. 5). This also means that tRNA\textsubscript{Ser} and tRNA\textsubscript{Ala} could potentially capture any other free codon. The anticodon nucleotide A35 in leucine tRNAs is a system-dependent identity determinant (e.g., it is a determinant in yeasts but not in human). This might explain the presence of many leucine tRNAs with noncognate anticodons in nematodes (Fig. 5; Hamashima et al. 2012). Because the anticodon identity determinants are not entirely conserved but could vary from species to species or taxon to taxon, it would

**Figure 5.** Compilation of nuclear codon reassignments. The scheme shows tRNA anticodon identity determinants plotted onto the genetic code. If mutated, Ala- and Ser-tRNAs (and in some systems also the Leu-tRNAs) should, in principle, be able to capture any other codon. Plotting all known cases of nuclear genetic code reassignments onto the codon table shows that most cases resulted from extending the decoding capabilities of near-cognate tRNAs. Reassignments reported for \textit{Mycoplasma capricolum}, which was thought to lack a dedicated tRNA for decoding CGG although still containing six CGG codons in its genome (Oba et al. 1991), and \textit{Mycococcus luteus}, which was thought to lack AUA and AGA codons (Kano et al. 1993), are not supported by whole-genome sequencing data (Young et al. 2010; Chu et al. 2011). The available data show that the \textit{Caenorhabditis elegans} GGG codons are translated as glycine in vivo and not as leucine (Hamashima et al. 2012).
be necessary to determine the orthogonal tRNA/aaRS pairs for each species to generate species-specific sets of potentially capturing tRNAs.

The second scenario is represented by the mitochondrial tRNA<sup>Ala</sup><sub>ATG</sub> in Eremothecium, which originated from a tRNA<sup>Thr</sup><sub>GAG</sub> by acquiring the acceptor stem identity determinant “G3:U70” (Supplemental Fig. S13; Ling et al. 2014). This tRNA<sup>Ala</sup><sub>ATG</sub> is charged by the AlaRS, which is nondiscriminative against the anticodon. The tRNA<sup>Thr</sup><sub>GAG</sub> itself originated from a tRNA<sup>His</sup><sub>UAG</sub> according to the fourth scenario by mutating the anticodon and the acceptor stem and by acquiring a dedicated ThrRS for correct charging (Pape et al. 1985; Su et al. 2011).

The third scenario, extending the aminoacylation potential of aaRS by removing tRNA recognition sites, is found, for example, in the reassignment of the UAA and UAG stop codons in the nucleic acids of ciliates and diplomonads (Fig. 5; Knight et al. 2001a; Lozupone et al. 2001). Here, the stop codons were captured by single-base mutated Gin- or Glu-tRNAs, and both the cognate tRNAs and the new tRNAs decoding stop codons are correctly charged (Hanyu et al. 1986; Sánchez-Silva et al. 2003). Although the details of the molecular mechanism are unknown, it is tempting to assume that the respective aaRSs do not discriminate the third position of the anticodon, similar to the bacterial and archaeal glutamyl-tRNA synthetases (Nureki et al. 2010). Reassignment of the stop codons might have happened according to the tRNA loss driven codon reassignment hypothesis by mutation of the single eukaryotic release factor eRF1 freeing respective stop codons. In contrast to bacteria that often have polycistrionic mRNAs, eukaryotic mRNAs are usually monocistronic. Thus, stop codon read-through of eukaryotic mRNAs does not impose any consequences other than elongation of protein tails by usually a few residues.

The identity determinants for mitochondrial tRNAs are largely unknown (Salinas-Giegé et al. 2015), and the origins of many of the tRNAs with altered anticodons have never been determined (Supplemental Fig. S13). Nevertheless, the current data suggest a close connection between the tRNAs capturing a free codon and the respective aaRSs being able to correctly charge the cognate and the newly assigned tRNAs.

### Predictions based on the tRNA loss driven codon reassignment theory

Our tRNA loss driven codon reassignment hypothesis presents several testable predictions that are mutually exclusive with the codon capture and ambiguous intermediate theories. We predict identification of (1) additional yeast species with tRNA<sup>Ser</sup><sub>Val</sub> recognition, (2) species with tRNA<sup>Ser</sup><sub>Val</sub> identification of (1) additional yeast species with tRNA<sup>Ser</sup><sub>Val</sub> recognition, and (3) additional yeast species with tRNA<sup>Ser</sup><sub>Val</sub> recognition. We predict the emergence of additional yeast species with tRNA<sup>Ser</sup><sub>Val</sub> recognition and ambiguous intermediate theories. We predict the emergence of additional yeast species with tRNA<sup>Ser</sup><sub>Val</sub> recognition and ambiguous intermediate theories.

### Methods

#### Growth and lysis of P. tannophilus NRRL Y-2460

*P. tannophilus* NRRL Y-2460 was obtained from ATCC (LGC Standards). Cells were grown in YPFD medium at 30°C, harvested by centrifugation (20 min at 5000g), and washed and resuspended in lysis buffer (50 mM HEPES at pH 6.8, 100 mM KCl). The cells were disrupted by three passages through a French press (20,000 lb/in<sup>2</sup>) at 4°C, and intact cells and the cell debris were removed by centrifugation (10 min at 15,000g). The supernatant was subjected to SDS-PAGE gel electrophoresis.

#### Genome annotation

The *Pachysolen* genome assembly (Liu et al. 2012) has been obtained from NCBI (GenBank accessions CAHV01000001–CAHV01000267). Gene prediction was done with AUGUSTUS (Stanke and Waack 2003) using the parameter “genemodel-complete,” the gene feature set of *C. albicans*, and the standard codon translation table. The gene prediction resulted in 5288 predicted proteins, out of which 4210 contain at least one CUG codon. For the mass spectrometry database search, the database was multiplied so that each new database contains the CUG codons translated by another amino acid.

#### Mass spectrometry analysis

SDS-PAGE–separated protein samples were processed as described previously (Shevchenko et al. 1996). The resuspended peptides in sample loading buffer (2% acetonitrile and 0.1% trifluoroacetic acid) were fractionated and analyzed by an online UltiMate 3000 RSLCnano HPLC system (Thermo Fisher Scientific) coupled online to the Q Exactive HF mass spectrometer (Thermo Fisher Scientific). Firstly, the peptides were desalted on a reverse-phase C18 precolumn (3 cm long, 100 µm inner diameter, 360 µm outer diameter) for 3 min. After 3 min the precolumn was switched online with the analytical column (30 cm long, 75 µm inner diameter) prepared in-house using ReproSil-Pur C18 AQ 1.9 µm reverse-phase resin (Dr. Maisch). The peptides were separated with a linear gradient of 5%–35% buffer (80% acetonitrile and 0.1% formic acid) at a flow rate of 300 nL/min (with back pressure 500 bars) over a 90-min gradient time. The precolumn and the column temperature were set to 50°C during the chromatography. The MS data were acquired by scanning the precursors in mass range from 350–1600 m/z at a resolution of 70,000 at m/z 200. Top 30 precursor ions were chosen for MS2 by using data-dependent acquisition (DDA) mode at a resolution of 15,000 at m/z 200 with maximum IT of 50 msec. Data analysis and search were performed using MaxQuant v.1.5.2.8 as search engine with 1% FDR against the *Pachysolen* genome annotation database as annotated above. The search parameters for searching the precursor and fragment ion masses against the database were as previously described (Oellerich et al. 2011) except that all peptides shorter than seven amino acids were excluded. To increase confidence in the amino acids translated from CUG codons, we determined the observed fragment ions around each CUG-encoded residue. Only amino acids with fragment ions at both sides of the amino acid, which allows the determination of the mass of the respective amino acid, were regarded as supported by the data. If fragment ions at both sides of the CUG-encoded residue are missing, the respective CUG translation can be misinterpreted because the potential post-translational modifications and chemical reactions (as result from the data generation process) at neighboring residues are not included in the database search.
tRNA phylogeny

tRNA genes in 60 sequenced yeast species and four *Schizosaccharomyces* species, which were used as outgroup (Supplemental Table S1; Mühlhausen and Kollmar 2014), were identified with tRNAscan (Lowe and Eddy 1997) using standard parameters. All genomes in which tRNAAGs were not found by tRNAscan, were searched with BLAST and respective tRNA reconstructions manually. This especially accounts for the many tRNAAGs having long introns (up to 287 bp). The intron-free tRNAAGs were aligned against all C. albicans cytoplasmic tRNAs to identify the closest related tRNA types for in-depth analysis. While the tRNALeuAGs and tRNASerAGs were easily identified, the Pachysolen tRNAAG grouped within the Candida albine and valine tRNAs. To finally resolve tRNA codon type relationships and reconstruct tRNAAG evolution, we increased sequence and taxonomic sampling. Therefore, we randomly selected three to 10 homologs from all leucine, serine, and alanine isoacceptor tRNAs from all 60 yeast species, as well as similar numbers of tRNAs from a selection of veline, phenylalanine, methionine, arginine, isoleucine, and threonine codon types. Identical tRNAAG sequences from gene duplications were removed, resulting in an alignment of 172 tRNA sequences (Supplemental Fig. S2). To refine the resolution of tRNA relationships within codon family boxes, we manually removed mitochondrial Leu-, Ser-, and Ala-tRNAs from the data sets and performed separate phylogenetic analyses of all NAG-tRNAs (320 leucine isoacceptor tRNAs), NGA-tRNAs (776 serine isoacceptor tRNAs), and NGC-tRNAs (824 alanine isoacceptor tRNAs) (Supplemental Figs. S7–S9). Sequence redundancy was removed using the CD-HIT suite (Li and Godzik 2006), generating reduced alignments of representative sequences of <95% identity (80 Leu-tRNAs, 76 Ser-tRNAs, and 70 Ala-tRNAs).

Phylogenetic trees were inferred using neighbor joining-based, Bayesian-based, and maximum likelihood-based methods as implemented in ClustalW v.2.1 (Chenna et al. 2003), Phase v. 2.0 (Jow et al. 2002; Huedelot et al. 2003) and FastTree v. 2.1.7 (Price et al. 2010), respectively. The most appropriate model of nucleotide substitution was determined with jModelTest v. 2.1.5 (Darriba et al. 2012). Accordingly, FastTree was run with the GTR model for estimating the proportion of invariable sites and the GAMMA model to account for rate heterogeneity. Bootstrapping in ClustalW and FastTree was performed with 1000 replications. Phase was used with a mixed model, the REV-T model for the loops, and the RNA7D-T model for the stem regions, which were given by a manually generated consensus tRNA secondary structure. Phase was run with 750,000 burn-in and 3 million sampling iterations, as well as a sampling period of 150 cycles. The phylogenetic network was generated with SplitsTree v.4.1.3.3 (Huson and Bryant 2006) using the neighbor-net method to identify alternative splits.

Data access

The mass spectrometry data from this study have been submitted to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE (Vizcaíno et al. 2016) partner repository with the data set identifier PXD003898.

Acknowledgments

We thank Prof. Dr. Christian Griesinger for his continuous generous support, Dr. Hans-Dieter Schmitt for his help in growing *Pachysolen*, Dr. Christof Lenz and M.Sc. Samir Karaca for help in analyzing the mass spectrometry data, and Prof. Dr. Lauren Hurst for helpful comments. This project was funded by the Deutsche Forschungsgemeinschaft DFG Grant KO 2251/13-1 to M.K. and a Synaptic Systems fellowship to S.M.

**Author contributions:** M.K. initiated the study. S.M. performed MS/MS data and phylogenetic analyses. P.F. prepared experimental samples. U.P. performed MS/MS experiments. H.U. was involved in MS/MS data interpretation. M.K. assembled, aligned, and analyzed tRNA sequences. S.M. and M.K. wrote the manuscript.

**References**


A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes

Stefanie Mühlhausen, Peggy Findeisen, Uwe Plessmann, et al.

*Genome Res.* 2016 26: 945-955 originally published online May 6, 2016
Access the most recent version at doi:10.1101/gr.200931.115

**Supplemental Material**

http://genome.cshlp.org/content/suppl/2016/05/27/gr.200931.115.DC1

**References**

This article cites 60 articles, 30 of which can be accessed free at:
http://genome.cshlp.org/content/26/7/945.full.html#ref-list-1

**Creative Commons License**

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

**Email Alerting Service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://genome.cshlp.org/subscriptions).

© 2016 Mühlhausen et al.; Published by Cold Spring Harbor Laboratory Press