The Evolution of Alternative Genetic Codes

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Committee on Higher Degrees in Systems, Synthetic, and Quantitative Biology

have examined a dissertation entitled

The Evolution of Alternative Genetic Codes

presented by Yekaterina Shulgina

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Andrew Murray

Signature _____

Typed name: Prof. Michael Desai

Signature _____

Typed name: Prof. Amy Lee

Date: November 18, 2022

The Evolution of Alternative Genetic Codes

A dissertation presented

by

Yekaterina Shulgina

to

The Committee on Higher Degrees in Systems Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Systems Biology

Harvard University

Cambridge, Massachusetts

November 2022

Dissertation Advisor: Professor Sean R. Eddy Yekaterina Shulgina

The Evolution of Alternative Genetic Codes

Abstract

The genetic code is so critical to the production of all proteins in the cell that it was once thought to be a frozen accident, incapable of evolving further. However, the discovery of alternative genetic codes in dozens of clades of bacteria, eukaryotes, archaea, organelles, and viruses has shown that it is flexible to some degree. In Chapter 1, I review major concepts in alternative genetic code evolution: the evolutionary trajectories by which alternative genetic codes can evolve, how changes to the translational machinery allow for codon reassignment, and potential selective pressures that could drive reassignment. In Chapters 2 and 3, I describe Codetta, a computational tool for predicting the genetic code from a nucleotide sequence. In Chapter 2, I use Codetta to screen over 250,000 bacterial and archaeal genomes and identify five bacterial clades using new genetic codes. All of the new reassignments affect arginine codons, highlighting a potential sensitivity of arginine codons to reassignment. In a clade of uncultivated *Enterosoma* bacteria, AGG has been reassigned from arginine to become the dominant methionine codon and appears to have evolved by a change in the aminoacylation of the original arginine tRNA. The other reassignments all affect the arginine codons CGA and/or CGG. These clades all have very low genomic GC content, an evolutionary force which likely drove GC-rich codons to rarity and allowed for their reassignment. The CGA and CGG reassignments are explored in more detail in Chapter 4, including close outgroup species with unusual tRNA features that may shed light on intermediate stages of reassignment.

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to thank Sean Eddy. I was lucky to have an advisor with whom I was so well aligned in scientific interests and values and who continues to be a role model for me. I especially thank Sean for believing in me when I was a stubborn and insecure young graduate student and for giving me so much independence in research. The self-confidence that I gained from driving a project from start to finish will stay with me for a very long time.

Lab members over the years helped made the Eddy lab a supportive and vibrant environment to do my PhD. Special thanks to Tom Jones for the hundreds of hours of impromptu science discussions, rooftop hijinks, and board games; to Harleen Saini for being my fellow "connoisseur of evil" and for all the runs along the Charles while the world was locked down; to Jenny Chen for always having her door open to talk about anything; to Mary Richardson for sharing our love of RNA and snowy mountains; and to everyone else who was there through the fun moments and the tough ones.

Thank you to my dissertation advisory committee—Andrew Murray, Michael Desai, and Stirling Churchman—for their wise advice during critical junctures in my PhD. In particular, thanks to Andrew for giving me a bench in his lab when I needed it and for being so invested in my progress since G1. Thank you to Amy Lee for serving on my defense committee.

I owe gratitude to the Systems Biology PhD program. The first year PQE1 requirement gave me the unique opportunity to look within and come up with a project that spoke to my intrinsic curiosity, propelling the rest of my PhD. Thank you to everyone who imperceptibly ensured that the PhD process ran smoothly, including Elizabeth Pomerantz, Colleen Frain, and Eddy lab admins Michelle Merry, Veronika Chudik, and Ariane Berelowitch.

During my stint in the wet lab, I was lucky to have Alicia Darnell as a mentor—a kind and brilliant scientist who made RNA biochemistry so fun that I decided to do it for my postdoc. Thank you to everyone on the fourth floor of Northwest Building for sharing lab equipment and helping me out when I was stuck, especially Alex Nguyen Ba for being incredibly knowledgeable and generous with his time.

This work would not have been possible without the researchers around the world who sequenced hundreds of thousands of genomes that I screened for new genetic codes. I have also whispered many thank yous to biochemists decades ago who studied the effects of tRNA mutations that I used to validate new genetic codes.

Outside of lab, I was fortunate to have a supportive network of friends. Most of my weekends were spent on some adventure or another in the White Mountains or elsewhere in New England, hiking, skiing, backpacking, canoeing, or biking with great company. In particular, thank you to Sophie Giguere and Monika O'Shea for their adventurous spirits.

Of course, I wouldn't be who I am today without my family. Thank you to my parents and siblings for teaching me to be always curious, critical, and questioning. Lastly, I would like to thank Kasun, who has been by my side through my entire development as a scientist. Thanks for making this journey a little bit lighter and happier.

Preface to Chapter 1

This chapter was written as a current review of the field, discussing the main themes of genetic code evolution. As such, it will function as an introduction to this thesis. Note that the publication Shulgina and Eddy (2021) deriving from Chapter 2 will be mentioned in this chapter retrospectively. Tables of all known alternative genetic codes can be found in Appendix A.

Chapter 1

Introduction

The flow of information from nucleic acid to protein sequences is fundamental to modern life on Earth. Codons in mRNA are decoded as specific amino acids or as stop signals in a predictable manner due to the coordinated action of the translational machinery—ribosomes, tRNAs, aminoacyl-tRNA synthetases, peptide release factors, and many other players. The resulting decoding rules are schematically represented by the genetic code table (Figure 1.1). The translational machinery consists of some of the most highly conserved RNAs and proteins, and the same genetic code is used by almost all known organisms. This is in part because evolving a change to even one of the 64 codons would require not only altering the translational machinery but also acquiring genome-wide compensatory mutations to continue producing the same protein products.

The existence of alternative genetic codes, first discovered in human mitochondria (Barrell *et al.*, 1979), shows that the genetic code has some capacity to evolve despite its critical role. In the past few decades, over seventy clades have been reported to have evolved a genetic code change, many in mitochondrial genomes but also in bacteria such as mycoplasmas, insect

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC		UCC		UAC		UGC	
UUA	Leu	UCA		UAA	Stop	UGA	Stop
UUG		UCG		UAG		UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	Gln	CGA	
CUG		CCG		CAG		CGG	
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	Lys	AGA	Arg
AUG	Met	ACG		AAG		AGG	
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC		GCC		GAC		GGC	
GUA		GCA		GAA	Glu	GGA	
GUG		GCG		GAG		GGG	

Figure 1.1: The standard genetic code

endosymbiotic bacteria, and members of the Candidate Phyla Radiation, single-celled eukaryotes ranging from yeasts to green algae to trypanosomes, methanogenic archaea, and even some viruses. Tables A.1-A.7 list all of the currently known alternative genetic codes, and Figure 1.2 shows their phylogenetic distributions. Most involve changes to just one or two codons, with the most extreme example in yeast mitochondria where six codons differ from the standard genetic code. Codon reassignments are typically identified by comparing coding sequence in the species of interest to homologous sequences from related species and examining if a given codon recurrently occurs at positions for an unexpected amino acid. Recent efforts to explore more of the microbial diversity by sequencing metagenomes, metatranscriptomes, and single-cell genomes have led to the discovery of many new genetic codes (Ivanova *et al.*, 2014; Swart *et al.*, 2016; Shulgina and Eddy, 2021; Borges *et al.*, 2022). Reassignments can be validated experimentally by comparing protein sequences obtained by direct protein sequencing methods like Edman degradation or proteomic mass spectrometry to their mRNA transcripts.

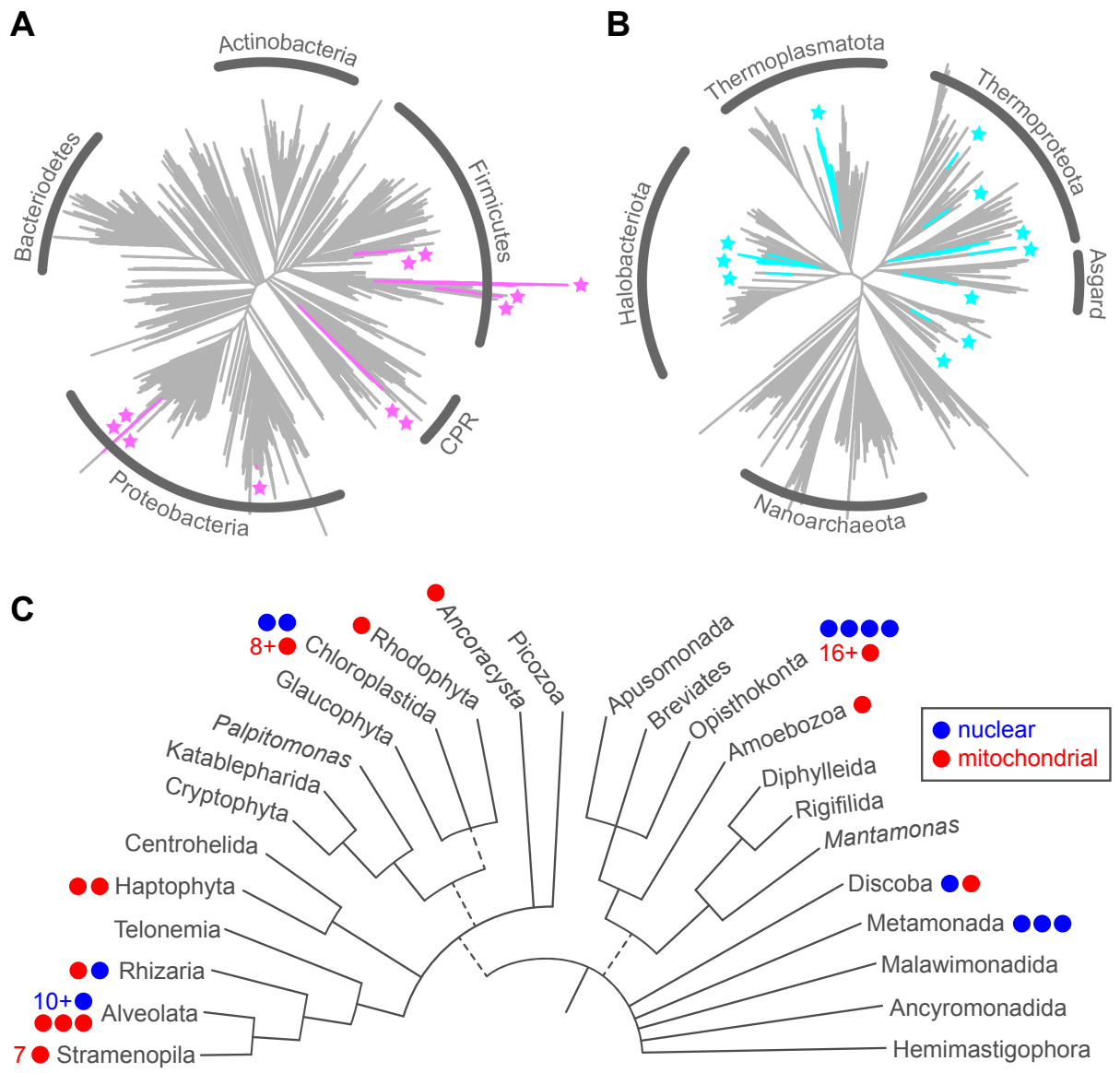

Figure 1.2: Broad phylogenetic distribution of all clades known to use alternative genetic codes in bacteria, archaea, and eukaryotes. (A) Unrooted bacterial phylogenetic tree from the Genome Taxonomy Database (GTDB, Parks *et al.* (2020)) with some major groups labelled and clades reported to use alternative genetic codes highlighted and marked with a star. A complete list can be found in Table A.1. (B) Unrooted archaeal phylogenetic tree from GTDB with some phyla labelled and clades reported to decode UAG as pyrrolysine (only codon reassignment known in archaea) highlighted and marked with a star. A complete list can be found in Table A.2. (C) Cladogram of major eukaryotic lineages, adapted from Burki *et al.* (2020). Groups that contain members that have evolved a codon reassignment are marked with a dot (blue for nuclear genetic codes and red for mitochondrial genetic codes), one for each independent codon change. Complete lists can be found in Table A.3 for eukaryotic nuclear reassignments and Tables A.4-A.7 for mitochondrial reassignments.

Variations to the genetic code have been observed even within a single organism. Some bacteriophages have been reported to use two different genetic codes to translate different genes, potentially as a way to tightly couple the lysis-lysogeny switch to the expression of a tRNA (Borges *et al.*, 2022). The bacterium *Acetohalobium arabaticum* expresses a gene cassette for the biosynthesis and incorporation of the non-canonical amino acid pyrrolysine at UAG stop codons selectively in the presence of a trimethylamine carbon source (Prat *et al.*, 2012). Some microbial eukaryotes including *Blastocrithidia* trypanosomes, an *Amoebophrya* dinoflagellate, and two groups of ciliates have dual-use codons interpreted as either stop or sense depending on their location within an mRNA transcript (Swart *et al.*, 2016; Záhonová *et al.*, 2016; Bachvaroff, 2019; Seah *et al.*, 2022). Meanwhile, yeasts in the genera *Saccharomycopsis* and *Ascoidea* appear to decode CUG codons as both serine and leucine in a stochastic manner by two competing tRNAs (Krassowski *et al.*, 2018; Mühlhausen *et al.*, 2018; Shulgina and Eddy, 2021). These dynamic reinterpretations of the genetic code force us to broaden perspective on what variation to the translational machinery is possible and tolerated. This may be especially instructive for efforts to engineer microbes with altered genetic codes for the incorporation of non-canonical amino acids, including ideas on how to free codons for reassignment and ways in which the translational machinery may be amenable to engineering (Kuo *et al.*, 2018; Kim *et al.*, 2022).

In this review, I begin with a general framework for thinking about codon reassignment, and then cover major questions of how and why the genetic code evolves: 1) what changes to the translational machinery lead to codon reassignment? 2) why are some codons more susceptible to reassignment? and 3) are alternative genetic codes advantageous or merely products of neutral evolution?

1.1 A framework for alternative genetic code evolution

What is an alternative genetic code?

I define an alternative genetic code as the reassignment of one or more of the 64 codons to a default sense or stop meaning that differs from the standard genetic code. "Default meaning" refers to how the translating ribosome would interpret the codon in the absence of additional contextual information. This is to distinguish codon reassignment from other codon-dependent signals for events like programmed ribosomal frameshifting, translation initiation, and incorporation of some non-canonical amino acids. An example of a codon reassignment is in mycoplasmas where the canonical stop codon UGA codes for tryptophan (Yamao *et al.*, 1985). The gene for release factor 2, which terminates translation at UGA in bacteria, is no longer present and UGA is instead read by a tryptophan tRNA that is cognate (i.e. has an anticodon that can decode UGA) (Yamao *et al.*, 1985; Inagaki *et al.*, 1993). Thus, mycoplasmas decode UGA as tryptophan without any additional signals. In contrast, the selective insertion of selenocysteine at UGA stop codons in many species would not be considered to be a codon reassignment because it requires the presence of a structural mRNA element; without it, the default meaning of UGA is stop. (Gonzalez-Flores *et al.*, 2013). By the same logic, pyrrolysine incorporation at UAG codons in some methanogenic archaea and bacteria is a codon reassignment because no additional sequence context is needed (Gaston *et al.*, 2011). A default meaning could also be ambiguous, for example how *Ascoidea* yeast translate CUG as serine and leucine by competing tRNA species (Mühlhausen *et al.*, 2018), and this can be thought of as an intermediate stage in genetic code evolution.

Gains, losses, and codon substitutions

For a codon reassignment to evolve, not only does the translational machinery need to mutate to change how the codon is interpreted during translation, but coding regions must also adapt to remove the codon from positions that cannot tolerate the new meaning. Three events must occur to reassign a codon, building upon the gain-loss framework of Sengupta and Higgs (2005):

1. **Gain** of the ability to translate the codon in the new way
2. **Loss** of the ability to translate the codon in the former way
3. **Substitutions** to remove the codon from positions that cannot tolerate the new translation

The first two events—gain and loss—require changes to the translational machinery. Gain events lead to the codon being decoded in a new way. For stop codons, this usually means changing release factor specificity and for sense codons, altering tRNA aminoacylation or codon recognition. Loss events involve either deletion of a component of the translational machinery or mutation that disrupts functionality.

The order of the gain and loss events leads to different evolutionary intermediate states (Figure 1.3A). If gain occurs first then the codon is translated ambiguously as two meanings for a period of time; this is called the 'ambiguous intermediate' model (Schultz and Yarus, 1994). The most well-known examples of extant ambiguous intermediates are in yeast, where the CUG codon is decoded as both serine and leucine in *Candida albicans* at a ratio of about 97% serine to 3% leucine by tRNA mischarging (Gomes *et al.*, 2007) and in *Ascoidea asiatica* at a ratio of about 50-50 by two different tRNAs (Mühlhausen *et al.*, 2018). Other examples of am-

biguity in translation can also be thought of as intermediates, even if they are not associated with a codon reassignment. For example, some *Streptomyces* plant pathogens have a specialized proline tRNA and prolyl-tRNA synthetase pair that may be used to mistranslate alanine codons as proline (Vargas-Rodriguez *et al.*, 2021); *Mycoplasma mobile* mischarges leucine tRNAs as isoleucine, valine, and methionine at a higher rate due to a missing leucyl-tRNA synthetase editing domain (Li *et al.*, 2011); in *Bacillus subtilis*, the UGA stop codon is read through by the near-cognate tryptophan tRNA at a rate of about 6% (Lovett *et al.*, 1991; Matsugi *et al.*, 1998). The evolutionary trajectories have no intrinsic directionality, so an intermediate could revert back to the original translation instead of progressing to a complete codon reassignment.

Alternatively, if loss occurs before gain then there is a period where the codon cannot be translated efficiently, which has been named the 'unassigned codon' or 'tRNA loss driven reassignment' model (Figure 1.3A) (Sengupta and Higgs, 2005; Mühlhausen *et al.*, 2016). When the function of a cognate tRNA or release factor for a codon is impaired or completely lost, the codon can still be inefficiently translated by near-cognate tRNAs or resolved by other mechanisms (Roy *et al.*, 2015; Ma *et al.*, 2018). There are several known examples of codons that have lost the machinery for any efficient translation. In human mitochondria, the codons AGA and AGG lack a cognate tRNA or release factor and instead lead to ribosome pausing and frameshifting (Temperley *et al.*, 2010). Other mitochondria rely on codon-anticodon mispairing to translate codons lacking cognate tRNAs—in echinoderm mitochondria, decoding of the AUA isoleucine codon relies on a G-A mispair with the isoleucine $tRNA_{GAU}$ (Yokobori *et al.*, 2001); similarly, in *Drosophila* mitochondria, AGA is read by a serine $tRNA_{GCU}$, also requiring a G-A mispair (Tomita *et al.*, 1999a). In bacteria, *Mycoplasma capricolum* lacks a cognate tRNA to decode the arginine codon CGG, which is used only six times in coding sequences (Oba

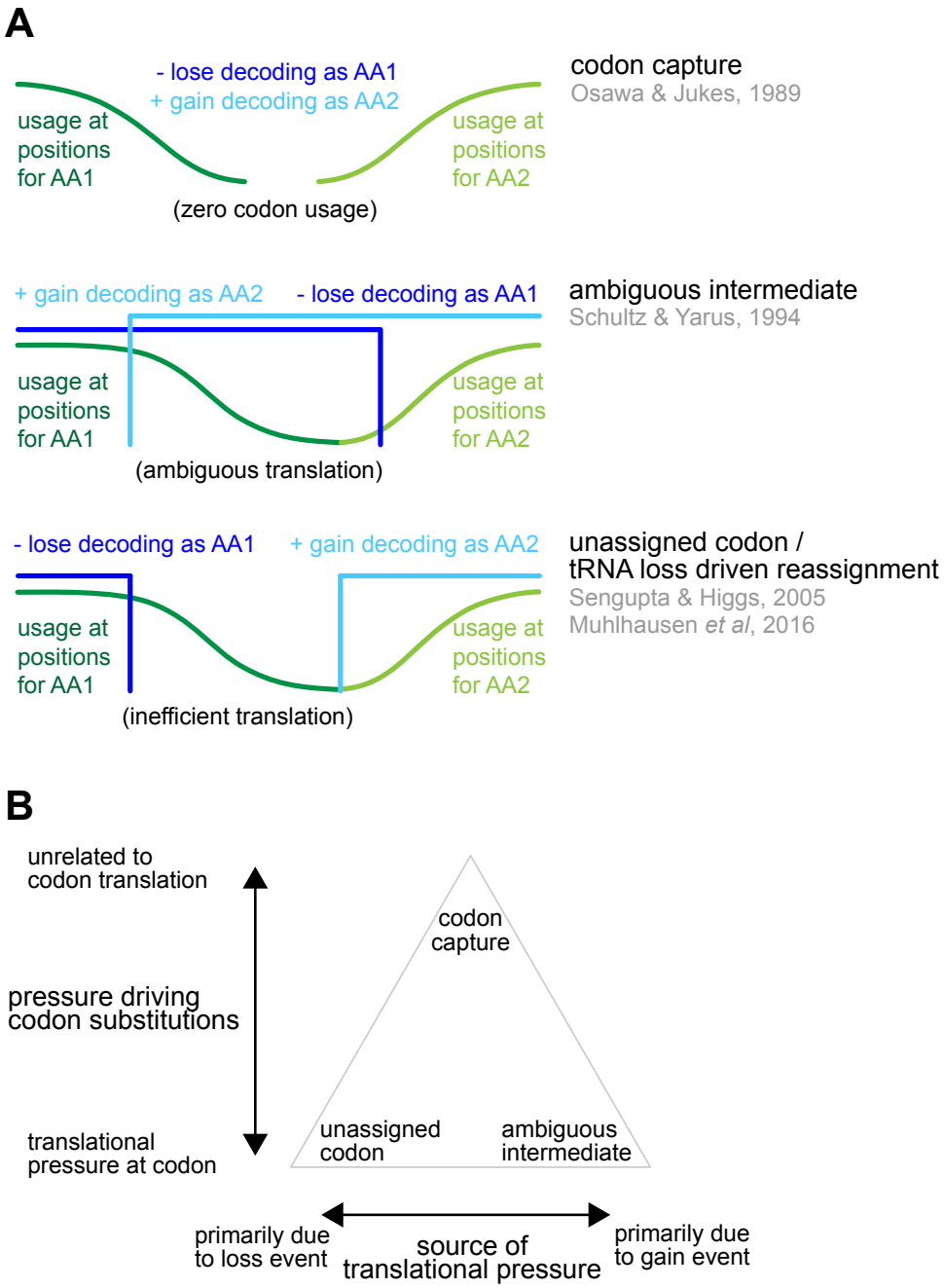

Figure 1.3: (A) The three main evolutionary models of codon reassignment, represented as a sense codon reassignment from amino acid 1 (AA1) to amino acid 2 (AA2). The evolutionary trajectories differ in the order of the gain event (light blue), loss event (dark blue), and codon substitutions (green) which affects the nature of the codon reassignment intermediate state. (B) The evolutionary trajectories also differ in the nature of the pressure driving the codon substitutions. Reassignments can be split on two axes: whether the main pressure driving the genome-wide codon substitutions is unrelated to the reassignment or due to changes to the translation of the codon and, if so, then whether it is driven primarily by the gain event or the loss event. The named evolutionary trajectories represent "corner cases" on the triangle, and, in nature, evolutionary trajectories can involve multiple forces and fall somewhere within the triangle.

et al., 1991; Chu *et al.*, 2011; Kollmar and Mühlhausen, 2017b), instead relies on a small fraction of arginine $tRNA_{ACG}$ that escapes modification and remains with an unusual unmodified A34 wobble base to decode CGG (Yokobori *et al.*, 2013). For decades, *Micrococcus luteus* was thought to lack tRNAs to decode AGA and AUA (Kano *et al.*, 1993) until the cognate tRNA genes were found when the genome was sequenced (Young *et al.*, 2010). There are other bacterial genomes where tRNA genes required for reading certain codons have not yet been found and the codon may potentially be unassigned, such as the CGA codon in some Gracilibacteria and the CGG codon in *Clostridioides difficile* (Shulgina and Eddy, 2021).

The compensatory substitution barrier

Soon after the standard genetic code was first determined in the 1960's, Francis Crick posited that evolving any change "would be highly disadvantageous unless accompanied by many simultaneous mutations to correct the 'mistakes' produced by altering the code" (Crick, 1968). This is the third event in our framework—the genome-wide codon substitutions that must occur to counteract the potentially deleterious effect of mutating the translational machinery on existing coding sequences. While gain or loss events can be as simple as single mutations to tRNAs or release factors, there is no way to shortcut the massive removal of the reassigned codon from positions that cannot tolerate the new meaning. This is the main evolutionary barrier to codon reassignment.

What sorts of evolutionary pressures could drive these codon substitutions? On one extreme is the 'codon capture' model (Figure 1.3A) (Osawa and Jukes, 1989). Here, the codon is first driven to zero-usage prior to reassignment by forces unrelated to its translation, such as an extreme bias in genomic GC content disfavoring some codons in favor of synonymous alternatives or pressure to reduce the genome size which may drive very rare codons to extinc-

tion. Once the codon is gone from coding regions, it is free to be 'captured' by a new meaning. In this trajectory, the codon substitutions occur before any gain or loss event. On the other hand, if the gain or loss were to occur while the codon was still abundantly used, then the ambiguous or inefficient translation of the codon in the intermediate state would itself incur a selective pressure for substitutions away from the codon, gradually clearing the codon from many coding positions (Massey *et al.*, 2003; Mühlhausen *et al.*, 2016).

In nature, real trajectories probably involve a combination of these pressures; for example, a codon might be first driven to rarity by biased genomic GC content and then removed from the remaining positions by translational pressure caused by the reassignment itself. From the perspective of codon substitutions, we can imagine all possible trajectories falling within a triangle where the "corner cases" are the three main evolutionary models (Figure 1.3B). The vertical axis represents whether the codon substitutions were predominantly driven by unrelated forces prior to reassignment or by translational pressure caused by the reassignment process itself. The horizontal axis separates the 'ambiguous intermediate' and 'unassigned codon' models based on the nature of the translational pressure, whether is it primarily due to the loss or gain event. The 'codon capture' model has no horizontal axis because once a codon is driven to extinction the gain or loss events do not lead to any further codon substitutions.

One intuition gained from this framework is that codon reassignments are more likely to evolve when the compensatory substitution barrier is lessened. For instance, if a codon is first driven to rarity by unrelated forces (in a 'codon capture'-like scenario), then much of the burden of codon reassignment is already completed and the barrier to overcoming the effects of gain or loss are much smaller. In bacteria, almost all known alternative genetic codes occur in clades with reduced genome sizes and low genomic GC content, forces that may drive

codons like the UGA stop codon and CGG arginine codon to rarity and "prime" them for reassignment (Ivanova *et al.*, 2014; Shulgina and Eddy, 2021). The codon substitution barrier may also explain why the greatest diversity of codon reassignments is found in mitochondria (Figure 1.2). The nature of their reduced genomes—ranging from just two protein-coding genes in some apicomplexan mitochondria (Flegontov *et al.*, 2015) to a maximum of around 70 protein-coding genes in jakobid mitochondria (Burger *et al.*, 2013)—means that few codon substitutions are needed to adapt to a codon reassignment. Another implication is that mutations that reduce the translational accuracy or efficiency of a codon may preempt reassignment by initiating synonymous substitutions away from that codon. Also, reassigning a single codon should have a smaller burden than reassigning multiple at the same time because fewer compensatory substitutions need to occur. As a result, we observe more reassignments at codons that are translated by release factors or tRNAs that are not responsible for translating other codons as well (discussed in more detail in Section 1.3).

1.2 What changes to the translational machinery can lead to codon reassignment?

The genetic code is implemented when recognition of a specific codon in the decoding center of the ribosome is linked to an action in the peptidyl transferase center—either incorporation of a specific amino acid or termination. This could involve changes to any number of molecular players, including tRNAs, release factors, aminoacyl-tRNA synthetases, tRNA-modifying enzymes, elongation factors, the ribosome itself, and potentially others as well. The molecular details of how the translational machinery works together to translate codons can help us understand which reassignments are simpler to evolve, and conversely, codon reassignments

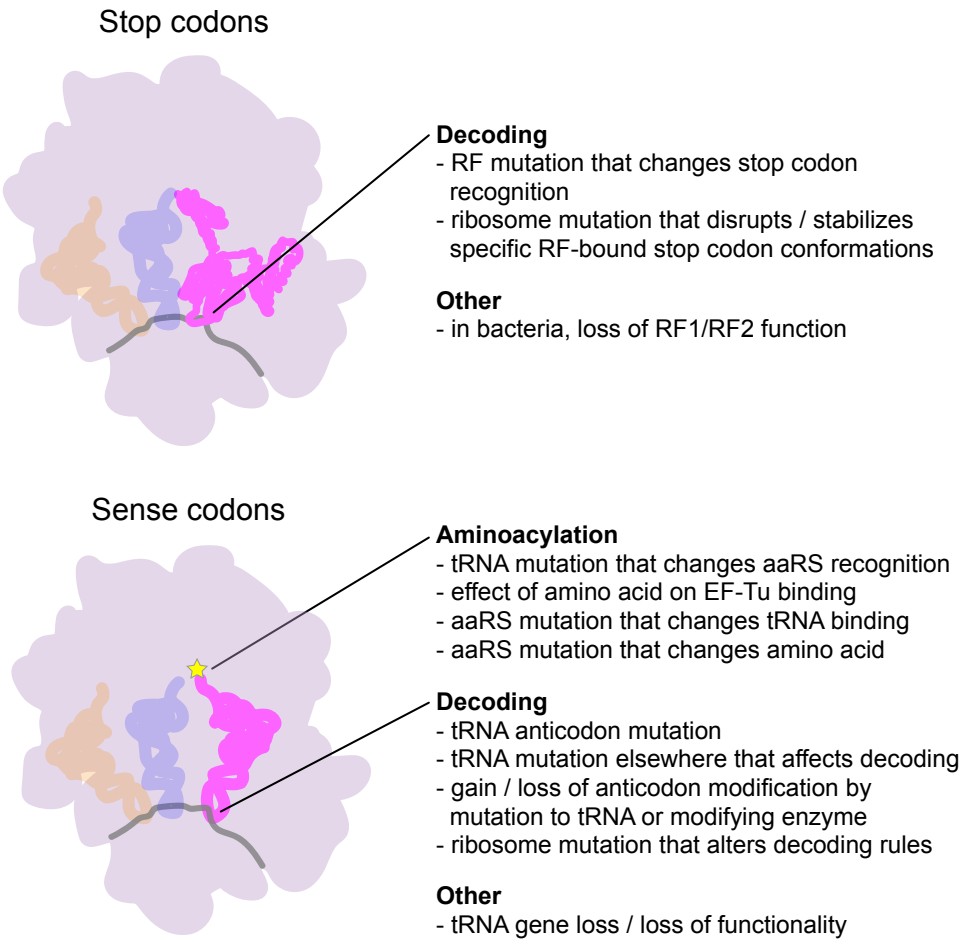

Figure 1.4: Overview of processes that could change the translational meaning of a codon.

can help us understand what sort of changes are possible to the translational machinery. In this section, I will review how stop and sense codons are translated by release factors and tRNAs, describe changes to the existing translational machinery that could potentially lead to a codon reassignment, and describe the changes that have been observed in nature.

Overview of translation termination

Release factors (RFs) are enzymes that terminate translation at stop codons. Termination involves the action of a class 1 release factor which binds the stop codon and cleaves the peptidyl-tRNA bond (in bacteria RF1 and RF2, Laurberg *et al.* (2008); Weixlbaumer *et al.* (2008);

in eukaryotes eRF1, Brown *et al.* (2015a); in archaea aRF1, Kobayashi *et al.* (2012)) and the action of a class 2 release factor, an elongation factor homolog that either stimulates peptidyl-tRNA bond cleavage or promotes release factor dissociation after peptide release (Jackson *et al.*, 2012; Kobayashi *et al.*, 2012; Adio *et al.*, 2018). Like tRNAs, class 1 release factors sample the codon in the ribosomal A site and have a decoding domain which forms specific interactions with the stop codon to discriminate against sense codons (Laurberg *et al.*, 2008; Weixlbaumer *et al.*, 2008; Brown *et al.*, 2015a). Upon recognition of a stop codon (and GTP hydrolysis by the class 2 release factor in eukaryotes and archaea), the release factor catalytic residues are positioned within the peptidyl transferase center of the ribosome and hydrolyze the peptidyl-tRNA bond to release the nascent peptide (Schmeing and Ramakrishnan, 2009). Since the mechanism of termination differs between bacteria and eukaryotes/archaea, I will discuss them separately.

Translation termination in bacteria

In bacteria, the three stop codons are recognized by two different class 1 release factors. RF1 recognizes UAA and UAG in a unique conformation where an RF1 histidine residue intercalates between codon positions 2 and 3 (Laurberg *et al.*, 2008). The three stop codon nucleotides are specified by extensive hydrogen bonding interactions with RF1 and additional packing by hydrophobic residues (Laurberg *et al.*, 2008). The RF1-bound stop codon is further stabilized by stacking and phosphate backbone interactions with the 16S ribosomal RNA nucleotides G530, A1492 and A1493, which monitor the codon-anticodon helix during sense codon translation (Laurberg *et al.*, 2008). The homologous RF2 recognizes UAA and UGA in a very similar conformation with slightly different hydrogen bonding requirements (Weixlbaumer *et al.*, 2008).

There are three possible ways that termination can be gained at sense codon (sense-to-stop reassignment) or lost at an existing stop codon (stop-to-sense reassignment):

1. *Release factor mutations that change stop codon recognition.* The most straightforward way to change the set of stop codons recognized by a release factor would be to mutate release factor residues that affect stop codon binding. Recognition of a canonical stop codon could be lost in a stop-to-sense reassignment by mutating residues involved in binding and stabilizing a specific stop codon. For sense-to-stop reassignments, it's hard to predict which release factor mutations could lead to recognition of a sense codon because the rules for stop codon recognition are not explicitly defined, in contrast to sense codon recognition by tRNAs which follow the logic of base pairing. It is unclear if recognition of some sense codons is mutationally more accessible. In bacteria, termination at sense codons does occur (about 3-6 orders of magnitude less efficiently than at stop codons) and there is huge variation across sense codons suggesting that some sense codons might be easier targets for reassignment (Freistroffer *et al.*, 2000). For example, termination at UAU by RF1 or UGG by RF2 is an order of magnitude more efficient than at any other sense codon (Freistroffer *et al.*, 2000). Altered specificity can be accomplished with few mutations—in one case, a single mutation to the body of the RF2, far from the decoding residues, expands recognition to all three stop codons (Ito *et al.*, 1998).

2. *Ribosome mutations that disrupt/stabilize a release factor-bound codon.* Since rRNA residues are involved in stabilizing the release factor-bound stop codon (Laurberg *et al.*, 2008; Weixlbaumer *et al.*, 2008), changes to the ribosome can hypothetically allow or disrupt termination at various codons. For example, in the process of post peptidyl transfer quality control, RF2 can terminate translation at sense codons if there is a mispaired

tRNA in the P site of the ribosome (Petropoulos *et al.*, 2014). This indicates that there are altered ribosome conformations (such as that caused by mispairing at the P-site codon) that change release factor specificity. Thus, the set of stop codons could also be hypothetically changed by mutations to the ribosome. However, genetic code changes due to an altered ribosome would be surprising since such changes would likely affect the translation of sense codons as well.

3. *Loss of RF1 or RF2 function.* Since bacteria have two class 1 release factors with different specificities, loss of UAG or UGA recognition could come from release factor gene loss (or mutations that impair function). Loss of UAA recognition cannot be mediated by loss of one of the release factors and would require changes to RF1 and RF2 or the ribosome.

Examples of stop codon loss or gain in bacteria

In bacteria, stop codon loss is only known for UGA, which is reassigned to tryptophan in mycoplasmas and several insect endosymbiotic bacteria and to glycine in Absconditabacteria and Gracilibacteria (Table A.1). Consistent with mechanism 3, UGA reassignment is always accompanied by the concomitant loss of RF2, leaving the remaining stop codons UAA and UAG to be recognized by RF1 (Himmelreich *et al.*, 1996; McCutcheon *et al.*, 2009; McCutcheon and Moran, 2010; Bennett and Moran, 2013; Salem *et al.*, 2017; Campbell *et al.*, 2013). Mitochondrial and plastid genomes may also encode organellar versions of RF1 and RF2, and reassignments of UGA are typically correlated with a loss in RF2 with few exceptions (Duarte *et al.*, 2012). UAG loss has not yet been observed in wild bacteria—methanogenic bacteria that incorporate pyrrolysine at UAG codons continue to terminate at UAG (Zhang *et al.*, 2005). However, engineering efforts to expand the genetic code to incorporate non-canonical amino

acids in *Escherichia coli* have successfully reassigned UAG by deleting the RF1 gene, introducing a desired tRNA that can decode UAG, and removing UAG codons where readthrough is deleterious (Mukai *et al.*, 2010; Lajoie *et al.*, 2013).

Stop codon gain has only been reported in a few mitochondria. For decades, the AGA and AGG codons in human mitochondria (arginine codons in the standard genetic code) were thought to be stop codons because they lacked cognate tRNAs and were found at the end of two reading frames and nowhere else (Anderson *et al.*, 1981). It was later discovered that these codons actually have no assigned meaning and are resolved by a ribosomal -1 frameshift directly into a stop codon (Temperley *et al.*, 2010). Currently, sense-to-stop reassignments are reported only in Labyrinthulea (slime net) mitochondria and green algal mitochondria (Tables A.6 and A.7). In Labyrinthulea mitochondria, the canonical leucine codon UUA is reassigned to stop and a subgroup may also terminate at AGA and AGG codons (Wideman *et al.*, 2020; Žihala *et al.*, 2020). These reassignments are supported by the lack of a cognate tRNA, abundant usage of these codons at typical termination positions in alignments, and substitutions and deletions in the mitochondrial RF2 gene of extremely conserved residues known to be involved in stop codon recognition (Žihala *et al.*, 2020). Green algal mitochondria have two independent sense-to-stop changes—in *Pycnococcus* mitochondria, the canonical leucine codons UUA and UUG are stop codons, and in Scenedesmaceae and Sphaeropleales mitochondria, the canonical serine codons UCA and UCG are stop codons (Turmel *et al.*, 2010; Fučíková *et al.*, 2014; Žihala and Eliáš, 2019). Similar to the Labyrinthulea reassignments, these are supported by the lack of a cognate tRNA, common occurrence at termination positions, and sequence variations to the mitochondrial RF1 (Žihala and Eliáš, 2019). Despite this supporting evidence, new stop codons require experimental demonstration of termination to definitively exclude

the possibility of being unassigned codons that are resolved by an alternative mechanism (as was the case for AGA/AGG in human mitochondria).

Translation termination in eukaryotes & archaea

In eukaryotes and archaea, a single class 1 release factor (eRF1/aRF1) is responsible for recognizing all three stop codons (Jackson *et al.*, 2012; Kobayashi *et al.*, 2012), relying on a clever molecular mechanism for discrimination (Brown *et al.*, 2015a). The A-site codon is drawn into an unusual compacted conformation where uridine at the first codon position (i.e. U1) is specified by multiple hydrogen bonding interactions with the release factor and the subsequent two codon nucleotides are oriented to form a purine stack with 18S rRNA nucleotide A1825 (Brown *et al.*, 2015a). This narrows down the range of possibilities to the stop codons UAA, UAG, UGA, and the tryptophan codon UGG. The release factor residue Glu55 approaches the stacked 2nd and 3rd codon bases from the Hoogsteen edge to form either one or two hydrogen bonds with adenines at these positions; whereas the G2 and G3 of UGG would be repulsed (Brown *et al.*, 2015a). Additional interactions specifically stabilize UAG and UGA (Brown *et al.*, 2015a). Mutagenesis studies of human eRF1 established the importance of many of the same residues in stop codon recognition (Frolova *et al.*, 2002; Kolosov *et al.*, 2005; Wong *et al.*, 2012; Kryuchkova *et al.*, 2013).

Similar to bacteria, termination could occur at sense codons by altering eRF1/aRF1 or the ribosome, although neither has been reported. Molecular dynamics simulations suggest that certain eRF1 mutations could lower the free energy barrier for binding to near-cognate sense codons (Lind *et al.*, 2017). In some eukaryotes, eRF1 is present in redundant copies of unknown function indicating a potential for divergence in specificity (Liang *et al.*, 2001; Kervestin *et al.*, 2002; Chapman and Brown, 2004). In contrast to bacteria, termination at a

stop codon cannot be lost by deleting a release factor gene because all three stop codons are decoded by the same release factor. Stop codon loss in eukaryotes and archaea requires mutations to the release factor or ribosome.

Stop codon reassignments in eukaryotes & archaea

Reassignments of stop codons to amino acids are the most frequently reported genetic code changes in eukaryotic nuclear genomes and in archaea (Tables A.2 and A.3). For some stop reassignments, eRF1 no longer binds the reassigned stop codon—for example in *Euplotes* ciliates, which have reassigned UGA to cysteine, eRF1 can only recognize UAA and UAG as stop codons in a mammalian *in vitro* translation system (Kervestin *et al.*, 2001). Some eukaryotes with stop codon reassignments have variations to highly conserved release factor residues known to be involved in stop codon recognition (Lozupone *et al.*, 2001; Inagaki *et al.*, 2002). Experiments introducing such variations into the human and yeast eRF1 result in skewed stop codon specificity consistent with the alternative genetic code (Seit-Nebi *et al.*, 2002; Eliseev *et al.*, 2011a; Conard *et al.*, 2012). However, we still lack a first-principles understanding how eRF1 mutations alter stop codon recognition, perhaps because changes to the eRF1 sequence affect the overall structure of the decoding domain in a complex and interdependent way.

Some eukaryotes and archaea with stop codon reassignments have not completely lost recognition of the reassigned stop codon—eRF1 in the ciliate *Blepharisma japonicum* still recognizes UGA in a mammalian *in vitro* translation system, albeit less efficiently than UAA and UGA, despite having reassigned UGA to tryptophan (Eliseev *et al.*, 2011b). Similarly, aRF1 from the archaeon *Methanosarcina barkeri*, which has reassigned UAG to pyrrolysine, retains a low level of termination activity against UAG in the same *in vitro* system (Alkalaeva *et al.*, 2009). It possible that these codons are ambiguously translated *in vivo* or that inefficient ter-

mination is outcompeted by the cognate tRNA to the point where translation is essentially unambiguous. Another possibility, which has been proposed for eukaryotes with dual-use stop codons (Swart *et al.*, 2016; Heaphy *et al.*, 2016; Záhonová *et al.*, 2016), is that efficient termination has become dependent on another factor such that termination only occurs in specific mRNA contexts. In the ciliate *Condylostoma magnum*, all three stop codons are interpreted as sense codons unless within about 30-60 nt of poly(A) tail (Swart *et al.*, 2016; Heaphy *et al.*, 2016). It has been proposed that poly(A)-binding protein stimulates termination, perhaps by activating the class 2 release factor (Ivanov *et al.*, 2016; Alkalaeva and Mikhailova, 2017). In this situation, a stop codon can be reassigned to an amino acid regardless of the stop codon specificity of eRF1, as long as the codon does not occur in the mRNA context where termination is efficient.

Overview of sense codon translation

tRNAs are central to the translation of sense codons. When a tRNA, in complex with elongation factor (EF) Tu (in bacteria; EF1A in eukaryotes and archaea) and GTP, samples the ribosomal A site, the anticodon is positioned in the ribosome decoding center where it discriminates between mRNA sense codons by base pairing (Dever *et al.*, 2018; Rodnina, 2018). Correct codon-anticodon pairing propagates conformational changes within the tRNA and ribosome leading to GTP hydrolysis by EF-Tu and accommodation of the tRNA completely into the A site (Dever *et al.*, 2018; Rodnina, 2018). The 3' end of the tRNA, which is attached to a defined amino acid, is positioned in the peptidyl transferase center, primed for addition to the nascent peptide chain (Dever *et al.*, 2018; Rodnina, 2018).

In a codon reassignment, loss of the ability to translate a sense codon as a particular amino acid could come from breaking any number of steps required for productive translation,

ranging from simply deleting the tRNA gene to impaired aminoacylation to loss of tRNA nucleotide modifications required to stabilize the 3-dimensional fold of the tRNA. On the other hand, gaining the ability to translate a codon as a new amino acid could come from only two processes: altering *decoding* (i.e. codon-anticodon interactions) such that a new tRNA now recognizes the codon or altering *aminoacylation* such that a new amino acid is ligated to the tRNA reading the codon. I will expand on each of these processes.

Mechanism of decoding by tRNAs

In the ribosomal A site, the codon-anticodon helix is monitored by rRNA nucleotides A1492, A1493 and G530 (*E. coli* numbering) which bind the minor groove and allow only strict Watson-Crick base pairing of codon positions 1 and 2 with the tRNA anticodon (Ogle *et al.*, 2001). Codon position 3 is not monitored as closely which allows for wobble decoding. This set of expanded base pairing rules between codon position 3 and tRNA position 34 includes not only typical G-U wobble base pairs but also unique rules, many of which involve modified tRNA nucleotides (Figure 1.5A). Most tRNAs can decode multiple codons so only 22-46 tRNA anticodons are usually used to decode all sense codons (Grosjean *et al.*, 2010).

Figure 1.5A is an overview of all possible tRNA wobble decoding rules. Different decoding strategies are employed for family boxes (where all four XXN codons are the same amino acid), split boxes (where XXY and XXR are different amino acids), and the isoleucine AUN codon box (which is split 3:1 with methionine) (Figure 1.5B). Position 34 modifications that allow for the greatest degree of wobble, like inosine (I34) in eukaryotes and 5-hydroxyuridine derivatives (xo^{5}U34) in bacteria, are used in family boxes where all codons have the same meaning (Bjork, 1994; Grosjean *et al.*, 2010). Wobble modifications that restrict decoding, like 5-methyluridine derivatives (xm^{5}U34) and queuosine (Q34), are typically found in split boxes

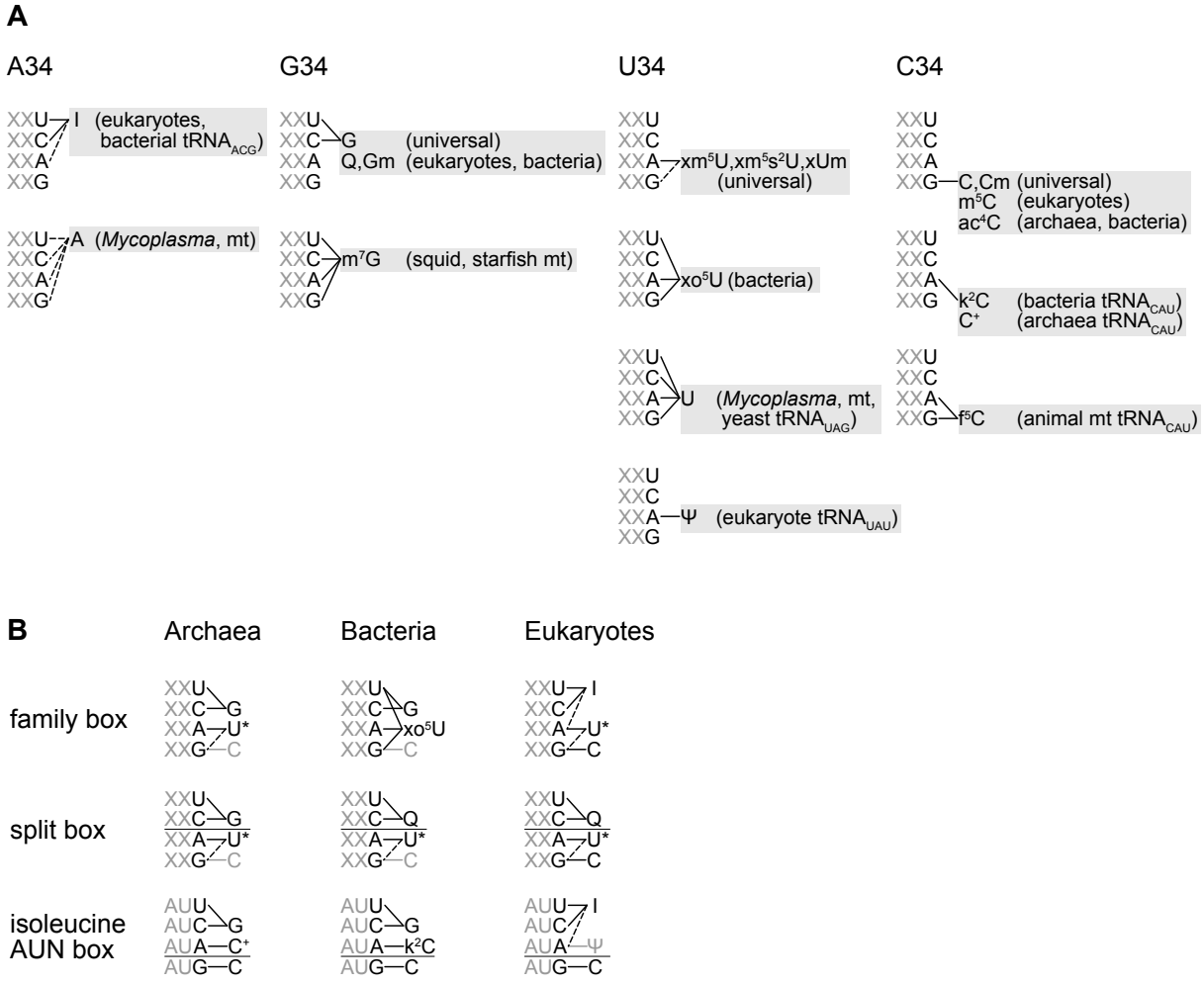

Figure 1.5: Wobble decoding rules between codon position 3 and tRNA position 34. (A) For each genomically encoded nucleotide at tRNA position 34, a summary of all possible decoding rules. Dashed lines represent decoding interactions that are presumed to be weaker. (B) The typical tRNA decoding schemes for family boxes, split boxes, and the isoleucine AUN box in the three domains of life. Interactions in gray are optionally present. Nucleotide abbreviations: I, inosine; Q, queuosine; Gm, 2′-O-methylguanosine; m^{7}G, N7-methylguanosine; U*, uridine derivative; xm^{5}U, 5-methyluridine derivative; xm^{5}s^{2}U, 5-methyl-2-thiouridine derivative; xUm, 2′-O-methyluridine derivative; xo^{5}U, 5-hydroxyuridine derivative; Ψ, pseudouridine; Cm, 2′-O-methylcytidine; m^{5}C, 5-methylcytidine; ac^{4}C, N(4)-acetylcytidine; k^{2}C, lysidine; C$^+$, agmatidine; f^{5}C, 5-formylcytidine. References: Bjork (1994); Curran (1998); Grosjean *et al.* (2010); Mandal *et al.* (2010); Wolff *et al.* (2020).

to avoid crosstalk between the XXY and XXR codons (Bjork, 1994; Curran, 1998; Grosjean *et al.*, 2010). The isoleucine AUN box features some unique wobble modifications like lysidine (k^2C34), agmatidine (C^+34), and pseudouridine ($\Psi 34$) to read the AUA codon as isoleucine and not methionine (Yokoyama *et al.*, 1989; Mandal *et al.*, 2010; Szweykowska-Kulinska *et al.*, 1994). Many modifications are phylogenetically restricted, resulting in different decoding strategies in archaea, bacteria, and eukaryotes (Figure 1.5B; Bjork (1994); Grosjean *et al.* (2010)). Mitochondria and *Mycoplasma* bacteria have some of the most unusual decoding schemes, sometimes using an unmodified U34 or A34 tRNA to read all codons in a family box (Sibler *et al.*, 1986; Andachi *et al.*, 1989) or unique wobble modifications found nowhere else such as N7-methylguanosine (m^7G) in squid and starfish mitochondria (Tomita *et al.*, 1998; Matsuyama *et al.*, 1998) and 5-formylcytidine (f^5C) in animal mitochondria (Takemoto *et al.*, 2009).

How the decoding rules impact codon reassignment

One consequence of the existing decoding rules is that there is no way to discriminate between U- and C-ending codons, so those codons must always be reassigned together (Figure 1.5A). Indeed, the only known reassignments of any U- or C-ending codon are in yeast mitochondria, where the entire leucine CUN codon box is reassigned to either threonine or alanine (Li and Tzagoloff, 1979; Ling *et al.*, 2014), and in sponge mitochondria, where the entire arginine CGN codon box is reassigned to glycine (Lavrov *et al.*, 2013). Similarly, the A- and G-ending codons should have their meanings linked when a U34 tRNA is present. These codons are often reassigned together, as we see in the reassignments of the UAR stop codons in multiple eukaryotic lineages and in Labyrinthulean mitochondria (Table A.3; Wideman *et al.* (2020)), the reassignment of CGR arginine codons in Absconditabacteria (Shulgina and Eddy, 2021), and reassignments of the AGR arginine codons in multiple mitochondrial lineages (Tables A.4,

A.6, and A.7).

Reassignments affecting fewer codons are more likely to evolve, since they typically require fewer compensatory codon substitutions. Within the known decoding rules, there are only three ways to reassign one and only one codon:

1. *Using a C34 tRNA to decode the G-ending codon.* G-ending codons are some of the most frequently reassigned codons. Since an unmodified C34 tRNA only reads a G-ending codon, the meaning of this codon can be altered by mutating this tRNA. The canonical stop codon UAG can be reassigned this way, especially if release factor recognition of UAG is lost. In some methanogenic archaea and bacteria, UAG is reassigned to pyrrolysine by a specialized $tRNA_{CUA}$ (Srinivasan *et al.*, 2002), and likewise in the microbial eukaryotes *Iotanema spirale* and Rhizarian exLh, the UAG codon is reassigned to either glutamine or leucine, respectively, also by a $tRNA_{CUA}$ (Pánek *et al.*, 2017). Reassigning a G-ending sense codon is more complicated because U34 tRNAs that are typically used to decode the A-ending codon can often read G-ending codons by wobble, potentially leading to ambiguous translation (Figure 1.5A). Eukaryotes avoid this problem in most family boxes by using an I34 tRNA to read the U-, C-, and A-ending codons (Figure 1.5B). If the now-redundant U34 tRNA is lost, then the G-ending codon is decoded by its own C34 tRNA and can be cleanly reassigned. Indeed, the only sense codon reassignment in eukaryotes uses this mechanism—in some yeast, the leucine codon CUG is reassigned to serine or alanine by a $tRNA_{CAG}$ while the CUU, CUC, and CUA codons are read by an I34 tRNA (Krassowski *et al.*, 2018). This mechanism is also used in bacterial reassignments of the CGG arginine codon (Shulgina and Eddy, 2021). The CGN codon box is an exception in bacteria, using a eukaryotic-like decoding scheme where an I34 tRNA

reads CGU, CGC, and CGA, leaving CGG to be read by its own $tRNA_{CCG}$ (Grosjean *et al.*, 2010). Once again, the G-ending codon is unlinked by the usage of an I34 tRNA elsewhere in the codon box, and can be reassigned without affecting the translation of other codons. This may partially explain why CGG is the most frequently reassigned sense codon in bacteria (Figure 1.6, Table A.1).

2. *Using restrictive modifications to decode the A-ending codon.* The meanings of the A- and G-ending codons are typically linked because most U34 modifications can recognize both codons. However, these codons are split occasionally—in some *Enterosoma* bacteria, the canonical arginine codon AGG is reassigned to methionine by a $tRNA^{Met}_{CCU}$ while AGA is still decoded as arginine by a $tRNA^{Arg}_{UCU}$, which may be able to misread AGG (Shulgina and Eddy, 2021). In Gracilibacteria and Absconditabacteria, the stop codon UGA is reassigned to glycine by a $tRNA^{Gly}_{UCA}$ (Rinke *et al.*, 2013; Campbell *et al.*, 2013), which may be able to misread the UGG tryptophan codon. Similarly, in *Euplotes* ciliates, UGA is read as cysteine by a $tRNA^{Cys}_{UCA}$ (Turanov *et al.*, 2009). These organisms might minimize misreading of the G-ending codon by using a U34 modification that severely limits G-ending codon recognition like 5-methylaminomethyluridine (Spanjaard *et al.*, 1990). Alternatively, the G-ending codon could be effectively unambiguous if the C34 tRNA outcompetes the U34 tRNA for reading the G-ending codon, especially if the U34 tRNA is expressed as lower levels (Mukai *et al.*, 2015).

   The only known modifications that can read an A-ending codon uniquely are k^2C34, C^+34, and Ψ34 which are used to read the AUA isoleucine codon in bacteria, archaea, and some eukaryotes, respectively (Yokoyama *et al.*, 1989; Mandal *et al.*, 2010; Szweykowska-Kulinska *et al.*, 1994). Thus, AUA could be reassigned separately from

other AUN codons by changing how the modified tRNA is aminoacylated; however AUA reassignments have not been observed, apart from AUA to methionine (described below). These modifications could also hypothetically be co-opted by a different tRNA which would require expanding the substrate specificity of the modifying enzymes.

3. *Expanding an existing codon box to include the A-ending codon.* The last way to reassign only one codon is by expanding an existing codon box. This mechanism is limited to UGA joining the tryptophan codon UGG to form a two-codon box, AUA joining the methionine codon AUG to form a two-codon box, or A-ending codons in a split box joining the XXY half of the split box (for example, AGA joining AGR serine codons).

   In bacteria, eukaryotes, and mitochondria, this occurs in the frequent reassignment of UGA to tryptophan, which creates a UGR tryptophan two-codon box. This can be achieved by mutating the typical tryptophan $tRNA_{CCA}$ anticodon to $tRNA_{UCA}$ which can read both UGA and UGG as tryptophan, and adopting the typical U34 modification for split codon boxes (Andachi *et al.*, 1989; Martin *et al.*, 1990; Sakurai *et al.*, 2005). A similar scenario happens in the reassignment of the isoleucine codon AUA to methionine in animal mitochondria. The canonical methionine codon AUG is decoded by a $tRNA^{Met}_{CAU}$. One could imagine expanding decoding to AUA by mutating the anticodon to $tRNA_{UAU}$ and using the typical split box U34 modification—indeed, this is what happens in tunicate mitochondria (Suzuki *et al.*, 2011). However, most other animal mitochondria instead have adopted a new position 34 modification, f^5C, which can read both AUA and AUG as methionine (Takemoto *et al.*, 2009). Animal mitochondria also have several examples of XXY two codon boxes expanding to include the XXA codon: in *Cephalodiscus* (hemichordate) mitochondria the stop codon UAA has been co-opted

to be a tyrosine codon, joining the canonical tyrosine codons UAU and UAC (Li *et al.*, 2019); in echinoderm and Platyhelminthes mitochondria, the AAA lysine codon has been reassigned to asparagine, expanding the asparagine AAY codon box to include AAA (Himeno *et al.*, 1987; Telford *et al.*, 2000); and in some hemichordate and arthropod mitochondria, the AGA arginine codon has been reassigned to serine, merging with the AGY serine codon box (Tomita *et al.*, 1999a; Perseke *et al.*, 2011). Unusually, these reassignments appear to use a G34 tRNA to read the A-ending codon. It seems that decoding of A-ending codons by unmodified G34 tRNAs may be possible in mitochondria, perhaps aided by additional anticodon loop mutations (Tomita *et al.*, 1999b; Watanabe and Yokobori, 2011; Li *et al.*, 2019), and has also been reported for decoding of the AUA isoleucine codon in echinoderm mitochondria (Yokobori *et al.*, 2001). Expansion of XXY codon boxes to include the A-ending codon could also hypothetically evolve in bacteria and eukaryotic nuclear genomes as well by co-opting the I34 modification to read the U-, C-, and A-ending codons and leaving the C-ending codon to be read by a C34 tRNA.

Mutations that affect tRNA decoding

Changes to how a codon is decoded by tRNAs can evolve by several routes. The simplest is by direct mutation to the anticodon of an existing tRNA or a duplicated paralog, leading to a new codon being recognized by that tRNA. Early work on mutation suppression found that a missense mutation of a glycine GGA codon to an arginine AGA codon can be rescued (i.e. suppressed) by a mutation of the $tRNA^{Gly}_{UCC}$ anticodon to $tRNA_{UCU}$, creating a "suppressor tRNA" for GGA to AGA missense mutations (Roberts and Carbon, 1974). The suppressor $tRNA_{UCU}$ can presumably decode AGA while still being charged with glycine. Analogous missense suppressor tRNAs were reported for mutations of the glycine GGG codon to glutamine GAG (Hill

et al., 1974) and of the glycine GGU codon to cysteine UGU (Carbon and Fleck, 1974), showing that recognition of a new codon could be achieved by a variety of anticodon mutations. Anticodon mutations have likely been a part of many natural codon reassignments; however, it is hard to distinguish whether a tRNA gene involved in reassignment evolved by anticodon mutation or by a change in tRNA aminoacylation without tracing the evolutionary history to an outgroup. In Mycoplasmatales and Entomoplasmatales, the canonical stop codon UGA is decoded as tryptophan by a $\text{tRNA}^{\text{Trp}}_{\text{UCA}}$ which is genomically encoded directly adjacent to a $\text{tRNA}^{\text{Trp}}_{\text{CCA}}$ gene, suggesting that they may have evolved by tandem gene duplication and subsequent anticodon mutation (Yamao *et al.*, 1988; Citti *et al.*, 1992).

Another way to create a new tRNA anticodon is by altering tRNA intron splicing. Some tRNAs have an intron that must be removed, typically in between position 37 and 38 of the anticodon loop (Schmidt and Matera, 2020). In *Candida* yeasts that have reassigned CUG from leucine to serine, the new $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ is thought to have evolved by an insertion in the anticodon loop of the intron-containing $\text{tRNA}^{\text{Ser}}_{\text{AGA}}$ and a shift in the intron splice sites, creating a new anticodon (Yokogawa *et al.*, 1992). Other anticodon loop mutations can also alter decoding—in *Saccharomyces* yeast mitochondria, the canonical CUN leucine codons are decoded as threonine by an unusual threonine tRNA with an 8 nt anticodon loop (Li and Tzagoloff, 1979). Furthermore, mutations to other parts of the tRNA body may affect decoding in unexpected ways. A classic example is the Hirsh suppressor, an *E. coli* tRNA found to suppress UGA nonsense mutations by a mutation in the D-arm of a $\text{tRNA}^{\text{Trp}}_{\text{CCA}}$ (Hirsh, 1971).

The decoding capacity of a tRNA can also be changed via the wobble position 34 modification. Mutations to the tRNA itself could abolish or introduce recognition by a tRNA nucleotide modifying enzyme, or mutations to the modifying enzyme could change which

tRNAs are modified. The loss of a modification could manifest as either a gain or a loss of codon decoding by a tRNA, depending on the modification. For example, if animal mitochondria were to lose the f^5C34 modification on $tRNA^{Met}_{CUA}$, then the capacity to translate AUA as methionine would be lost. In contrast, if a $tRNA^{Arg}_{UCU}$ were to lose a restrictive U34 modification, then it could misread the AGU and AGC serine codons as arginine and initiate reassignment of those codons by the ambiguous intermediate mechanism. One interesting example of codon reassignment through tRNA modification is in *Leishmania* mitochondria, which have reassigned UGA from stop to tryptophan. *Leishmania* mitochondria do not encode their own tRNAs but instead import the cytoplasmic $tRNA^{Trp}_{CCA}$ and modify the anticodon to $tRNA^{Trp}_{UCA}$ within the mitochondrion to expand tryptophan decoding to UGA (Alfonzo *et al.*, 1999). Other mitochondria appear to have invented new modifications to enable codon reassignments, although its possible that the new modifications emerged as an adaptation to improve translation efficiency once the codon reassignment has established. These include f^5C34 in animal mitochondria for reading AUA as methionine (Takemoto *et al.*, 2009) and m^7G34 in echinoderm and squid mitochondria for reading AGA and AGG as serine (Tomita *et al.*, 1998; Matsuyama *et al.*, 1998).

The most extreme way to alter mRNA decoding by tRNAs is through changes to the ribosome. Changes that cause widespread mistranslation of many codons are unlikely to persist. More subtle changes to the relative positioning of the A-site tRNA anticodon and mRNA codon could hypothetically permit new types of wobble decoding at codon position 3. In *Mycoplasma mobile*, the three isoleucine codons (AUU, AUC, and AUA) are decoded by a $tRNA^{Ile}_{UAU}$ with an unmodified U34, differing from the typical bacterial strategy for decoding this codon box (Figure 1.5B). Using an *in vitro* assay to measure tRNA binding to ribosomes

with either an AUA or AUG codon in the A site, Taniguchi *et al.* (2013) showed that when using *E. coli* ribosomes, the $tRNA^{Ile}_{UAU}$ bound to both AUA and AUG codons, but when using *M. mobile* ribosomes the $tRNA^{Ile}_{UAU}$ specifically bound AUA and not AUG. Differences in $tRNA^{Ile}_{UAU}$ decoding behavior may stem from differences between the ribosomes, indicating the potential for ribosome mutations to affect decoding rules.

Aminoacylation and tRNA identity

After a tRNA is transcribed and processed into a mature tRNA, it is aminoacylated by an aminoacyl-tRNA synthetase (aaRS). aaRSs have a binding pocket that specifically binds and activates one amino acid, and a tRNA binding surface which recognizes the set of tRNAs destined for the specific amino acid (called isoacceptor tRNAs) through unique sequence and structural features (called tRNA identity elements) (Ibba and Söll, 2000). There is typically one aaRS for each amino acid; an exception being that many bacteria and archaea lack an asparaginyl- or glutaminyl-tRNA synthetase (i.e. AsnRS, GlnRS) and instead attach aspartate or glutamate to the tRNA and subsequently convert it to the correct amino acid by amidation (Ibba *et al.*, 2000).

Much of what is known about the tRNA identity elements used by each aaRS comes from *in vitro* experiments measuring the effect of tRNA substitutions on the kinetics of aminoacylation (Pallanck *et al.*, 1994; Giegé *et al.*, 1998). For most aaRSs, the identity elements are concentrated in the tRNA acceptor stem, especially the "discriminator base" N73, and the tRNA anticodon loop, in particular the anticodon itself (Pallanck *et al.*, 1994). tRNA features that enable recognition by an aaRS are called positive determinants. An example of a strong positive determinant is the anticodon CAU for *E. coli* MetRS. Changes to the $tRNA^{Met}_{CAU}$ anticodon, especially at C34 and U36, decrease the rate of *in vitro* methionylation by over four orders of

magnitude (Schulman and Pelka, 1983). Furthermore, introducing the CAU anticodon into a $tRNA^{Val}$ makes it an excellent MetRS substrate comparable to $tRNA^{Met}_{CAU}$ (Schulman and Pelka, 1988), indicating that features elsewhere on the tRNA body are not as important for MetRS recognition. Other features, called negative determinants, exist to prevent mischarging by a competing aaRS (Giegé *et al.*, 1998). For example, the m^1G37 modification of the *E. coli* $tRNA^{Asp}$ is not needed for recognition by AspRS *in vitro* but serves to abolish recognition by ArgRS (Perret *et al.*, 1990). tRNA identity elements usually do not depend on nucleotide modification (Giegé *et al.*, 1998) outside of a few notable exceptions (like m^1G37 in *E. coli* $tRNA^{Asp}$) (Muramatsu *et al.*, 1988; Perret *et al.*, 1990; Sylvers *et al.*, 1993; Suzuki *et al.*, 1997). Another example is in bacteria, where the $tRNA^{Ile}_{CAU}$ has a lysidine (k^2C) modification at the wobble position which enables it to decode AUA instead of the methionine codon AUG. The k^2C34 modification is required for recognition by the IleRS and also serves as a negative determinant against MetRS (Muramatsu *et al.*, 1988). Crystal structures of tRNA-bound aaRSs have confirmed that many experimentally-derived tRNA identity elements are directly contacted by aaRS residues (Giegé *et al.*, 1998; Ibba and Söll, 2000). For example, an important glutamine identity element in *E. coli* is a weak first base pair in the acceptor stem (Celis *et al.*, 1973; Jahn *et al.*, 1991). Consistent with the biochemical evidence, the crystal structure shows that GlnRS binds the tRNA in a conformation where the first acceptor stem base pair is forced apart by a leucine residue (Rould *et al.*, 1989). Some identity elements are evolutionarily highly conserved. The most well-known example is the G3-U70 acceptor stem pair, which is the predominant alanine identity element in all three domains of life (even present in mitochondria but shifted one position) (Hou and Schimmel, 1989; Lovato *et al.*, 2001). Other identity elements change across the phylogeny (Shiba *et al.*, 1994; Stehlin *et al.*, 1998), presumably

reflecting co-evolution between the isoacceptor tRNAs and the aaRS tRNA binding surface.

These findings are complemented by *in vivo* experiments that use a reporter protein with an in-frame stop codon to assay what amino acid is incorporated by different suppressor tRNAs (i.e. tRNAs that can decode stop codons) (Normanly *et al.*, 1990; Pallanck *et al.*, 1994). These experiments report on the outcome of all aaRSs competing to charge the suppressor tRNA, which is dependent on the cellular aaRS concentrations and can lead to a measurable level of suppression even with impaired aminoacylation kinetics (Sherman *et al.*, 1992). This mimics what would happen in a real-world codon reassignment; however, the results are limited to tRNAs that decode stop codons. The most surprising finding is that suppressor tRNAs for all twenty amino acids could not be created by simply mutating the anticodon (Normanly *et al.*, 1990). In particular, UAG suppressors generated from $tRNA^{fMet}$, $tRNA^{Trp}$ and $tRNA^{Ile}$ lead to glutamine insertion and UAG suppressors generated from $tRNA^{Met}$, $tRNA^{Asp}$, $tRNA^{Arg}$, $tRNA^{Thr}$, $tRNA^{Ile}$, and $tRNA^{Val}$ lead to lysine insertion (Normanly *et al.*, 1990). Since most aaRSs rely on the anticodon as a positive determinant, it's possible that mutating the anticodon to generate a suppressor tRNA may abolish recognition and the orphaned tRNAs are charged to some degree by GlnRS or LysRS because of some underlying affinity for the UAG suppressor anticodon (Pallanck *et al.*, 1994). Alternatively, GlnRS has been proposed to have some general non-specificity and may aminoacylate tRNAs that no other aaRS accepts (Sherman *et al.*, 1994). Furthermore, tRNAs aminoacylated with glutamine bind to EF-Tu more tightly, so glutamine incorporation may also reflect selection by EF-Tu on the pool of aminoacylated tRNAs (LaRiviere *et al.*, 2001).

Mutations that affect tRNA aminoacylation

The gain or loss of translation as a certain amino acid can come from changes to how the cognate tRNA is aminoacylated. The simplest route would be tRNA mutations that introduce or disrupt positive or negative determinants for an aaRS. There are several examples of this occurring in nature. In *Candida* yeasts that have reassigned the CUG codon from leucine to serine, the new $tRNA^{Ser}_{CAG}$ is mischarged with leucine about 3% of the time (Suzuki *et al.*, 1997). Recognition by the LeuRS *in vivo* and *in vitro* comes from the presence of the leucine identity element m^1G37 in the anticodon loop (Suzuki *et al.*, 1997). In uncultivated *Enterosoma* bacteria, the arginine codon AGG has been reassigned to methionine and is decoded by a $tRNA^{Met}_{CCU}$ (Shulgina and Eddy, 2021). By tracing the syntenic position of the $tRNA_{CCU}$ to outgroup species that translate AGG as arginine, it appears that methionine decoding is in part due to tRNA mutations that shifted aminoacylation from ArgRS to MetRS, such as loss of the arginine identity element A20 in the D loop and gain of methionine identity elements in the acceptor stem (Shulgina and Eddy, 2021). Another example of tRNA identity switching is in yeast mitochondria—in Saccharomycetaceae, the entire leucine CUN codon box is reassigned to threonine by a unique $tRNA^{Thr}$ (Li and Tzagoloff, 1979). However, the genus *Eremothecium* has reassigned CUN once more to alanine using a $tRNA^{Ala}$ that is very similar in sequence to $tRNA^{Thr}$ but now contains the strong alanine identity element G3-U70 in the acceptor stem (Ling *et al.*, 2014).

A change in tRNA charging could also come from mutations to aaRS binding of tRNAs. This has not been demonstrated to underlie any natural codon reassignments, but is hypothetically possible, based on examples of aaRS determinants changing across the phylogeny (Shiba *et al.*, 1994; Stehlin *et al.*, 1998; Taniguchi *et al.*, 2013). In the lab, mutants of *E. coli* GlnRS

have been selected that can mischarge a $tRNA^{Tyr}$ *in vivo* and have single amino acid substitutions at the GlnRS-tRNA interface (Perona *et al.*, 1989). Mutations could also affect the amino acid specificity of the aaRS and thus change which amino acid is incorporated at codons. In *Mycoplasma*, the LeuRS lacks an editing domain, resulting in misincorporation of isoleucine, valine, and methionine at leucine codons (Li *et al.*, 2011). Also, much effort has been put into designing aaRSs that take new amino acid substrates, as part of attempts to engineer an expanded genetic code (Perona and Hadd, 2012). For a codon reassignment, changes to amino acid specificity would have to occur in a duplicated aaRS gene in order to leave some codons for the original amino acid unaffected.

Even if a tRNA is successfully aminoacylated as a new amino acid, this does not guarantee incorporation into proteins. This is because aminoacylated tRNAs must form a ternary complex with EF-Tu (in bacteria; EF1A in eukaryotes and archaea) and GTP (Dever *et al.*, 2018; Rodnina, 2018). The binding affinity of EF-Tu to aminoacylated tRNAs is an additive contribution of amino acid binding and tRNA binding such that naturally occurring amino acid-tRNA pairs bind EF-Tu with similar tightness (LaRiviere *et al.*, 2001). In a codon reassignment, novel combinations of amino acids and tRNAs might have overall EF-Tu binding that is too weak, such that the ternary complex fails to form, or too strong, such that tRNA accommodation is impaired. Thus, altering tRNA aminoacylation might also require compensatory changes to the tRNA sequence (or EF-Tu) to modulate ternary complex formation (Uhlenbeck and Schrader, 2018).

1.3 Why do some codon reassignments convergently re-appear?

Even a cursory examination of the list of alternative genetic codes reveals that some codon reassignments appear much more frequently than others. For example, the reassignment of UGA to tryptophan has evolved at least 22 times, and most known codon changes affect either stop codons or arginine codons (Figure 1.6). Meanwhile, most of the 64 codons have no reported reassignments, including any proline, valine, glutamate, and histidine codons as an example (Figure 1.6). Of course, reassignments of these codons may yet be discovered. Sense codon reassignments are more difficult to detect from a genome sequence than stop codon changes because they don't lead to the appearance of in-frame "stop" codons in coding regions but instead appear as a subtle pattern of amino acid substitutions. Computational approaches have been developed to predict the genetic code from nucleotide sequence (Telford *et al.*, 2000; Abascal *et al.*, 2006b; Dutilh *et al.*, 2011; Mühlhausen and Kollmar, 2014b; Noutahi *et al.*, 2017; Shulgina and Eddy, 2021), including one tool which can be used at scale to analyze tens of thousands of genomes (Shulgina and Eddy, 2021). However, a comprehensive screen for alternative genetic codes has only been performed for bacterial and archaeal genomes (Shulgina and Eddy, 2021), yeast nuclear genomes (Mühlhausen and Kollmar, 2014a; Krassowski *et al.*, 2018; Shulgina and Eddy, 2021), green algal and animal mitochondrial genomes (Noutahi *et al.*, 2017, 2019; Žihala and Eliáš, 2019), green algal nuclear genomes (Cocquyt *et al.*, 2010; Li *et al.*, 2021), and marine microbial eukaryotes (Swart *et al.*, 2016; Heaphy *et al.*, 2016). For other parts of the tree of life, alternative genetic codes are reported in a potentially biased way and may not accurately represent the true genetic code diversity.

Setting aside potential observation bias, there may be biological reasons for why some codon reassignments may evolve more readily, ranging from genome evolutionary forces to

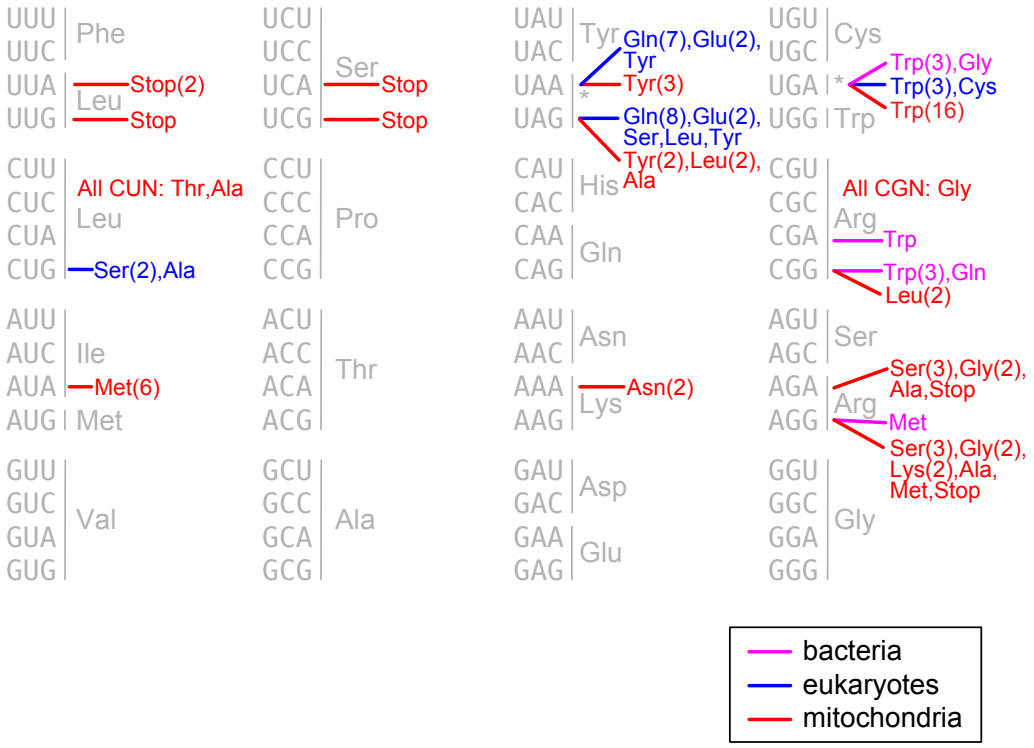

Figure 1.6: The standard genetic code annotated with all of the codon reassignments reported in bacteria, eukaryotes, and mitochondria. For each codon reassignment, a conservative estimate for how many times the reassignment has evolved is provided in parentheses, based on data and references in Tables A.1-A.7.

constraints imposed by the existing translational machinery. In this section, I will delve into each of these potential reasons that may favor the emergence of some codon reassignments over others.

Codon reassignment due to horizontal transfer

A codon reassignment may appear to have emerged multiple times "independently" if it evolved once but was then horizontally transferred into other organisms. A clear example of this is the case of pyrrolysine. Pyrrolysine is genomically encoded at UAG positions in many methanogenic archaea and bacteria (Table A.2) by its own $tRNA^{Pyl}_{CUA}$ and PylRS and is used at three positions in methylamine methyltransferases (Gaston *et al.*, 2011) . All of the

pyrrolysine biosynthesis and encoding machinery is contained within a single gene cluster and the sporadic phylogenetic distribution of pyrrolysine usage is almost certainly a result of horizontal gene transfer of the gene cluster and not convergent evolution (Longstaff *et al.*, 2007; Fournier and Andam, 2012; Guo *et al.*, 2022).

Other reassignments could also transfer horizontally. Coding regions would still need to undergo compensatory substitutions to adapt to the new codon translation, and loss of the original meaning must occur for unambiguous reassignment. However, for all other codon reassignments—outside of viral genetic codes co-evolving to match that of their hosts (Shackelton and Holmes, 2008)—independent evolution is assumed due to the absence of evidence for horizontal transfer.

Some codons may be targets for reassignment due to genome evolutionary forces

Since the main barrier to codon reassignment are the substitutions to remove the codon from positions that cannot tolerate the new translation, some codons may be reassigned more often because they tend to become rare for reasons unrelated to their translation. In the standard genetic code, the number of codons allocated to an amino acid is roughly proportional to the usage of that amino acid in proteins (King and Jukes, 1969). Some amino acids, in particular arginine, deviate from this trend and have a disproportionate number of codons given their usage (King and Jukes, 1969; Jukes *et al.*, 1975; Holmquist, 1978), so each codon for these amino acids has lower usage on average. Usage of some synonymous codons is further disfavored by pressures like biased genomic GC content (Osawa *et al.*, 1988). When the rate of GC to AT mutations does not equal the reverse rate (Sueoka, 1962; Lind and Andersson, 2008), mutations to some synonymous codons are favored over others (Osawa *et al.*, 1988). Amino acids whose codons cover a wider range of GC contents (like arginine and leucine) may have a greater

difference in usage between the favored and disfavored synonymous codon. The rarest codons may be more vulnerable to reassignment—low GC content together with genome reduction is thought to play a major role in the reassignments of UGA stop codons and CGA and CGG arginine codons in low GC bacteria (Osawa and Jukes, 1989; Rinke *et al.*, 2013; Shulgina and Eddy, 2021) and UAG and UGA stop codons, CGG arginine codons, and CUN leucine codons in low GC mitochondria (Sengupta *et al.*, 2007; Noutahi *et al.*, 2019) and may explain why these codons are repeatedly reassigned in general. Similarly, AU-rich codons like AUA isoleucine, UUR leucine, and AGR arginine codons could be targets for reassignment in high GC genomes (Ohama *et al.*, 1990a).

The background frequencies of nucleotide mutations could bias which reassignments evolve. Single nucleotide mutations are more likely to generate nonsense and missense mutations from certain source codons (Miller *et al.*, 1978). For example, a nonsense UGA stop codon is more likely to originate from a tryptophan UGG codon by a single transition mutation than from a GCC alanine codon by three transversions. Thus, reassignments of UGA could favor tryptophan over other amino acids, in part to rescue pseudogenes generated from nonsense mutations of UGG tryptophan codons (or another codon depending on the spectrum of nucleotide mutations). This sort of process might also bias sense codon reassignments. For example, high genomic GC content might drive substitutions of lysine AAA and AAG codons to arginine AGA and AGG codons at residues that can tolerate either positively charged amino acid. In this situation, AGR codons already occur at former lysine residues so potential reassignments of AGR codons might be biased towards lysine rather than other amino acids.

Some codon reassignments are more likely due to tRNA decoding

Reassignments that affect fewer codons are likely easier to evolve because they require fewer compensatory codon substitutions. Due to how codons are recognized by release factors and tRNAs, some codons are simpler to reassign individually than others. This is discussed at length in the Section 1.2, but will be briefly reiterated here. In bacteria, UAG and UGA stop codons are simpler to reassign than UAA because termination at UAG and UGA can be lost by deleting one of the class 1 release factor genes, while UAA reassignment requires changing release factor specificity for stop codons. Consistent with this prediction, almost all known stop codon reassignments in bacteria and mitochondria affect either UGA or UAG (Figure 1.6, Tables A.1, A.4-A.7). For sense codons, changing the meaning of just one codon is limited to G-ending codons by a G34 tRNA, A-ending codons by a U34 tRNA with a very restrictive wobble modification, or by expansion of an existing codon box. All known reassignments of only one sense codon, across all domains of life, fall into one of these categories (Tables A.1-A.7). In particular, UGA to tryptophan is by far the most common reassignment and can be explained by the concurrent ease of losing UGA decoding by RF2 deletion (in bacteria and mitochondria) and gain of translation as tryptophan by a single nucleotide change to the anticodon of $tRNA^{Trp}_{CCA}$. Similarly, AUA isoleucine codons are recurrently reassigned to methionine to create a two-codon box, and AAA lysine and AGA arginine codons are merged with the accompanying XXY two-codon boxes in mitochondria.

tRNA decoding may also bias which amino acid a codon is reassigned to. Any particular anticodon is more likely to originate by a single nucleotide change from an already similar anticodon, potentially biasing reassignments to amino acids decoded by near-cognate tRNAs. Indeed, most known codon reassignments are a single nucleotide substitution away from other

codons for the new amino acid (Figure 1.6); however, this pattern could also be explained by patterns of missense and nonsense mutations (described in the previous subsection) or the anticodon preferences of aminoacyl-tRNA synthetases (described in the next subsection).

The background rate of decoding errors might also bias reassignments to amino acids read by near-cognate tRNAs. The original proposal of the ambiguous intermediate model suggested that reassignments might be initiated by a low level of ambiguous decoding by near-cognate tRNAs, which then primes reassignment to those amino acids (Schultz and Yarus, 1994, 1996). In support of this idea, in eukaryotes, stop codons are read through most often by the near-cognate $tRNA^{Trp}$, $tRNA^{Arg}$, and $tRNA^{Cys}$ for the UGA stop codon and by the near-cognate $tRNA^{Gln}$, $tRNA^{Tyr}$, and $tRNA^{Lys}$ for UAA and UAG stop codons (Roy *et al.*, 2015), mirroring the most frequent stop codon reassignments in eukaryotes (Figure 1.6). Furthermore, in the reassignment of UGA to tryptophan in *Condylostoma magnum*, the cognate $tRNA^{Trp}_{UCA}$ cannot be found in the genome assembly or by tRNA sequencing suggesting that decoding by the near-cognate $tRNA^{Trp}_{CCA}$ may persist (Swart *et al.*, 2016).

Codon reassignments may be constrained by aaRS anticodon specificity

Many aaRSs use the tRNA anticodon as an identity element for aminoacylation (Pallanck *et al.*, 1994; Giegé *et al.*, 1998). Therefore, not every amino acid identity - tRNA anticodon combination may be possible without additional changes to the aaRS. Codon reassignments may be biased by which new tRNA anticodons an aaRSs could accept as a substrate, with the nature of the restriction dependent on which anticodon nucleotides are used as identity elements and the degree of importance. All known codon reassignments involve a single nucleotide change from the canonical codons for the new amino acid, with the following exceptions: CUG to serine and alanine in yeast nuclear genomes, CUN to threonine and alanine in yeast

mitochondria, and UAG, AGA, and AGG to alanine in green algal mitochondria (Figure 1.6). Incidentally, AlaRS and SerRS are some of the few aaRSs that do not use the anticodon as an identity element (Giegé *et al.*, 1998), which may explain why they are enriched in this set of exceptions. Since these aaRSs should be more amenable to accepting new anticodons, we may expect to see more reassignments of other codons to these amino acids (Kollmar and Mühlhausen, 2017a).

Furthermore, some aaRSs may have a "hidden specificity" for some non-cognate anticodons, based on shared anticodon identity elements. This idea comes from the work of Normanly *et al.* (1990), which showed that UAG suppressor tRNAs derived from some tRNAs ended up being charged with glutamine or lysine instead of the expected amino acid. It appears that if a UAG suppressor $tRNA_{CUA}$ has lost recognition by the original aaRS due to the anticodon change, it may be secondarily picked up by GlnRS or LysRS based on some underlying affinity for the CUA anticodon, which is similar to the typical CUG glutamine and CUU lysine anticodons. In eukaryotes, reassignments of the UAR stop codons are heavily biased towards glutamine (Table A.3), perhaps reflecting a preference for these anticodons by GlnRS. Furthermore, other codons may have their own set of "second best" aaRSs which are likely to takeover aminoacylation of tRNAs if recognition by the original aaRS is lost. This may potentially explain the frequency of reassignments to amino acids whose typical codons are one nucleotide away. Because tRNA aminoacylation is the outcome of a multi-way competition between all aaRSs, it is hard to predict exactly how an arbitrary tRNA will be charged without a more complete understanding of tRNA identity elements, the cellular concentrations of each aaRS, and the effect of EF-Tu binding affinity for different amino acids and tRNAs.

1.4 Are codon reassignments selectively advantageous?

So far, I have reviewed the mechanisms of codon reassignment, but have not yet addressed the most basic question—why would the genetic code change at all? Indeed, the standard genetic code is extremely optimized to minimize the biochemical consequences of nucleotide mutations and mistranslation (Freeland and Hurst, 1998), and many alternative genetic codes are worse than the standard genetic code by these measures (Freeland *et al.*, 2000). One possibility is that alternative genetic codes are not selected upon but are merely consequences of neutral evolution. This may well be the case for codon reassignments that evolve by a "codon capture"-like trajectory, where a codon falls from usage by genetic drift in genomes with biased GC content or reduced size (Osawa and Jukes, 1989). Low codon usage may release purifying selection on accurate and efficient translation of the codon, allowing for mutations to accumulate in the translational machinery that may change the meaning of the codon. In contrast, evolutionary trajectories that involve a substantial ambiguous or unassigned intermediate state must either be driven by positive selection for the reassignment or be in a regime where selection is inefficient.

A codon reassignment could be strongly advantageous if it imparts new biochemical capabilities. This is likely true in the case of the UAG reassignment to pyrrolysine, which is used by methanogenic archaea and bacteria in the active sites of some methyltransferases to correctly position the methylamine substrate (Krzycki, 2004), and may explain why this reassignment has readily horizontally transferred across many methanogens despite presumably causing readthrough at UAG stop codons (Zhang *et al.*, 2005; Guo *et al.*, 2022).

All other known reassignments are limited to the 20 canonical amino acids. It has been proposed that ambiguous translation of a codon may be selectively advantageous and this

may propel some reassignments by the "ambiguous intermediate" mechanism. Studies in yeast have shown that expressing a tRNA that causes mistranslation can improve survival in a variety of environmental conditions, perhaps by preemptively triggering a stress response (Santos *et al.*, 1999). Ambiguity in translation also diversifies the proteome, which could potentially enable adaptation to new environments akin to "mutator" alleles with high rates of DNA mutation (Taddei *et al.*, 1997).

A change in the genetic code may make coding sequences from other organisms unintelligible, potentially limiting the ability of the reassigned organism to partake in horizontal gene transfer. On the other hand, this may also protect the reassigned organism from viruses, selfish genetic elements, and other agents that hijack the host's translational machinery (Shackelton and Holmes, 2008). Indeed, a major reason to engineer organisms with alternative genetic codes is to protect industrial cultures from decimation by viruses (Kuo *et al.*, 2018). A strain of *E. coli* engineered to have two codon reassignments has been shown to be resistant to infection by a wide range of phage (Nyerges *et al.*, 2022). However, natural alternative genetic codes likely co-evolve together with their viruses, allowing time to adapt (Shackelton and Holmes, 2008; Taylor *et al.*, 2013). Conversely, some phage have been reported to use different genetic codes from their hosts, especially in genes expressed in the lytic stage (Borges *et al.*, 2022). This may benefit the phage by disfavoring translation of host genes and tightly controlling the switch from lysogeny to lysis on a translational level (Ivanova *et al.*, 2014; Borges *et al.*, 2022). A codon reassignment might also result from a pressure to eliminate a specific tRNA—the repeated reassignment of CUG in yeasts has been proposed to be driven by selection to avoid usage of $\mathrm{tRNA}^{\mathrm{Leu}}_{\mathrm{CAG}}$ which may have been specifically degraded by a tRNA endonuclease toxin of a yeast killer plasmid; however, no such killer plasmid is currently known to exist

(Krassowski *et al.*, 2018).

In mitochondria, some codon reassignments have been proposed to have a selective advantage because they reduce the number of tRNA genes in the genome (Andersson and Kurland, 1991, 1995). Mitochondria are under strong pressure to reduce their genome size—animal mitochondrial genomes only encode 13 protein-coding genes in addition to the tRNAs and rRNAs needed for translation (Wolstenholme, 1992). Thus, codon reassignments which allow the genetic code to be read by fewer tRNAs can help further reduce the genome size. For example, reassignments of UGA from stop to tryptophan allows the deletion of RF2, reassignment of the AUA isoleucine codon to methionine allows for a single tRNA to read both codons, and reassignments of AGR arginine codons to serine allow for the entire AGN codon box to be read by a single serine tRNA (Andersson and Kurland, 1991, 1995). Other mitochondrial codon reassignments do not result in net loss of a tRNA and may have evolved due to other reasons.

Lastly, a codon reassignment may be selectively advantageous if it can rescue a deleterious nonsense or missense mutation, such that the fitness advantage outweighs the cost of ambiguous translation at other positions for the codon. This may seem like a crude solution to rescue a mutation; however, from the earliest days of molecular biology, selection for mutants that can suppress a nonsense or missense mutation have yielded tRNA mutants (Gorini, 1970). It may be possible that natural codon reassignments may also be initially spurred by the ability to rescue specific mutations.

1.5 Conclusion

At first glance, the existence of alternative genetic codes is surprising and counterintuitive—the genetic code dictates how every single protein is interpreted from genetic information, so any change should be intolerably disruptive. We can better understand how and why alternative genetic codes evolve by keeping several main principles in mind. First, the primary barrier to codon reassignment are the compensatory substitutions needed to correct protein sequences that cannot tolerate the change in codon meaning. These substitutions can be overcome by neutral evolution in some scenarios or by selection against usage of that codon. This principle helps us understand in what situations a codon reassignment may be more likely to evolve, the sorts of evolutionary pressures that may drive these substitutions, and which codons may be more susceptible to these pressures. There are many reasons why alternative genetic codes could evolve in the wild but they are hard to test. In bacteria, the occurrence of codon reassignments in reduced, GC biased genomes points to neutral "codon capture"-like processes being predominant. In mitochondria, reduced genome sizes likely enable reassignments as well, in addition to pressure to reduce the number of tRNA genes. It is less well understood why codon reassignments evolve in eukaryotes, where losing a codon due to neutral processes is less likely. A more complete distribution of codon reassignments in eukaryotes and in depth study of each reassignment may help shed light on this question.

Focusing on the molecular details of how the translational machinery decodes stop and sense codons can help us understand what changes are required for different codon reassignments. Codon reassignments that can be implemented with minimal changes to the translational machinery should be more likely to evolve. For example, due to existing tRNA decoding rules, some codons are easier to assign separate meanings, and due to the tRNA features

used for discrimination by aaRSs, reassignment to some amino acids may be more likely than others. Known codon reassignments are largely consistent with our understanding of what changes should be possible to the translational machinery. An incomplete understanding of how changes to release factors affect stop codon recognition limits our ability to predict which sense-to-stop reassignment could hypothetically evolve. Likewise, an incomplete understanding of how tRNA mutations will affect the outcome of competition between all aminoacyl-tRNA synthetases limits our ability to predict which amino acid assignments are more likely. A better understanding of these processes will help us understand why certain codon reassignments evolve more readily, and conversely, studying codon reassignments can illuminate how the translational machinery capable of changing. New genetic codes continue to push the boundaries on what was thought to be possible in translation—from multiple genetic codes used in a single phage genome (Ivanova *et al.*, 2014; Borges *et al.*, 2022) to the complete loss of dedicated stop codons in ciliates (Swart *et al.*, 2016). As more of the microbial diversity is uncovered by sequencing efforts, it will be interesting to see if we find new genetic codes that violate any other expectations. These examples could provide a deeper understanding of the extremely conserved process of translation and new avenues for engineering.

Preface to Chapter 2

Chapter 2 was published separately and reprinted here with minor modifications to update taxon names and figure references. Supplementary figures can be found in Appendix B. Supplementary files and source data files can be downloaded at `http://eddylab.org/publications/Shulgina21/Shulgina21-supplement.tar.gz`.

Shulgina, Y. and Eddy, S. R. (2021). A computational screen for alternative genetic codes in over 250,000 genomes. *eLife*, **10**, e71402.

Chapter 2

A computational screen for alternative genetic codes in over 250,000 genomes

2.1 Abstract

The genetic code has been proposed to be a “frozen accident", but the discovery of alternative genetic codes over the past four decades has shown that it can evolve to some degree. Since most examples were found anecdotally, it is difficult to draw general conclusions about the evolutionary trajectories of codon reassignment and why some codons are affected more frequently. To fill in the diversity of genetic codes, we developed Codetta, a computational method to predict the amino acid decoding of each codon from nucleotide sequence data. We surveyed the genetic code usage of over 250,000 bacterial and archaeal genome sequences in GenBank and discovered five new reassignments of arginine codons (AGG, CGA, and CGG), representing the first sense codon changes in bacteria. In a clade of uncultivated Bacilli in the genus *Enterosoma*, the reassignment of AGG to become the dominant methionine codon likely

evolved by a change in the amino acid charging of an arginine tRNA. The reassignments of CGA and/or CGG were found in genomes with low GC content, an evolutionary force which likely helped drive these codons to low frequency and enable their reassignment.

2.2 Introduction

The genetic code defines how mRNA sequences are decoded into proteins. The ancient origin of the standard genetic code is reflected in its near-universal usage, once proposed to be a "frozen accident" that is too integral to the translation of all proteins to change (Crick, 1968). However, the discovery of alternative genetic codes in over 30 different lineages of bacteria, eukaryotes, and mitochondria over the past four decades has made it clear that the genetic code is capable of evolving to some degree (Knight *et al.*, 2001a; Kollmar and Mühlhausen, 2017b).

The first alternative genetic codes were discovered by comparing newly sequenced genomes to amino acid sequences obtained by direct protein sequencing. Nonstandard codon translations were found this way in human mitochondria (Barrell *et al.*, 1979), *Candida* yeasts (Kawaguchi *et al.*, 1989), green algae (Schneider *et al.*, 1989), and *Euplotes* ciliates (Meyer *et al.*, 1991). Some reassignments of stop codons to amino acids were detected from DNA sequence alone, based on the appearance of in-frame stop codons in critical genes (Yamao *et al.*, 1985; Caron and Meyer, 1985; Cupples and Pearlman, 1986; Keeling and Doolittle, 1996; McCutcheon *et al.*, 2009; Campbell *et al.*, 2013; Záhonová *et al.*, 2016). As DNA sequence data has accumulated faster than direct protein sequences, computational methods have been developed to predict the genetic code from DNA sequence. The core principle of most methods is to align genomic coding regions to homologous sequences in other organisms (creating

multiple sequence alignments) and then to tally the most frequent amino acid aligned to each of the 64 codons. This approach led to the discovery of new genetic codes in screens of ciliates (Swart *et al.*, 2016; Heaphy *et al.*, 2016), yeasts (Riley *et al.*, 2016; Krassowski *et al.*, 2018), green algal mitochondria (Noutahi *et al.*, 2019; Žihala and Eliáš, 2019), invertebrate mitochondria (Telford *et al.*, 2000; Abascal *et al.*, 2006a; Li *et al.*, 2019), and stop codon reassignments in metagenomic data (Ivanova *et al.*, 2014) and the development of software for specific phylogenetic groups (Abascal *et al.*, 2006b; Mühlhausen and Kollmar, 2014b; Noutahi *et al.*, 2017). Some approaches, such as FACIL (Dutilh *et al.*, 2011), have expanded phylogenetic breadth by using profile Hidden Markov model (HMM) representations of conserved proteins from phylogenetically diverse databases such as Pfam (El-Gebali *et al.*, 2019). However, a systematic survey of genetic code usage across the tree of life has not yet been possible. Existing methods are generally either 1) phylogenetically restricted to clades where multiple sequence alignments can be built for a predetermined set of proteins or 2) lacking sufficiently robust and objective statistical footing to enable a large-scale screen with high accuracy.

A potentially incomplete set of alternative genetic codes limits our ability to understand the evolutionary processes behind codon reassignment. One open question is why some codon reassignments reappear independently. Reassignment of the stop codons UAA and UAG to glutamine is the most common change in eukaryotic nuclear genomes, appearing at least five independent times (Schneider *et al.*, 1989; Keeling and Doolittle, 1996; Keeling and Leander, 2003; Karpov *et al.*, 2013; Swart *et al.*, 2016). In bacteria, all of the known changes reassign the stop codon UGA to either glycine in the Absconditabacteria and Gracilibacteria (Campbell *et al.*, 2013; Rinke *et al.*, 2013) or tryptophan in the Mycoplasmatales, Entomoplasmatales (Bové, 1993), and several insect endosymbiotic bacteria (McCutcheon *et al.*, 2009;

McCutcheon and Moran, 2010; Bennett and Moran, 2013; Salem *et al.*, 2017). These recurring changes may reflect constraints imposed by the existing translational machinery. The mechanism of codon reassignment may involve changes to tRNA anticodons or tRNA wobble nucleotide modifications (which together dictate anticodon-codon pairing), aminoacyl-tRNA synthetase recognition of cognate tRNAs, release factor binding of stop codons, among others, each of which may bias which reassignments are easier to evolve. However, without a complete picture of genetic code diversity, it is hard to disentangle patterns of codon reassignment from observation bias. For instance, in-frame stop codons caused by a stop codon reassignment may be more easily detectable than a subtle change in amino acid conservation indicative of a sense codon reassignment.

Another open question is how a new codon meaning can evolve without disrupting the translation of most proteins. Reassigning a codon leads to the incorporation of the incorrect amino acid at all preexisting codon positions (Crick, 1968). Three evolutionary models differ in the pressure driving substitutions to remove the codon from positions that cannot tolerate the new translation. In the 'codon capture' model, the codon is first driven to near-extinction by pressures unrelated to reassignment, such as biased genomic GC content or genome reduction, which then minimizes the impact of reassignment on protein translation (Osawa and Jukes, 1989). This model was first proposed for the reassignment of the stop codon UGA to tryptophan in *Mycoplasma capricolum*, whose low genomic GC content (25% GC) in combination with small genome size (1 Mb) was thought to have driven the stop codon UGA to extremely low usage in favor of UAA and allowed 'capture' of UGA by a tryptophan tRNA (Bové, 1993; Osawa and Jukes, 1989). For larger nuclear genomes, other models have been proposed where codon usage changes occur concurrently with, and are driven by, changes

in decoding capability. In the 'ambiguous intermediate' model, a codon is decoded stochastically as two different meanings in an intermediate step of codon reassignment, and this translational pressure induces codon substitutions at positions where ambiguity is deleterious (Schultz and Yarus, 1994; Massey *et al.*, 2003). Extant examples of ambiguous translation support the plausibility of this model, such as yeasts that translate the codon CUG as both leucine and serine by stochastic tRNA charging (Gomes *et al.*, 2007) or by competing tRNA species (Mühlhausen *et al.*, 2018). Alternatively, the 'tRNA loss driven reassignment' model proposes an intermediate stage where a codon cannot be translated efficiently, perhaps due to tRNA gene loss or mutation, creating pressure for synonymous substitutions specifically away from that codon, allowing it to be captured later by a different tRNA (Mühlhausen *et al.*, 2016; Sengupta and Higgs, 2005). These three models are not mutually exclusive and substitutions at the reassigned codon can occur due to a combination of these pressures.

Here, we describe Codetta, a computational method for predicting the genetic code which can scale to analyze thousands of genomes. We perform the first survey of genetic code usage in all bacterial and archaeal genomes, reidentifying all known codes in the dataset and discovering the first examples of sense codon changes in bacteria. All five reassignments affect arginine codons (AGG, CGA, and CGG) and provide clues to help us understand how alternative genetic codes evolve.

2.3 Results

2.3.1 Codetta: a computational method to infer the genetic code

We developed Codetta, a computational method that takes DNA or RNA sequences from a single organism and predicts an amino acid translation for each of the 64 codons. Codetta can

analyze sequences from all domains of life, including bacteria, archaea, eukaryotes, organelles, and viruses, and the ability to confidently predict codon decodings depends on having protein-coding regions with recognizable homology. The general idea is to align the input nucleotide sequence to probabilistic profiles of conserved protein domains in order to obtain, for each of the 64 codons, a set of profile positions aligned to that codon. Each profile position has twenty probabilities describing the expected amino acid. For each of the 64 codons, we aggregate over the set of aligned profile positions to infer the single most likely amino acid decoding of the codon. Most previous approaches for genetic code prediction use the same basic idea (Telford *et al.*, 2000; Abascal *et al.*, 2006b; Dutilh *et al.*, 2011; Mühlhausen and Kollmar, 2014b; Swart *et al.*, 2016; Heaphy *et al.*, 2016; Riley *et al.*, 2016; Krassowski *et al.*, 2018; Noutahi *et al.*, 2019), typically aligning the input sequence to multiple sequence alignments and using a simple rule to select the best amino acid for each codon.

With Codetta, we extend this idea to systematic high-throughput analysis by using a probabilistic modeling approach to infer codon decodings, and by taking advantage of the large collection of probabilistic profiles of conserved protein domains (profile HMMs) in the Pfam database (El-Gebali *et al.*, 2019). Profile HMMs are built from multiple sequence alignments, and the emission probabilities at each consensus column are estimates of the expected amino acid frequencies. The Pfam database contains over 17,000 profile HMMs of conserved protein domains from all three domains of life, which are expected to align to about 50% of coding regions in a genome (El-Gebali *et al.*, 2019). We align Pfam profile HMMs to a six-frame standard genetic code translation of the input DNA/RNA sequence using the HMMER `hmmscan` program (Eddy, 2011) (Figure 2.1A). Since we rely on a preliminary standard code translation, conserved protein domains could fail to align in organisms using radically different genetic

codes. In the set of statistically significant `hmmscan` alignments (E-value $< 10^{-10}$), we make the simplifying approximation of considering each aligned consensus column independently, so the alignments are viewed as a set of pairwise associations between a codon Z (64 possibilities) and a consensus column of a Pfam domain profile (denoted C, an index identifying a Pfam consensus column).

From these data, we infer each of the 64 codons one at a time (Figure 2.1B). For a codon Z (e.g. UGA), the observed data $\vec{C}^Z$ are a set of N consensus columns C_i^Z ($i = 1..N$) that associate to Z in the provisional alignments. We model the main data-generative process abstractly, imagining that each column C_i^Z was drawn from the pool of all possible consensus columns by codon Z which is translated as an unknown amino acid A. Each column has an affinity for codon Z proportional to the column's emission probability for the amino acid A, $P(A|C)$. A consensus column strongly conserved for a particular amino acid A will tend to only associate with codons that translate to A; moreover, consensus columns weakly conserved for A may also associate with probability proportional to their conservation for A. Thus this abstract matching process generates an observed C_i^Z column association with the codon Z (translated as amino acid A) with probability

$$P(C_i^Z|A) = \frac{P(A|C_i^Z)P(C_i^Z)}{P(A)}.$$

Here $P(A|C_i^Z)$ is the emission probability for amino acid A at the Pfam consensus column C_i^Z. $P(A)$ is the average emission probability for amino acid A over the pool of all possible consensus columns C, which we take to be all columns aligned to the target genome in order to better reflect genome-specific biases in amino acid usage.

Given the data $\vec{C}^Z$ and this abstract generative model, we infer the most likely decoding M

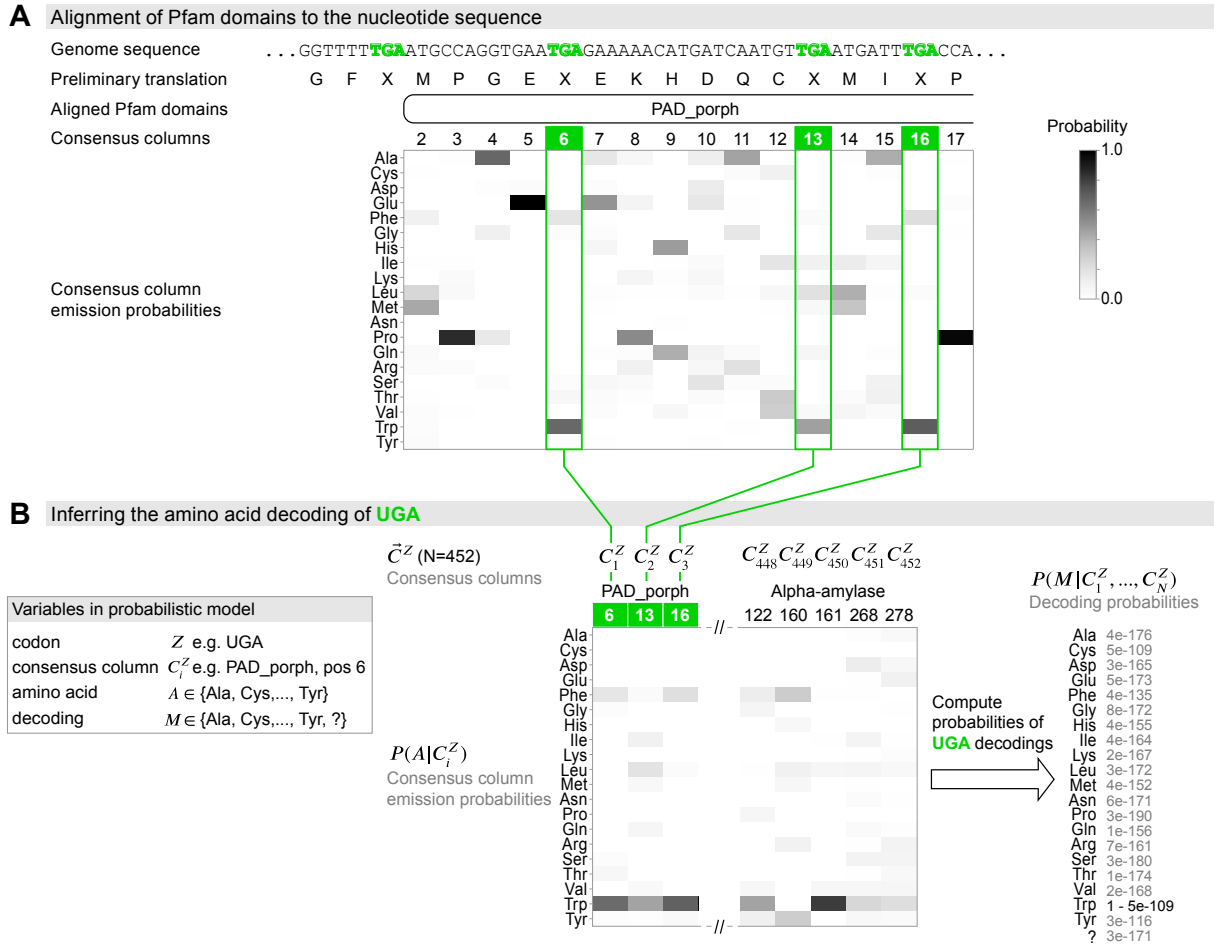

Figure 2.1: Schematic of the genetic code inference method implemented in Codetta. (A) A fragment of the *Mycoplasma capricolum* genome is used to demonstrate alignment of a Pfam domain (PAD_porph) to a preliminary standard code translation of the input DNA sequence (one of six frames shown). All canonical stop codons, including UGA (TGA in genome sequence, reassigned to tryptophan in *M. capricolum*), are translated as 'X' in the preliminary standard code translation which `hmmscan` (program used to align Pfam domains) treats as an unknown amino acid. Each consensus column in the PAD_porph domain has a characteristic emission probability for each of the twenty canonical amino acids, represented by a heatmap. (B) Pfam consensus columns aligning to UGA codons across the entire genome comprise the $\vec{C}^Z$ set for UGA (N=452 Pfam consensus columns). The Pfam emission probabilities $P(A|C_i^Z)$ for all 452 aligned consensus columns are used to compute the decoding probabilities $P(M|\vec{C}^Z)$. The most likely amino acid translation of UGA is inferred to be tryptophan, with decoding probability greater than the cutoff of 0.9999.

Appendix B, Figure B.1– Estimates of error rate and power of Codetta inference.

for codon Z out of 21 possibilities $M \in \{\text{Ala}, \text{Cys}, ..., \text{Tyr}, ?\}$ (Figure 2.1B). The $M = ?$ model of non-specific translation draws columns randomly and serves to catch codons that do not encode a specific amino acid, such as stop codons and ambiguously translated codons. For a given decoding M, the probability of the observed columns $\vec{C}^Z$ is then:

$$P(\vec{C}^Z|M) = \begin{cases} \prod_{i=1}^{N} \frac{P(A=M|C_i^Z)P(C_i^Z)}{P(A=M)} & \text{if } M \in \{\text{Ala}, \text{Cys}, ..., \text{Tyr}\} \\ \prod_{i=1}^{N} P(C_i^Z) & \text{if } M = ? \end{cases}$$

Setting the prior probability of each decoding, $P(M)$, to be uniform, we compute the probability of the decoding M as:

$$P(M|\vec{C}^Z) = \frac{P(\vec{C}^Z|M)}{\sum_{M'} P(\vec{C}^Z|M')}$$

We assign an amino acid translation to a codon if it attains a decoding probability above some threshold (typically 0.9999). We assign a '?' if no amino acid decoding satisfies the probability threshold (including the case where '?' itself has high probability). A '?' assignment tends to happen if the codon is rare, with few aligned Pfam consensus columns on which to base the inference, or if the codon is ambiguously translated such that no single amino acid model reaches high probability. Because we do not model stop codons explicitly, we expect '?' to be the inferred meaning since stop codons ideally would have few or no aligned Pfam consensus columns.

To assess how many columns in $\vec{C}^Z$ are needed for reliable codon assignment, we constructed synthetic $\vec{C}^Z$ datasets ranging from 1 to 500 consensus columns by subsampling the Pfam consensus columns aligned to each of the 61 sense codons in the *Escherichia coli* genome.

We calculated the per-codon error rate (fraction of samples predicting the incorrect amino acid) and the per-codon power (fraction of samples predicting the correct amino acid) using a probability threshold of 0.9999 (Figure B.1). Lack of an amino acid inference ('?') was considered neither an error nor a correct prediction. Per-codon error rates were < 0.00002 for all sizes of $\vec{C}^Z$. Depending on the codon, we found that about 8 to 34 aligned consensus columns suffice for >95% power to infer the correct amino acid. Accuracy may differ in real genomes for various biological reasons, but these results gave us confidence to proceed.

2.3.2 Genetic code prediction of 462 yeast species confirms known distributions of CUG reassignment

We further validated Codetta on the budding yeasts (Saccharomycetes, 462 sequenced species) which vary in their translation of CUG as either serine, leucine, or alanine depending on the species (Mühlhausen *et al.*, 2016; Krassowski *et al.*, 2018; Mühlhausen *et al.*, 2018). In some CUG-Ser clade species, such as *Candida albicans*, CUG codons are stochastically decoded as a mix of serine (97%) and leucine (3%) because the CUG-decoding tRNA_{CAG} is aminoacylated by both the seryl- and leucyl-tRNA synthetases (Suzuki *et al.*, 1997; Gomes *et al.*, 2007). Codetta is not designed to predict ambiguous decoding and is expected to assign either the dominant amino acid or a '?' in cases like *C. albicans*.

For 453 species, the predicted CUG translation was consistent with the known phylogenetic distribution of CUG reassignments (Figure 2.2A). This includes *C. albicans*, which was predicted to use the predominant serine translation (Gomes *et al.*, 2007). For the remaining nine species, Codetta did not put a high probability on any amino acid decoding of CUG (inferred meaning of '?'). Two of these species—*Babjeviella inositovora* and *Cephaloascus fragrans*—are basal members of the CUG-Ser clade. Both of these genomes contain a CUG-decoding

$tRNA_{CAG}$ gene with features of serine identity (see Materials and methods) and *B. inositovora* has previously been shown to translate CUG codons primarily as serine by whole proteome mass spectrometry (Krassowski *et al.*, 2018; Mühlhausen *et al.*, 2018), suggesting that CUG is decoded as serine in these species. Codetta did not infer an amino acid for CUG because the aligned Pfam consensus columns were not consistently conserved for a single amino acid (Figure B.2).

The other seven species without an inferred amino acid for CUG all belong to the closely-related genera *Ascoidea* and *Saccharomycopsis* (four additional species in these clades were predicted to translate CUG as serine). Analysis of tRNA genes revealed that 10 out of 11 species in this clade encode two types of $tRNA_{CAG}$ genes, one predicted to be serine-type and one leucine-type, suggesting that CUG may be ambiguously translated as both serine and leucine via competing tRNAs in some of these species (Figure 2.2B). We used Northern blotting to assay the expression of both $tRNA_{CAG}$ genes in some of these species under a variety of growth conditions (data not shown), but could detect reliable expression of both serine- and leucine-type $tRNA_{CAG}$ genes only in *Saccharomycopsis malanga* (only the serine $tRNA_{CAG}$ could be detected in other species) (Figure 2.2C). To determine whether both tRNAs are aminoacylated, we performed acid urea PAGE Northern blotting which separates aminoacylated and deacylated tRNAs but does not identify the charged amino acid. We found that both serine and leucine *S. malanga* $tRNA_{CAG}$ are predominantly charged in cells (Figure 2.2C), likely partaking in the translation of CUG codons. If CUG is indeed translated ambiguously in this clade, it would explain why Codetta did not place a high probability on any single amino acid decoding for some species.

The existence of serine and leucine $tRNA_{CAG}$ genes in some *Ascoidea* and *Saccharomycopsis*

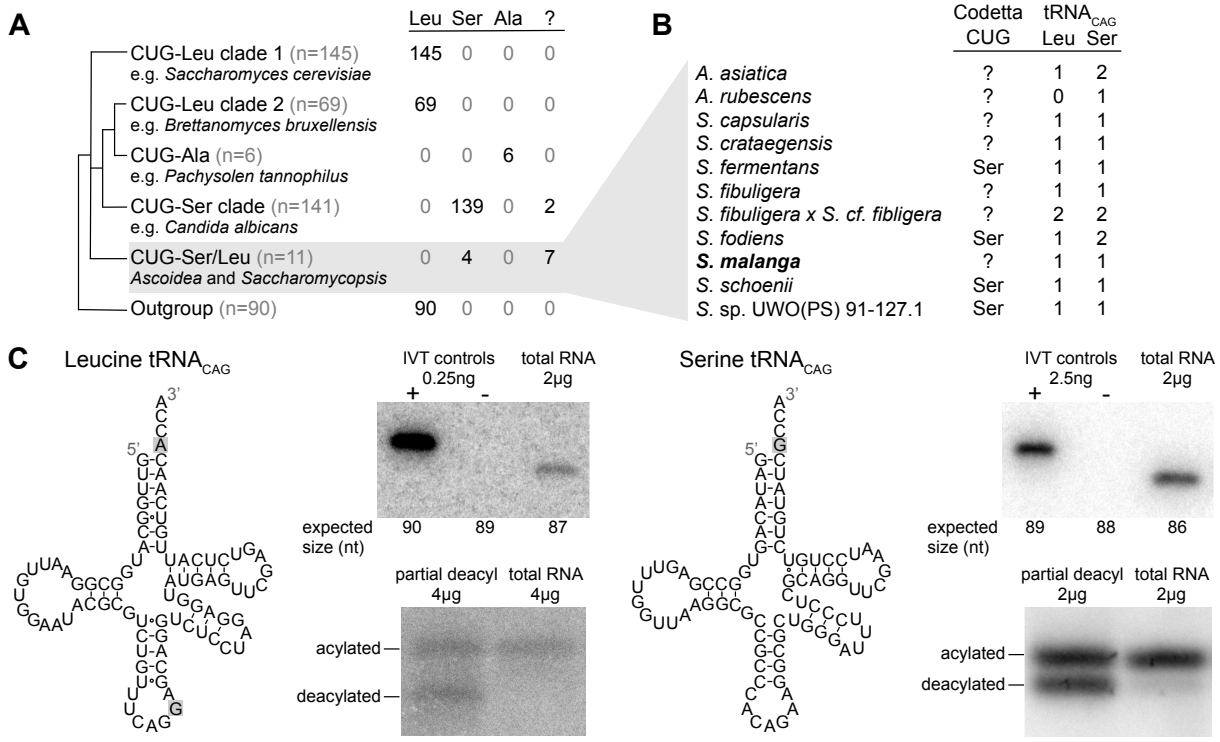

Figure 2.2: (A) CUG translation inferred by Codetta of 462 Saccharomycetes species, grouped by phylogenetic clade. Cladogram was adapted from Shen *et al.* (2018). Phylogenetic placement of CUG-Leu/Ser clade is unresolved, thus the three-way branch. (B) Codetta CUG inference and number of $tRNA_{CAG}$ genes in *Ascoidea* and *Saccharomycopsis* genomes. $tRNA_{CAG}$ genes were identified using tRNAscan-SE 2.0 and were classified as being serine-type or leucine-type based on the presence of tRNA identity elements. (C) Northern blotting to confirm expression and charging of leucine and serine $tRNA_{CAG}$ genes in *S. malanga*. Probable secondary structures of the two *S. malanga* $tRNA_{CAG}$ are shown with features used for leucine/serine classification highlighted in gray. In the tRNA expression blots, *in vitro* transcribed (IVT) versions of the target tRNA (+ control) and the most similar other tRNA (- control, as determined by sequence homology with the probe) were used as controls for probe specificity. In the tRNA charging blots, a partial deacylation control was used to help visualize the expected band sizes for acylated and deacylated versions of the probed tRNA.

Source data 1– Table of all analyzed yeast genomes with phylogenetic grouping and Codetta CUG inference.
Source data 2– Table of *in vitro* transcription DNA template sequences.
Appendix B, Figure B.2– Distribution of Pfam consensus column amino acid support for CUG codons in *B. inositovora* and *C. fragrans*.

yeasts was reported by Krassowski *et al.* (2018) and Mühlhausen *et al.* (2018) while we were conducting experiments. Ambiguous translation of CUG was demonstrated in *A. asiatica* (Mühlhausen *et al.*, 2018); however, for *S. malanga* only expression of the serine $tRNA_{CAG}$ could be detected (Krassowski *et al.*, 2018) and incorporation of predominantly serine at protein positions encoded by CUG (Mühlhausen *et al.*, 2018). In contrast to these studies, we used a saturated growth condition where the leucine $tRNA_{CAG}$ seems to be more highly expressed. While we did not quantify the relative expression of the two $tRNA_{CAG}$ in *S. malanga*, a visual comparison of the band intensities in Figure 2.2C suggests that the expression of the leucine $tRNA_{CAG}$ is at least ten times less than the serine $tRNA_{CAG}$ even in the saturated growth condition.

These results show that Codetta can correctly infer canonical and non-canonical codon translations and can flag unusual situations such as ambiguous translation even though it assumes unambiguous translation. All of the remaining 63 codons were inferred to use the expected translation in all species, with the following exceptions. In three species belonging to a lineage of *Hanseniaspora* with low genomic GC content (Steenwyk *et al.*, 2019), the arginine codons CGC and/or CGG had a '?' inference due to few (<20) aligned Pfam consensus columns presumably due to rare usage of those codons. In eight other species, either the stop codon UAG or UGA was inferred to code for tryptophan due to some (<23) aligned Pfam consensus columns. We could not find any nuclear suppressor tRNA genes, and we believe these inferences are due to the erroneous alignment of Pfam domains to in-frame stop codons in pseudogenes. In-frame stop codons do not appear randomly within pseudogenes but instead are most likely to result from single nucleotide transitions from certain codons (such as the UGG tryptophan codon).

2.3.3 Computational screen of bacterial and archaeal genomes finds previously known alternative genetic codes

To explore the diversity of genetic codes in bacterial and archaeal genomes, we used Codetta to analyze 251,571 assembled genomes from GenBank, including partial assemblies and those derived from single-cell genomics and metagenomic assembly. Summaries of our analysis (Table 2.1 and Table 2.2) are shown for a subset of the results, dereplicated to reduce the over-representation of frequently sequenced organisms by selecting a single assembly for each species-level NCBI taxonomic ID (48,693 unique species: 46,384 bacteria, 2,309 archaea). Results for the full dataset and the dereplicated subset are available in Supplementary File 1.

To see if our screen recovered known alternative genetic codes, we collated a comprehensive literature summary of all bacterial and archaeal clades known to use alternative genetic codes (Table 2.1) and layered it over the NCBI taxonomy, annotating all remaining organisms with the standard genetic code. This resulted in a genetic code annotation for each species. For most species using known alternative genetic codes in our dataset, our predictions at the reassigned codon agreed with the the expected amino acid translation (Table 2.1). There were no instances of reassigned codons predicted to translate as an unexpected amino acid, but there were a few cases of reassigned UGA codons which had no amino acid meaning inferred ('?' inference).

Since the uninferred codons could represent a lack of sensitivity by Codetta, we looked more closely at these examples. In the Mycoplasmatales and Entomoplasmatales, which are believed to translate the canonical stop codon UGA as tryptophan (Bové, 1993), eight species had no inferred amino acid meaning for UGA due to fewer than 4 aligned Pfam consensus columns. All of these genomes lack a UGA-decoding $tRNA^{Trp}_{UCA}$ gene and all but one instead

contain a release factor 2 gene (which terminates translation at UGA). Five of these species are included in the Genome Taxonomy Database (GTDB) (Parks *et al.*, 2020), a comprehensive phylogeny of over 190,000 bacterial and archaeal genomes, where they are grouped into a different order (GTDB order RF39). We therefore attribute at least 5 (and perhaps all 8) as a taxonomic misannotation in the NCBI database, and we believe that UGA is a stop codon in these species. In the Gracilibacteria, which are believed to translate the stop codon UGA as glycine (Rinke *et al.*, 2013), two species had no inferred amino acid meaning for UGA. Neither genome contained the expected UGA-decoding $\text{tRNA}^{\text{Gly}}_{\text{UCA}}$ gene and both instead encoded a release factor 2 gene, supporting that UGA is a stop codon and not a glycine codon in these species. Indeed, one of these species is included in the GTDB and is grouped in a different order than the other UGA-reassigned Gracilibacteria and Absconditabacteria.

Across the 48,693 genomes (dereplicated to one assembly per species), we predicted the amino acid translation of a total of 2,970,497 individual sense codons (roughly 61 times the number of genomes), with 99.79% of the predictions consistent with the expected amino acid (similar proportion across bacteria and archaea) (Table 2.2). About 0.19% of sense codons had a '?' inference, demonstrating that entire genomes contain more than enough information to infer the amino acid translation of most sense codons. Unexpected amino acid meanings were predicted for 612 sense codons. These are candidates for new codon reassignments, but could also include inference errors. For stop codons, 99.80% out of total of 145,855 stop codons across the dereplicated bacterial and archaeal genomes had no inferred amino acid meaning, as expected. 290 bacterial stop codons and 9 archaeal stop codons were inferred to translate as an amino acid, adding to our list of candidate new genetic codes.

Phylogenetic distribution	NCBI taxids	Ref	N species	Codon reassignment	*Reassigned codon* Expected amino acid	Uninferred ('?')
Entomoplasmatales & Mycoplasmatales	186328, 264638, 2085	[1,2]	199	UGA Stop→W	191	8
Hodgkinia cicadicola	573658	[3]	1	UGA Stop→W	1	0
Nasuia deltocephalinicola	1160784	[4]	1	UGA Stop→W	1	0
Zinderia insecticola	884215	[5]	1	UGA Stop→W	1	0
Stammera capleta	2608262	[6]	1	UGA Stop→W	1	0
Gracilibacteria	363464	[7]	15	UGA Stop→G	13	2
Absconditabacteria	221235	[8]	6	UGA Stop→G	6	0

Table 2.1: A summary of all bacterial clades previously known to use a codon reassignment. For each clade, the NCBI taxonomic IDs (taxids) shown most closely correspond to the known phylogenetic distribution from the literature. For each codon reassignment, we show the number of sequenced species analyzed by Codetta and how many were inferred to use the expected amino acid or had no inferred amino acid. None of the analyzed species belonging to reassigned clades were predicted to use an unexpected amino acid at the reassigned codon. [1] Bové (1993), [2] Volokhov *et al.* (2007), [3] McCutcheon *et al.* (2009), [4] Bennett and Moran (2013), [5] McCutcheon and Moran (2010), [6] Salem *et al.* (2017), [7] Rinke *et al.* (2013), [8] Campbell *et al.* (2013)

		Bacteria 46,384 species		Archaea 2,309 species	
Sense	Total (N codons x N species)	2,829,648		140,849	
	Expected amino acid	2,823,497	99.78%	140,631	99.85%
	Other amino acid	612	0.02%	0	0.00%
	Uninferred ('?')	5,539	0.20%	218	0.15%
Stop	Total (N codons x N species)	138,928		6,927	
	Amino acid	290	0.21%	9	0.13%
	Uninferred ('?')	138,638	99.79%	6,918	99.87%

Table 2.2: A summary of codon inferences from the set of genomes analyzed by Codetta, dereplicated to one assembly per species. The Codetta inference for each codon is compared against a genetic code annotation derived by layering the known bacterial genetic codes in Table 2.1 over the NCBI taxonomy. Reassigned stop codons are included with sense codons. Values can be calculated from Supplementary File 1.

2.3.4 Validation of candidate new alternative genetic codes

To prioritize high-confidence novel genetic codes, we gathered additional evidence by examining 1) the translational components (tRNA and/or release factor genes) involved in the reassignment, 2) the usage of the reassigned codon, including manual examination of alignments of highly conserved single-copy genes, and 3) the phylogenetic extent of the proposed reassignment. Since many candidate genetic codes were found in uncultivated clades with only rough taxonomic classification on NCBI, we explored phylogenetic relationships using the Genome Taxonomy Database (GTDB) (Parks *et al.*, 2020). The GTDB is a phylogeny of over 190,000 archaeal and bacterial genomes, providing provisional domain-to-species phylogenetic classifications for uncultivated as well as established clades. A list of all candidate novel genetic codes can be found in Supplementary File 2. We focused on the candidate codon reassignments with the highest degree of additional evidence and attempted to characterize common sources of error. The set of lower-confidence candidates may still include additional real codon reassignments requiring further validation.

The most common error was the inference of AGA and/or AGG arginine codons as coding for lysine, occurring in 567 bacterial species. Almost all of the AGA- and AGG-decoding tRNAs found in these genomes were consistent with arginine identity (based on the arginine identity elements A/G73 and A20), supporting that AGA and AGG are arginine codons in the majority of these species. The unusually high GC content of these genomes (ranging between 0.52 - 0.77, median 0.68) suggests that the source of the lysine inference may come from high GC content-driven nonsynonymous substitutions of the AAA and AAG lysine codons to AGA and AGG arginine codons at protein residues that can tolerate either positively-charged amino acid. As a result, AGA and AGG codons would consistently appear at residues conserved for

lysine in other species, which Codetta would mistake for the signature of codon reassignment. Genomic GC content has long been correlated with greater arginine usage and lower lysine usage (Sueoka, 1961; Lightfield *et al.*, 2011), possibly due to substitutions between the aforementioned lysine and arginine codons (Knight *et al.*, 2001b). This error could be mitigated in future analyses by using profile HMMs built from sequences that match the analyzed genome in GC content or amino acid composition.

Some erroneous stop codon inferences resulted from genome contamination by organisms with known stop codon reassignments. We suspected contamination when the Pfam consensus columns aligned to a stop codon were only present in a limited part of the genome and confirmed the origin of these regions by homology search of the genes containing the in-frame stop codons. We have found examples of predicted stop reassignments in *Sulfolobus* assemblies caused by contamination with UGA-recoding *Mycoplasma* contigs, in an alphaproteobacteria assembly caused by contamination with UAA- and UAG-recoding ciliate contigs, in *Chloroflexi* assemblies caused by contamination with UGA-recoding Absconditabacteria contigs, and in others.

We found five clades using candidate novel alternative genetic codes with additional computational evidence, such as tRNA genes consistent with the new translation. All five new genetic codes involve the reassignment of arginine codons, representing the first sense codon reassignments in bacteria.

2.3.5 Reassignment of the canonical arginine codon AGG to methionine in a clade of uncultivated Bacilli

Eight bacterial genomes were inferred to translate AGG, a canonical arginine codon, as methionine. All eight genomes were assembled from fecal metagenomes of baboons or humans

(Parks *et al.*, 2017; Almeida *et al.*, 2019) and have only coarse-grained NCBI genome classification as uncultured Bacillales or Mollicutes bacteria. The GTDB assigns these eight genomes to a three species clade within the genus *Enterosoma* (family CAG-288, order RFN20, class Bacilli), of which all other species were inferred to translate AGG as arginine (Figure 2.3A).

In each of the reassigned *Enterosoma* genomes, the AGG inference by Codetta is based on a sufficiently large number of aligned Pfam consensus columns (over 2,200 compared to an average of about 1,800 for each of the other 60 sense codons) from over 480 different Pfam domains. Figure 2.3B shows an example multiple sequence alignment of uridylate kinase, a single-copy conserved bacterial gene, from the reassigned species, outgroup genomes, and several more distantly related bacteria. In the alignment, AGG codons are used interchangeably with AUG methionine codons within the reassigned clade and tend to occur at positions conserved for methionine and other nonpolar amino acids in the other species.

In the reassigned *Enterosoma* clade, AGG is the dominant methionine codon with a usage of 209-235 per 10,000 codons in Pfam alignments, outnumbering the canonical methionine codon AUG (59-69 per 10,000 codons) (Figure 2.3A). The process of codon reassignment involves genome-wide codon substitutions to remove the reassigned codon from positions that cannot tolerate the new amino acid, leading to depressed codon usage. High usage of AGG in the reassigned clade suggests that this is an established codon reassignment that has had time to rebound in frequency through synonymous substitutions with the standard AUG methionine codon. In many outgroup genomes, AGG is a rare arginine codon (Figure 2.3A).

Escape from viral infection has been put forth as a potential selective pressure for the evolution of alternative genetic codes, although viruses are also known to infect some alternative genetic code organisms such as *Mycoplasma* and mitochondria (Shackelton and Holmes, 2008).

We inferred the genetic code of phage genomes assembled by Al-Shayeb *et al.* (2020) from the same baboon fecal metagenomic dataset as some reassigned *Enterosoma* genomes. Two phage assemblies were predicted to translate AGG as methionine (assemblies GCA_902730795.1 and GCA_902730815.1). The assemblies do not contain genes for the AGG-decoding $tRNA_{CCU}$, so the phage presumably rely on the host tRNAs for translation. Thus, some phage may have adapted to the AGG translation as methionine in the reassigned *Enterosoma*.

We used tRNAscan-SE 2.0 (Chan *et al.*, 2021) to determine which tRNAs are available to decode AGG in the reassigned and outgroup genomes (Figure 2.3A). Some tRNA genes are missing, possibly due to the incomplete nature of some metagenome-assembled genomes as indicated by low genome completeness estimates. The cognate tRNA for the AGG codon, $tRNA_{CCU}$, from the reassigned *Enterosoma* clade has features of methionine identity (including an A73 discriminator base and G2:C71 and C3:G70 base pairs in the acceptor stem) and lacks the important arginine identity element A20 in the D loop (Meinnel *et al.*, 1993; Giegé *et al.*, 1998), supporting translation of AGG as methionine (Figure 2.3C). *In vitro* experiments have shown that anticodon mutations to $tRNA^{Met}_{CAU}$ disrupt recognition by the *E. coli* methionyl-tRNA synthetase (MetRS); however, the C35 change necessary to decode the AGG codon affects the least critical anticodon nucleotide (Schulman and Pelka, 1983). To see if any compensatory changes have occurred in the MetRS from the reassigned *Enterosoma* clade to accommodate recognition of a new anticodon, we compared the predicted MetRS sequence to that of *Aquifex aeolicus* (crystal structure in complex with $tRNA_{CAU}$, PDB 2CSX/2CT8, Nakanishi *et al.* (2005)). The three residues that contact the anticodon nucleotides in *A. aeolicus*, which are conserved through all domains of life (Nakanishi *et al.*, 2005), also remain unchanged in the reassigned clade MetRS sequence (Figure B.3).

Figure 2.3: Reassignment of AGG from arginine to methionine in a clade of uncultivated *Enterosoma*. (A) GTDB phylogenetic tree of the *Enterosoma* AGG→Met clade and closest outgroup genomes, with the annotated NCBI genome name shaded according to the GTDB CheckM estimated genome completeness. GTDB genus *Enterosoma* (UBA7642) corresponds to species #1-9 and GTDB family CAG-288 corresponds to species #1-16. For each genome, the Codetta AGG inference is indicated by colored circles (red: arginine, blue: methionine). The presence of tRNA genes is also indicated by filled circles for $tRNA_{UCU}$ and $tRNA_{CCU}$, colored by the predicted amino acid charging based on known identity elements (see Materials and methods), or a white circle if no tRNA gene could be detected. The lines connecting codons and anticodons represent the likely wobble decoding capabilities, with dashed lines representing likely weaker interactions. Codon usage is the frequency per 10,000 codons aligned to Pfam domains. (B) Multiple sequence alignment of uridylate kinase (BUSCO POG091H02JZ) from the reassigned species, selected outgroup species, and four more distantly related bacteria (*Bacillus subtilis*, *Nostoc punctiforme*, *Chlamydia caviae*, and *E. coli*). All AGGs are represented by $\boldsymbol{\alpha}$. Alignment regions containing multiple nearby AGG positions in the reassigned species are shown. (C) A comparison of the AGG-decoding $tRNA_{CCU}$ in the *Enterosoma* AGG → Met clade (identical sequence in all genomes) and in an outgroup genome (#6, uncultured Clostridiales). tRNA sequence features involved in methionine identity in the reassigned clade $tRNA_{CCU}$ and arginine identity in the outgroup tRNA are highlighted (Meinnel *et al.*, 1993; Giegé *et al.*, 1998), with nucleotide numbering following the convention of Sprinzl *et al.* (1998). The C35 anticodon nucleotide in the *Enterosoma* AGG→Met clade tRNA is highlighted in gray because it does not match the A35 methionine identity element. (D) The genomic context surrounding the $tRNA_{CCU}$ gene in a member of the *Enterosoma* AGG→Met clade (#2, Bacillales UBA4682) and in an outgroup species (#6, uncultured Clostridiales). Gene lengths and intergenic distances are drawn proportionally, with the number of base pairs between each gene indicated below.

Source data 1– Table of genome accessions, Codetta AGG inference, tRNA gene presence, codon usage, and genome GC content for the reassigned AGG→Met *Enterosoma* and outgroup species.
Appendix B, Figure B.3– Alignment of MetRS sequences.
Appendix B, Figure B.4– Local genomic context of $tRNA_{CCU}$ in *Enterosoma* AGG→Met clade and outgroup genomes.

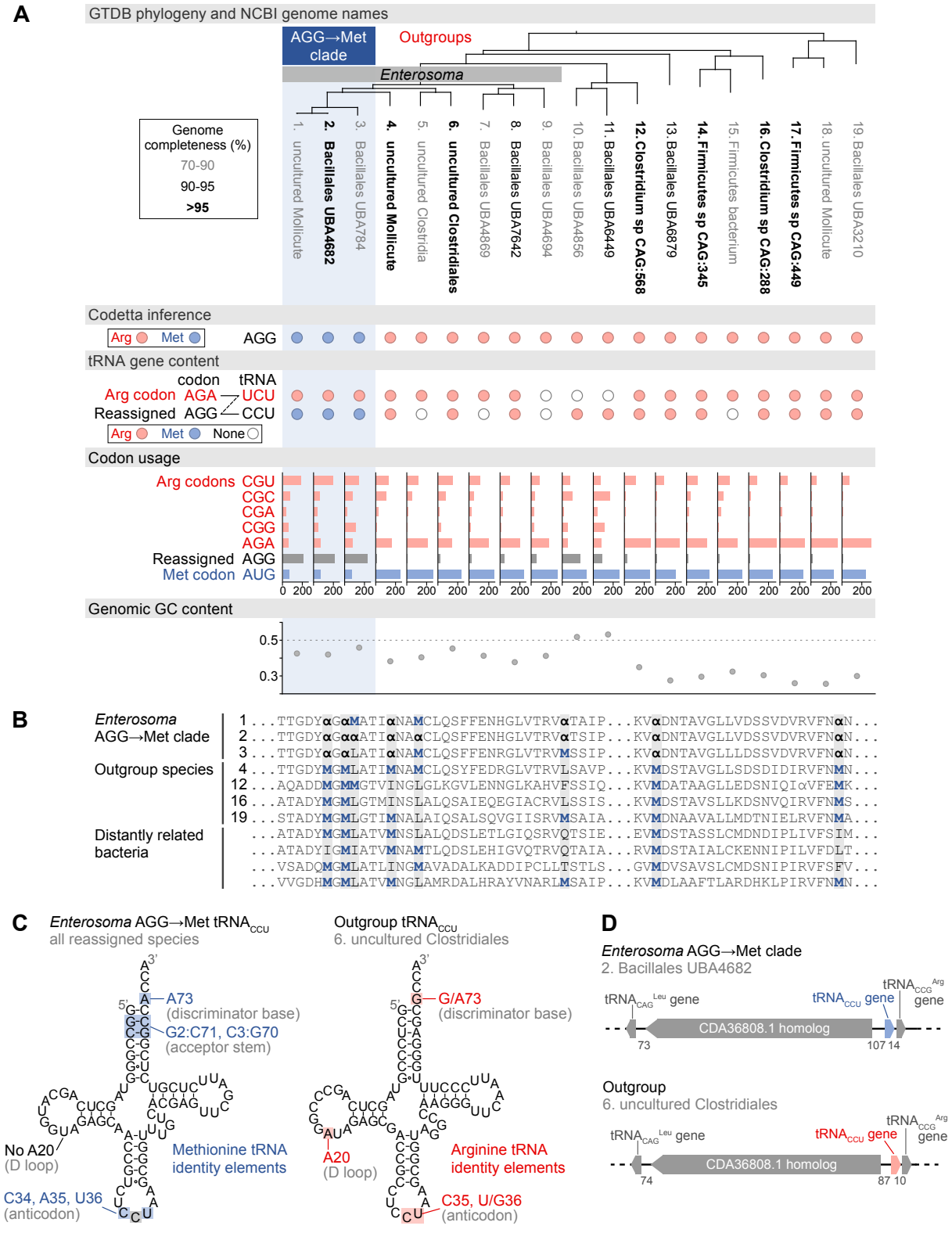

Figure 2.3 (Continued)

The outgroup genomes contain a $tRNA_{CCU}$ with features of arginine identity (including a G73 discriminator base and A20 in the D loop). The genomic context of the $tRNA_{CCU}$ is similar in many reassigned clade and outgroup genomes, flanked by a $tRNA^{Arg}_{CCG}$ immediately downstream and a homolog of GenBank protein CDA36808.1 upstream (Figure 2.3D, Figure B.4). This implies that the reassigned and outgroup $tRNA_{CCU}$ evolved from the same ancestral tRNA gene, and the reassigned methionine $tRNA_{CCU}$ likely emerged through a change in aminoacylation of an arginine $tRNA_{CCU}$ rather than through duplication and anticodon mutation of a methionine tRNA.

The reassigned *Enterosoma* genomes use an arginine-type $tRNA_{UCU}$ to decode the unaffected AGA arginine codon. Depending on the post-transcriptional modification of the U34 anticodon nucleotide, the arginine $tRNA_{UCU}$ could recognize AGG via wobble and potentially cause ambiguous translation. In *E. coli*, the U34 of $tRNA_{UCU}$ is modified to 5-methylaminomethyluridine (Sakamoto *et al.*, 1993) which primarily decodes the AGA codon with a low level of AGG recognition (Spanjaard *et al.*, 1990). Mukai *et al.* (2015) demonstrated that it is possible to engineer separate decodings for AGA and AGG in *E. coli* by reducing expression level of the $tRNA_{UCU}$ to the point where decoding of AGG by $tRNA_{UCU}$ is presumably insignificant in competition with the cognate $tRNA_{CCU}$. In most outgroup genomes, AGA is the dominant arginine codon, while in the reassigned clade the preferred arginine codon is CGU (Figure 2.3A), which may indicate reduced demand and expression of $tRNA_{UCU}$ to avoid ambiguous translation of AGG. A similar potential for ambiguous translation due to U34 wobble exists with the previously known decoding of UGA as glycine and UGG as tryptophan in Absconditabacteria and Gracilibacteria (Campbell *et al.*, 2013; Rinke *et al.*, 2013).

2.3.6 Reassignments of arginine codons CGA and CGG occur in clades with low genomic GC content

The remaining four clades with codon reassignments supported by additional computational evidence all affect the arginine codons CGA and/or CGG (Figure 2.4). Three clades are in the phylum Firmicutes: the genus *Peptacetobacter* is predicted to translate CGG as glutamine, a clade of uncultivated Bacilli in the GTDB genus *Onthovivens* (same order as *Enterosoma*) is predicted to translate CGG as tryptophan, and members of the genus *Anaerococcus* are also predicted to translate CGG as tryptophan. The fourth clade is Absconditabacteria (also known as Candidate Division SR1, part of the Candidate Phyla Radiation), which is predicted to have reassigned CGA and CGG both to tryptophan, in addition to the already known reassignment of UGA from stop to glycine. See Chapter 4 for detailed analysis and discussion of evidence in support of each reassignment.

In contrast to the reassignment of AGG to become the dominant methionine codon (described in the previous section), these CGA/CGG reassignments resemble earlier stages of codon reassignment where the reassigned codon has not yet rebounded in frequency through synonymous substitutions with the new amino acid. Due to the rarity of the reassigned CGA/CGG codons, these predictions are based on fewer aligned Pfam consensus columns and may be more prone to error. As a check for each reassignment, we looked for examples of the reassigned codon in conserved regions of single-copy gene alignments and found multiple supporting positions for all reassigned codons except the extremely rare CGG codon in *Anaerococcus*. We also looked for tRNA genes with an anticodon and amino acid identity elements consistent with the reassignment, and found consistent tRNAs for all clades except for *Peptacetobacter* whose CGG-decoding $tRNA_{CCG}$ resembles neither an arginine or glutamine isotype.

While amino acid conservation at the reassigned codon and sequence-based prediction of tRNA charging may lend support to a predicted codon reassignment, only experimental confirmation can establish how the reassigned codons are translated *in vivo* and whether there is ambiguous translation. In particular, *Anaerococcus* and *Peptacetobacter* include culturable species and may be experimentally confirmed in the future.

The four CGA/CGG candidate reassignments share several features that suggest common evolutionary forces at play. Most notable is the very low genomic GC content of the reassigned clades (0.26-0.38, Figure 2.4A). In all four clades, the usage of GC-rich CGN-box codons—including CGA and CGG—is depressed and arginine residues are primarily encoded by AGA codons (Figure 2.4B). In the three Firmicute CGG reassignments, CGG is an extremely rare codon (codon usage <6 per 10,000 in aligned Pfam domains for all species). In the Absconditabacteria, CGG also tends to be quite rare (<7 per 10,000 in all but one species) with CGA slightly more abundant (<37 per 10,000 in all species). In one Absconditabacteria (assembly GCA_002791215.1), the frequency of both CGA and CGG approaches the frequency of the canonical tryptophan codon UGG, consistent with a more advanced stage of codon reassignment (usage of CGA and CGG is 30 and 24 per 10,000, compared to 35 for UGG). Low genomic GC content is thought to be created by mutational bias in favor of AT nucleotides, causing a gradual shift towards synonymous codons with lower GC compositions (Knight *et al.*, 2001b; Muto and Osawa, 1987). This may have helped disfavor usage of CGA and/or CGG prior to reassignment, lessening the impact of changing the codon meaning.

The tRNAs used to decode the CGN codon box may have also influenced the reassignment of CGA and CGG codons. A shared feature of the three Firmicute CGG reassignments is that the $tRNA_{UCG}$ is missing (Figure 2.4B), presumably lost prior to or during the reassignment

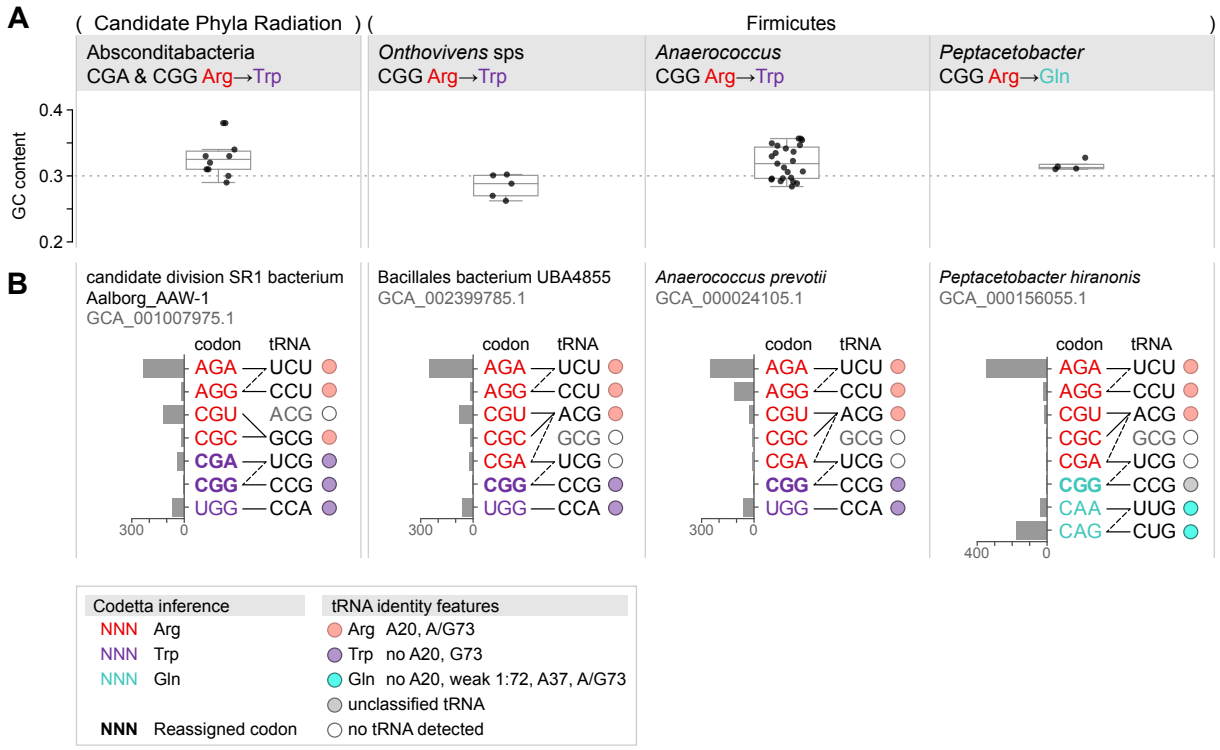

Figure 2.4: Summary of GC content, codon usage, and tRNA genes of four CGA and/or CGG reassignments. (A) Distribution of genomic GC contents across all species in the reassigned clades. (B) For each reassigned clade, we selected a representative species to show codon usage and tRNA decoding ability. Codon usage is plotted for the reassigned codon and for all other codons of the original and new amino acids in usage per 10,000 in Pfam alignments. Codons are colored by their Codetta inference and reassigned codons are bolded. The lines connecting codons and tRNA anticodons represent the likely wobble decoding capabilities, with dashed lines representing likely weaker interactions. The anticodon ACG is presumed to be modified to ICG, and UCG is presumed to be modified in a way that restricts wobble to CGA and CGG, but could potentially recognize CGU and CGC as well depending on the true modification state. Anticodons in gray font are not expected to be found in the respective clade. Presence of tRNA genes is indicated by filled circles, colored by the predicted amino acid charging based on the identity elements in the key.

Source data 1– Table for each CGA/CGG reassignment containing genome accessions, Codetta CGA/CGG inference, tRNA gene presence, codon usage, and genome GC content for the reassigned clade and outgroup species.

of CGG. If the $tRNA_{UCG}$ were present, it would likely recognize both CGA and CGG via wobble which would complicate assigning different amino acid meanings to those two codons. In the absence of $tRNA_{UCG}$, CGA (along with CGU and CGC) is presumably decoded by a $tRNA^{Arg}_{ICG}$ (derived by deamination of $tRNA^{Arg}_{ACG}$, the only widespread instance of inosine tRNA wobble in bacteria). This leaves CGG to be decoded solely by a $tRNA_{CCG}$ (Figure 2.4B). In this situation, CGG is one of a few codons in the genetic code decoded by a single dedicated tRNA, potentially facilitating codon reassignment since the translational meaning of CGG can now be altered independently of neighboring codons. The inosine wobble modification is not used by some deeply branching bacteria (Rafels-Ybern *et al.*, 2018), and the $tRNA^{Arg}_{ACG}$ gene appears to be lacking in the Candidate Phyla Radiation, including Absconditabacteria. Instead, these bacteria use a $tRNA^{Arg}_{GCG}$ to decode CGU and CGC, and rely on a $tRNA_{UCG}$ and $tRNA_{CCG}$ to recognize CGA and CGG (Figure 2.4B). Since the ability of $tRNA_{UCG}$ to decode CGA and CGG makes it difficult to split the translational meanings of the two codons, it may explain why both CGA and CGG are reassigned to tryptophan together in the Absconditabacteria.

For some of these reassignments, close outgroup species may shed light on potential intermediate stages of codon reassignment (described in more detail in Chapter 4). The CGG reassignment in the Absconditabacteria may extend to members of the sister clade Gracilibacteria—some Gracilibacteria were predicted to translate CGG as tryptophan, while others translate CGG as arginine (Figure 4.1). This may reflect a complicated history of CGG reassignment and possible reversion to arginine translation. For the CGG reassignment in *Peptacetobacter*, the closest sister group (which includes the pathogen *Clostridioides difficile*) has extremely rare usage of CGG (<1 per 10,000 in aligned Pfam domains in all but two species) and appears to lack any tRNA capable of decoding CGG by standard codon-anticodon pairing

rules (Figure 4.11). This may resemble an intermediate stage in codon reassignment before the ability to translate CGG as a new amino acid is gained, similar to the unassigned CGG codon in *Mycoplasma capricolum* (Oba *et al.*, 1991). In *Anaerococcus*, all species contain a CGG-decoding $tRNA_{CCG}$ with features of tryptophan identity (Figure 4.5). Unexpectedly, members of an outgroup genus *Finegoldia* also have a $tRNA_{CCG}$ with features of tryptophan identity (CGG inferred to be '?' by Codetta). It is unclear if the $tRNA_{CCG}$ genes in these two clades share an evolutionary history or represent independent events.

2.4 Discussion

We present a method for computationally inferring the genetic code that can scale to analyze hundreds of thousands of genomes which we call Codetta. We validate Codetta on the well-studied reassignments of CUG in yeasts and rediscover the likely ambiguous translation of CUG as serine and leucine in some *Ascoidea* and *Saccharomycopsis* species. We conduct the first systematic survey of genetic code usage across the majority of sequenced organisms, analyzing all sequenced bacteria and archaea (over 250,000 assemblies). The five new alternative genetic codes described here substantially expand the known diversity of codon reassignments in bacteria. Now, in addition to reassignments of the stop codon UGA to tryptophan or glycine, we have the first sense codon reassignments in bacteria, affecting the arginine codons AGG, CGA, and CGG. Two reassignments occur in culturable bacteria—in *Anaerococcus* and *Peptacetobacter*—and could be experimentally confirmed in the future, for example by proteomic mass spectrometry.

Since Codetta selects the most likely amino acid translation among the twenty canonical amino acids, some types of codon reassignments may be missed. We cannot predict reassign-

ment to a non-canonical amino acid—for such codons, Codetta would pick the non-specific model or an amino acid that is used similarly in other species. We also cannot directly detect ambiguous translation, which may represent an important stage in codon reassignment. However, the failure to infer an amino acid translation despite a significant number of aligned Pfam consensus columns may hint at ambiguous translation, as was the case for CUG in *Ascoidea* and *Saccharomycopsis*. Since we do not model translational initiation and termination, we cannot detect the use of new start and stop codons or context-dependent stop codons that also possess an amino acid meaning, known to occur in some eukaryotes (Swart *et al.*, 2016; Heaphy *et al.*, 2016; Záhonová *et al.*, 2016).

Expanding our analysis to eukaryotic, organellar, and viral genomes will help fill in the diversity of alternative genetic codes, but poses additional challenges. Since we align profile HMMs to a six-frame translation of the entire genome, the pervasive pseudogenes in many eukaryotic genomes will likely increase the rate of incorrect codon inferences by having sufficient homology for alignment but enough accumulated mutations to cause incorrect pairing of codons to consensus columns. Smaller scale surveys of eukaryotic genetic code diversity have focused on transcriptomic datasets (Swart *et al.*, 2016; Heaphy *et al.*, 2016), which may alleviate this problem. Some viral and organellar genomes have very few protein-coding genes which may limit the ability to confidently infer the entire genetic code. One strategy is to improve sensitivity at the cost of generalizability by using clade-specific profile HMMs instead of Pfam, which may increase the proportion of aligned coding sequence. Another challenge in some organellar genomes is extensive mRNA editing (Gray, 1996; Alfonzo *et al.*, 1997), which violates our assumption that the genomic codon sequence represents the mRNA sequence and may require analyzing the edited transcriptome to ensure correct correspondence of codons

to profile HMM positions.

In the 'codon capture' model of codon reassignment, genome-wide pressures such as biased GC content or genome reduction drive a codon to near extinction such that the codon can acquire a new tRNA decoding with a minimal effect on translation (Osawa and Jukes, 1989). Most UGA reassignments in bacteria occur in clades with very low genomic GC content, which is thought to have reduced UGA to very low usage in favor of the stop codon UAA. This includes the Mycoplasmatales and Entomoplasmatales (0.24-0.39 GC) (Jukes, 1985; Bové, 1993), Absconditabacteria and Gracilibacteria (0.21-0.53 GC, Figure 4.1) (Campbell *et al.*, 2013; Rinke *et al.*, 2013), and most insect endosymbiotic bacterial reassignments (0.13-0.17 GC, except for *Hodgkinia cicadicola* with 0.28-0.58 GC) (McCutcheon and Moran, 2010; Bennett and Moran, 2013; Salem *et al.*, 2017; Campbell *et al.*, 2015). The CGA and/or CGG reassignments described here similarly exhibit low genomic GC content (0.26-0.38) and very rare usage of GC-rich codons including CGA and CGG. A codon does not need to completely disappear for reassignment to be facilitated by rare codon usage, and it is likely that a brief period of translational ambiguity or inefficiency helps drive the remaining codon substitutions. We posit that, in bacteria, reduction in codon usage driven by genome-wide processes plays a major role in enabling codon reassignment and may explain why codon reassignments repeatedly evolve in clades such as Firmicutes (known for their low genomic GC content) and lifestyles such as endosymbiosis (which is often accompanied by genome reduction and low GC content) (McCutcheon and Moran, 2011).

All five of the new reassignments affect arginine codons (AGG, CGA, and CGG). While these are the first instances of arginine codon reassignment in non-organellar genomes, several arginine reassignments are known in mitochondria: in various metazoan mitochondria the

codons AGA and AGG have been reassigned to serine, glycine, and possibly stop and AGG has been reassigned to lysine (Knight *et al.*, 2001a; Abascal *et al.*, 2006a), and in various green algal mitochondria AGG has been reassigned to alanine and methionine and CGG to leucine (Noutahi *et al.*, 2019). Arginine codons have several unique features that may predispose them to codon reassignment. First, across the tree of life, arginine has an over-representation of codons in the genetic code relative to usage in proteins (Jukes *et al.*, 1975; King and Jukes, 1969), contributing to rare usage of the least favored arginine codon. Second, the six arginine codons range from one to three GC nucleotides in composition (only equaled by leucine), which may create greater bias in codon usage in response to genomic GC content than for amino acids with less GC variability in their codons. In organisms with small genomes, these features alone might make the rarest arginine codon very low in number and more susceptible than other codons to reassignment. The arginine codon CGG may be even more of a target for reassignment because, in most bacteria, the only widespread instance of inosine tRNA wobble is used to decode the CGU, CGC, and CGA arginine codons (Grosjean *et al.*, 2010). In the absence of a $tRNA_{UCG}$, CGG is decoded by a dedicated $tRNA_{CCG}$ and can be reassigned without affecting the translation of other codons.

Some codon reassignments have convergently reappeared across the tree of life: CGG to tryptophan in three bacterial clades described here, AGG to methionine in the clade of *Enterosoma* described here and in green algal mitochondria (Noutahi *et al.*, 2019), UGA to tryptophan in multiple bacterial, mitochondrial, and eukaryotic lineages (Knight *et al.*, 2001a), and others. Recurrent changes could reflect 1) a common evolutionary process, e.g. low GC content-driven reassignments disproportionately affecting codons sensitive to GC fluctuations, or 2) shared constraints imposed by conserved translational machinery, including tRNAs and

aminoacyl-tRNA synthetases. For example, the tRNA anticodon-codon pairing rules dictate that U- and C-ending codons cannot be assigned separate meanings, and indeed this has not been observed in any known genetic codes. This may explain why in low GC content genomes, we see reassignments of the arginine codon CGG but not the arginine codon CGC, which would have to be reassigned together with CGU. The selection of amino acid changes in the codon reassignments described here is not clearly explained by biochemical similarity (except possibly for the reassignment of CGG from arginine to glutamine). The amino acid choice may be related to the constraints on evolving new tRNA anticodons. Most of the changes described here (and indeed all of the changes known in bacteria) involve a single nucleotide difference from cognate anticodons: $tRNA_{CCU}$ in addition to $tRNA^{Met}_{CAU}$ for the AGG to methionine reassignment, $tRNA_{CCG}$ in addition to $tRNA^{Trp}_{CCA}$ for the CGG to tryptophan reassignments, and $tRNA_{CCG}$ in addition to $tRNA^{Gln}_{CUG}$ for the CGG to glutamine reassignment. Evolving a new anticodon through a single mutation may be more probable than through multiple mutations. However, the methionine $tRNA_{CCU}$ involved in the reassignment of AGG in a clade of *Enterosoma* appears to have evolved from an arginine $tRNA_{CCU}$ through mutations that altered aminoacyl-tRNA synthetase recognition, rather than by an anticodon mutation to a methionine $tRNA_{CAU}$ gene. Alternatively, this pattern could result from a limitation on the new anticodons that an aminoacyl-tRNA synthetase could accept, since most aminoacyl-tRNA synthetases use the anticodon in part to distinguish cognate and non-cognate tRNAs (Giegé *et al.*, 1998). Upon characterizing the diversity of genetic codes in other parts of the tree of life, we may discover that the general patterns and evolutionary pressures differ from bacteria, reflecting differences in translational machinery, lifestyle, or genome characteristics.

2.5 Materials and methods

Computational inference of the genetic code from nucleotide sequence

A preliminary translation of the input nucleotide sequences is produced by first breaking any long sequences into non-overlapping 100 Kb pieces (because of a limit on input protein sequence length for `hmmscan`), then translating into all six frames (as six polypeptide sequences) using the standard genetic code with stop codons translated as 'X'. A custom version of Pfam 32.0 profiles was produced from Pfam seed alignments using `hmmbuild --enone`, which turns off entropy weighting, resulting in emission probability parameters closer to the original amino acid frequencies in the input alignments. Significant homologous alignments were identified by searching each translated polypeptide against the custom Pfam database using `hmmscan` from HMMER 3.1b2 (Eddy, 2011) for domain hits with E-value $< 10^{-10}$.

Alignments were further filtered to remove uncertainly aligned consensus columns (posterior probabilities of alignment <95%). By default, no single Pfam consensus columns was allowed to account for more than 1% of total aligned consensus columns for a codon, in order to mitigate some artifacts such as repetitive pseudogene families in some genomes; when this happened, the number of codon positions aligned to that specific consensus column was downsampled to 1% of the total (if a codon was aligned to fewer than 100 Pfam consensus columns total, then each unique consensus columns was downsampled to 1 occurrence). We excluded hits to five classes of Pfam models including mitochondrial proteins, viral proteins, selenoproteins, pyrrolysine-containing proteins, and proteins belonging to transposons and other mobile genetic elements. These filtered sets of aligned consensus columns defined the input $\vec{C}$ sets for each codon. The equations from the main text are then used (in log-probability calculations for numerical stability) to infer $P(M|\vec{C}^Z)$ for each codon, with a default decoding

probability threshold of 0.9999.

The computational requirements are dominated by the `hmmscan` step, which takes about an hour on a single CPU core for a ∼12 Maa six-frame translation of a typical 6 Mb bacterial genome. We ran different genomes in parallel on a 30,000 core computing resource, the Harvard Cannon cluster. We implemented this method as Codetta v1.0, a Python 3 program that can be found at `https://github.com/kshulgina/codetta/releases/tag/v1.0`.

Measuring error rate and power on synthetic datasets

A six-frame translation of the *E. coli* O157:H7 str. Sakai genome (GCA_000008865.2) was searched against the custom `--enone` Pfam 32.0 profile database as described above. We generated 1,000 random subsamples each of 1 through 50, 100, and 500 aligned consensus columns per sense codon and inferred the most likely decoding as described above. A codon inference was considered “true" (T) if the correct amino acid meaning was inferred, “false" (F) if an incorrect amino acid meaning was inferred, and “uninferred" (U) if the non-specific decoding was most probable or if no model surpassed the model probability threshold. For a given model probability threshold, per-codon error rate is the fraction of samples with a false inference (F / (T + F + U)). Per-codon power is the fraction of samples with a true inference (T / (T + F + U)). Both values were evaluated individually for each sense codon and also aggregated across all sense codons.

Genetic code inference of archaeal and bacterial genomes

Assembly identifiers for all archaeal and bacterial genomes were downloaded from the NCBI Genome database on June 4th, 2020, and Codetta analysis was performed on all archaeal and bacterial genome assemblies. Genetic code inference results for all analyzed genomes can be

found in Supplementary File 1. A variety of additional files supporting new genetic codes are available at `https://github.com/kshulgina/ShulginaEddy_21_genetic_codes`.

We used the NCBI taxonomy database (downloaded on July 15th, 2020) to cross-reference all assemblies with taxonomic identifiers. All analyzed genome assemblies from GenBank are associated with a NCBI taxonomic ID (taxid). Because some of these taxids correspond to subspecies or strain-level designations, we assigned a species-level taxid to each assembly by iteratively stepping up the NCBI taxonomy until a species-level node was reached. To create a dereplicated dataset, we picked one genome assembly per NCBI species-level taxid. If multiple genome assemblies were associated with an NCBI species-level taxid, assemblies were sorted based on RefSeq category (reference, representative, or neither) and then genome completeness level and a single genome assembly was randomly selected from the highest ranked category.

Phage assemblies derived from the same metagenomic samples as the AGG-recoding *Enterosoma* were obtained by identifying the phage assemblies from Al-Shayeb *et al.* (2020) whose sample accessions were linked to the metagenomic sequencing experiments SRX834619, SRX834622, SRX834629, SRX834636, SRX834653, SRX834655, or SRX834666. Codetta analysis of the phage genomes was performed as described above.

Cross-referencing the NCBI taxonomy with known distributions of genetic code usage

A complete list of bacterial clades previously known to use alternative genetic codes was collated with corresponding references for genetic code discovery and taxonomic distribution (Table 2.1). For each clade, we determined a set of NCBI taxids best defining the phylogenetic extent of each reassigned clade. We used this to generate a curated genetic code annotation

for all NCBI species-level taxids: for the taxids defining each reassigned clade, all species-level child nodes were annotated with the alternative genetic code; all remaining species-level taxids were annotated with the standard genetic code.

We used the Genome Taxonomy Database (GTDB, version R05-RS95) (Parks *et al.*, 2020) to determine the phylogenetic placement of species that use candidate new genetic codes and to identify the most closely related outgroup species.

Identification of tRNA genes and other translational components

The tRNA gene content of genomes was determined by running tRNAscan-SE 2.0 (Chan *et al.*, 2021) with default settings and a tRNA model appropriate for the domain of life (i.e. option `-E` for eukaryotes, `-B` for bacteria, `-A` for archaea). To help ensure that tRNAs of interest were not missed, we also ran a low-stringency search with the general tRNA model and no cutoff score (options `-G -X 0`) and a search with previous version of tRNAscan-SE (option `-L`) and manually examined the outputs.

We searched bacterial genomes for release factor genes using the TIGRFAM 15.0 (Haft *et al.*, 2013) release factor 2 model (TIGR00020) and release factor 1 model (TIGR00019) and for the methionyl-tRNA synthetase gene using the TIGRFAM 15.0 model (TIGR00399) with `hmmscan` against a six-frame translation of the entire genome with default settings. Since the release factors are homologs, if the two models hit overlapping genomic coordinates, we kept the hit with the more significant E-value. Methionyl-tRNA synthetase alignments were generated using MAFFT v7.429 (Katoh and Standley, 2013) with default settings.

For the AGG arginine to methionine reassignment in a clade of *Enterosoma*, we classified $tRNA_{CCU}$ genes as being primarily arginine acceptors if the tRNA had A20 in the D loop and a A/G73 discriminator base, and primarily methionine acceptors if the tRNA had an A73

discriminator base and not A20 in the D loop (Giegé *et al.*, 1998). The weaker methionine identity elements G2:C70, C3:G69 in the acceptor stem were used to support the assignment (Meinnel *et al.*, 1993). In the reassignment of CGG to glutamine in *Peptacetobacter*, we classified tRNAs as arginine-type using the rules above, and as glutamine-type if the tRNA was missing arginine identity element A20 and contained the set of glutamine identity elements consisting of a weak 1:72 basepair, A37, and A/G73 (Jahn *et al.*, 1991). We took the additional glutamine identity elements G2:C71 and G3:C70 in the acceptor stem, G38 in the anticodon loop, and G10 in the D-stem as support for glutamine identity (Jahn *et al.*, 1991; Hayase *et al.*, 1992). For the reassignments of CGA and/or CGG arginine to tryptophan, we classified tRNAs as primarily arginine acceptors using the rules above, and provisionally as tryptophan acceptors if the tRNA had a G73 discriminator base and not A20 in the D loop (Giegé *et al.*, 1998). We considered the weak tryptophan identity element A/G1:U72 in the acceptor stem as support for tryptophan identity but did not require it (Himeno *et al.*, 1991). In the Absconditabacteria and Gracilibacteria, we classified $tRNA_{UCA}$ genes as glycine acceptors if the tRNA had G1:C72, C2:G71, G3:C70 in the acceptor stem and U73 discriminator base (Giegé *et al.*, 1998). We refrained from assigning identity if the tRNA did not fit the above patterns or if the D loop sequence was unusual such that it was unclear which nucleotide is N20. D loop and variable loop insertions were placed at positions following the convention of Sprinzl *et al.* (1998).

Multiple sequence alignment of BUSCO genes

For some candidate novel alternative genetic codes, we constructed multiple sequence alignments of conserved single-copy bacterial genes from the BUSCO database v3 (Waterhouse *et al.*, 2018). To identify orthologs of a BUSCO gene in a particular genome, we first created a dataset of putative protein sequences by translating all open reading frames longer than 50

codons using the inferred genetic code (assuming standard stop codons unless reassigned), with candidate reassigned codons translated as 'X'. Then, we queried each of the 148 bacterial BUSCO profile HMMs against all putative proteins using `hmmsearch` from HMMER 3.1b2 with default settings and an E-value cutoff of 10^{-13}, and picked the most significant hit if it also yielded a reciprocal best hit against the entire BUSCO profile HMM database using `hmmscan` with the same E-value cutoff. Multiple sequence alignments were generated using MAFFT v7.429 (Katoh and Standley, 2013) with default settings.

For the described novel genetic codes, BUSCO alignments containing the reassigned codon in the reassigned clade were individually inspected and alignments containing the reassigned codon at conserved positions in well-aligned regions were preferentially selected as example alignments.

Annotation of genomic context

To determine the genomic context surrounding the $tRNA_{CCU}$ gene in the *Enterosoma* genomes predicted to have reassigned AGG to methionine and in close outgroup genomes, we predicted tRNA and protein coding genes in the whole genome as described above. We annotated each putative protein coding gene with the reciprocal best hit homolog among annotated protein-coding genes in the outgroup assembly GCA_000434395.1 using `phmmer` from HMMER 3.1b2 with a 10^{-10} E-value cutoff.

Phylogenetic grouping and Codetta analysis of CUG usage by budding yeasts

For analysis of CUG translation in budding yeasts, we selected all genomes belonging to the class Saccharomycetes (NCBI taxid 4891), which represent 463 unique NCBI species taxids with at least one genome. The genomes were dereplicated to one assembly per species-level

taxid as described above. Yeast species were split into six taxonomic categories based on the "major clade annotation from the phylogenetic analysis by Shen *et al.* (2018) as follows: Outgroups (major clades: Lipomycetaceae, Trigonopsidaceae, Dipodascaceae/Trichomonascaceae, Alloascoideaceae, Sporopachydermia), CUG-Leu clade 1 (major clades: Phaffomycetaceae, Saccharomycodaceae, Saccharomycetaceae), CUG-Leu clade 2 (major clade: Pichiaceae), CUG-Ser (major clade: CUG-Ser1), CUG-Ala (major clade: CUG-Ala), and CUG-Ser/Leu (major clade: CUG-Ser2). Species that were not included in the analysis by Shen *et al.* (2018) were sorted into the same major clade as other members of their annotated genus on NCBI. A single species (*Candida* sp. JCM 15000) could not be placed into a category and was excluded from the analysis. The expected CUG translation for each clade follows Shen *et al.* (2018) and is consistent with other studies of CUG translation (Riley *et al.*, 2016; Krassowski *et al.*, 2018; Mühlhausen *et al.*, 2018). Genetic codes were predicted by Codetta as described above. A table describing all yeast genomes analyzed can be found in Figure 2.2–source data 1.

Identification of tRNA genes and isotype classification in yeasts

tRNA gene content of yeast genomes was determined using tRNAscan-SE 2.0 as described above. In eukaryotes, only leucine- and serine-tRNAs have a long (>12 nucleotide) variable loop so we used this feature to confirm the $tRNA_{CAG}$ identity as serine or leucine. In yeast, serine tRNAs typically have a conserved G73 discriminator base but can tolerate any nucleotide (Himeno *et al.*, 1997), while leucine tRNA identity is conferred by a A73 discriminator base and A35 and G37 in anticodon loop (Soma *et al.*, 1996). We categorized $tRNA_{CAG}$ genes as either serine-acceptors or leucine-acceptors based on the presence of these features. In some CUG-Ser clade species, serine CAG-tRNAs containing a G37 have been found to be charged with leucine at a low level (3%) (Suzuki *et al.*, 1997); for categorization purposes, we would

consider these tRNAs to be primarily serine-acceptors.

S. malanga growth and RNA extraction

S. malanga (NRRL Y-7175) was obtained from the Agricultural Research Service Culture Collection (Peoria, Illinois USA). Cells were inoculated into 5 mL of YPD liquid media (containing 1% yeast extract, 2% peptone, and 2% dextrose) from a colony on a YPD agar plate and grown to saturation for 4 days at 25°C on rotating wheel.

Total RNA was extracted in acidic conditions to preserve tRNA charging, following the steps outlined in Varshney *et al.* (1991) with the following modifications. Cells were harvested by centrifugation (5 minutes at 4,000 rpm at 4°C), resuspended in 500 μL ice cold buffer containing 0.3 M NaOAc pH 4.5 and 10 mM EDTA and added to 500 μL ice cold phenol:chloroform (pH 4.5) and 500 μL of 0.4-0.5μm acid washed glass beads for cell lysis. All RNA extraction steps were performed at 4°C. In the first round of extraction, cells were vortexed for 30 minutes, rested on ice for 3 minutes, centrifuged for 15 minutes at 20,000$\times g$, and the aqueous layer was transferred to 500 μL of phenol:chloroform (pH 4.5), which was subject to a second round of extraction (identical, except for 3 minute vortex). A last round of extraction was performed in 500 μL of chloroform with a 15 second vortex and 2 minute centrifugation. RNA in the aqueous phase was precipitated and resuspended in buffer containing 10 mM NaOAc pH 4.5 and 1 mM EDTA.

Northern blotting for tRNA expression

The single-stranded DNA probes used for detection of *S. malanga* $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ (5′ GAAATCCCAGCGCCTTCTGTGGGCGGCGCCTTAACCAAACTCGGC 3′) and *S. malanga* $\text{tRNA}^{\text{Leu}}_{\text{CAG}}$ (5′ TTGACAATGAGACTCGAACTCATACCTCCTAG 3′) were 5′ end-labelled with [γ-P^{32}]-ATP

by T4 polynucleotide kinase (New England Biosciences) and purified using ProbeQuant G-50 Micro Columns (GE Healthcare Life Sciences).

In vitro transcribed tRNAs were used as controls for probe specificity. For the *S. malanga* $tRNA^{Ser}_{CAG}$ probe, an *in vitro* transcribed version of the target $tRNA^{Ser}_{CAG}$ was used as a positive control and an *in vitro* transcribed version of $tRNA^{Ser}_{CGU}$ was used as a control for cross-hybridization. For the $tRNA^{Leu}_{CAG}$ probe, an *in vitro* transcribed version of the target $tRNA^{Leu}_{CAG}$ was used as a positive control and an *in vitro* transcribed version of $tRNA^{Leu}_{CAA}$ was used as a control for cross-hybridization. Cross-hybridization controls were selected by aligning the reverse complement of the probe sequence using MAFFT v7.429 (Katoh and Standley, 2013) with default settings to all tRNA genes in the *S. malanga* genome (found by tRNAscan-SE 2.0), and selecting the non-target tRNA with the highest pairwise alignment score. *In vitro* transcribed tRNAs were produced using the MAXIscript T7 Transcription Kit (Thermo) from a DNA template composed of a T7 promoter (5′ GATCTAATACGACTCACTATAGGGAGA 3′) followed by the tRNA sequence (Figure 2.2–source data 2). The resulting tRNA transcript has an additional six nucleotides of the promoter included at the 5′ end. CCA-tails were not included in the *in vitro* transcribed tRNA sequences.

Total RNA and *in vitro* transcribed controls for probe specificity were denatured in formamide buffer (Gel Loading Buffer II, Thermo) at 90°C for 5 minutes and electrophoretically separated on a 10% TBE urea gel (Novex). Gels were rinsed in 0.5x TBE and RNA was transferred onto a Hybond N+ membrane (GE Healthcare) in 0.5x TBE by semi-dry transfer (Bio-Rad Transblot) at 3 mA/cm^2 for 1 hr. Blots were crosslinked on each side using a Stratalinker UV crosslinker on the "auto-crosslink setting. Blots were prehybridized in PerfectHyb Plus Hybridization buffer (Sigma) at 64°C for 1 hour prior to incubation with the radiolabelled

DNA probe overnight. Blots were washed at 64°C twice in low stringency buffer (0.1% SDS, 2x SSC) for 15 minutes and once in high stringency buffer (0.1% SDS, 0.1x SSC) for 10 minutes, exposed on storage phosphor screens, and scanned using a Typhoon imager.

Acid urea PAGE Northern blotting for tRNA charging

For the partial deacylation control, total RNA was treated in 100 mM Tris pH 7.0 at 37°C for 30 minutes, quenched with an equal volume of buffer containing 50 mM NaOAc and 100mM NaCl, and precipitated. Electrophoresis on acid urea polyacrylamide gels was performed as described in Varshney *et al.* (1991). 4 μg of total RNA and partial deacylation control in acid urea sample buffer (0.1 NaOAc pH 4.5, 8M urea, 0.05% bromophenol blue, 0.05% xylene cyanol) were loaded onto a 0.4mm thick 6.5% polyacrylamide gel (SequaGel) containing 8M urea and 100mM NaOAc pH 4.5 and run for 18 hours at 450 V in 4°C with 100mM NaOAc pH 4.5 running buffer. The region between the two dyes corresponds to the tRNA size range, and was cut out and transferred onto a blot for probing following the same steps as above for Northern blotting.

Acknowledgments

We thank members of the Eddy lab for discussions and for comments on the manuscript, A. Murray and A. Darnell for advice and guidance on Northern blotting experiments, and R. Helmiss for feedback on data presentation. Computations were performed on the Cannon cluster, supported by the Harvard FAS Division of Science's Research Computing Group.

Additional files

Supplementary files and source data files can be obtained at `http://eddylab.org/publications/Shulgina21/Shulgina21-supplement.tar.gz`.

Supplementary files

- Supplementary File 1. Spreadsheet of predicted genetic codes by Codetta across all analyzed genome assemblies.
- Supplementary File 2. Spreadsheet of all candidate novel genetic codes predicted by Codetta across all analyzed genome assemblies.

Source data files

- Figure 2.2–source data 1. Table of all analyzed yeast genomes with phylogenetic grouping and Codetta CUG inference.
- Figure 2.2–source data 2. Table of *in vitro* transcription DNA template sequences.
- Figure 2.3–source data 1. Table of genome accessions, Codetta AGG inference, tRNA gene presence, codon usage, and genome GC content for the reassigned AGG→Met *Enterosoma* and outgroup species.
- Figure 2.4–source data 1. Table for each CGA/CGG reassignment containing genome accessions, Codetta CGA/CGG inference, tRNA gene presence, codon usage, and genome GC content for the reassigned clade and outgroup species.

Preface to Chapter 3

The work presented in Chapter 3 is in press and presented here with minor modifications. Supplementary text, tables, and figures are provided in Appendix C and supplementary files can be downloaded at `http://eddylab.org/publications/Shulgina22/Shulgina22-supplement.tar.gz`.

Shulgina, Y. and Eddy, S. R. (in press). Codetta: predicting the genetic code from nucleotide sequence. *Bioinformatics*. btac802.

Chapter 3

Codetta: predicting the genetic code from nucleotide sequence

3.1 Abstract

Summary: Codetta is a Python program for predicting the genetic code table of an organism from nucleotide sequences. Codetta can analyze an arbitrary nucleotide sequence and needs no sequence annotation or taxonomic placement. The most likely amino acid decoding for each of the 64 codons is inferred from alignments of profile hidden Markov models of conserved proteins to the input sequence.

Availability and implementation: Codetta 2.0 is implemented as a Python 3 program for MacOS and Linux, and is available from `http://eddylab.org/software/codetta/codetta2.tar.gz` and at `http://github.com/kshulgina/codetta`.

3.2 Introduction

The genetic code table defines how mRNA codons are interpreted into an amino acid sequence. Almost all of life uses the same decoding scheme; however, dozens of clades have evolved alternative genetic codes where the meaning of one or more codons is changed (Knight *et al.*, 2001a). Knowing the genetic code is necessary to correctly predict protein sequences, but common practice is to assume the standard genetic code. New alternative genetic codes continue to be reported (Swart *et al.*, 2016; Shulgina and Eddy, 2021; Borges *et al.*, 2022), especially as advances in metagenomic sequencing expand known microbial diversity. Predicted proteins from newly sequenced organisms are added to protein sequence databases which primarily consist of *in silico* translations assuming a genetic code, underscoring the need for a broadly usable genetic code prediction method.

Computational methods for genetic code prediction have been developed since the early 2000s (Telford *et al.*, 2000; Abascal *et al.*, 2006b; Dutilh *et al.*, 2011; Mühlhausen and Kollmar, 2014b; Noutahi *et al.*, 2017). The general approach is to align the input nucleotide sequence to profiles of conserved proteins, and then, for each codon, to tally the most frequently aligned amino acid. These methods have led to the discovery of new genetic codes, but have been limited in usage due to phylogenetic constraints or error rates incompatible with large-scale analyses (Noutahi *et al.*, 2017).

Codetta is a Python program for genetic code prediction that can scale to analyze large numbers of genomes. In a recent publication, we used an early version of Codetta to screen over 250,000 bacterial and archaeal genomes and found five clades using new genetic codes (Shulgina and Eddy, 2021). Codetta improves upon previous efforts by using a probability model with a robust statistical footing to infer the amino acid for each codon and by aligning

to profile hidden Markov models (HMMs) from large publicly available datasets. Here we describe Codetta 2.0, substantially revised for easier setup and usage with new options for using custom profile HMM databases and for simple parallelization on a computing cluster.

3.3 Usage

Codetta takes nucleotide sequences from a single organism as input (DNA or RNA; any level of assembly completion; no annotation needed) and predicts the genetic code from coding regions with recognizable homology. For each codon, the best amino acid meaning is selected; thus, Codetta can detect canonical stop and sense codons with new amino acid meanings (i.e. stop-to-sense and sense-to-sense reassignments). Codetta does not predict termination codons and cannot detect sense-to-stop reassignments.

Codetta performs three main steps, in order:

1. *Alignment.* A preliminary six-frame standard genetic code translation of the input sequence is aligned to profile HMMs of conserved proteins using the HMMER `hmmscan` program (Eddy, 2011). A typical source of profile HMMs is the Pfam database (Mistry *et al.*, 2020), but custom profile HMMs can be provided with the `--profiles` option. To avoid some sources of alignments that can confound genetic code inference, by default, we exclude Pfam domains that commonly pseudogenize (i.e. mobile genetic elements) or come from mitochondrial contigs; these can be optionally included.

2. *Processing.* Correspondence between codons and aligned profile HMM consensus columns is processed into a single file.

3. *Inference.* For each codon, the most likely amino acid is chosen based on a probability

model of how profile HMM consensus columns match with codons in the nucleotide sequence (Shulgina and Eddy, 2021). Codetta considers 20 amino acid models and one model of non-specific translation. If an amino acid model has probability above a threshold (default 0.9999), then the amino acid meaning is selected. Otherwise the codon is left uninferred ('?' output), which results from insufficient or ambiguous information. This is also the expected outcome for stop codons, which are not explicitly predicted by Codetta. Simulations show that observing a codon 10-30 times in Pfam alignments suffices to infer the correct amino acid (Shulgina and Eddy, 2021).

These three steps are bundled in the `codetta.py` program and are also provided as separate programs. More detail on usage options and a tutorial can be found in the associated README document.

The most computationally intensive step is the `hmmscan` alignment of profile HMMs. Analyzing a 4.6 Mb *Escherichia coli* genome with Pfam 35.0 as a profile HMM database takes about 30 minutes on a single 2.0 GHz CPU. Speed depends on the length and number of input sequences and on the size of the profile HMM database (Eddy, 2011). Codetta can parallelize the `hmmscan` processes across machines on a SLURM computing cluster with the `--parallelize_hmmscan` option.

Codetta occasionally makes incorrect predictions stemming from its underlying assumptions. Known sources of error include profile HMMs aligning to non-coding sequence, often at recent pseudogenes; extensive mRNA editing relative to the input sequence; and input sequences that differ significantly in amino acid composition from the profile HMMs (Shulgina and Eddy, 2021). Candidate new genetic codes should be validated, ideally experimentally.

3.4 Performance

Codetta was designed to predict the genetic code from nucleotide sequence for species with potentially no close taxonomic relatives. Most existing tools are clade-restricted (Abascal *et al.*, 2006b; Mühlhausen and Kollmar, 2014b) or require the input of a phylogenetic tree and annotated genes (Noutahi *et al.*, 2017). The only other method for arbitrary nucleotide sequences is FACIL (Dutilh *et al.*, 2011). FACIL similarly uses Pfam alignments and for each codon simply selects the most frequently aligned amino acid, filtering out low confidence predictions at a later step.

We benchmarked Codetta and FACIL on two test sets: 506 yeast nuclear genomes with three genetic codes varying at the CUG codon (13.8 Mb average genome size) and 82 mitochondrial genomes with 18 alternative genetic codes affecting 14 different codons (27 Kb average genome size). For the mitochondrial comparison, we used a custom-built profile HMM database of mitochondrial proteins to demonstrate usage of the custom profiles feature; we also repeated the Codetta analysis with Pfam. Complete results and a breakdown of per-codon prediction accuracy are provided in Supplementary Files and in Appendix C.

Genetic code prediction requires high sensitivity to find rare codon reassignments but also a high positive predictive value (PPV) to avoid overwhelming predictions with errors. To assess how well Codetta and FACIL predict codon reassignments, we identified codon predictions that differed from the standard genetic code (predicted codon reassignments) and compared to the annotated genetic code to determine if they are correct (Figure 3.1). For the yeast test set, Codetta detected almost all of the 174 reassigned CUG codons and had an overall PPV of 96.6%. The six stop-to-sense errors were likely due to alignment of Pfam models to pseudogenes. In contrast, FACIL missed most of the true CUG reassignments and half of the

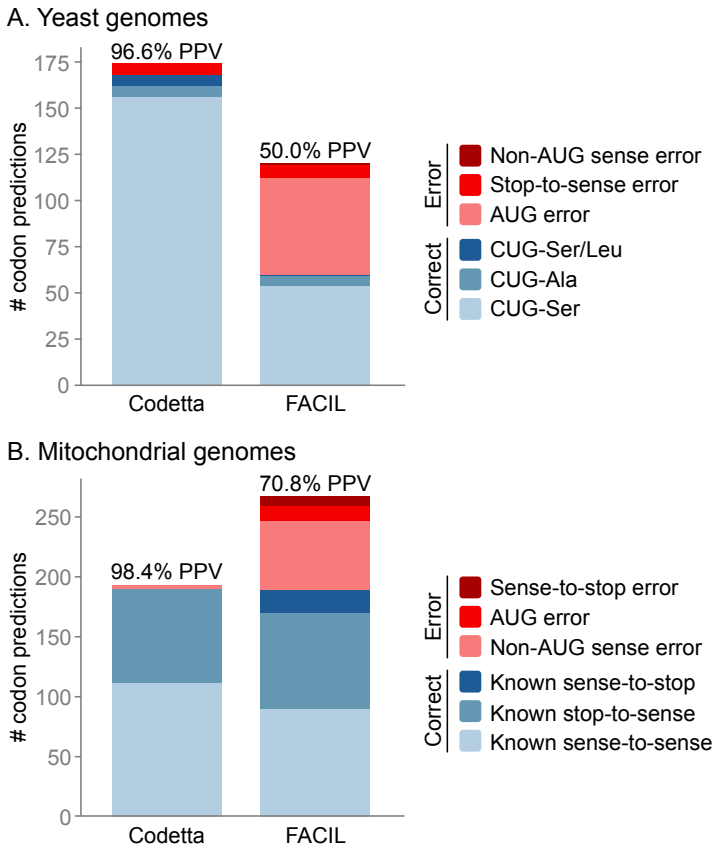

Figure 3.1: Breakdown of predicted codon reassignments for the yeast genomes test set (A) and mitochondrial genomes test set (B) with the associated positive predictive value (PPV).

predicted reassignments were errors (50.0% PPV), most commonly at AUG codons. For the mitochondrial test set, FACIL found a similar number of true reassignments but had a lower PPV (70.8%) than Codetta (98.4%). Since FACIL explicitly predicts stop codons, it was able to predict 19 true sense-to-stop reassignments missed by Codetta but also made 9 erroneous sense-to-stop predictions. In general, Codetta makes fewer sense codon errors because it leaves codons uninferred in the face of insufficient or conflicting information, which can be seen in the greater number of uninferred sense codons for mitochondria relative to FACIL (Table C.1). Predictions can be made less conservative by adjusting the probability threshold for predictions (`-r` option).

In the mitochondrial test set, Codetta predicted three sense codon reassignments that

we counted as errors in Figure 3.1. Two green algal mitochondrial genomes had the rare isoleucine codon AUA predicted as methionine and one chytrid mitochondrial genome had the lysine codon AAG predicted as methionine. Multiple sequence alignments of mitochondrial proteins show all three of the putative reassigned codons at conserved methionine positions (Figures C.2-C.4), suggesting these may be bona fide reassignments, pending additional confirmation.

3.5 Conclusion

In Shulgina and Eddy (2021), we analyzed all available bacterial and archaeal genomes, surpassing previous screens of genetic code diversity in scale. We have now extended Codetta 2.0 to be a well-documented and user-friendly tool with options to use custom profile HMM databases and to parallelize large analyses. As sequenced microbial diversity continues to grow, Codetta will allow confirming the genetic code to become a straightforward step in genome annotation, enabling discovery of new genetic codes and ensuring the accuracy of protein sequence databases.

3.6 Methods

Generating test sets for performance comparison

To compare performance on the budding yeasts, we downloaded all genomes in GenBank in the class Saccharomycetes (NCBI taxonomic identifier 4891) on December 2nd, 2021, representing 509 species (defined as unique NCBI species taxonomic IDs). Some species had multiple genome assemblies; in those cases, we sorted the assemblies by RefSeq category

(reference, representative, or neither) and then by genome completeness category (complete genome, chromosome, scaffold, or contig), and randomly selected a single assembly if more than one sorted into the highest level.

We annotated the expected genetic code following the “major clade" annotation in Shen *et al.* (2018), which is also consistent with previous studies of alternative genetic codes in yeast. Species belonging to the “major clade" CUG-Ser1 were annotated as translating CUG as serine (NCBI code 12, Alternative Yeast Nuclear Code), species in CUG-Ala were annotated as translating CUG as alanine (NCBI code 26, Pachysolen tannophilus Nuclear Code), and species in CUG-Ser2 were annotated as using both the standard genetic code and NCBI code 12 due to reported ambiguous translation of CUG (Mühlhausen *et al.*, 2018; Shulgina and Eddy, 2021). All remaining species were annotated as using the standard genetic code. If a species was not included in the Shen *et al.* (2018) analysis, then it was grouped in the same major clade as other members of its NCBI family or genus. Three species could not be placed into a major clade due to uncertain taxonomic position in NCBI annotation and were excluded from the analysis. A complete list of yeast genomes used in this analysis with the expected genetic codes can be found in Supplementary File 1.

For comparing performance on mitochondrial genomes, we collated a list of reported mitochondrial alternative genetic codes from the literature and a list of species reported to use each code. For evaluating performance, we excluded alternative genetic codes with disagreement in the literature, and those reported in only one species by a single reference. For each species, we selected a single mitochondrial genome from GenBank, picking the NCBI Reference Sequence if possible. A complete list of reported mitochondrial genetic codes and genomes used in this analysis can be found in Supplementary File 2.

Evaluating FACIL and Codetta performance

Complete results for both Codetta and FACIL can be found in Supplementary Files 3 and 4.

We compared performance against FACIL v1.0, run with default parameters (Dutilh *et al.*, 2011). Internally, FACIL uses Pfam 22.0 as the source of profile HMMs and HMMER 3.0 for profile HMM alignment.

For the yeast analysis, we ran Codetta 2.0 with default parameters (decoding probability threshold of 0.9999) using Pfam 35.0 as the source of profile HMMs and HMMER 3.3.2 for profile HMM alignment (Eddy, 2011; Mistry *et al.*, 2020). For the mitochondrial genome analysis with Codetta, we used a lower decoding probability threshold of 0.999 to improve sensitivity and a custom profile HMM database specific to mitochondrial proteins. A Codetta analysis using Pfam 35.0 was also performed with a decoding probability threshold of 0.999 and option `-m`, which prevents exclusion of Pfam domains commonly found in mitochondrial genomes.

The custom mitochondrial protein profile HMM database was made by taking a diverse representation of mitochondrial genomes—*Homo sapiens* (NC_012920.1), *Dictyostelium discoideum* (NC_000895.1), *Tetrahymena thermophila* (NC_003029.1), *Ancoracysta twista* (NC_036491.1), *Jakoba libera* (NC_021127.1), *Andalucia godoyi* (NC_021124.1), and *Kluyveromyces lactis* (NC_006077.1)—and generating multiple sequence alignments for some of the unique annotated protein-coding genes in each genome by searching with the HMMER `jackhmmer` program against the Swiss-Prot database (UniProt release 2020_05) (UniProt Consortium, 2021) with E-value inclusion thresholds tailored for each alignment. Profile HMMs were built from each alignment using the HMMER `hmmbuild` program with the `-enone` option to turn off entropy weighting. These profile HMMs were combined into a database, along with a set of Pfam domains commonly found in mitochondrial-encoded proteins. Since

Codetta only considers a single aligned consensus column for each codon position in the input sequence based on alignment E-value, inclusion of redundant Pfam domains serves to ensure that the maximum amount of protein-coding sequence is aligned by one profile HMM or another. The custom mitochondrial protein profile HMM database can be obtained at `http://eddylab.org/publications/Shulgina22/mito_profiles.tar.gz` .

Generating multiple sequence alignments of unexpected mitochondrial predictions

The mitochondrial genomes of *Pseudopediastrum boryanum* (KR026342.1), *Pedinomonas minor* (NC_000892.1), and *Rhizophydium* sp. 136 (NC_003053.1) had Codetta predictions at sense codons that differed from the known genetic code. To generate multiple sequence alignments of cytochrome c oxidase subunit 1 (COX1), NADH dehydrogenase subunit 4 (ND4), and NADH dehydrogenase subunit 5 (ND5), we first found the coding regions for these genes in the genomes of interest. For NC_000892.1 and NC_003053.1, we used the coding regions provided in the genome annotation. The genome KR026342.1 had no provided annotation, so the coding regions for these genes were found by aligning the corresponding model in the custom mitochondrial profile HMM database against a 6-frame genomic translation using the HMMER `hmmsearch` program with default settings, manually examining 100 nts upstream and downstream to identify the likely start and stop codons, and checking for group I or group II introns by aligning the region against Rfam 14.6 (Kalvari *et al.*, 2021) using Infernal 1.1.2 program `cmscan` (Nawrocki and Eddy, 2013).

An *in silico* translation of the predicted coding sequence (using the known genetic code for each organism, codons under question translated as X) was aligned to protein sequences obtained from *Homo sapiens* (NC_012920.1), *Drosophila melanogaster* (NC_024511.2), *Arabidopsis thaliana* (NC_037304.1), *Globisporangium ultimum* (NC_014280.1), *Jakoba libera* (NC_021127.1),

and *Leucocryptos marina* (NC_045933.1) using the alignment program MAFFT v7.407 with default settings (Katoh and Standley, 2013). Residues implicated in human disease were annotated according to Bridges *et al.* (2011) and (Lloyd and McGeehan, 2013).

These multiple sequence alignment are provided in Supplementary Files 5-7 and displayed in Figures C.2-C.4.

Supplementary files

Supplementary files can be obtained at `http://eddylab.org/publications/Shulgina22/Shulgina22-supplement.tar.gz`.

- Supplementary File 1. Spreadsheet with all genomes analyzed in the yeast test set and the annotated genetic code.
- Supplementary File 2. Spreadsheet of all alternative genetic codes in mitochondria, including list of genomes analyzed.
- Supplementary File 3. CSV file of yeast test set codon predictions for Codetta and FACIL.
- Supplementary File 4. CSV file of mitochondrial test set codon predictions for Codetta and FACIL.
- Supplementary File 5. Stockholm alignment of COX1 genes in mitochondria.
- Supplementary File 6. Stockholm alignment of ND4 genes in mitochondria.
- Supplementary File 7. Stockholm alignment of ND5 genes in mitochondria.

Funding

This work was supported by the National Human Genome Research Institute of the National Institutes of Health [F31-HG010984 to YS and R01-HG009116 to SRE]; and the Howard Hughes Medical Institute.

Chapter 4

Reassignment of CGA/CGG codons in four groups of bacteria

4.1 Introduction

In Chapter 2, we performed the first large-scale screen for new genetic codes across all available bacterial and archaeal genomes with Codetta. This led to the discovery of five bacterial clades using new alternative genetic codes. Surprisingly, all of the newly found codon reassignments affect arginine codons—AGG in one clade and CGA and/or CGG in four other clades. Despite both affecting arginine codons, the AGG reassignment and CGA/CGG reassignments differ significantly. The reassignment of AGG from arginine to methionine is a mature codon reassignment where AGG is abundantly used at methionine positions (Figure 2.3)—the codon substitutions to remove AGG from conserved arginine positions are a distant memory. In contrast, the CGA/CGG reassignments are very young, if not still midst the reassignment process. Across all four clades, CGA and CGG are incredibly rare codons,

with typical CGG usage less than 5 per 10,000 codons in Pfam alignments. As such, each of these predicted reassignments is more tentative. Furthermore, all four clades have very low genomic GC content, a genome-wide force which likely drove CGA and CGG codons to rarity in favor of AU-rich arginine codons like AGA and facilitated reassignment. In this chapter, I will delve into more detail about each of the CGA/CGG reassignments, including additional computational evidence in support of the predictions and whether we can learn anything about how and why these reassignments evolved.

4.2 Reassignments of UGA, CGA, and CGG in Absconditabacteria and Gracilibacteria

Two clades in the Candidate Phyla Radiation (CPR)—Absconditabacteria (SR1) and Gracilibacteria (BD1-5)—were previously described to use an alternative genetic code where the canonical stop codon UGA is translated as glycine (Campbell *et al.*, 2013; Rinke *et al.*, 2013). I confirmed using Codetta that genomes annotated as Absconditabacteria and Gracilibacteria on NCBI were predicted to translate UGA as glycine by Codetta (see Table 2.1). Some of these genomes were additionally predicted to have reassignments of the canonical arginine codons CGA and/or CGG to tryptophan.

Species phylogenetic tree

Since many genomes from the uncultivated CPR have only broad phylogenetic annotation in NCBI, I investigated the phylogenetic distribution of this reassignment on the Genome Taxonomy Database (GTDB), which has a phylogenetic tree and taxonomy for nearly all bacterial genomes (Parks *et al.*, 2020). In Figure 4.1, I show a part of the GTDB phylogenetic tree that

Figure 4.1: Phylogenetic tree from GTDB showing the Absconditabacteria, Gracilibacteria, and closest outgroup genomes (Peregrinibacteria and Peribacteria), each species indicated by a number which can be cross-referenced with genome information (Supplementary Files). Asterisks indicate genomes the most complete genomes, which have CheckM estimated genome completeness from GTDB >75% and the entire minimal set of 22 required tRNAs (excluding CGN-decoding tRNAs). For each species, the inferred translation of the three reassigned codons (UGA, CGA, and CGG) by Codetta is indicated by colored circles (red: arginine, purple: tryptophan, yellow: glycine, white: uninferred). The presence of release factor and tRNA genes that recognize the UGR- and CGN-codons is also indicated by filled circles, colored black if present and white is absent for release factor 2 and according to the predicted amino acid charging based on identity elements for tRNAs (see Methods). Anticodons in gray font are typically not found in CPR genomes. The lines connecting codons and tRNA anticodons represent the likely wobble decoding capabilities, with dashed lines representing weaker interactions. For the anticodon UCG, U34 is presumed to be modified in a way that restricts wobble to CGA and CGG, at least in the Absconditabacteria, but could potentially recognize CGU and/or CGC depending on the true modification state. The remaining anticodons are not expected to be modified in a way that alters which codons are recognized. The codon usage for the reassigned codons CGA and CGG, the arginine codons AGA, AGG, CGU, and CGC, and the tryptophan codon UGG is the frequency per 10,000 codons aligned to Pfam domains. Genomic GC content is calculated over the entire genome.

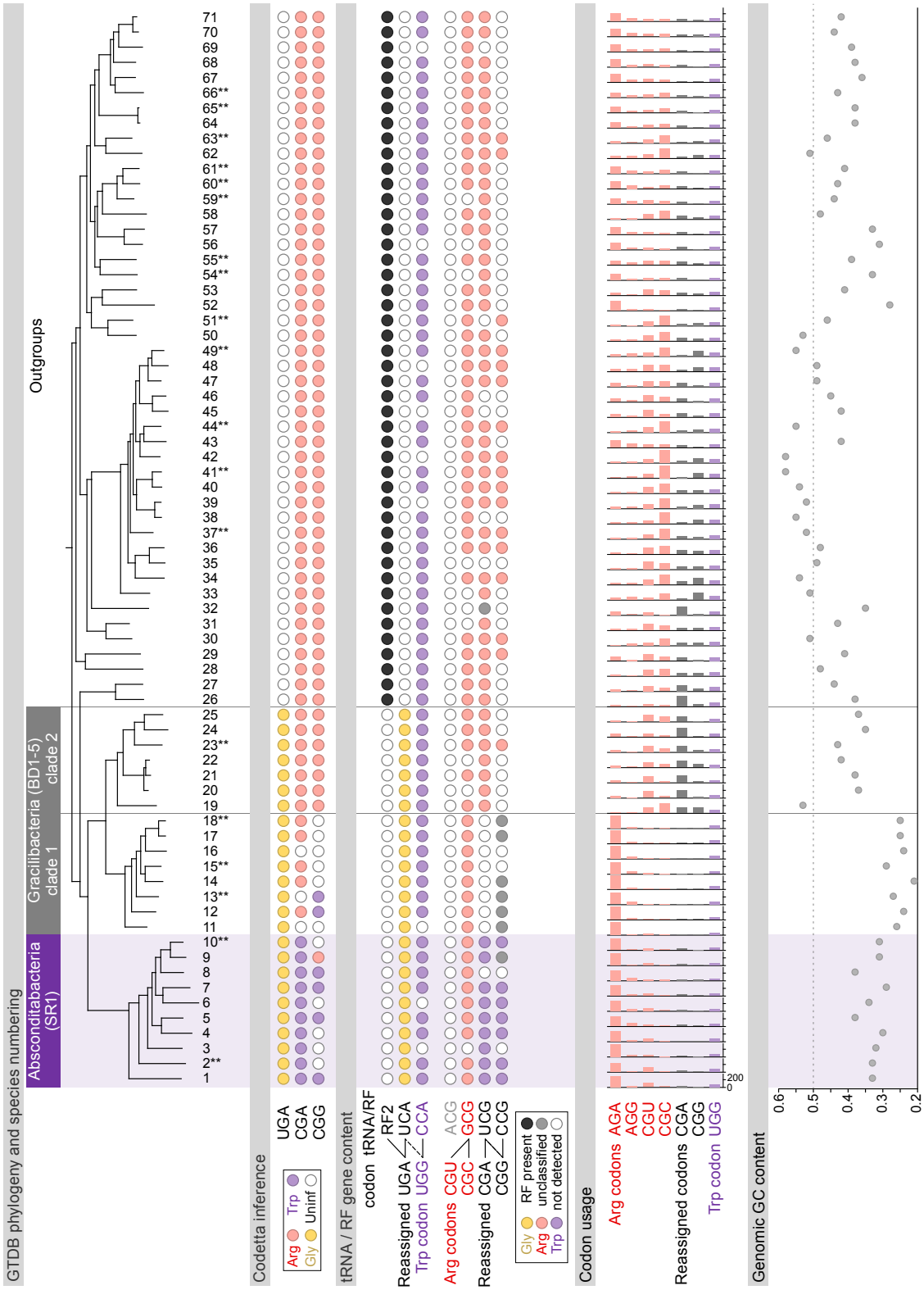

Figure 4.1 (Continued)

includes all members of the Absconditabacteria (GTDB order o__Absconditabacterales, 10 species clusters), the sister group Gracilibacteria (GTDB order o__BD1-5, 15 species clusters), and all outgroup species out to Peregrinibacteria and Peribacteria (46 species clusters).

In order to help gauge whether a tRNA gene is absent from a genome or whether it is missing due to incomplete genome assembly, I marked the most complete genomes on the tree. CheckM estimates of genome completeness (provided on GTDB for each genome; based on presence of conserved protein sets) tend to underestimate the completeness of genomes with lineage-specific gene loss, which affects CPR genomes (Parks *et al.*, 2015). Therefore, I additionally tabulated the presence of a minimal set of 22 required tRNA genes (excluding CGN-decoding tRNAs) found by tRNAscan-SE 2.0 (see Methods). In Figure 4.1, I marked the genomes which contained all 22 required tRNA genes and had CheckM completeness estimate of 75% or greater (maximum observed was 90%).

Reassignment of UGA to glycine in both Absconditabacteria and Gracilibacteria

On the GTDB phylogenetic tree (Figure 4.1), all members of the Absconditabacteria and Gracilibacteria were inferred by Codetta to use the known reassignment of the stop codon UGA as glycine, while none of the outgroup genomes were predicted to have a UGA reassignment. Consistent with a UGA reassignment, all Absconditabacteria and Gracilibacteria genomes do not have an release factor 2 (RF2) gene which terminates translation at UGA codons and most instead have a glycine-type $tRNA_{UCA}$. The absence of a $tRNA_{UCA}$ in some genomes may be due to the incomplete nature of most of these genomes, since all were either assembled from metagenomic data or derived from single-cell genome sequencing. None of the outgroup genomes encoded a $tRNA_{UCA}$ and all instead had an RF2 gene.

Reassignment of CGA and CGG to tryptophan in Absconditabacteria

In the Absconditabacteria clade on the GTDB phylogeny (Figure 4.1), all 10 species were predicted by Codetta to translate CGA as tryptophan and CGG as either tryptophan, uninferred, or arginine (in one species). The one species (#9 on tree) predicted to translate CGG as arginine may represent a true reversion of CGG back to arginine or could be an error caused by assembly contamination with standard genetic code sequences.

Figure 4.2 shows four alignments of conserved single-copy bacterial genes across Absconditabacteria, Gracilibacteria, and outgroup species. The four alignments show that multiple CGA and CGG positions across all 10 Absconditabacteria genomes are used interchangeably with the canonical tryptophan codon UGG within the Absconditabacteria and align to positions often conserved for tryptophan in the Gracilibacteria and outgroup species. In contrast, outgroup CGA and CGG codons occur at positions broadly conserved for arginine. The situation in Gracilibacteria will be discussed in the next section.

The tRNAs present in Absconditabacteria genomes are overall consistent with the inferred codon translations (Figure 4.1). The $tRNA_{CCG}$ and $tRNA_{UCG}$ genes, which decode CGA and CGG codons, were found to be predominantly tryptophan-type (classified by G73 discriminator base and the absence of A20), supporting the inferred translation for most Absconditabacteria genomes. Some genomes were missing tRNAs, however this is likely due to the incomplete nature of genomes from uncultivated bacteria. Figure 4.3A shows the sequence and likely secondary structure of the Absconditabacteria $tRNA_{UCG}$ and $tRNA_{CCG}$ from GCA_001007975.1 (candidate division SR1 bacterium Aalborg_AAW-1, #2 on tree), highlighting sequence features used for classification as tryptophan-type.

One exception was the $tRNA_{CCG}$ from GCA_002414185.1 (Patescibacteria group bacterium

Figure 4.2: Multiple sequence alignments of undecaprenyl diphosphate synthase (BUSCO POG091H00BZ), tRNA (guanine-N1)-methyltransferase (BUSCO POG091H01WE), Peptidase M50 (BUSCO POG091H0131), and DNA-directed RNA polymerase subunit beta (BUSCO POG091H02K5) from Absconditabacteria, Gracilibacteria clades 1 and 2, and selected outgroup species. Alignment regions containing CGA (α) or CGG (γ) positions are shown, with columns containing CGA or CGG in Absconditabacteria sequences or CGG in Gracilibacteria clade 1 sequences highlighted with an asterisk on the top, and columns containing CGA or CGG in Gracilibacteria clade 2 and outgroup sequences or CGA in Gracilibacteria clade 1 sequences highlighted with an asterisk below.

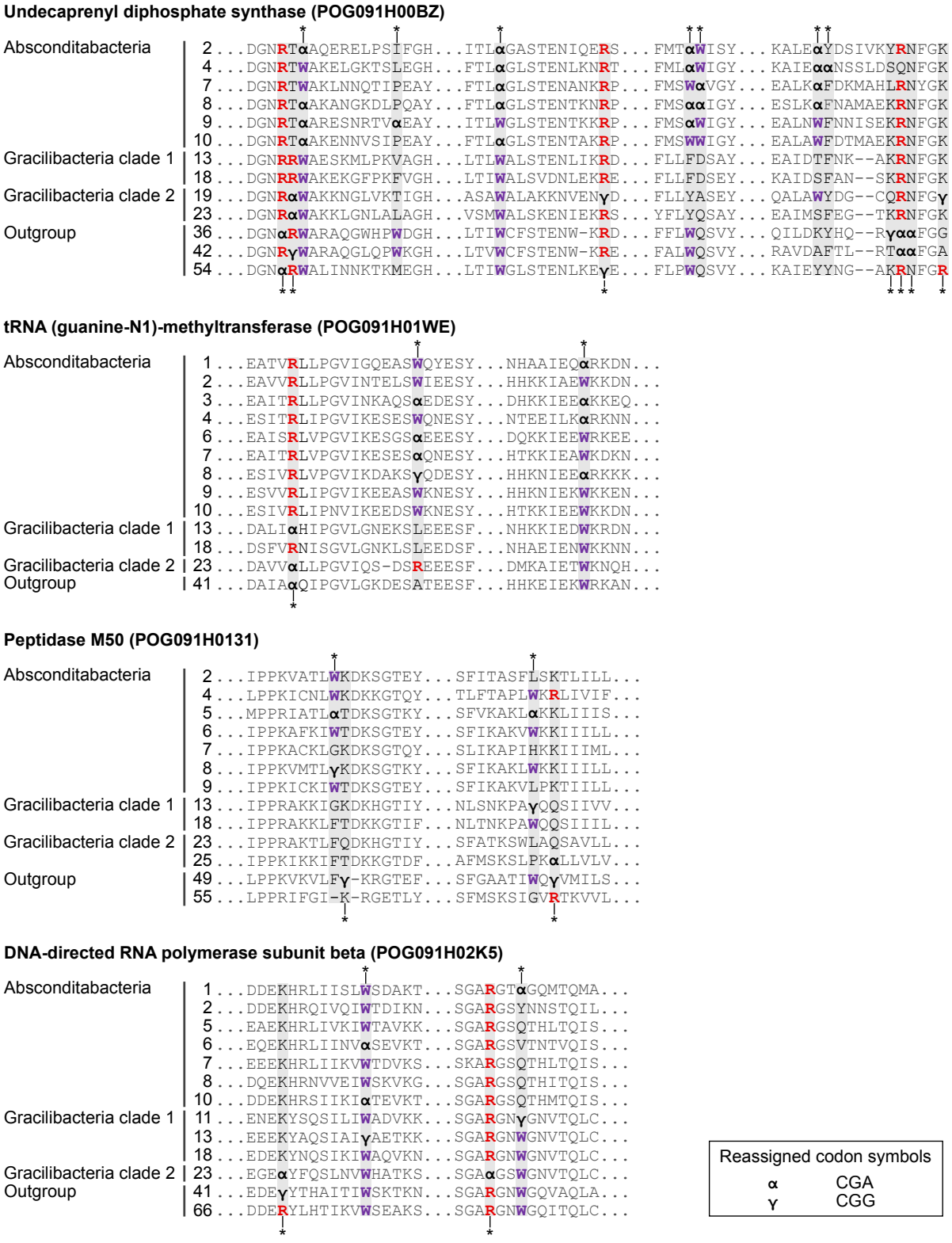

Figure 4.2 (Continued)

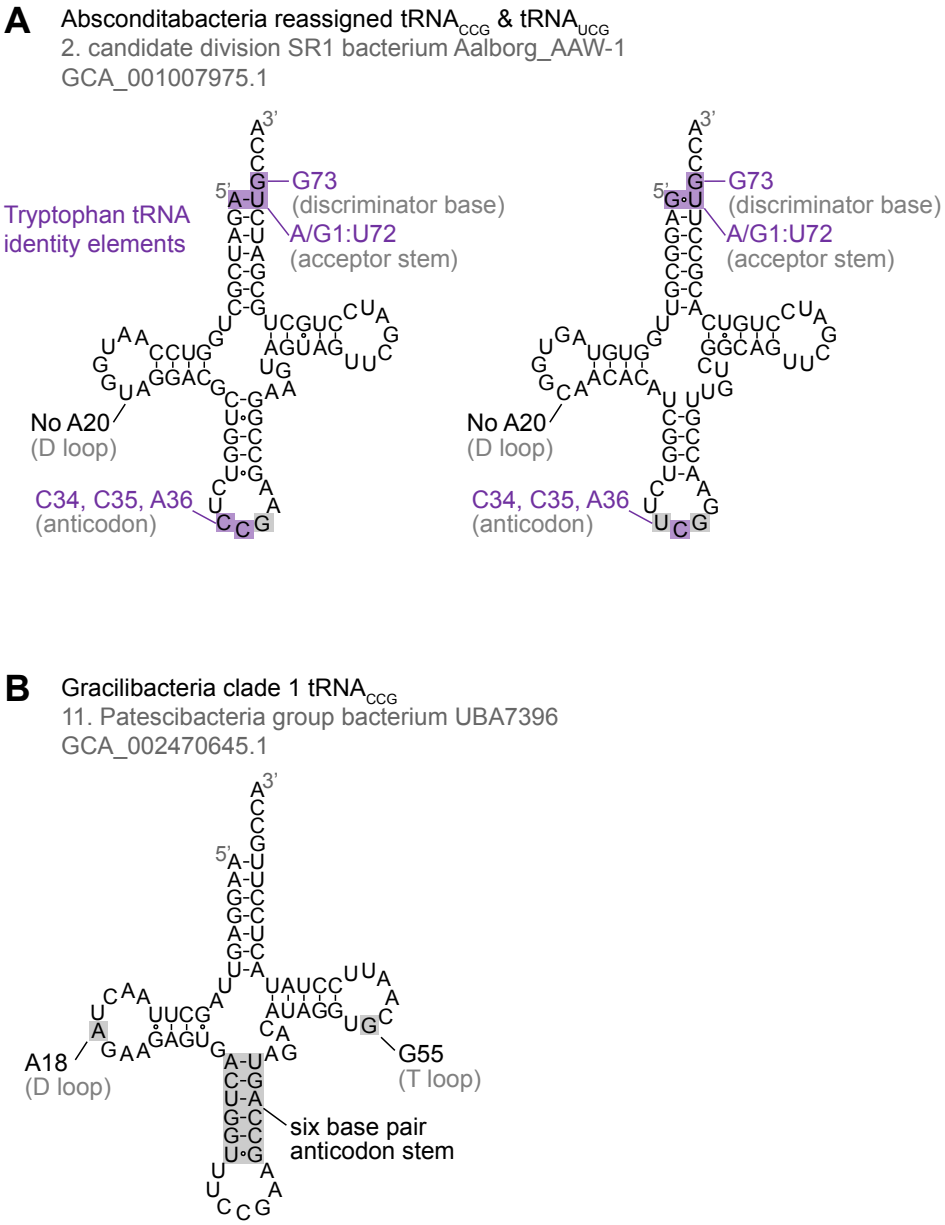

Figure 4.3: (A) The CGA- and CGG-decoding tRNAs (UCG and CCG anticodons) from species #2 (candidate division SR1 bacterium Aalborg_AAW-1, GCA_001007975.1). tRNA sequence features involved in tryptophan identity are highlighted (Giegé *et al.*, 1998; Himeno *et al.*, 1991), with nucleotide numbering following the convention of Sprinzl *et al.* (1998). Nucleotides highlighted in gray do not match the expected identity element. (B) The CGG-decoding tRNA$_{CCG}$ in Gracilibacteria clade 1 (species #11; Patescibacteria group bacterium UBA7396) with unusual structural features highlighted in gray.

UBA5124; #9 on tree), which is the only Absconditabacteria genome with an arginine translation inferred for CGG by Codetta. The $tRNA_{CCG}$ could not be classified as either arginine- or tryptophan-type because it has a A73 discriminator base, which supports arginine identity and is inconsistent with tryptophan identity, but has a C20 in the D loop instead of the A20 that would be expected of arginine tRNAs. It is possible that this unusual tRNA could support translation of CGG as arginine to some extent in this species.

In a phylogenetic tree built from arginine and tryptophan tRNA genes from across the Absconditabacteria, Gracilibacteria, and outgroup species, all $tRNA_{CCG}$ and $tRNA_{UCG}$ genes from Absconditabacteria (including GCA_002414185.1, #9 on species tree) cluster within the clade of $tRNA^{Trp}_{CCA}$ genes from across the three groups (Figure 4.4). I compared the likelihood of two phylogenetic models, one where the Absconditabacteria $tRNA_{CCG}$ and $tRNA_{UCG}$ sequences are constrained to cluster with the tryptophan $tRNA_{CCA}$ genes against another model where they are constrained to cluster with the arginine tRNAs. The log2 ratio of the likelihoods is 43, very strongly favoring the tryptophan model. This indicates that the Absconditabacteria $tRNA_{CCG}$ and $tRNA_{UCG}$ genes are more similar in sequence to $tRNA^{Trp}_{CCA}$ genes, not arginine tRNA genes, and may serve as tryptophan tRNAs in translation.

In Absconditabacteria, both CGA and CGG tend to be rare with codon usages of 14-37 per 10,000 codons in regions aligned to Pfam domains for CGA and 1-24 per 10,000 codons for CGG (Figure 4.1). At the extreme, Absconditabacteria genomes where CGG was either uninferred by Codetta or inferred to code for arginine had the lowest CGG codon usage (<3 per 10,000 codons) resulting in fewer than 10 total Pfam consensus columns aligning to CGG. In contrast, other Absconditabacteria genomes such as GCA_002791215.1 (candidate division SR1 bacterium CG_4_9_14_3_um_filter_40_9; #8 on tree) had CGA and CGG codon usages (30

Figure 4.4: An unrooted phylogenetic tree of arginine, tryptophan, and reassigned tRNA sequences from Absconditabacteria, Gracilibacteria, and outgroup genomes, generated by a Bayesian approach implemented in MrBayes 3.2.7a. Values below each internal branch indicate the posterior probability of each clade. Since this is an unrooted tree, this value represents the probability of the sequences on either side of the branch form two distinct clades. Branch length scale in the bottom right is in expected substitutions per site. Sequence labels follow the format "species number - clade _ tRNA anticodon". This tree includes sequences from $tRNA_{CCG}$ and $tRNA_{UCG}$ in the Absconditabacteria and $tRNA_{CCG}$ in Gracilibacteria clade 1 (potentially involved in codon reassignments and indicated with a gray circle), the arginine $tRNA_{UCU}$, $tRNA_{CCU}$, and $tRNA_{GCG}$ (along with $tRNA_{UCG}$ in all Gracilibacteria and outgroups and $tRNA_{CCG}$ in Gracilibacteria clade 2 and outgroups) indicated by a red circle, and the tryptophan $tRNA_{CCA}$ from all groups indicated by a purple circle.

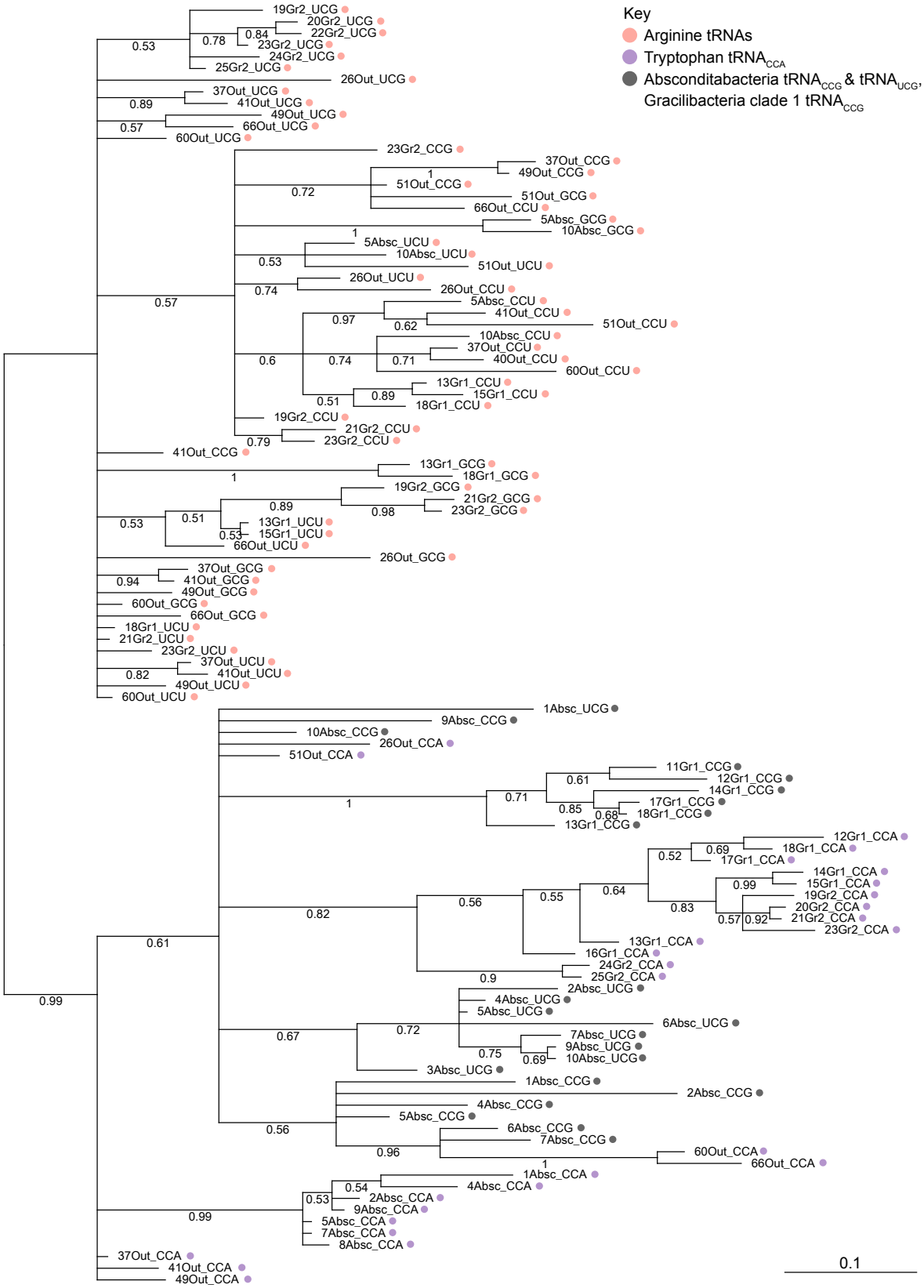

Figure 4.4 (Continued)

and 24 per 10,000) approaching that of the standard tryptophan codon UGG (35 per 10,000). The overall low usage of CGN codons in Absconditabacteria may be tied to the low GC content of the clade, which ranges between 0.29-0.38. Due to low GC content, many CGA and CGG codons were likely replaced by AT-rich alternatives, such as AGA and AGG, and this may have facilitated reassignment of CGA and CGG by reducing the number of compensatory substitutions needed to adapt to the new translation. Low GC content has been previously suggested to have played a role in the reassignment of UGA from stop to glycine in this clade by disfavoring UGA stop codon usage in favor of UAA and by creating a less GC-rich codon for glycine (which typically uses GGN codons) (Rinke *et al.*, 2013).

Translation of CGA and CGG in Gracilibacteria

The Gracilibacteria genomes on the GTDB phylogeny (Figure 4.1) split into two clades (dubbed here “clade 1" and “clade 2") that differ in codon usage, genomic GC content, Codetta CGA/CGG inference, and tRNA decoding capability. The 8 species in Gracilibacteria clade 1 have very low genomic GC content (ranging between 0.21-0.29) and extremely low usage of CGA and CGG codons (<7 per 10,000 for all genomes). Consequently, CGA and CGG codons are uninferred by Codetta for many of these species due to lack of aligned Pfam consensus columns. When inferred, CGA is predicted to be an arginine codon while CGG is predicted to code for tryptophan. In contrast, the 7 species in Gracilibacteria clade 2 have higher genomic GC content (ranging between 0.35-0.53), CGA and CGG are abundantly used codons (33-254 per 10,000). Both CGA and CGG are inferred by Codetta to be arginine codons in all Gracilibacteria clade 2 species, as in all outgroup genomes on the tree.

In the example alignments of conserved single-copy genes (Figure 4.2), CGA occurs at conserved arginine positions in members of both Gracilibacteria clade 1 and clade 2 (only

one aligned CGA position in Gracilibacteria clade 1 is shown). In Gracilibacteria clade 2, CGG also occurs at positions conserved for arginine or other positively-charged amino acids, consistent with the Codetta prediction. In Gracilibacteria clade 1, three aligned CGG positions appear in the alignments; two of them coming from GCA_002435385.1 (Patescibacteria group bacterium UBA6489; #13 on tree). These CGG positions are encoded by the tryptophan codon UGG in other Gracilibacteria clade 1 genomes, and are sometimes more broadly conserved for tryptophan in Absconditabacteria and outgroup genomes. This supports possible translation of CGG as tryptophan in some members of Gracilibacteria clade 1.

The tRNAs to decode CGA and CGG in Gracilibacteria support the divide between the clade 1 and clade 2 (Figure 4.1). Clade 2 genomes all contain an arginine-type $tRNA_{UCG}$, which would translate both CGA and CGG codons as arginine, and typically no $tRNA_{CCG}$ which would decode CGG. In contrast, none of the clade 1 genomes contain a $tRNA_{UCG}$. It is possible that I failed to find this tRNA due to a long intron or incomplete genome assembly, but if it is indeed missing, then there does not exist any tRNA that recognizes CGA by conventional anticodon pairing rules in Gracilibacteria clade 1. Interestingly, some animal mitochondria use a G34 tRNA translate an A-ending codon, and this is often accompanied by an unusual substitution of U33 to C33 in the anticodon loop (Tomita *et al.*, 1999b; Li *et al.*, 2019). U33 is considered to be an invariable tRNA position and is is critical for establishing the U-turn motif which positions the anticodon for interaction with the codon (Ashraf *et al.*, 1999). In Gracilibacteria clade 1, the $tRNA^{Arg}_{GCG}$ also has a C33 nucleotide (Supplementary Files). None of the Gracilibacteria clade 2, Absconditabacteria, or outgroup $tRNA^{Arg}_{GCG}$ share this substitution. Thus it is possible that C33 helps the tRNA anticodon loop adopt a conformation more favorable for decoding A-ending codons, and this strategy may be used by the

$tRNA^{Arg}_{GCG}$ in Gracilibacteria clade 1 to translate CGA as arginine.

A highly unusual $tRNA_{CCG}$ in Gracilibacteria clade 1

In Gracilibacteria clade 1, 6 of the 8 genomes contain a CGG-decoding $tRNA_{CCG}$. I could not confidently assign either arginine or tryptophan isotype to these tRNAs because they have the minor tryptophan identity element A1:U72, a G73 discriminator which is compatible with either arginine or tryptophan isotype, and an unusual D loop sequence that makes it unclear which nucleotide is at position 20 (important for ascertaining arginine identity) (Figure 4.3B). In the phylogenetic tree of arginine and tryptophan tRNAs in Absconditabacteria, Gracilibacteria, and outgroups (Figure 4.4), the unusual $tRNA_{CCG}$ sequences from Gracilibacteria clade 1 form a clade within the cluster of tryptophan tRNAs—this includes the canonical $tRNA^{Trp}_{CCA}$ from all groups and the reassigned $tRNA_{UCG}$ and $tRNA_{CCG}$ genes from Absconditabacteria—and not with the arginine tRNAs from all groups. I compared the likelihoods of two phylogenetic models, one where the Gracilibacteria clade 1 $tRNA_{CCG}$ sequences are constrained to form a clade with all tryptophan tRNAs (including the reassigned $tRNA_{CCG}$ and $tRNA_{UCG}$ tRNAs from Absconditabacteria) against another model where the Gracilibacteria clade 1 $tRNA_{CCG}$ sequences are constrained to group with all arginine tRNAs. The log2 ratio of these two likelihoods is 3.9, indicating support for Gracilibacteria $tRNA_{CCG}$ grouping with tryptophan tRNA sequences. All of this suggests that the clade 1 $tRNA_{CCG}$ genes are more similar in sequence to tryptophan tRNAs and not arginine tRNAs, and may have been ancestrally derived from a tryptophan tRNA.

Additionally, the $tRNA_{CCG}$ genes in Gracilibacteria clade 1 have unique features almost never seen in other tRNAs. Figure 4.3B shows the sequence and likely secondary structure of the $tRNA_{CCG}$ from GCA_002470645.1 (#11 on species tree; Patescibacteria group bacterium

UBA7396). The anticodon stem is six base pairs long, as opposed to the typical five base pair stem, which to my knowledge has only been observed in pyrrolysine tRNAs (Srinivasan *et al.*, 2002). Furthermore, this tRNA has variation to the tertiary contact between the D loop and TΨC loop which serves as the linchpin of the tRNA 3-dimensional L-shape. Typically, the invariant D loop nucleotides G18 and G19 form base pairs with TΨC loop nucleotides Ψ55 and C56 (Doyon *et al.*, 2004). In the Gracilibacteria clade 1 $tRNA_{CCG}$, the non-Watson-Crick G18-Ψ55 pair is replaced by either A18-G55 or C18-U55 (in one genome; unknown if modified to Ψ55). Interestingly, the A18-G55 pair has been previously found in a mutagenesis screen of active D loop/T loop variants in *E. coli* and is thought to form a A-G trans Watson-Crick / sugar edge interaction, which should be structurally interchangeable with the typical G18-Ψ55 pair (Doyon *et al.*, 2004). A18-G55 has also been reported to occur in several fungal mitochondrial tRNAs (Dirheimer *et al.*, 1994) and in a tmRNA in oomycete mitochondria (Hafez *et al.*, 2013). To my knowledge, this is the first time this alternative interaction has been reported in a cytosolic tRNA.

Summary of evidence for CGA and CGG translation in Gracilibacteria

All together, the evidence paints a picture where Gracilibacteria clade 2 species decode CGA and CGG as arginine, whereas Gracilibacteria clade 1 genomes might treat CGA as an arginine codon and CGG as a tryptophan codon; however, it remains to be seen how these codons are decoded in living cells. It is unclear from tRNA gene sequence alone how the CGG-decoding $tRNA_{CCG}$ in group 1 is aminoacylated. Charging with tryptophan would be consistent with the usage of CGG at a few conserved tryptophan residues in the conserved single-copy gene alignments and placement of $tRNA_{CCG}$ with tryptophan tRNAs on a tRNA phylogeny. It is also possible that this unusual tRNA may be charged with arginine or even recognized by

multiple aminoacyl-tRNA synthetases and ambiguously charged. If group 1 Gracilibacteria indeed do not have tRNA capable of reading CGA, CGA might be inefficiently decoded by the arginine $tRNA_{GCG}$ (decoding by G-A mispair is thought to occur in echinoderm and *Drosophila* mitochondria (Yokobori *et al.*, 2001; Tomita *et al.*, 1999a)).

Models of alternative genetic code evolution in Absconditabacteria and Gracilibacteria

I will outline two models of how the multiple codon reassignments may have evolved in Absconditabacteria and Gracilibacteria. These are two examples of evolutionary histories that are consistent with the data presented in this section and do not encompass all possibilities. Both models hinge on the the GTDB phylogenetic tree being correct in placing both clades of Gracilibacteria as a sister group to the Absconditabacteria and assume that Gracilibacteria clade 1 decodes CGG as tryptophan.

Both models presume that the ancestor of Absconditabacteria and Gracilibacteria had low genomic GC content which reduced the usage of GC-rich codons such as the UGA stop codon (in favor of UAA) and CGN arginine codons (in favor of AGA and AGG). This released purifying selection on the accurate translation of UGA and CGN codons, and the translation of UGA was captured by duplication and anticodon mutation of a glycine tRNA.

The first model assumes that in the intermediate stage, CGG was decoded ambiguously. A $tRNA^{Trp}_{CCG}$ emerged, most likely through duplication and anticodon mutation of a $tRNA^{Trp}_{CCA}$. This $tRNA^{Trp}_{CCG}$ may have just appeared by chance and persisted due to the low selection for arginine translation of the rare CGG codon. Alternatively, it is possible that the CGG reassignment was triggered by the reassignment of UGA to glycine—the new $tRNA^{Gly}_{UCA}$ may misread the canonical tryptophan codon UGG as glycine at a low rate due to wobble decoding, which

may have favored the appearance of an additional tryptophan codon at CGG. Regardless, the ancestor decoded the CGA arginine codon using a $tRNA^{Arg}_{UCG}$, which can also recognize the CGG codon. In the presence of the new $tRNA^{Trp}_{CCG}$, this would likely lead to some degree of ambiguous translation at CGG. Ambiguous translation at CGG was resolved in three different ways in the three descendant lineages. In the Absconditabacteria, the $tRNA^{Arg}_{UCG}$ was lost and replaced by a $tRNA^{Trp}_{UCG}$, so both CGA and CGG are translated as tryptophan such that both codons are decoded unambiguously. In the Gracilibacteria clade 1, the $tRNA^{Arg}_{UCG}$ was lost and to this day there is no cognate tRNA for CGA. In the Gracilibacteria clade 2, the $tRNA^{Trp}_{CCG}$ was lost and CGG reverted to being an arginine codon.

The second model assumes that in the intermediate stage both CGA and CGG codons were unassigned, inspired by the tRNA loss-driven reassignment model (Mühlhausen *et al.*, 2016). Once more, the low GC content has driven the CGA and CGG codons to rarity in the ancestor. The $tRNA^{Arg}_{UCG}$ that decodes both CGA and CGG is lost such that both codons are unassigned, and they are subsequently captured by three separate tRNA decoding schemes in the three descendant lineages. In Absconditabacteria, both CGA and CGG are captured by tryptophan-derived tRNAs. In Gracilibacteria clade 1, only CGG as been captured, by a tryptophan tRNA independently from Absconditabacteria. Two independent reassignment of CGG may not be so far fetched since CGG is reassigned to tryptophan in two independent Firmicute lineages (*Onthovivens* spp and *Anaerococcus*). In Gracilibacteria clade 2, both codons are captured a arginine-derived $tRNA_{UCG}$.

Both models are consistent with the tRNA phylogeny, which suggests that the $tRNA_{CCG}$ from Absconditabacteria and Gracilibacteria clade 1 are both related to $tRNA^{Trp}_{CCA}$ and that the $tRNA^{Arg}_{UCG}$ genes in Gracilibacteria clade 2 are related to other arginine tRNAs. Alternative

evolutionary trajectories would require more complex patterns of CGA and CGG reassignment and reversion, assuming the GTDB phylogenetic tree is correct. It is interesting that the Absconditabacteria and Gracilibacteria also have a reassignment of the canonical stop codon UGA to glycine, which raises questions about whether the UGA and CGA/CGG reassignments happened contemporaneously (and some Gracilibacteria later reverted to the original arginine translation for CGA and CGG), enabled by the same fluctuation in genomic GC content which has been proposed to be a force leading to the UGA reassignment (Rinke *et al.*, 2013), or whether putative misreading of the tryptophan UGG codon by the $tRNA^{Gly}_{UCA}$ (for which there is no evidence) may have provided pressure for the subsequent evolution of two new tryptophan codons.

4.3 Reassignment of CGG to tryptophan in *Anaerococcus*

In the analysis by Codetta of all available bacterial genome assemblies (Chapter 2), four genomes annotated as belonging to the genus *Anaerococcus* were predicted to translate the canonical arginine codon CGG as tryptophan. Three of these genomes are included in the GTDB phylogeny, where they belong to two species clusters within the genus g__Anaerococcus (family f__Helcococcaceae, order o__Tissierellales, class c__Clostridia) which is comprised of 23 species clusters in total.

Species phylogenetic tree

I investigated the phylogenetic extent of the predicted reassignment and other supporting features on the GTDB bacterial phylogeny. In Figure 4.5A, I show a part of the GTDB phylogenetic tree that includes all members of the GTDB family f__Helcococcaceae, including the

genus *Anaerococcus* (23 species) and outgroup species (22 species). In order to help gauge whether a tRNA gene is truly not present in a genome or whether it missing due to incomplete genome assembly, I marked the most complete genomes on the tree (Figure 4.5A), based on a CheckM completeness estimate of 98% or greater and the presence of a minimal set of 22 tRNA genes (excluding CGN-decoding tRNAs) that are required in all bacteria.

Additional evidence for the reassignment of CGG to tryptophan in *Anaerococcus*

In *Anaerococcus* (Figure 4.5A, species #1-23), two species were inferred to translate CGG as tryptophan, two species were inferred to translate CGG as arginine, and the remaining 19 species did not have an inferred meaning for CGG due to fewer than 18 aligned Pfam consensus columns at CGG codons. Reanalysis of the GTDB *Anaerococcus* clade and outgroups using Codetta 2.0 (updated version) and a more expansive profile HMM database (combination of Pfam 35.0 and NCBIFAM) led to predicted tryptophan decoding of CGG for half of the *Anaerococcus* species (data not shown). I refer to all of *Anaerococcus* as the reassigned clade based on the presence of a tryptophan-type $tRNA_{CCG}$ in all species (see below). All of the outgroup species were inferred to translate CGG as either arginine or the meaning was left uninferred in species where CGG is very rare (<9 aligned Pfam consensus columns).

I constructed multiple sequence alignments of conserved single-copy bacterial genes from the BUSCO dataset across *Anaerococcus* species and the outgroup species to find examples of CGG occurring at conserved positions for some amino acid. Unfortunately, CGG is so rare in *Anaerococcus* that only a single CGG position could be found in a well-aligned region of a BUSCO gene alignment. In *Anaerococcus nagyae* (#14 on tree, GCA_003433955.1) a single CGG occurs at position conserved for tryptophan in other *Anaerococcus* and outgroup species in an alignment of DNA ligase (Figure 4.5B).

Figure 4.5: (A) Phylogenetic tree from GTDB showing *Anaerococcus* and closest outgroup genomes, each species indicated by a number which can be cross-referenced with genome information (Supplementary Files). Double asterisks indicate genomes the most complete genomes, which have CheckM estimated genome completeness from GTDB >98% and the entire minimal set of 22 required tRNAs (excluding CGN-decoding tRNAs). For each species, the inferred translation of the reassigned codon CGG by Codetta is indicated by colored circles (red: arginine, purple: tryptophan, white: uninferred). The presence of tRNA genes that recognize the CCA and CGN-codons is also indicated by filled circles, colored according to the predicted amino acid charging based on identity elements for tRNAs (see Methods). Gray box outline highlights the $tRNA_{CCG}$ in *Finegoldia*. Anticodons in gray font are typically not found in Firmicutes. The lines connecting codons and tRNA anticodons represent the likely wobble decoding capabilities, with dashed lines representing weaker interactions. The anticodon ACG is presumed to be modified to ICG. The U34 of anticodon UCG is presumed to be modified in a way that restricts decoding to CGA and CGG, but could potentially recognize CGU and/or CGC depending on the true modification state. The remaining anticodons are not expected to be modified in a way that alters which codons are recognized. The codon usage for the reassigned codon CGG, the arginine codons AGA, AGG, CGU, CGC, and CGA, and the tryptophan codon CCA is the frequency per 10,000 codons aligned to Pfam domains. Genomic GC content is calculated over the entire genome. (B) Multiple sequence alignment of DNA ligase (BUSCO POG091H024G) from the *Anaerococcus* species and selected outgroup species. An alignment region containing CGG (**ɑ**) at conserved positions is shown, with a column with CGG in a single *Anaerococcus* species highlighted. (C) The CGG-decoding $tRNA_{CCG}$ from species #7 (*Anaerococcus* sp. Marseille-P3915, GCA_900258475.1). tRNA sequence features involved in tryptophan identity are highlighted (Giegé *et al.*, 1998; Himeno *et al.*, 1991), with nucleotide numbering following the convention of Sprinzl *et al.* (1998). Nucleotides highlighted in gray do not match the expected identity element.

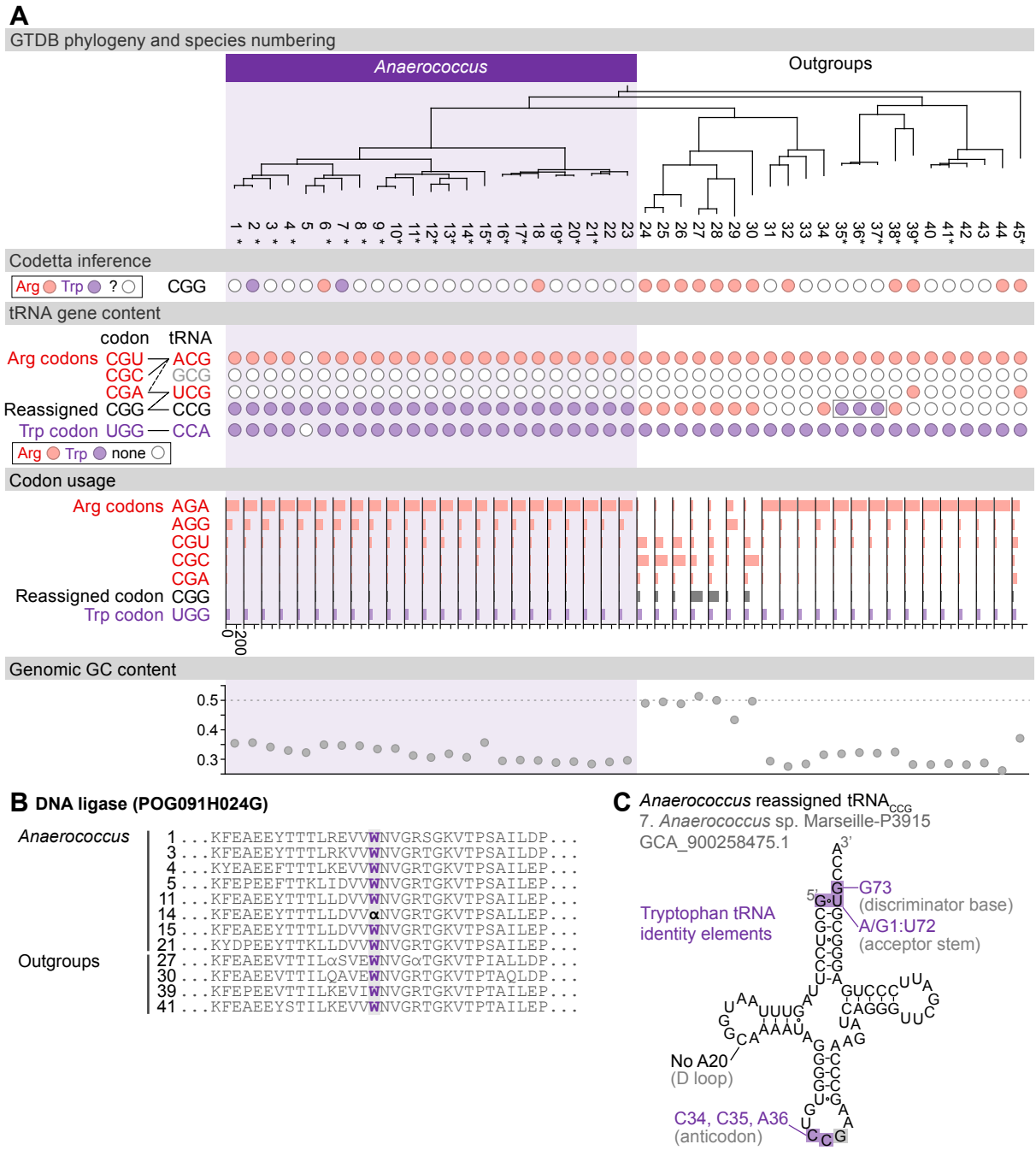

Figure 4.5 (Continued)

In *Anaerococcus*, all 23 species contain a CGG-decoding $tRNA_{CCG}$ which does not appear to be an arginine tRNA due to lack of A20 in the D loop, and contains elements supportive of tryptophan identity such as a G73 discriminator base and G1:U72 in the acceptor stem. Figure 4.5C shows the sequence and likely secondary structure of the $tRNA_{CCG}$ from *Anaerococcus* sp. Marseille-P3915 (#7 on tree, GCA_900258475.1), highlighting sequence features used for classification as tryptophan-type. None of the *Anaerococcus* genomes have a $tRNA_{UCG}$ and instead presumably decode the arginine codons CGU, CGC, and CGA with a $tRNA^{Arg}_{ACG}$ (A presumably modified to I).

To further explore the origin of the $tRNA_{CCG}$ in *Anaerococcus*, I built a tRNA phylogenetic tree from sequences of arginine and tryptophan tRNAs from *Anaerococcus* and outgroup species (Figure 4.6). In this phylogenetic tree, the *Anaerococcus* $tRNA_{CCG}$ sequences form a clade (along with a $tRNA_{CCG}$ gene from outgroup species #34) that is not placed within the cluster of arginine tRNAs, but instead falls within a three-way multifurcation at the base of the cluster of tryptophan tRNAs. I compared the likelihood of two phylogenetic models, one where the *Anaerococcus* $tRNA_{CCG}$ sequences are constrained to cluster with the tryptophan $tRNA_{CCA}$ sequences against another model where they are constrained to cluster with the arginine tRNAs. The log2 ratio of the likelihoods is 2.3 in favor of the tryptophan model. This indicates that the *Anaerococcus* $tRNA_{CCG}$ genes are more similar in sequence to $tRNA^{Trp}_{CCA}$ than to arginine tRNAs.

CGG is a very rare codon in all *Anaerococcus* species, with usage no greater than 1.5 per 10,000 codons in aligned Pfam domains. The very low genomic GC content in the clade, ranging between 0.28-0.36, may have contributed to the low codon usage of CGG and the GC-rich arginine codons CGC, CGA, and AGG in favor of AGA (which is used at 196-278

per 10,000 codons).This would have facilitated the reassignment of CGG by minimizing the number of substitutions needed to adapt to the new translation. This codon usage pattern is repeated in other low GC content genomes in the outgroup, while outgroup genomes with higher GC content (>0.40) favor other arginine codons such as AGG, CGU, CGG, or CGG.

In summary, two *Anaerococcus* species were predicted to translate CGG as tryptophan by Codetta, while most species did not have an inferred amino acid for CGG and two species were predicted to translate CGG as arginine. The two species predicted to decode CGG as arginine may be real or may stem from some degree of assembly contamination with standard genetic code-using sequences. Overall, CGG is very rare in this genus (possibly due to the low genomic GC content), and many species had CGG uninferred because there were too few aligned Pfam consensus columns for Codetta to make a confident amino acid prediction. Despite the sparse distribution of CGG prediction as tryptophan by Codetta, all *Anaerococcus* species contain a $tRNA_{CCG}$ gene that is consistent with a tryptophan isotype.

Possible CGG reassignment in the outgroup genus *Finegoldia*

The outgroup consists of all remaining 22 species in the family f__Helcococcaceae. All outgroup species were predicted to translate CGG as either arginine or were uninferred (Figure 4.5A). All outgroup species presumably use a $tRNA^{Arg}_{ACG}$ (A modified to I) to decode the arginine codons CGU, CGC, and CGA. In some species, I could not find a CGG-decoding tRNA, which may be due to incomplete genome assembly, tRNA gene detection failure, or a true lack of CGG-decoding capability.

Outgroup species in the genus *Finegoldia* (species #35-37 on tree) have a CGG-decoding $tRNA_{CCG}$ that lacks the arginine identity element A20 and contains elements consistent with tryptophan decoding including a G73 discriminator base and A1:U72 in the acceptor stem

Figure 4.6: An unrooted phylogenetic tree of arginine, tryptophan, and reassigned tRNA sequences from *Anaerococcus* and outgroup genomes, generated by a Bayesian approach implemented in MrBayes 3.2.7a. Values below each internal branch indicate the posterior probability of each clade. Since this is an unrooted tree, this value represents the probability of the sequences on either side of the branch form two distinct clades. Branch length scale in the bottom right is in expected substitutions per site. Sequence labels follow the format "species number - clade _ tRNA anticodon". This tree includes sequences from $tRNA_{CCG}$ from *Anaerococcus* and *Finegoldia* (potentially involved in the CGG reassignment and indicated with a gray circle), the arginine $tRNA_{UCU}$, $tRNA_{CCU}$, $tRNA_{ACG}$, and $tRNA_{UCG}$ (and $tRNA_{CCG}$ from outgroup species) indicated by a red circle, and the tryptophan $tRNA_{CCA}$ from all groups indicated by a purple circle.

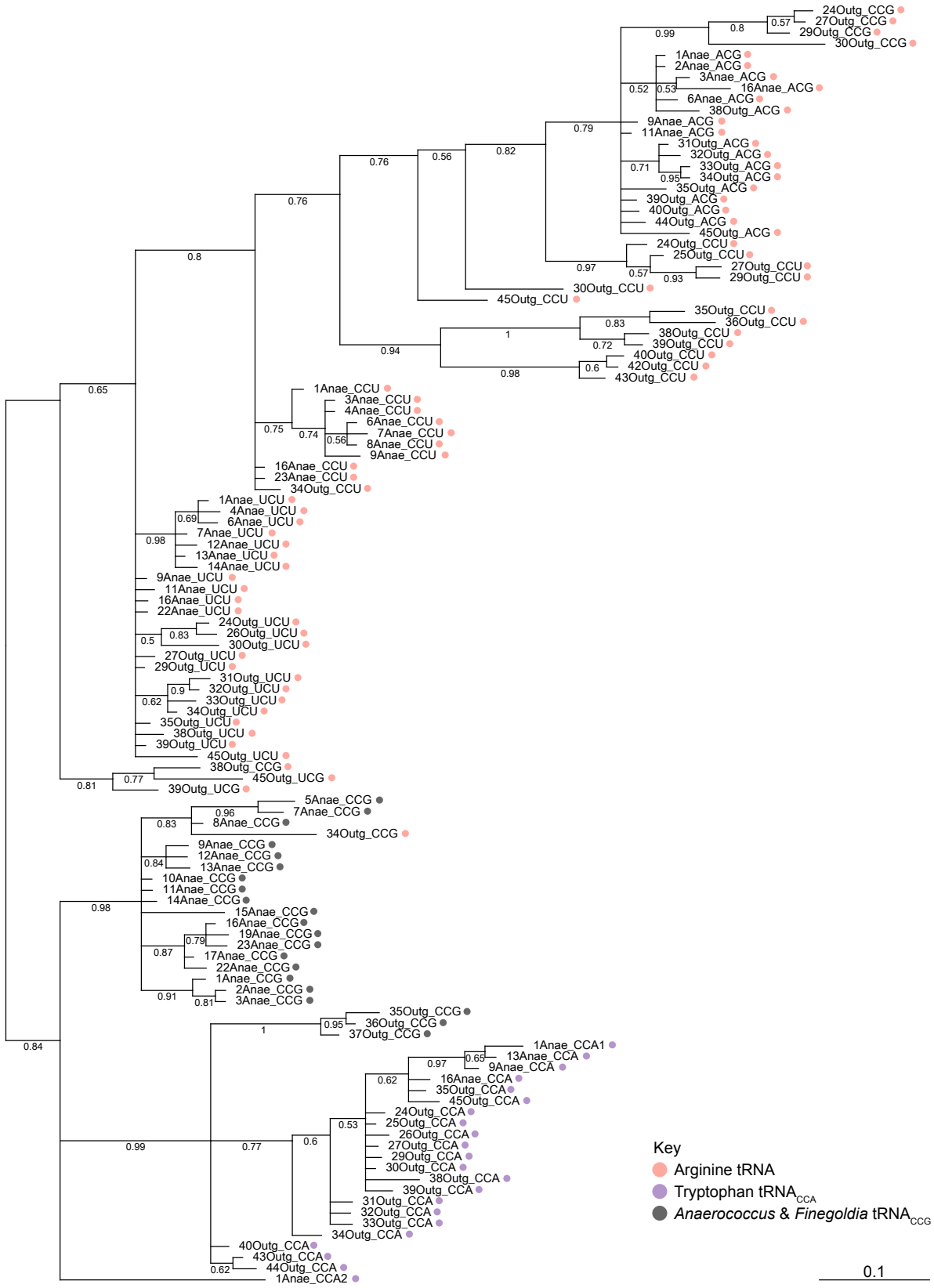

Figure 4.6 (Continued)

(Figure 4.5A, gray box outline). On the tRNA phylogenetic tree, *Finegoldia* $tRNA_{CCG}$ sequences form a clade within other tryptophan tRNAs (Figure 4.6). I compared the likelihood of two phylogenetic models, one where the *Finegoldia* $tRNA_{CCG}$ sequences are constrained to cluster with the tryptophan tRNAs and *Anaerococcus* $tRNA_{CCG}$ against a model where the *Finegoldia* $tRNA_{CCG}$ sequences are constrained to cluster within arginine tRNAs. The log2 ratio of the likelihoods was 24.3, indicating very strong support for the tryptophan model. This suggests that the *Finegoldia* $tRNA_{CCG}$ is more similar in sequence to and may have been derived from a tryptophan tRNA. CGG is very rare in this genus (fewer than 0.5 per 10,000 codons in Pfam alignments) and there were insufficient aligned Pfam consensus columns for Codetta to predict an amino acid inference, and no CGGs could be found in well-aligned regions of BUSCO alignments.

In an *ex post facto* reanalysis of all *Anaerococcus* and outgroup genomes with an updated version of Codetta and a more comprehensive profile HMM database, three out of 37 *Finegoldia* genomes predicted to translate CGG as tryptophan based on an exceedingly small number (3-7) of aligned consensus columns (data not shown). Thus, both the tentative Codetta prediction and tRNA gene analysis support a potential CGG codon reassignment to tryptophan in *Finegoldia*. This might represent a very early stage in a reassignment of CGG to tryptophan. It is unclear whether the *Finegoldia* and *Anaerococcus* $tRNA_{CCG}$ genes and putative reassignments share an evolutionary history or are independently derived.

Possible model for the evolution of a CGG reassignment in *Anaerococcus*

The results described above can be consolidated to propose a codon capture-like model of how the reassignment of CGG from arginine to tryptophan may have evolved in *Anaerococcus*. The ancestor of *Anaerococcus* likely had very low genomic GC content as do all extant species

in the genus. This likely pushed the usage of arginine codons to shift away from the GC-rich codons like CGC and CGG and towards AGA, leading to very low usage of CGG. This released purifying selection on the accurate and efficient translation of CGG, and at some point the CGG-decoding tRNA (either $tRNA^{Arg}_{UCG}$ or $tRNA^{Arg}_{CCG}$) was lost and a tryptophan-type $tRNA_{CCG}$ was gained, likely by duplication and divergence of a $tRNA^{Trp}_{CCA}$. The existence of an inosine-wobble $tRNA^{Arg}_{ICG}$ allowed the arginine translation of CGA to be unaffected by reassignment of CGG. Over time, CGG codons began to synonymously substitute with the UGG tryptophan codon leading to the appearance of CGG at conserved tryptophan residues.

4.4 Reassignment of CGG to tryptophan in *Onthovivens* spp

In the analysis by Codetta of all available bacterial genome assemblies (Chapter 2), 5 genomes broadly annotated as “uncultured Mollicutes" and “Bacillales sp" were predicted to translate the canonical arginine codon CGG as tryptophan. All of these genomes are included in the GTDB phylogeny, where they belong to 4 species clusters that form a monophyletic clade within the GTDB genus g__Onthovivens (family f__CAG-826, order o__RFN20, class c__Bacilli). The genus *Enterosoma*, which contains the AGG to methionine reassignment, also belongs to the order o__RFN20, but to a different family.

Species phylogenetic tree

I investigated the phylogenetic extent of the predicted reassignment and other supporting features on the GTDB bacterial phylogeny. In Figure 4.7, I show a part of the GTDB phylogenetic tree that includes all members of the GTDB family f__CAG-826 which includes the reassigned clade (5 species) and outgroup species (45 species). I designated the reassigned

Figure 4.7: Phylogenetic tree from GTDB showing the *Onthovivens* spp CGG→Trp clade and closest outgroup genomes, each species indicated by a number which can be cross-referenced with genome information (Supplementary Files). Double asterisks indicate genomes the most complete genomes, which have CheckM estimated genome completeness from GTDB >95% and the entire minimal set of 22 required tRNAs (excluding CGN-decoding tRNAs). For each species, the inferred translation of the reassigned codon CGG by Codetta is indicated by colored circles (red: arginine, purple: tryptophan, white: uninferred). The presence of tRNA genes that recognize the CCA and CGN-codons is also indicated by filled circles, colored according to the predicted amino acid charging based on identity elements for tRNAs (see Methods). Anticodons in gray font are typically not found in Firmicutes. The lines connecting codons and tRNA anticodons represent the likely wobble decoding capabilities, with dashed lines representing weaker interactions. The anticodon ACG is presumed to be modified to ICG. The U34 of anticodon UCG is presumed to be modified in a way that restricts decoding to CGA and CGG, but could potentially recognize CGU and/or CGC depending on the true modification state. The remaining anticodons are not expected to be modified in a way that alters which codons are recognized. The codon usage for the reassigned codon CGG, the arginine codons AGA, AGG, CGU, CGC, and CGA, and the tryptophan codon CCA is the frequency per 10,000 codons aligned to Pfam domains. Genomic GC content is calculated over the entire genome.

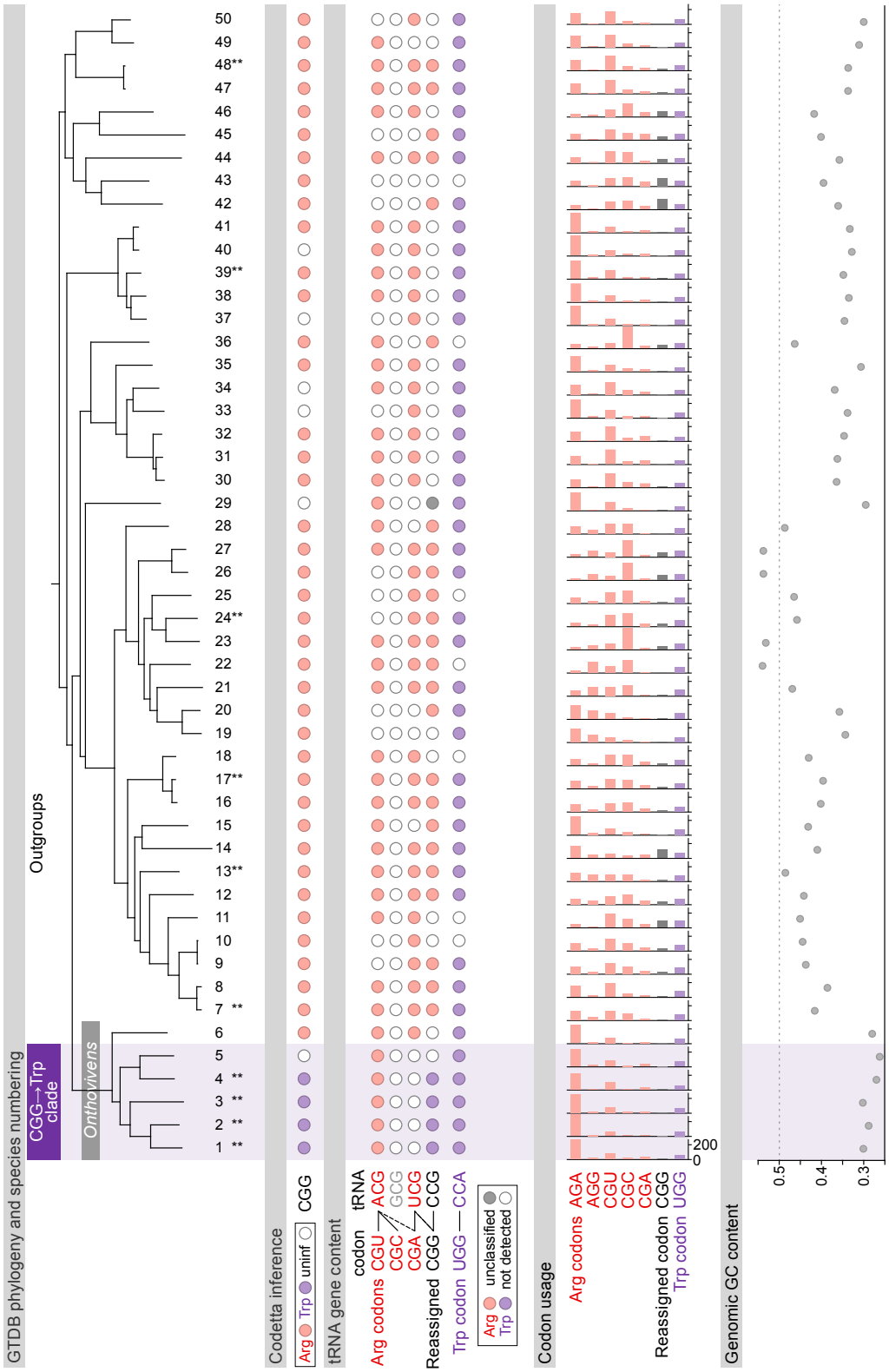

Figure 4.7 (Continued)

clade as the branch of the tree containing all 4 species inferred to translate CGG as Trp plus one additional species that has CGG meaning uninferred (species s__UBA4855 sp002438785) due to the tree topology. The reassigned clade is a subgroup of the genus *Onthovivens*, so I refer to the reassignment as affecting "*Onthovivens* spp" to indicate that it does not affect the whole genus.

In order to help assess whether a tRNA gene is used by members of a clade or whether it missing due to incomplete genome assembly, I marked the most complete genomes on the tree (Figure 4.7), which had a CheckM completeness estimate of 95% or greater and contained a minimal set of 22 tRNA genes (excluding CGN-decoding tRNAs) that are required in all bacteria.

Additional evidence for the reassignment of CGG to tryptophan in *Onthovivens* spp

In the reassigned clade (Figure 4.7, species #1-5), four species were inferred to translate CGG as tryptophan; one species did not have an inferred meaning for CGG (species s__UBA4855 sp002438785) due to fewer than 2 aligned Pfam consensus columns at CGG codons. All of the outgroup species were inferred to translate CGG as either arginine or the meaning was left uninferred in a few cases where CGG is very rare (<10 aligned Pfam consensus columns).

Figure 4.8 shows five example alignments of conserved single-copy bacterial genes across the reassigned clade and outgroup species. The five aligned genes show individual CGG positions in reassigned species #1,3,4 at residues that are primarily conserved for tryptophan in other members of the reassigned clade and outgroups. This suggests that CGG can be used interchangeably with the tryptophan codon UGG at conserved tryptophan residues in these species. These alignments do not help determine how CGG is translated in the uninferred

species (s__UBA4855 sp002438785, #5 on tree), as I could not find instances of CGG in aligned BUSCO genes. For the nearest outgroup species (s__UBA4855 sp002451465, #6 on tree), in the CTP synthase alignment, there is a single CGG position at a conserved arginine residue, supporting canonical arginine translation of CGG in this species.

In the reassigned clade, the four species that were inferred to translate CGG as tryptophan contain CGG-decoding $tRNA_{CCG}$ gene (Figure 4.7). This tRNA does not appear to be an arginine tRNA due to lack of A20 in the D loop and contains elements supportive of tryptophan identity such as a G73 discriminator base and A/G1:U72 in the acceptor stem. Figure 4.9 shows the sequence and likely secondary structure of the $tRNA_{CCG}$ from GCA_900540395.1 (uncultured Mollicutes bacterium, #1 on tree), highlighting sequence features used for classification as tryptophan-type. The uninferred member of the reassigned clade (s__UBA4855 sp002438785, #5 on tree) does not contain a CGG-decoding tRNA (either $tRNA_{CCG}$ or $tRNA_{UCG}$) in any of the 3 genomes assigned to the species, indicating either incomplete genome assembly, tRNA gene detection failure, or a bona fide inability to translate CGG with a cognate tRNA.

In tRNA phylogenetic tree built from arginine and tryptophan tRNAs from the reassigned clade and outgroup species, the $tRNA_{CCG}$ sequences from the reassigned clade do not cluster with the outgroup $tRNA_{CCG}$ genes among the other arginine tRNAs, but instead branch right outside of the tryptophan $tRNA_{CCA}$ sequences (Figure 4.10). I compared the likelihood of two phylogenetic models, one where the reassigned $tRNA_{CCG}$ sequences are constrained to cluster with the tryptophan tRNAs against another model where they are constrained to cluster with the arginine tRNAs. The log2 ratio of the likelihoods is 4.5, in support of the tryptophan model and consistent with the presence of tryptophan identity elements.

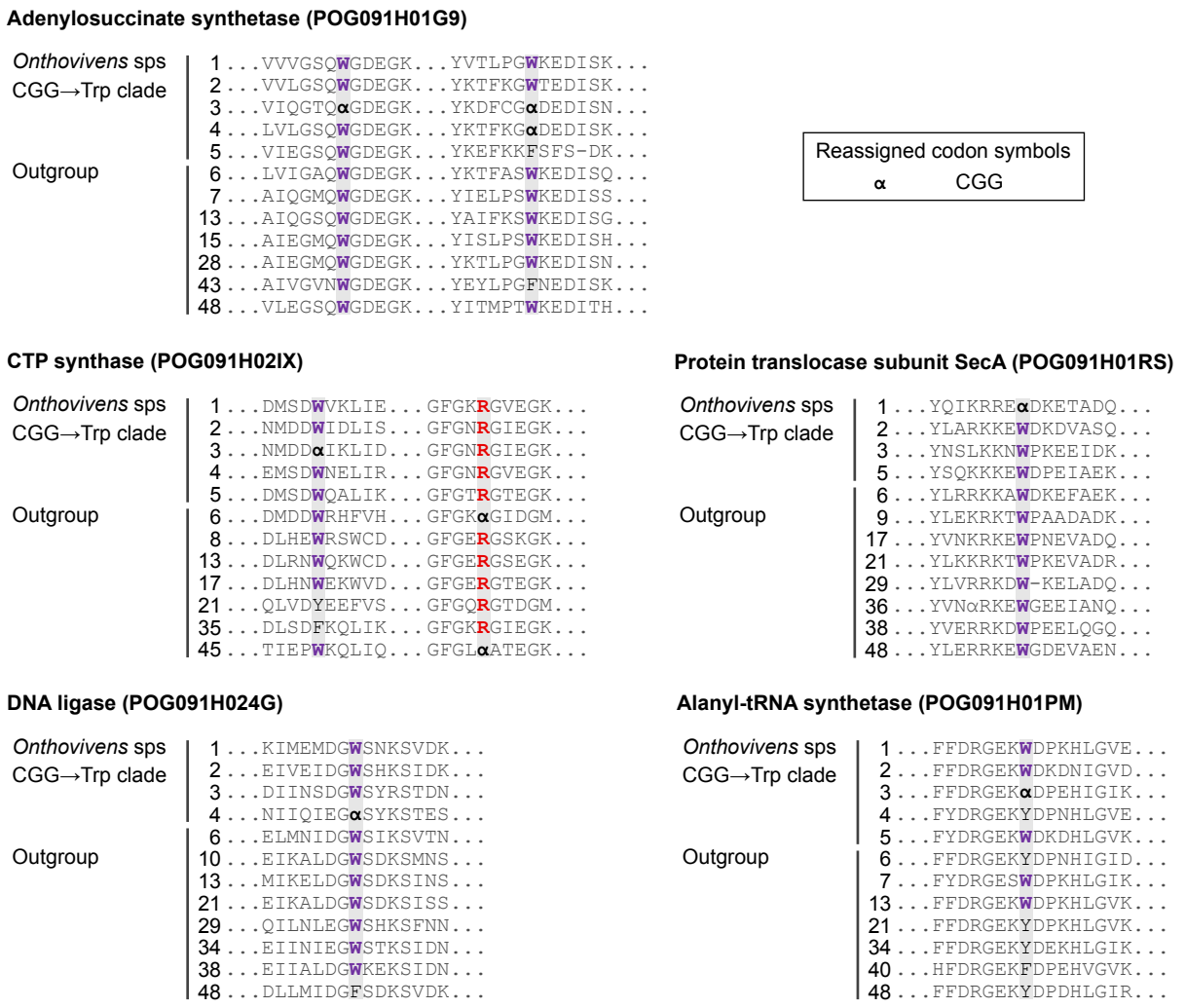

Figure 4.8: Multiple sequence alignments of adenylosuccinate synthetase (BUSCO POG091H01G9), CTP synthase (BUSCO POG091H02IX), protein translocase subunit SecA (BUSCO POG091H01RS), DNA ligase (BUSCO POG091H024G), and alanyl-tRNA synthetase (BUSCO POG091H01PM) from the *Onthovivens* spp CGG→Trp clade and selected outgroup species. Alignment regions containing CGG (**α**) at conserved positions are shown, with columns with CGG in *Onthovivens* spp CGG→Trp clade and the closest outgroup (species #6) sequences highlighted.

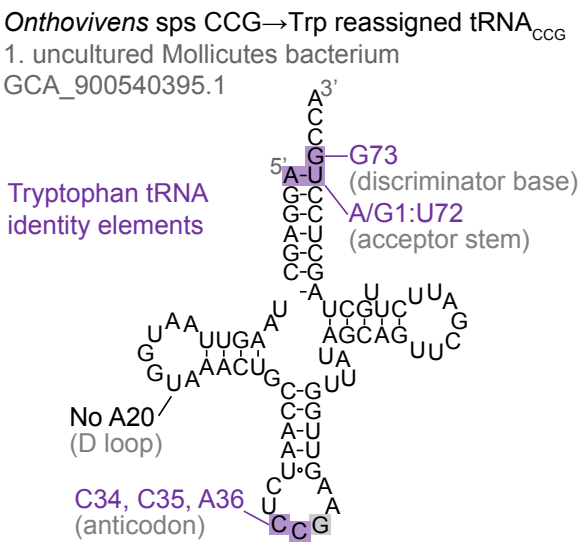

Figure 4.9: The CGG-decoding tRNA$_{CCG}$ from species #1 (uncultured Mollicutes bacterium, GCA_900540395.1). tRNA sequence features involved in tryptophan identity are highlighted (Giegé *et al.*, 1998; Himeno *et al.*, 1991), with nucleotide numbering following the convention of Sprinzl *et al.* (1998). Nucleotides highlighted in gray do not match the expected identity element.

None of the reassigned species have a tRNA$_{UCG}$ and instead presumably decode the arginine codons CGU, CGC, and CGA with a tRNA$^{Arg}_{ACG}$ (A presumably modified to I) (Figure 4.7). This is in contrast to many of the outgroup species (focusing on the more complete genomes as representatives), which appear to rely on a tRNA$^{Arg}_{ACG}$ gene to decode CGU, CGC, and CGA and a tRNA$^{Arg}_{UCG}$ to decode CGA and CGG, with an optional CGG-decoding tRNA$^{Arg}_{CCG}$ present in some species.

CGG is a very rare codon in all members of the GTDB genus g__UBA4855 (reassigned species #1-5 and outgroup species #6 on tree), with usage no greater than 2.5 per 10,000 codons in aligned Pfam domains (Figure 4.7). The very low genomic GC content in the clade, ranging between 0.26-0.30, may have contributed to the low codon usage of CGG and the GC-rich arginine codons CGC, CGA, and AGG (for each, usage <31 per 10,000) in favor of AGA (which is used at 220-292 per 10,000 codons). This would have facilitated reassignment by minimizing the number of substitutions needed to adapt to the new translation. This codon

Figure 4.10: An unrooted phylogenetic tree of arginine, tryptophan, and reassigned tRNA sequences from *Onthovivens* spp CGG→Trp clade and outgroup genomes, generated by a Bayesian approach implemented in MrBayes 3.2.7a. Values below each internal branch indicate the posterior probability of each clade. Since this is an unrooted tree, this value represents the probability of the sequences on either side of the branch form two distinct clades. Branch length scale in the bottom right is in expected substitutions per site. Sequence labels follow the format "species number - clade _ tRNA anticodon". This tree includes sequences from $tRNA_{CCG}$ from *Onthovivens* spp CGG→Trp clade (involved in the CGG reassignment and indicated with a gray circle), the arginine $tRNA_{UCU}$, $tRNA_{CCU}$, $tRNA_{ACG}$, and $tRNA_{UCG}$ (and $tRNA_{CCG}$ from outgroup species) indicated by a red circle, and the tryptophan $tRNA_{CCA}$ from all groups indicated by a purple circle.

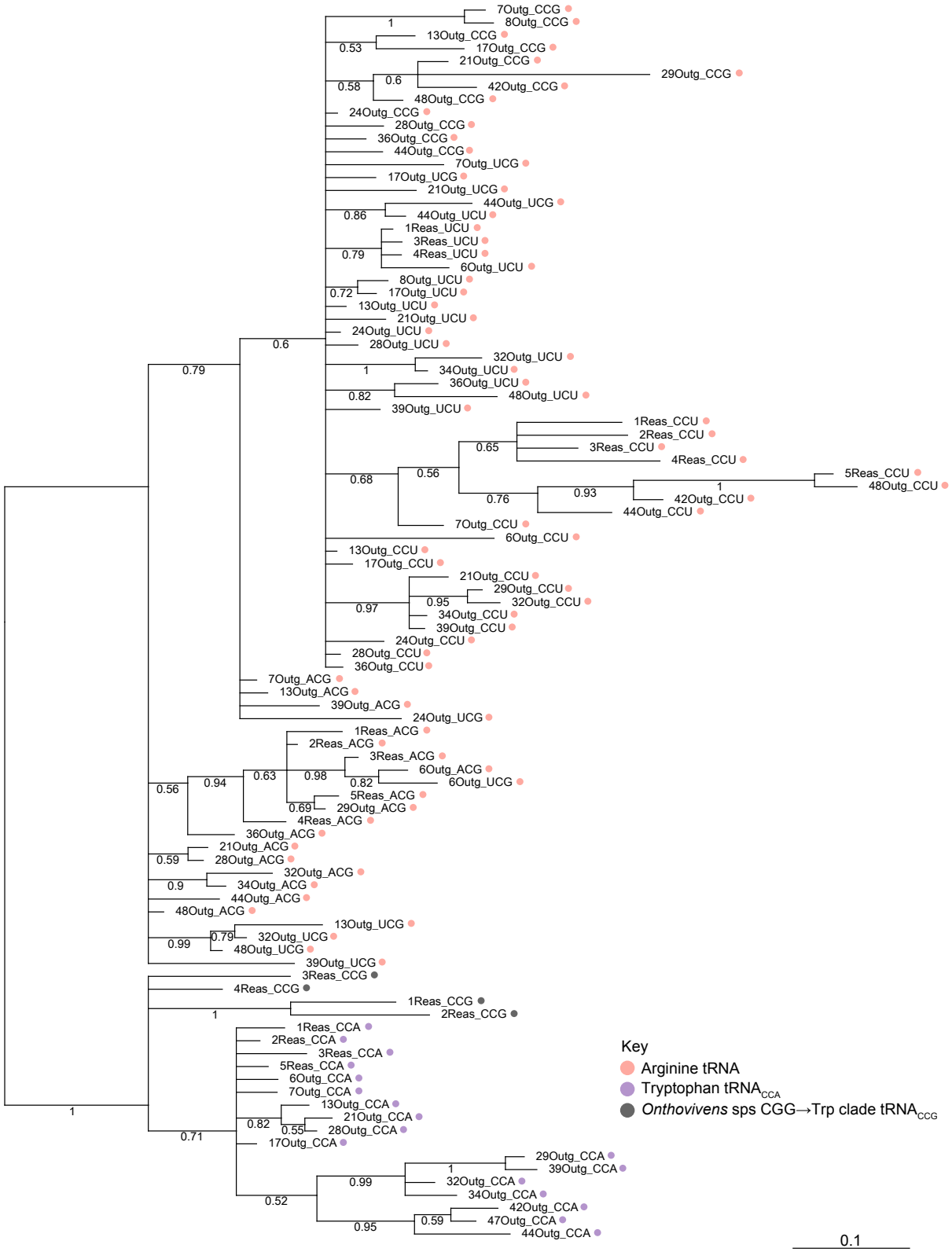

Figure 4.10 (Continued)

usage pattern is repeated in other low GC content genomes in the outgroup, while outgroup genomes with higher GC content (>0.40) favor CGU and CGC.

In summary, the appearance of CGG at conserved tryptophan positions in proteins has led to CGG being inferred as a tryptophan codon in reassigned species by Codetta, and is supported usage of CGG in conserved tryptophan positions in multiple sequence alignments of conserved single-copy genes. CGG is decoded in this clade by a $tRNA_{CCG}$ that is consistent with a tryptophan isotype. s__UBA4855 sp002438785 (species #5 on tree) did not have an inferred CGG meaning by Codetta and does not have a $tRNA_{CCG}$, but is included in the reassigned clade because it is placed in the same clade as the reassigned species according to the GTDB tree. The entire reassigned clade plus the closest outgroup species s__UBA4855 sp002451465 (species #6 on tree) form the genus *Onthovivens* on GTDB, which features low genomic GC content and extremely low codon usage of CGG.

Possible model for the evolution of a CGG reassignment in *Onthovivens* spp

The results described above can be collated to propose a possible model of how the reassignment of CGG from arginine to tryptophan evolved in members of the genus *Onthovivens*. The *Onthovivens* ancestor likely had very low genomic GC content as do all extant species. This may have caused the arginine codon bias to shift away from the GC-rich codons such as CGC and CGG and towards AGA, causing CGG to become a very rare codon. This released purifying selection on the accurate and efficient translation of CGG, and at some point, the CGA- and CGG-decoding $tRNA^{Arg}_{UCG}$ was lost and a tryptophan-type $tRNA_{CCG}$ was gained, most likely via duplication and divergence of a $tRNA^{Trp}_{CCA}$. The existence of an inosine-wobble $tRNA^{Arg}_{ICG}$ allowed the arginine translation of CGA to be unaffected by reassignment of CGG. Over time, CGG codons began to synonymously substitute with the UGG tryptophan codon

leading to the appearance of CGG at conserved tryptophan residues. The lack of a $tRNA_{UCG}$ or a $tRNA_{CCG}$ in s__UBA4855 sp002438785 (#5 on tree) might indicate that the new $tRNA_{CCG}$ was lost, but I cannot definitively assert that these tRNAs are absent as the genome assemblies are not 100% complete.

4.5 Reassignment of CGG to glutamine in *Peptacetobacter*

In the analysis by Codetta of all sequenced bacterial genomes (Chapter 2), I noticed that six genomes were predicted to translate the canonical arginine codon CGG as glutamine. These six genomes included uncultivated Clostridia assembled from metagenomic sequences and culturable species such as *Peptacetobacter hiranonis*. Four of these genomes are included in the GTDB phylogeny, where they are assigned to three species clusters within the GTDB genus g__Peptacetobacter (family f__Peptostreptococcaceae, order o__Peptostreptococcales, class c__Clostridia), which additionally includes *Peptacetobacter hominis* (Codetta did not infer an amino acid translation for CGG in this species).

Species phylogenetic tree

I investigated the phylogenetic extent of the predicted reassignment and other supporting features on the GTDB bacterial phylogeny. In Figure 4.11, I show a part of the GTDB phylogenetic tree that includes all members of the GTDB genus *Peptacetobacter* (4 species clusters) and outgroup species including the remaining species in the family Peptostreptococcaceae and the families Peptoclostridiaceae and Filifactoraceae (52 species clusters).

To help gauge whether a tRNA gene is truly absent from a genome or whether it missing due to incomplete genome assembly, I marked the most complete genome assemblies on

Figure 4.11: Phylogenetic tree from GTDB showing *Peptacetobacter* and closest outgroup genomes, each species indicated by a number which can be cross-referenced with genome information (Supplementary Files). Double asterisks indicate genomes the most complete genomes, which have CheckM estimated genome completeness from GTDB >98% and the entire minimal set of 22 required tRNAs (excluding CGN-decoding tRNAs), while single asterisks indicate genomes which have CheckM estimated genome completeness from GTDB >98% and >18 out of 22 required tRNAs. For each species, the inferred translation of the reassigned codon CGG by Codetta is indicated by colored circles (red: arginine, light blue: glutamine, white: uninferred). The presence of tRNA genes that recognize the CAR- and CGN-codons is also indicated by filled circles, colored according to the predicted amino acid charging based on based on identity elements for tRNAs (see Methods). Gray box outline highlights the inability to detect any CGG-decoding tRNA in the Peptostreptococcaceae (species #5-34). Anticodons in gray font are typically not found in Firmicutes. The lines connecting codons and tRNA anticodons represent the likely wobble decoding capabilities, with dashed lines representing weaker interactions. The anticodon ACG is presumed to be modified to ICG, and UUG is presumed to be modified in a way that restricts wobble to CGA and CGG. The U34 of anticodon UCG is presumed to be modified in a way that restricts decoding to CGA and CGG, but could potentially recognize CGU and/or CGC depending on the true modification state. The remaining anticodons are not expected to be modified in a way that alters which codons are recognized. The codon usage for the reassigned codon CGG, the arginine codons AGA, AGG, CGU, CGC, and CGA, and the glutamine codons CAA and CAG is the frequency per 10,000 codons aligned to Pfam domains. Genomic GC content is calculated over the entire genome.

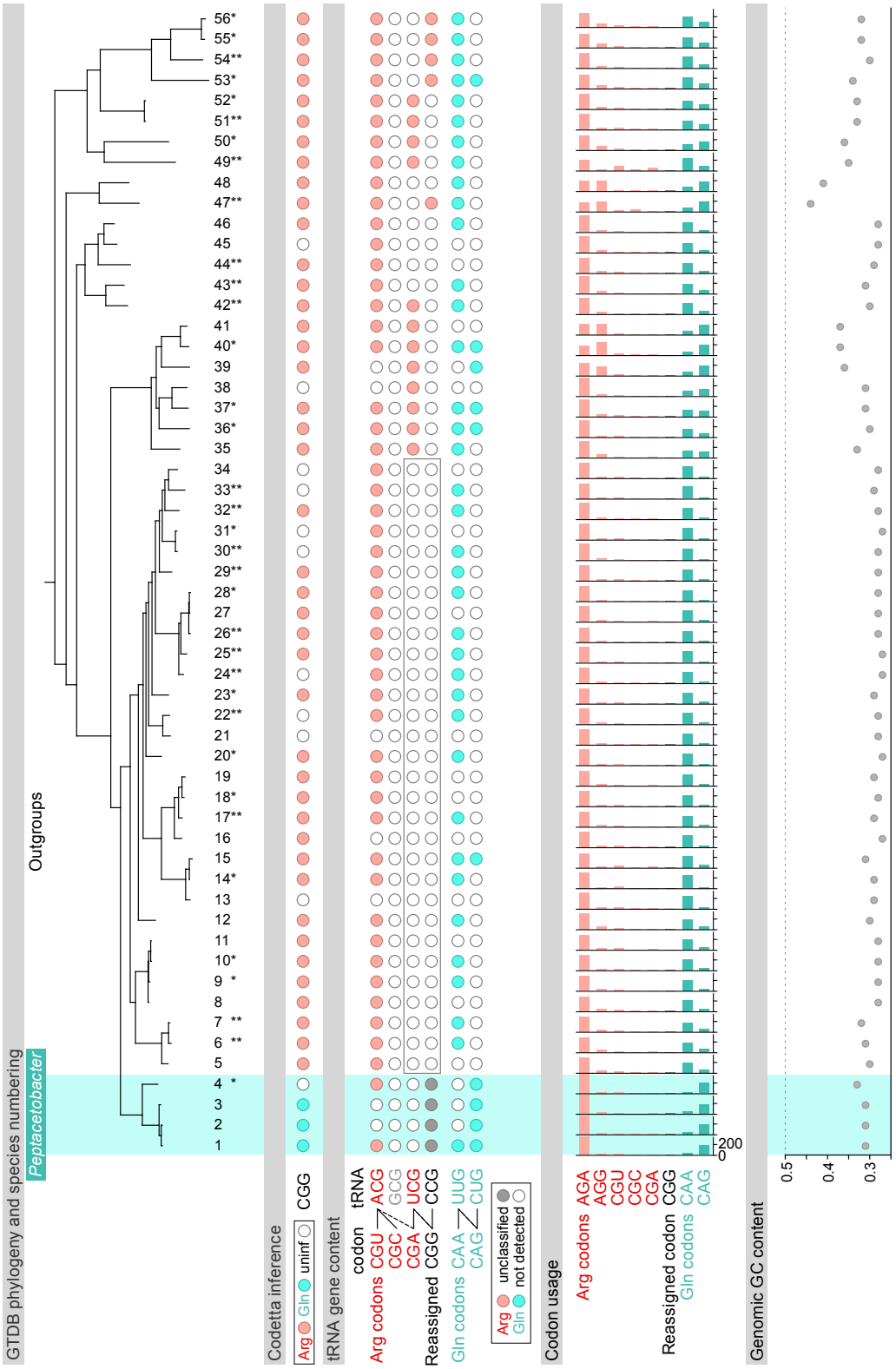

Figure 4.11 (Continued)

the tree (Figure 4.11), which had a CheckM completeness estimate of 98% or greater and contained the entire minimal set of 22 tRNA genes (excluding CGN-decoding tRNAs) that are required in all bacteria (see Methods). I noticed that some members of the reassigned clade and outgroup had very high CheckM completeness scores (>98%) but were missing more than half of the required tRNA set. I attribute this due to the fact that tRNA genes are known to occur in large gene clusters in Firmicutes (Tran *et al.*, 2015) so it is possible for an assembly to be missing a tRNA cluster (and therefore a large fraction of tRNAs) without affecting the CheckM completeness significantly (which is based on the presence of non-clustered protein coding genes). As an example of the highly clustered nature of tRNA genes in this clade, the genome of *Peptostreptococcus stomatis* DSM 17678 (GCA_000147675.2, #40 on tree) was found to have 86 tRNA genes by tRNAscan-SE 2.0, 71 of which are located in one of two clusters (one 4kb cluster of 42 tRNAs and one 8kb cluster of 29 tRNAs).

Additional evidence for the reassignment of CGG to glutamine in *Peptacetobacter*

In the *Peptacetobacter* clade on the GTDB tree (Figure 4.11), three species were inferred to translate CGG as glutamine and one species did not have an inferred meaning for CGG (*Peptacetobacter hominis*, GCA_006861675.1). One non-representative genome for *Peptacetobacter hiranonis* (species #1 on tree) was inferred to translate CGG as arginine; however, I suspect the signal may be coming from contaminating contigs (CheckM contamination estimate is 5%). All of the outgroup species were inferred to translate CGG as either arginine or the meaning was left uninferred in a few cases where CGG is very rare (<14 aligned Pfam consensus columns).

Figure 4.12 shows four example alignments of conserved single-copy bacterial genes across *Peptacetobacter* and outgroup species. The four aligned genes contain several positions encoded by CGG in *Peptacetobacter* genomes and these residues are primarily conserved for glutamine

in other species. The occurrence of CGG at conserved glutamine positions supports CGG being used as a glutamine codon all *Peptacetobacter* species, even *Peptacetobacter hominis* (species #4 on tree) which had CGG uninferred by Codetta.

In all *Peptacetobacter* genomes, CGG is decoded by a $tRNA_{CCG}$. I did not classify this tRNA as an arginine tRNA due to lack of A20 in the D loop. This tRNA also lacks many glutamine identity elements such as a weak 1:72 base pair in the acceptor stem. Therefore, I could not determine which aminoacyl-tRNA synthetase recognizes the $tRNA_{CCG}$ from the tRNA sequence alone. Figure 4.13 shows the sequence and likely secondary structure of the $tRNA_{CCG}$ from *Peptacetobacter hiranonis* DSM 13275 (#1 on tree, GCA_000156055.1), highlighting sequence features examined for classification as glutamine-type. It is possible that the glutaminyl-tRNA synthetase (GlnRS) still recognizes the $tRNA_{CCG}$ despite missing many of the known identity elements because the relaxed specificity of GlnRS towards non-cognate tRNAs (Sherman *et al.*, 1994). For example, in *E. coli*, many amber (UAG) suppressor tRNAs derived from non-glutamine tRNAs insert glutamine at UAG, even if they lack glutamine identity elements (Normanly *et al.*, 1990). The amino acid charging of $tRNA_{CCG}$ could be experimentally confirmed in the future for culturable species such as *Peptacetobacter hiranonis*.

In a tRNA phylogeny built from arginine and glutamine tRNA sequences from *Peptacetobacter* species and outgroups, the *Peptacetobacter* $tRNA_{CCG}$ branches outside of either the cluster of arginine or glutamine tRNAs (Figure 4.14). I compared the likelihood of two phylogenetic models, one where the *Peptacetobacter* $tRNA_{CCG}$ sequences are constrained to cluster with the glutamine tRNAs against another model where they are constrained to cluster with the arginine tRNAs. The log2 ratio of the likelihoods is 1.5, indicating more-or-less even support for both models. This would be consistent with a non-$tRNA^{Gln}$ origin for the *Peptace-*

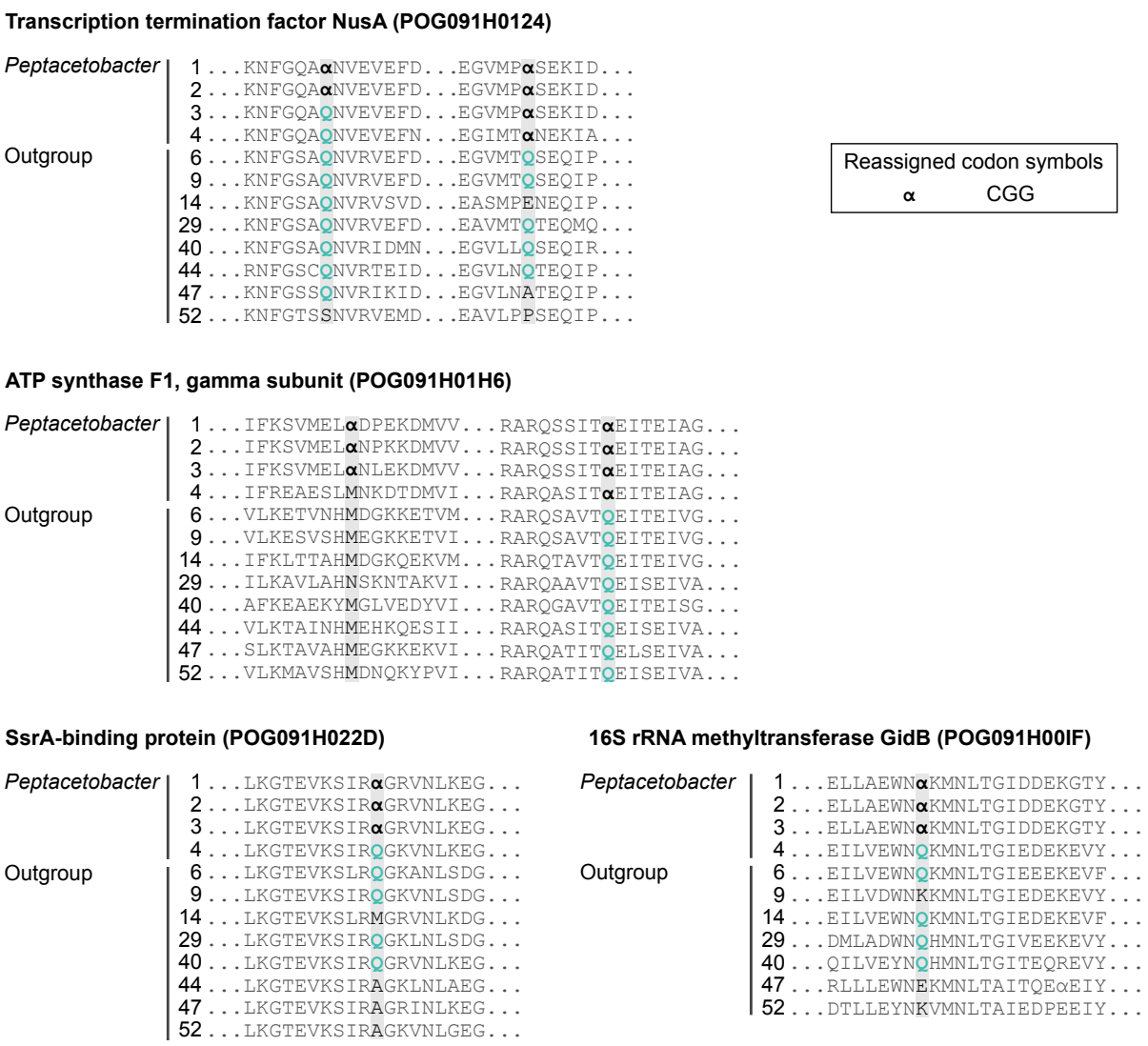

Figure 4.12: Multiple sequence alignments of transcription termination factor NusA (BUSCO POG091H0124), ATP synthase F1, gamma subunit (BUSCO POG091H01H6), SsrA-binding protein (BUSCO POG091H022D), and 16S rRNA methyltransferase GidB (BUSCO POG091H00IF) from *Peptacetobacter* and selected outgroup species. Alignment regions containing CGG (**α**) positions in Clostridium_U are shown.

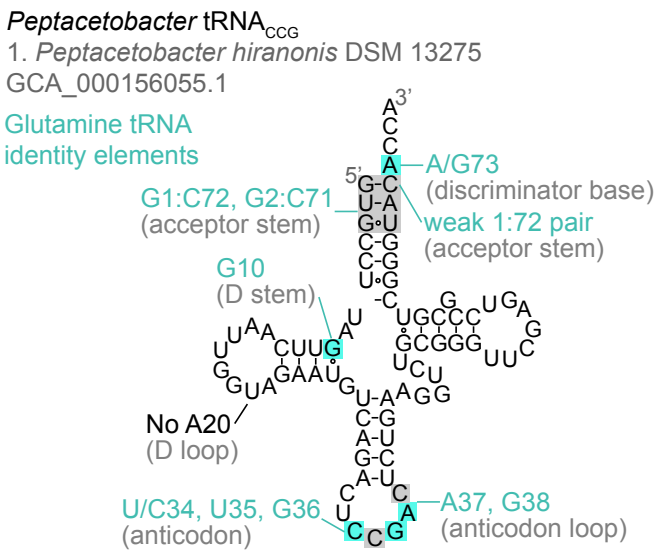

Figure 4.13: The CGG-decoding tRNA$_{CCG}$ from species #1 (*Peptacetobacter hiranonis* DSM 13275, GCA_000156055.1). tRNA sequence features involved in glutamine identity are highlighted (Jahn *et al.*, 1991; Hayase *et al.*, 1992), with nucleotide numbering following the convention of Sprinzl *et al.* (1998). Nucleotides highlighted in gray do not match the expected identity element.

tobacter tRNA$_{CCG}$.

In some *Peptacetobacter* species, the tRNA genes needed to read the arginine codons CGA, CGC, and CGU (tRNA$_{ACG}$ and/or tRNA$_{UCG}$) and the glutamine codon CAA (tRNA$_{UUG}$) cannot be found (Figure 4.11). This may be due to incomplete genome assembly because some members of *Peptacetobacter* do have these tRNA genes, and all of the four *Peptacetobacter* species are missing at least a few other required tRNA genes.

CGG is a rare codon in all members of *Peptacetobacter* (Figure 4.11), with usage of 4-6 per 10,000 codons in aligned Pfam domains. The low genomic GC content in *Peptacetobacter* species, ranging between 0.31-0.33, may have contributed to the low codon usage of CGG and the GC-rich arginine codons CGC, CGA, and AGG (each of these three codons has usage less than 22 per 10,000 codons) in favor of AGA (which is used at 314-355 per 10,000 codons). If low genomic GC content favored substitutions away from CGG prior to reassignment, this

Figure 4.14: An unrooted phylogenetic tree of arginine, glutamine, and reassigned tRNA sequences from *Peptacetobacter* and outgroup genomes, generated by a Bayesian approach implemented in MrBayes 3.2.7a. Values below each internal branch indicate the posterior probability of each clade. Since this is an unrooted tree, this value represents the probability that the sequences on either side of the branch form two distinct clades. Branch length scale in the bottom right is in expected substitutions per site. Sequence labels follow the format "species number - clade _ tRNA anticodon". This tree includes sequences from $tRNA_{CCG}$ from *Peptacetobacter* species (involved in the CGG reassignment and indicated with a gray circle), the arginine $tRNA_{UCU}$, $tRNA_{CCU}$, $tRNA_{ACG}$, and $tRNA_{UCG}$ (and $tRNA_{CCG}$ from outgroup species) indicated by a red circle, and the glutamine $tRNA_{UUG}$ and $tRNA_{CUG}$ from all groups indicated by a light blue circle.

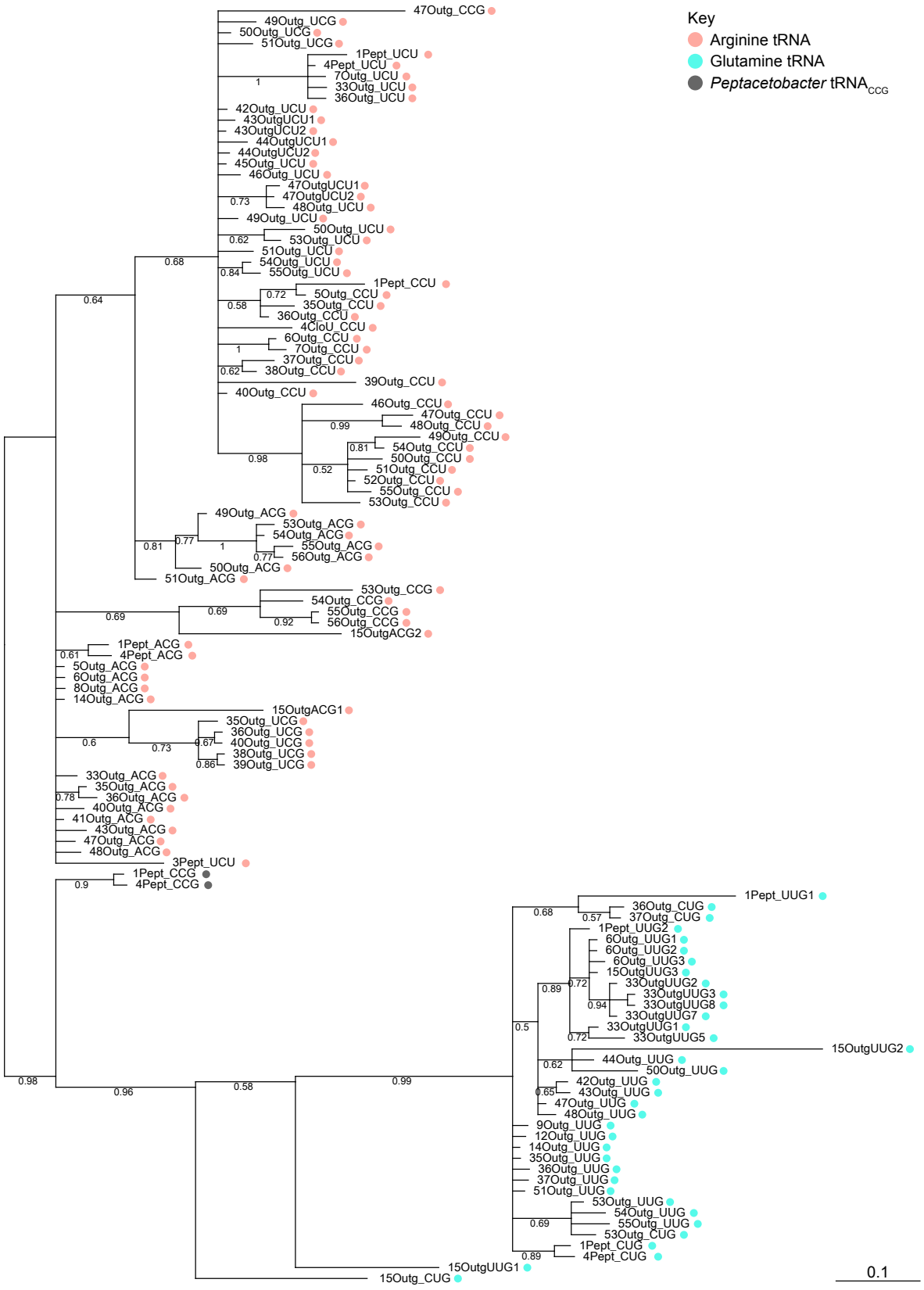

Figure 4.14 (Continued)

would have facilitated reassignment by minimizing the number of substitutions needed to adapt to the new translation.

In summary, the appearance of CGG at conserved glutamine positions in proteins has led to CGG being inferred as a glutamine codon in *Peptacetobacter* species by Codetta, and is supported by the usage of CGG at conserved glutamine positions in multiple sequence alignments of conserved single-copy genes. CGG is decoded in this clade by an unusual $tRNA_{CCG}$ whose isotype cannot be predicted by examining the tRNA sequence. *Peptacetobacter hominis* (GCA_006861675.1, species #4 on tree) did not have an inferred CGG meaning by Codetta despite 145 Pfam consensus columns aligned to CGG; however, the top amino acid model for CGG is glutamine (model probability of 0.984, below the default Codetta threshold for reporting). This might possibly reflect a low level of assembly contamination, occurrence of CGG at weakly conserved positions, or possibly ambiguous translation of CGG as more than one amino acid.

Potential loss of CGG decoding in some outgroup species

In *Peptacetobacter*, CGG is a rare codon, with codon usage of 4-6 per 10,000 codons in regions aligned to Pfam domains. In other members of the Peptostreptococcaceae (species #5-41 on tree) CGG is even rarer, occurring <2 times per 10,000 codons in all but two species. The overall low usage of CGG codons in the Peptostreptococcaceae may be tied to the low GC content of the clade, which ranges between 0.27-0.38 (Figure 4.11).

In the sister group to *Peptacetobacter* within the Peptostreptococcaceae (species #5-34 on tree, genera *Clostridioides*, *Asaccharospora*, *Intestinibacter*, *Terrisporobacter*, *Paeniclostridium*, *Paraclostridium*, and *Romboutsia*), I could not find any tRNA capable of decoding CGG (Figure 4.11, gray box outline). I wanted to confirm that this is not due to incomplete genome assembly, so I

analyzed the tRNA gene content of all 2,114 genome assemblies belonging to the 29 species in this clade (predominantly consisting of 2,014 *Clostridioides difficile* assemblies, species #8-11 on tree). None of the 2,114 genome assemblies contained a $tRNA_{CCG}$ and only a single assembly contained a $tRNA_{UCG}$ gene. The possibility remains that this tRNA cannot be found because it has a highly divergent sequence or contains an unusually long intron.

If a cognate tRNA for CGG is truly missing in this clade, then the few CGG codons in coding regions may be decoded inefficiently by the $tRNA^{Arg}_{ACG}$ or by a non-cognate tRNA. The bacterium *Mycoplasma capricolum* uses only a $tRNA^{Arg}_{ACG}$ to decode the entire arginine CGN codon box, which it accomplishes by leaving some fraction of the A34 wobble positions unmodified (typically deaminated to inosine) which can read all four CGN codons, albeit inefficiently (Yokobori *et al.*, 2013). The lack of a tRNA to decode CGG in most Peptostreptococcaceae species, in conjunction with low GC content, may also contribute to the extremely low usage of CGG in this family.

Possible model for the evolution of a CGG reassignment in *Peptacetobacter*

The results described above can be pulled together to paint a possible model of how the reassignment of CGG from arginine to glutamine evolved in *Peptacetobacter*. In the ancestor of the Peptostreptococcaceae, CGG was likely already a very rare codon due to low genomic GC content. The rarity of CGG may have released purifying selection on the efficiency and fidelity of CGG translation. At some point, the tRNA responsible for decoding CGG was lost and replaced by the $tRNA_{CCG}$ found in *Peptacetobacter* genomes presently. It would be parsimonious if the loss of the original CGG-decoding tRNA occurred first, prior to the divergence with the sister clade (species #5-34 on tree) which appear to lack a CGG-decoding tRNA, and the gain of the $tRNA_{CCG}$ occurred in the ancestor of *Peptacetobacter*.

One additional observation is that *Peptacetobacter* species favor the CAG glutamine codon over the CAA glutamine codon, in contrast to other Peptostreptococcaceae species. The two canonical glutamine codons are decoded by $tRNA^{Gln}_{UUG}$ and optionally $tRNA^{Gln}_{CUG}$. It is worth considering whether the reassignment of CGG to glutamine instead of to another amino acid may have been initiated by low genomic GC content-driven non-synonymous mutations from CUG to CGG introducing CGG into previously conserved glutamine positions, or possibly by mispairing of the $tRNA^{Gln}_{CUG}$ with CGG by a G-U wobble at the second anticodon position.

4.6 Conclusion

In this chapter, I provide further support for four reassignments of the CGA and/or CGG codons, originally predicted by Codetta in a screen of most available bacterial genomes (Chapter 2). This includes evidence of the reassigned codon being used at conserved positions for the predicted amino acid in multiple sequence alignments of conserved single-copy genes, features of the reassigned tRNA sequence that are consistent with known identity elements for the predicted amino acid, and an investigation of the reassigned codon in the nearest outgroup species. While computational support is compelling, only direct experimental evidence can definitively confirm these codon reassignments. Since Codetta selects the single best amino acid translation for each codon, some of these reassigned codons may be ambiguously translated in reality. Codetta also selects between the twenty canonical amino acids, it is also possible for the reassigned codons to actually translate as a non-canonical amino acid has a substitution pattern similar to either tryptophan or glutamine (depending on the reassignment). Two of these clades—*Anaerococcus* and *Peptacetobacter*—have culturable representatives, so the codon reassignments can be experimentally confirmed by determining the protein sequence of

a CGG-containing transcript, either by mass spectrometry or Edman degradation. The other two reassignments in Absconditabacteria and *Onthovivens* spp are only known from genome sequences of uncultured bacteria. Experimental support for these reassignments could be provided by showing that the reassigned aminoacyl-tRNA synthetase can aminoacylate the reassigned tRNA *in vitro* or by expressing the aminoacyl-tRNA synthetase and tRNA in *E. coli* and demonstrating that the reassigned tRNA can rescue missense mutations mirroring the reassignment. Analysis of close outgroup species also found possible reassignments of CGG to tryptophan in Gracilibacteria clade 1 (outgroup to Absconditabacteria) and in *Finegoldia* (outgroup to *Anaerococcus*), which would almost certainly require experimental validation.

I noticed that many of the close outgroup species have unusual decoding of the CGN codon box which might reflect exposure to shared evolutionary forces with the reassigned clade. As discussed at length in Chapter 2, all of the clades with CGA and/or CGG reassignments also have very low genomic GC content, which may have driven the GC-rich arginine codons like CGA and CGG to very rare usage in favor of more AT-rich synonyms like AGA and enabled reassignment. Many of the close outgroup species also share low genomic GC content and very low usage of CGN arginine codons such as CGA and CGG. Presumably due to relaxed selection on the accurate and efficient translation of these codons, we also see unusual features in the tRNAs for the CGN codon box. In Peptostreptococcaceae species (outgroups to the *Peptacetobacter* reassignment of CGG to glutamine), only a single $tRNA^{Arg}_{ACG}$ is available to decode the entire very rare CGN codon box. The $tRNA_{UCG}$ and/or $tRNA_{CCG}$ that are typically present were likely able to be lost due to the extremely rare usage of these codons, either caused by the very low genomic GC content shared with *Peptacetobacter* or possibly translational pressure from the first stages of CGG codon reassignment. In *Finegoldia*

(outgroup to *Anaerococcus*), it appears that a concurrent CGG to tryptophan reassignment may be in the early stages of emerging. It would be interesting to understand whether this is a convergent reassignment that was enabled the same pressures that drove CGG to rarity in the ancestor of *Anaerococcus*, or whether the common ancestor of *Anaerococcus* and *Finegoldia* had some degree of CGG translation as tryptophan which was lost in other outgroups. Finally, Gracilibacteria clade 1 (outgroup of Absconditabacteria) has a slew of unusual features in the CGN-decoding tRNAs, including a C33 in the anticodon loop of the $tRNA^{Arg}_{GCG}$ which may permit extended wobble to read the CGA codon as arginine, and a six base pair anticodon stem and non-canonical TΨC loop - D loop interaction in the $tRNA_{CCG}$ which might partake in reassigning CGG to tryptophan. Once more, it is unclear whether these deviations in CGN decoding are caused by background factors shared with Absconditabacteria or whether the common ancestor was in some stage of codon reassignment that was lost in other outgroups.

The genetic code, and translation in general, is typically thought of as a system heavily optimized for efficient operation and minimal error. These codon reassignments highlight how in the right conditions, codon translation can be inefficient, noisy, liable to mutations to "invariant" features, and change even on relatively short evolutionary distances. By studying organisms that have evolved exceptions to canonical translation, we stand to gain a deeper and more holistic understanding of this otherwise ancient, conserved, and unchanging biological process.

4.7 Methods

Genetic code inference of bacterial genomes with Codetta was performed as part of the larger screen in Chapter 2; detailed methods are described there. Other methods, such as tRNA

and release factor gene prediction, manual tRNA isotype classification, and generation of single-copy gene multiple sequence alignments are also described in Chapter 2. The following methods are unique to this chapter.

Genome completeness estimate by tRNA presence

In addition to the CheckM completeness score that is provided for every genome in GTDB, for CGA and/or CGG codon reassignments I additionally assessed genome completeness by tabulating the presence of a set of required tRNA genes found by tRNAscan-SE 2.0 (Chan *et al.*, 2021) (top isotype score of >35 or general model score >50). This is a minimal set of 22 tRNA anticodons that are required for the ability to decode all sense codons (excluding CGN) in bacteria, comprised of: Phe GAA, Leu UAA and UAG, Ile GAU, Ile/Met CAU, Val UAC, Ser UGA and GCU, Pro UGG, Thr UGU, Ala UGC, Tyr GUA, His GUG, Gln UUG, Asn GUU, Lys UUU, Asp GUC, Glu UUC, Cys GCA, Arg UCU, Gly UCC, Gly/Trp CCA. I excluded tRNAs that are involved in the decoding of the CGN-box which includes the reassigned codons.

tRNA phylogeny

Phylogenetic trees were inferred from manually-created tRNA alignments of tRNAs that decode the reassigned codon and a selection of tRNAs for the original and new amino acid from the reassigned clade and outgroups. Trees were inferred using a Bayesian approach implemented in MrBayes 3.2.7a (Ronquist and Huelsenbeck, 2003) run for 40 million generations with sampling every 500 generations with default burn-in. I excluded columns corresponding to the anticodon and to the 3′ half of stems (to remove the correlated information in base paired stem regions). The remaining columns were partitioned into stem and loop regions and were modeled by separately parameterized GTR substitution models with gamma-distributed

rate variation across sites and a proportion of invariable sites. To estimate the marginal likelihood of specific phylogenetic models where the tree topology is constrained, a constraint was specified that forced a specific partition of sequences and the likelihood was estimated via the stepping-stone sampling approach implemented in MrBayes 3.2.7a run for 40 million generations (50 steps of 799,680 generations, default burn-in).

Supplementary files

Supplementary files can be obtained at `https://github.com/kshulgina/ShulginaEddy_21_genetic_codes`. For each reassignment, I provide:

- An Excel spreadsheet with information on the reassigned and close outgroup genomes, including assembly accessions, tRNA gene presence, codon usage, and GC content.
- Stockholm alignment of tRNA genes used to build the tRNA phylogenies.
- Stockholm alignments of BUSCO genes shown in alignment figures.

Acknowledgments

Thanks to Eric Westhof for stimulating discussions about the Gracilibacteria clade 1 $tRNA_{CCG}$ and for bringing my attention to some of the unusual features that I had not previously noticed.

Chapter 5

Conclusion

Early in graduate school, I read several papers that spurred my interest in alternative genetic codes. In one paper, Mühlhausen *et al.* (2016) showed that yeasts in the genus *Pachysolen* translate the leucine codon CUG as alanine, adding to the already-known reassignment of CUG to serine in other yeasts and raising questions about why CUG is repeatedly affected. In another paper, Swart *et al.* (2016) found a truly unprecedented genetic code in the ciliate *Condylostoma magnum*, where all three stop codons have been reassigned to amino acids, only functioning as stop codons in certain mRNA contexts. Both discoveries made me wonder—how and why do codon reassignments like these evolve? And could there be other (and perhaps even more unusual) variations out there?

These papers came after about a fifteen year lull in the alternative genetic code field. Much of what we know about alternative genetic codes came from research in the 1980's to early 2000's (reviewed in Chapter 1). New genetic codes led to a deeper understanding of the process of translation—from the discovery of new types of tRNA wobble modifications in mitochondria that widened the known possibilities of tRNA decoding (Watanabe and Yokobori,

2011) to the discovery of the 22nd amino acid pyrrolysine (Srinivasan *et al.*, 2002), kickstarting the field of genetic code expansion. The recent resurgence in research on alternative genetic codes was in part enabled by the maturing of sequencing technology, making it feasible and cost-effective to sequence the diversity of life. Improvements in metagenomic and metatranscriptomic sequencing of unculturable samples and subsequent computational assembly of single organism genomes and transcriptomes started shedding light on the so-called "microbial dark matter" (Venter *et al.*, 2004; Tyson *et al.*, 2004; Simon and Daniel, 2011; Parks *et al.*, 2017). Massive metagenomic and metatranscriptome sequencing efforts, such as the Tara Oceans project (Sunagawa *et al.*, 2015), the Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling *et al.*, 2014), work by the Banfield lab (Tyson *et al.*, 2004; Brown *et al.*, 2015b; Al-Shayeb *et al.*, 2020), and many others led to the discovery of many uncultivated microbial lineages, including entire bacterial and archaeal phyla (Hug *et al.*, 2016). Furthermore, early analyses revealed that some of the newly sequenced organisms use alternative genetic codes (Rinke *et al.*, 2013; Ivanova *et al.*, 2014). Thus, by the time I began my PhD, there was already a solid foundation in understanding how alternative genetic codes may evolve and a recent boom in sequences from previously unseen parts of the tree of life. It was an opportune time to take on the challenge of sifting through thousands of genomes to look for new variations in the genetic code.

In this thesis, I developed a computational approach called Codetta, which can predict the genetic code from DNA or RNA sequences from a single organism. Codetta improves upon previous approaches to predict the genetic code and has sufficient speed and low false positive rate to enable screening whole domains of life (Chapter 3). In a screen of all available bacterial and archaeal genomes, we found five bacterial clades using new alternative genetic

codes (Chapter 2). These new genetic codes are the first sense codon changes found in bacteria, highlighting the fact that known reassignments may be biased towards stop codon changes because they are simpler to detect. Surprisingly, all of the new reassignments affected arginine codons, which may be a common target for reassignment due to the tendency of some arginine codons to become rare and due to how these codons are decoded by tRNAs. With a more complete view of bacterial genetic code diversity, we could see that most bacterial genetic code variations were associated with low genomic GC content and reduced genome size, suggesting that “codon capture"-like processes might be dominant reasons why codon reassignments evolve in bacteria. Once we have a similarly complete distribution of genetic codes in other domains of life, we may find that codon reassignments may tend to evolve in other domains for different predominant reasons. Individually, each of the new reassignments revealed traces of evolutionary events and changes to the translational machinery that led to the codon reassignment, some persisting in closely related outgroup species (Chapter 4). Each of the new genetic codes and other exceptions that were found should be confirmed experimentally, especially in *Anaerococcus* and *Peptacetobacter* which have culturable representatives.

Codetta has some limitations, and future work can improve on these aspects. One restriction is that Codetta only predicts amino acid meanings and does not predict which codons are stop codons. This is because Codetta relies on alignment of Pfam models, which typically span a single protein domain. Stop codon prediction depends on mapping the location of the termination codon, which requires full length gene models; otherwise stop codons cannot be reliably distinguished from extremely rare sense codons. Furthermore, Codetta assumes that the entire genome uses the same genetic code. Since the discovery of multiple translational meanings within a single organism, either by stochastic decoding like in some yeasts

(Mühlhausen *et al.*, 2018), mRNA context-dependent decoding like in some ciliates (Swart *et al.*, 2016; Heaphy *et al.*, 2016; Seah *et al.*, 2022), or condition-specific reassignment like in some phage and pyrrolysine-using bacteria (Borges *et al.*, 2022; Prat *et al.*, 2012), it has become clear that this is not always a correct assumption. Thus it could be interesting to develop a way to infer codon ambiguity or different translational meanings at different codon positions.

We focused on bacterial and archaeal genomes for the initial Codetta screen for several practical reasons: 1) bacterial and archaeal genomes tend to have fewer pseudogenes which results in fewer Codetta errors, 2) we have a rich knowledge base about bacterial tRNA identity elements, which simplifies confirming predicted genetic codes, and 3) the existence of complete phylogenies of almost all sequenced bacterial and archaeal genomes on the Genome Taxonomy Database (Parks *et al.*, 2020). Other alternative genetic codes likely exist, waiting to be found, in organellar genomes, eukaryotic genomes, viral genomes, and, of course, new bacterial and archaeal genomes that continue to be sequenced. These should be screened using Codetta. Once we have a reasonably complete set of alternative genetic codes, we can ask general questions in a more precise way—how many organisms use alternative genetic codes? How often do they evolve? What evolutionary trajectories do most codon reassignments appear to take? Which codon reassignments are the most frequent and why?

As they have in the past, alternative genetic codes continue to expand our understanding of translation and raise questions about what else is possible. How much capacity does the genetic code have to evolve? To what extent does the translational machinery constrain the types of codon reassignments that can evolve? Can translation termination be completely decoupled from stop codons? Do existing tRNA-codon wobble rules represent the limit of codon discrimination that is physically possible or merely what the standard genetic code

requires? By studying exceptions to normally invariant rules, we gain a richer understanding of the fundamental limitations of translation and how the translational machinery can be potentially engineered to design orthogonal translation systems and expanded genetic codes.

References

Abascal, F., Posada, D., Knight, R. D. and Zardoya, R. (2006a). Parallel evolution of the genetic code in arthropod mitochondrial genomes. *PLoS Biology*, **4**, e127.

—, Zardoya, R. and Posada, D. (2006b). GenDecoder: Genetic code prediction for metazoan mitochondria. *Nucleic Acids Research*, **34**, W389–W393.

Adio, S., Sharma, H., Senyushkina, T., Karki, P., Maracci, C., Wohlgemuth, I., Holtkamp, W., Peske, F. and Rodnina, M. V. (2018). Dynamics of ribosomes and release factors during translation termination in *E. coli*. *eLife*, **7**, e34252.

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., Sharrar, A., Jaffe, A. L., Nelson, T. C., Kantor, R., Keren, R., Lane, K. R., Farag, I. F., Lei, S., Finstad, K., Amundson, R., Anantharaman, K., Zhou, J., Probst, A. J., Power, M. E., Tringe, S. G., Li, W.-J., Wrighton, K., Harrison, S., Morowitz, M., Relman, D. A., Doudna, J. A., Lehours, A.-C., Warren, L., Cate, J. H. D., Santini, J. M. and Banfield, J. F. (2020). Clades of huge phages from across Earth's ecosystems. *Nature*, **578**, 425–431.

Alfonzo, J. D., Blanc, V., Estévez, A. M., Rubio, M. A. T. and Simpson, L. (1999). C to U editing of the anticodon of imported mitochondrial $tRNA^{Trp}$ allows decoding of the UGA stop codon in *Leishmania tarentolae*. *The EMBO Journal*, **18**, 7056–7062.

—, Thiemann, O. and Simpson, L. (1997). The mechanism of U insertion/deletion RNA editing in kinetoplastid mitochondria. *Nucleic Acids Research*, **25**, 3751–3759.

Alkalaeva, E., Eliseev, B., Ambrogelly, A., Vlasov, P., Kondrashov, F. A., Gundllapalli, S., Frolova, L., Söll, D. and Kisselev, L. (2009). Translation termination in pyrrolysine-utilizing archaea. *FEBS Letters*, **583**, 3455–3460.

— and Mikhailova, T. (2017). Reassigning stop codons via translation termination: How a few eukaryotes broke the dogma. *Bioessays*, **39**.

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D. and Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, **568**, 499–504.

Andachi, Y., Yamao, F., Muto, A. and Osawa, S. (1989). Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *Mycoplasma capricolum*. resemblance to mitochondria. *Journal of Molecular Biology*, **209**, 37–54.

Anderson, S., Bankier, A., Barrell, B., de Bruijn, M., Coulson, A., Drouin, J., Eperon, I., Nierlich, D., Roe, B., Sanger, F., Schreier, P., Smith, A., Staden, R. and Young, I. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.

Andersson, S. G. E. and Kurland, C. G. (1991). An extreme codon preference strategy: Codon reassignment. *Molecular Biology and Evolution*, **8**, 530–544.

— and — (1995). Genomic evolution drives the evolution of the translation system. *Biochemistry and Cell Biology*, **73**, 775–787.

Ashraf, S. S., Ansari, G., Guenther, R., Sochacka, E., Malkiewicz, A. and Agris, P. F. (1999). The uridine in "U-turn": Contributions to tRNA-ribosomal binding. *RNA*, **5**, 503–511.

Bachvaroff, T. R. (2019). A precedented nuclear genetic code with all three termination codons reassigned as sense codons in the syndinean *Amoebophrya* sp. ex *Karlodinium veneficum*. *PLoS One*, **14**, e0212912.

Barrell, B. G., Bankier, A. T. and Drouin, J. (1979). A different genetic code in human mitochondria. *Nature*, **282**, 189–194.

Beagley, C. T., Okimoto, R. and Wolstenholme, D. R. (1998). The mitochondrial genome of the sea anemone *Metridium senile* (cnidaria): Introns, a paucity of tRNA genes, and a near-standard genetic code. *Genetics*, **148**, 1091–1108.

Benne, R., De Vries, B. F., Van den Burg, J. and Klaver, B. (1983). The nucleotide sequence of a segment of *Trypanosoma brucei* mitochondrial maxi-circle DNA that contains the gene for apocytochrome b and some unusual unassigned reading frames. *Nucleic Acids Research*, **11**, 6925–6941.

Bennett, G. M. and Moran, N. A. (2013). Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution*, **5**, 1675–1688.

Bjork, G. R. (1994). Biosynthesis and function of modified nucleosides. In *tRNA: Structure, Biosynthesis, and Function*, ASM Press, pp. 165–205.

Boore, J. L., Daehler, L. L. and Brown, W. M. (1999). Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). *Molecular Biology and Evolution*, **16**, 410–418.

Borges, A. L., Lou, Y. C., Sachdeva, R., Al-Shayeb, B., Penev, P. I., Jaffe, A. L., Lei, S., Santini, J. M. and Banfield, J. F. (2022). Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. *Nature Microbiology*, **7**, 918–927.

Borrel, G., Gaci, N., Peyret, P., O'Toole, P. W., Gribaldo, S. and Brugère, J.-F. (2014). Unique characteristics of the pyrrolysine system in the 7th order of methanogens: Implications for the evolution of a genetic code expansion cassette. *Archaea*, **2014**, 374146.

Bové, J. M. (1993). Molecular features of mollicutes. *Clinical Infectious Diseases*, **17 Suppl 1**, S10–31.

Boyen, C., Leblanc, C., Bonnard, G., Grienenberger, J. M. and Kloareg, B. (1994). Nucleotide sequence of the cox3 gene from *Chondrus crispus*: evidence that UGA encodes tryptophan and evolutionary implications. *Nucleic Acids Research*, **22**, 1400–1403.

—, Oudot, M.-P. and Goer, S. L. D. (2001). Origin and evolution of plastids and mitochondria: The phylogenetic diversity of algae. *Cahiers de Biologie Marine*, **42**, 11–24.

Bridges, H. R., Birrell, J. A. and Hirst, J. (2011). The mitochondrial-encoded subunits of respiratory complex I (NADH:ubiquinone oxidoreductase): Identifying residues important in mechanism and disease. *Biochemical Society Transactions*, **39**, 799–806.

Brown, A., Shao, S., Murray, J., Hegde, R. S. and Ramakrishnan, V. (2015a). Structural basis for stop codon recognition in eukaryotes. *Nature*, **524**, 493–496.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H. and Banfield, J. F. (2015b). Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, **523**, 208–211.

Burger, G., Forget, L., Zhu, Y., Gray, M. W. and Lang, B. F. (2003). Unique mitochondrial genome architecture in unicellular relatives of animals. *PNAS*, **100**, 892–897.

—, Gray, M. W., Forget, L. and Lang, B. F. (2013). Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biology and Evolution*, **5**, 418–438.

—, Plante, I., Lonergan, K. M. and Gray, M. W. (1995). The mitochondrial DNA of the amoeboid protozoon *Acanthamoeba castellanii*: complete sequence, gene content and genome organization. *Journal of Molecular Biology*, **245**, 522–537.

—, Zhu, Y., Littlejohn, T. G., Greenwood, S. J., Schnare, M. N., Lang, B. F. and Gray, M. W. (2000). Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. *Journal of Molecular Biology*, **297**, 365–380.

Burki, F., Roger, A. J., Brown, M. W. and Simpson, A. G. B. (2020). The new tree of eukaryotes. *Trends in Ecology and Evolution*, **35**, 43–55.

Campbell, J. H., O'Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., Soll, D. and Podar, M. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *PNAS*, **110**, 5540–5545.

Campbell, M. A., Van Leuven, J. T., Meister, R. C., Carey, K. M., Simon, C. and McCutcheon, J. P. (2015). Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*. *PNAS*, **112**, 10192–10199.

Carbon, J. and Fleck, E. W. (1974). Genetic alteration of structure and function in glycine transfer RNA of *Escherichia coli*: Mechanism of suppression of the tryptophan synthetase A78 mutation. *Journal of Molecular Biology*, **85**, 371–391.

Caron, F. and Meyer, E. (1985). Does *Paramecium primaurelia* use a different genetic code in its macronucleus? *Nature*, **314**, 185–188.

Celis, J. E., Hooper, M. L. and Smith, J. D. (1973). Amino acid acceptor stem of *E. coli* suppressor $tRNA^{Tyr}$ is a site of synthetase recognition. *Nature New Biology*, **244**, 261–264.

Chan, P. P., Lin, B. Y., Mak, A. J. and Lowe, T. M. (2021). tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, **49**, 90779096.

Chapman, B. and Brown, C. (2004). Translation termination in *Arabidopsis thaliana*: Characterisation of three versions of release factor 1. *Gene*, **341**, 219–225.

Chu, Y., Gao, P., Zhao, P., He, Y., Liao, N., Jackman, S., Zhao, Y., Birol, I., Duan, X. and Lu, Z. (2011). Genome sequence of *Mycoplasma capricolum* subsp. *capripneumoniae* strain M1601. *Journal of Bacteriology*, **193**, 6098–6099.

Citti, C., Maréchal-Drouard, L., Saillard, C., Weil, J. H. and Bové, J. M. (1992). *Spiroplasma citri* UGG and UGA tryptophan codons: Sequence of the two tryptophanyl-tRNAs and organization of the corresponding genes. *Journal of Bacteriology*, **174**, 6471–6478.

Clary, D. O. and Wolstenholme, D. R. (1985). The mitochondrial DNA molecule of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution*, **22**, 252–271.

Cocquyt, E., Gile, G. H., Leliaert, F., Verbruggen, H., Keeling, P. J. and De Clerck, O. (2010). Complex phylogenetic distribution of a non-canonical genetic code in green algae. *BMC Ecology and Evolution*, **10**, 327.

Conard, S. E., Buckley, J., Dang, M., Bedwell, G. J., Carter, R. L., Khass, M. and Bedwell, D. M. (2012). Identification of eRF1 residues that play critical and complementary roles in stop codon recognition. *RNA*, **18**, 1210–1221.

Crick, F. H. (1968). The origin of the genetic code. *Journal of Molecular Biology*, **38**, 367–379.

Cupples, C. G. and Pearlman, R. E. (1986). Isolation and characterization of the actin gene from *Tetrahymena thermophila*. *PNAS*, **83**, 5160–5164.

Curran, J. F. (1998). Modified nucleosides in translation. In *Modification and Editing of RNA*, ASM Press, pp. 493–516.

De Koning, A. P., Noble, G. P., Heiss, A. A., Wong, J. and Keeling, P. J. (2008). Environmental PCR survey to determine the distribution of a non-canonical genetic code in uncultivable oxymonads. *Environmental Microbiology*, **10**, 65–74.

Dever, T. E., Dinman, J. D. and Green, R. (2018). Translation elongation and recoding in eukaryotes. *Cold Spring Harbor Perspectives in Biology*, **10** (8), a032649.

Dirheimer, G., Keith, G., Dumas, P. and Westhof, E. (1994). Primary, secondary, and tertiary structures of tRNAs. In *tRNA: Structure, Biosynthesis, and Function*, ASM Press, pp. 93–126.

Doyon, F. R., Zagryadskaya, E. I., Chen, J. and Steinberg, S. V. (2004). Specific and non-specific purine trap in the t-loop of normal and suppressor tRNAs. *Journal of Molecular Biology*, **343**, 55–69.

Duarte, I., Nabuurs, S. B., Magno, R. and Huynen, M. (2012). Evolution and diversification of the organellar release factor family. *Molecular Biology and Evolution*, **29**, 3497–3512.

Durrheim, G. A., Corfield, V. A., Harley, E. H. and Ricketts, M. H. (1993). Nucleotide sequence of cytochrome oxidase (subunit III) from the mitochondrion of the tunicate *Pyura stolonifera*: Evidence that AGR encodes glycine. *Nucleic Acids Research*, **21**, 3587–3588.

Dutilh, B. E., Jurgelenaite, R., Szklarczyk, R., van Hijum, S. A. F. T., Harhangi, H. R., Schmid, M., de Wild, B., Francoijs, K. J., Stunnenberg, H. G., Strous, M., Jetten, M. S. M., Op den Camp, H. J. M. and Huynen, M. A. (2011). FACIL: Fast and accurate genetic code inference and logo. *Bioinformatics*, **27**, 1929–1933.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

Ehara, M., Hayashi-Ishimaru, Y., Inagaki, Y. and Ohama, T. (1997). Use of a deviant mitochondrial genetic code in yellow-green algae as a landmark for segregating members within the phylum. *Journal of Molecular Evolution*, **45**, 119–124.

—, Inagaki, Y., Watanabe, K. I. and Ohama, T. (2000). Phylogenetic analysis of diatom coxi genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. *Current Genetics*, **37**, 29–33.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. and Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, **47**, D427–D432.

Eliseev, B., Kryuchkova, P., Alkalaeva, E. and Frolova, L. (2011a). A single amino acid change of translation termination factor eRF1 switches between bipotent and omnipotent stop-codon specificity. *Nucleic Acids Research*, **39**, 599–608.

Eliseev, B. D., Alkalaeva, E. Z., Kryuchkova, P. N., Lekomtsev, S. A., Wang, W., Liang, A.-H. and Frolova, L. Y. (2011b). Translation termination factor eRF1 of the ciliate *Blepharisma japonicum* recognizes all three stop codons. *Molecular Biology of the Cell*, **45**, 614.

Flegontov, P., Michálek, J., Janouškovec, J., Lai, D.-H., Jirků, M., Hajdušková, E., Tomčala, A., Otto, T. D., Keeling, P. J., Pain, A., Oborník, M. and Lukeš, J. (2015). Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. *Molecular Biology and Evolution*, **32**, 1115–1131.

Fournier, G. P. and Andam, C. P. (2012). Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. *Origins of Life and Evolution of Biospheres*, **41**, 621–632.

Fox, T. D. (1987). Natural variation in the genetic code. *Annual Reviews of Genetics*, **21**, 67–91.

Freeland, S. J. and Hurst, L. D. (1998). The genetic code is one in a million. *Journal of Molecular Evolution*, **47**, 238–248.

—, Knight, R. D., Landweber, L. F. and Hurst, L. D. (2000). Early fixation of an optimal genetic code. *Molecular Biology and Evolution*, **17**, 511–518.

Freistroffer, D. V., Kwiatkowski, M., Buckingham, R. H. and Ehrenberg, M. (2000). The accuracy of codon recognition by polypeptide release factors. *PNAS*, **97**, 2046–2051.

Frolova, L., Seit-Nebi, A. and Kisselev, L. (2002). Highly conserved NIKS tetrapeptide is functionally essential in eukaryotic translation termination factor eRF1. *RNA*, **8**, 129–136.

Fučíková, K., Lewis, P. O., González-Halphen, D. and Lewis, L. A. (2014). Gene arrangement convergence, diverse intron content, and genetic code modifications in mitochondrial genomes of Sphaeropleales (Chlorophyta). *Genome Biology and Evolution*, **6**, 2170–2180.

Futo, S., Seto, Y., Mitsuse, S., Mori, Y., Suzuki, T. and Kawai, K. (1995). Molecular cloning of a 46-kilodalton surface antigen (p46) gene from *Mycoplasma hyopneumoniae*: Direct evidence of CGG codon usage for arginine. *Journal of Bacteriology*, **177**, 1915–1917.

Garey, J. R. and Wolstenholme, D. R. (1989). Platyhelminth mitochondrial DNA: Evidence for early evolutionary origin of a $tRNA^{Ser}_{AGN}$ that contains a dihydrouridine arm replacement loop, and of serine-specifying AGA and AGG codons. *Journal of Molecular Evolution*, **28**, 374–387.

Gaston, M. A., Jiang, R. and Krzycki, J. A. (2011). Functional context, biosynthesis, and genetic encoding of pyrrolysine. *Current Opinion in Microbiology*, **14**, 342–349.

Giegé, R., Sissler, M. and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research*, **26**, 5017–5035.

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

Gonzalez-Flores, J. N., Shetty, S. P., Dubey, A. and Copeland, P. R. (2013). The molecular biology of selenocysteine. *Biomolecular Concepts*, **4**, 349–365.

Gorini, L. (1970). Informational suppression. *Annual Review of Genetics*, **4**, 107–134.

Gornik, S. G., Flores, V., Reinhardt, F., Erber, L., Salas-Leiva, D. E., Douvropoulou, O., Lassadi, I., Einarsson, E., Mörl, M., Git, A., Stadler, P. F., Pain, A. and Waller, R. F. (2022). Mitochondrial genomes in *Perkinsus* decode conserved frameshifts in all genes. *Molecular Biology and Evolution*, **39**, msac191.

Gray, M. W. (1996). RNA editing in plant organelles: A fertile field. *PNAS*, **93**, 8157–8159.

—, Lang, B. F., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T. G., Plante, I., Rioux, P., Saint-Louis, D., Zhu, Y. and Burger, G. (1998). Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Research*, **26**, 865–878.

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

Guan, Y., Haroon, M. F., Alam, I., Ferry, J. G. and Stingl, U. (2017). Single-cell genomics reveals pyrrolysine-encoding potential in members of uncultivated archaeal candidate division MSBL1. *Environmental Microbiology Reports*, **9**, 404–410.

Guo, L.-T., Amikura, K., Jiang, H.-K., Mukai, T., Fu, X., Wang, Y.-S., O'Donoghue, P., Söll, D. and Tharp, J. M. (2022). Ancestral archaea expanded the genetic code with pyrrolysine. *Journal of Biological Chemistry*, p. 102521.

Hafez, M., Burger, G., Steinberg, S. V. and Lang, B. F. (2013). A second eukaryotic group with mitochondrion-encoded tmRNA: In silico identification and experimental confirmation. *RNA Biology*, **10**, 1117–1124.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K. and Beck, E. (2013). TIGRFAMs and genome properties in 2013. *Nucleic Acids Research*, **41**, D387–95.

Hanke, A., Hamann, E., Sharma, R., Geelhoed, J. S., Hargesheimer, T., Kraft, B., Meyer, V., Lenk, S., Osmers, H., Wu, R., Makinwa, K., Hettich, R. L., Banfield, J. F., Tegetmeyer, H. E. and Strous, M. (2014). Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between alpha- and gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. *Frontiers of Microbiology*, **5**, 231.

Hanyu, N., Kuchino, Y., Nishimura, S. and Beier, H. (1986). Dramatic events in ciliate evolution: Alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* $\text{tRNAs}^{\text{Gln}}$. *The EMBO Journal*, **5**, 1307–1311.

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.

Hatanaka, A., Umeda, N., Yamashita, S. and Hirazawa, N. (2007). Identification and characterization of a putative agglutination/immobilization antigen on the surface of *Cryptocaryon irritans*. *Parasitology*, **134**, 1163–1174.

Hayase, Y., Jahn, M., Rogers, M. J., Sylvers, L. A., Koizumi, M., Inoue, H., Ohtsuka, E. and Söll, D. (1992). Recognition of bases in *Escherichia coli* tRNA(Gln) by glutaminyl-tRNA synthetase: A complete identity set. *The EMBO Journal*, **11**, 4159–4165.

Hayashi-Ishimaru, Y., Ehara, M., Inagaki, Y. and Ohama, T. (1997). A deviant mitochondrial genetic code in prymnesiophytes (yellow-algae): UGA codon for tryptophan. *Current Genetics*, **32**, 296–299.

—, Ohama, T., Kawatsu, Y., Nakamura, K. and Osawa, S. (1996). UAG is a sense codon in several chlorophycean mitochondria. *Current Genetics*, **30**, 29–33.

Heaphy, S. M., Mariotti, M., Gladyshev, V. N., Atkins, J. F. and Baranov, P. V. (2016). Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. *Molecular Biology and Evolution*, **33**, 2885–2889.

Heckman, J. E., Sarnoff, J., Alzner-DeWeerd, B., Yin, S. and RajBhandary, U. L. (1980). Novel features in the genetic code and codon reading patterns in *Neurospora crassa* mitochondria based on sequences of six mitochondrial tRNAs. *PNAS*, **77**, 3159–3163.

Helftenbein, E. (1985). Nucleotide sequence of a macronuclear DNA molecule coding for alpha-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a translation termination codon. *Nucleic Acids Research*, **13**, 415–433.

Hill, C. W., Combriato, G. and Dolph, W. (1974). Three different missense suppressor mutations affecting the $tRNA_{GGG}^{Gly}$ species of *Escherichia coli*. *Journal of Bacteriology*, **117**, 351–359.

Himeno, H., Hasegawa, T., Asahara, H., Tamura, K. and Shimizu, M. (1991). Identity determinants of *E. coli* tryptophan tRNA. *Nucleic Acids Research*, **19**, 6379–6382.

—, Masaki, H., Kawai, T., Ohta, T., Kumagai, I., Miura, K.-i. and Watanabe, K. (1987). Unusual genetic codes and a novel gene structure for $tRNA^{Ser}_{AGY}$ in starfish mitochondria DNA. *Gene*, **56**, 219–230.

—, Yoshida, S., Soma, A. and Nishikawa, K. (1997). Only one nucleotide insertion to the long variable arm confers an efficient serine acceptor activity upon *Saccharomyces cerevisiae* tRNA(Leu) in vitro. *Journal of Molecular Biology*, **268**, 704–711.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. and Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research*, **24**, 4420–4449.

Hirsh, D. (1971). Tryptophan transfer RNA as the UGA suppressor. *Journal of Molecular Biology*, **58**, 439–458.

Holmquist, R. (1978). Evaluation of compositional nonrandomness in proteins. *Journal of Molecular Evolution*, **11**, 349–360.

Horowitz, S. and Gorovsky, M. A. (1985). An unusual genetic code in nuclear genes of *Tetrahymena*. *PNAS*, **82**, 2452–2455.

Hou, Y. M. and Schimmel, P. (1989). Evidence that a major determinant for the identity of a transfer RNA is conserved in evolution. *Biochemistry*, **28**, 6800–6804.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C. and Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, **1**, 16048.

Ibba, M., Becker, H. D., Stathopoulos, C., Tumbula, D. L. and Söll, D. (2000). The adaptor hypothesis revisited. *Trends in Biochemical Sciences*, **25**, 311–316.

— and Söll, D. (2000). Aminoacyl-tRNA synthesis. *Annual Review of Biochemistry*, **69**, 617–650.

Inagaki, Y., Bessho, Y. and Osawa, S. (1993). Lack of peptide-release activity responding to codon UGA in *Mycoplasma capricolum*. *Nucleic Acids Research*, **21**, 1335–1338.

—, Blouin, C., Doolittle, W. F. and Roger, A. J. (2002). Convergence and constraint in eukaryotic release factor 1 (eRF1) domain 1: The evolution of stop codon specificity. *Nucleic Acids Research*, **30**, 532–544.

—, Ehara, M., Watanabe, K. I., Hayashi-Ishimaru, Y. and Ohama, T. (1998). Directionally evolving genetic code: The UGA codon from stop to tryptophan in mitochondria. *Journal of Molecular Evolution*, **47**, 378–384.

Ito, K., Uno, M. and Nakamura, Y. (1998). Single amino acid substitution in prokaryote polypeptide release factor 2 permits it to terminate translation at all three stop codons. *PNAS*, **95**, 8165–8169.

Ivanov, A., Mikhailova, T., Eliseev, B., Yeramala, L., Sokolova, E., Susorov, D., Shuvalov, A., Schaffitzel, C. and Alkalaeva, E. (2016). PABP enhances release factor recruitment and stop codon recognition during translation termination. *Nucleic Acids Research*, **44**, 7766–7776.

Ivanova, N. N., Schwientek, P., Tripp, H. J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N. C. and Rubin, E. M. (2014). Stop codon reassignments in the wild. *Science*, **344**, 909–913.

Jackson, R. J., Hellen, C. U. T. and Pestova, T. V. (2012). Termination and post-termination events in eukaryotic translation. *Advances in Protein Chemistry and Structural Biology*, **86**, 45–93.

Jacob, J. E. M., Vanholme, B., Van Leeuwen, T. and Gheysen, G. (2009). A unique genetic code change in the mitochondrial genome of the parasitic nematode *Radopholus similis*. *BMC Research Notes*, **2**, 192.

Jahn, M., Rogers, M. J. and Söll, D. (1991). Anticodon and acceptor stem nucleotides in $tRNA^{Gln}$ are major recognition elements for *E. coli* glutaminyl-tRNA synthetase. *Nature*, **352**, 258–260.

James, C. M., Ferguson, T. K., Leykam, J. F. and Krzycki, J. A. (2001). The amber codon in the gene encoding the monomethylamine methyltransferase isolated from *Methanosarcina barkeri* is translated as a sense codon. *Journal of Biological Chemistry*, **276**, 34252–34258.

Janouškovec, J., Tikhonenkov, D. V., Burki, F., Howe, A. T., Rohwer, F. L., Mylnikov, A. P. and Keeling, P. J. (2017). A new lineage of eukaryotes illuminates early mitochondrial genome reduction. *Current Biology*, **27**, 3717–3724.e5.

Jukes, T. H. (1985). A change in the genetic code in *Mycoplasma capricolum*. *Journal of Molecular Evolution*, **22**, 361–362.

—, Holmquist, R. and Moise, H. (1975). Amino acid composition of proteins: Selection against the genetic code. *Science*, **189**, 50–51.

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, R. D., Bateman, A. and Petrov, A. I. (2021). Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, **49**, D192–D200.

Kano, A., Ohama, T., Abe, R. and Osawa, S. (1993). Unassigned or nonsense codons in *Micrococcus luteus*. *Journal of Molecular Biology*, **230**, 51–56.

Karpov, S. A., Mikhailov, K. V., Mirzaeva, G. S., Mirabdullaev, I. M., Mamkaeva, K. A., Titova, N. N. and Aleoshin, V. V. (2013). Obligately phagotrophic aphelids turned out to branch with the earliest-diverging fungi. *Protist*, **164**, 195–205.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J. and Iwasaki, S. (1989). The codon CUG is read as serine in an asporogenic yeast *Candida cylindracea*. *Nature*, **341**, 164–166.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. B., Schroeder, D. C., Simpson, A. G. B., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A. and Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biology*, **12**, e1001889.

— and Doolittle, W. F. (1996). A non-canonical genetic code in an early diverging eukaryotic lineage. *The EMBO Journal*, **15**, 2285–2290.

— and — (1997). Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Molecular Biology and Evolution*, **14**, 895–901.

— and Leander, B. S. (2003). Characterisation of a non-canonical genetic code in the oxymonad *Streblomastix strix*. *Journal of Molecular Biology*, **326**, 1337–1349.

Kervestin, S., Frolova, L., Kisselev, L. and Jean-Jean, O. (2001). Stop codon recognition in ciliates: *Euplotes* release factor does not respond to reassigned UGA codon. *EMBO Reports*, **2**, 680–684.

—, Garnier, O. A., Karamyshev, A. L., Ito, K., Nakamura, Y., Meyer, E. and Jean-Jean, O. (2002). Isolation and expression of two genes encoding eukaryotic release factor 1 from *Paramecium tetraurelia*. *Journal of Eukaryotic Microbiology*, **49**, 374–382.

Kim, S., Yi, H., Kim, Y. T. and Lee, H. S. (2022). Engineering translation components for genetic code expansion. *Journal of Molecular Biology*, **434**, 167302.

King, J. L. and Jukes, T. H. (1969). Non-Darwinian evolution. *Science*, **164**, 788–798.

Knight, R. D., Freeland, S. J. and Landweber, L. F. (2001a). Rewiring the keyboard: Evolvability of the genetic code. *Nature Reviews Genetics*, **2**, 49–58.

—, — and — (2001b). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, **2**, research0010.10010.13.

Kobayashi, K., Saito, K., Ishitani, R., Ito, K. and Nureki, O. (2012). Structural basis for translation termination by archaeal RF1 and GTP-bound EF1α complex. *Nucleic Acids Research*, **40**, 9319–9328.

Kollmar, M. and Mühlhausen, S. (2017a). How tRNAs dictate nuclear codon reassignments: Only a few can capture non-cognate codons. *RNA Biology*, **14**, 293–299.

— and Mühlhausen, S. (2017b). Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. *Bioessays*, **39**.

Kolosov, P., Frolova, L., Seit-Nebi, A., Dubovaya, V., Kononenko, A., Oparina, N., Justesen, J., Efimov, A. and Kisselev, L. (2005). Invariant amino acids essential for decoding function of polypeptide release factor eRF1. *Nucleic Acids Research*, **33**, 6418–6425.

Krassowski, T., Coughlan, A. Y., Shen, X.-X., Zhou, X., Kominek, J., Opulente, D. A., Riley, R., Grigoriev, I. V., Maheshwari, N., Shields, D. C., Kurtzman, C. P., Hittinger, C. T., Rokas, A. and Wolfe, K. H. (2018). Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. *Nature Communications*, **9**, 1887.

Kryuchkova, P., Grishin, A., Eliseev, B., Karyagina, A., Frolova, L. and Alkalaeva, E. (2013). Two-step model of stop codon recognition by eukaryotic release factor eRF1. *Nucleic Acids Research*, **41**, 4573–4586.

Krzycki, J. A. (2004). Function of genetically encoded pyrrolysine in corrinoid-dependent methylamine methyltransferases. *Current Opinion in Chemical Biology*, **8**, 484–491.

Kuo, J., Stirling, F., Lau, Y. H., Shulgina, Y., Way, J. C. and Silver, P. A. (2018). Synthetic genome recoding: New genetic codes for new features. *Current Genetics*, **64**, 327–333.

Laforest, M. J., Roewer, I. and Lang, B. F. (1997). Mitochondrial tRNAs in the lower fungus *Spizellomyces punctatus*: tRNA editing and UAG 'stop' codons recognized as leucine. *Nucleic Acids Research*, **25**, 626–632.

Lajoie, M. J., Rovner, A. J., Goodman, D. B., Aerni, H.-R., Haimovich, A. D., Kuznetsov, G., Mercer, J. A., Wang, H. H., Carr, P. A., Mosberg, J. A., Rohland, N., Schultz, P. G., Jacobson, J. M., Rinehart, J., Church, G. M. and Isaacs, F. J. (2013). Genomically recoded organisms expand biological functions. *Science*, **342**, 357–360.

LaRiviere, F. J., Wolfson, A. D. and Uhlenbeck, O. C. (2001). Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. *Science*, **294**, 165–168.

Laurberg, M., Asahara, H., Korostelev, A., Zhu, J., Trakhanov, S. and Noller, H. F. (2008). Structural basis for translation termination on the 70S ribosome. *Nature*, **454**, 852–857.

Lavrov, D. V., Pett, W., Voigt, O., Wörheide, G., Forget, L., Lang, B. F. and Kayal, E. (2013). Mitochondrial DNA of *Clathrina clathrus* (Calcarea, Calcinea): six linear chromosomes, fragmented rRNAs, tRNA editing, and a novel genetic code. *Molecular Biology and Evolution*, **30**, 865–880.

Li, L., Boniecki, M. T., Jaffe, J. D., Imai, B. S., Yau, P. M., Luthey-Schulten, Z. A. and Martinis, S. A. (2011). Naturally occurring aminoacyl-tRNA synthetases editing-domain mutations that cause mistranslation in *Mycoplasma* parasites. *PNAS*, **108**, 9378–9383.

Li, M. and Tzagoloff, A. (1979). Assembly of the mitochondrial membrane system: Sequences of yeast mitochondrial valine and an unusual threonine tRNA gene. *Cell*, **18**, 47–53.

Li, X., Hou, Z., Xu, C., Shi, X., Yang, L., Lewis, L. A. and Zhong, B. (2021). Large phylogenomic data sets reveal deep relationships and trait evolution in chlorophyte green algae. *Genome Biology and Evolution*, **13**, evab101.

Li, Y., Kocot, K. M., Tassia, M. G., Cannon, J. T., Bernt, M. and Halanych, K. M. (2019). Mitogenomics reveals a novel genetic code in hemichordata. *Genome Biology and Evolution*, **11**, 29–40.

Liang, A., Brünen-Nieweler, C., Muramatsu, T., Kuchino, Y., Beier, H. and Heckmann, K. (2001). The ciliate *Euplotes octocarinatus* expresses two polypeptide release factors of the type eRF1. *Gene*, **262**, 161–168.

Lightfield, J., Fram, N. R. and Ely, B. (2011). Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One*, **6**, e17677.

Lind, C., Oliveira, A. and Åqvist, J. (2017). Origin of the omnipotence of eukaryotic release factor 1. *Nature Commununications*, **8**, 1425.

Lind, P. A. and Andersson, D. I. (2008). Whole-genome mutational biases in bacteria. *PNAS*, **105**, 17878–17883.

Ling, J., Daoud, R., Lajoie, M. J., Church, G. M., Söll, D. and Lang, B. F. (2014). Natural reassignment of CUU and CUA sense codons to alanine in *Ashbya* mitochondria. *Nucleic Acids Research*, **42**, 499–508.

Lloyd, R. E. and McGeehan, J. E. (2013). Structural analysis of mitochondrial mutations reveals a role for bigenomic protein interactions in human disease. *PLoS One*, **8**, e69003.

Longstaff, D. G., Larue, R. C., Faust, J. E., Mahapatra, A., Zhang, L., Green-Church, K. B. and Krzycki, J. A. (2007). A natural genetic code expansion cassette enables transmissible biosynthesis and genetic encoding of pyrrolysine. *PNAS*, **104**, 1021–1026.

Lovato, M. A., Chihade, J. W. and Schimmel, P. (2001). Translocation within the acceptor helix of a major tRNA identity determinant. *The EMBO Journal*, **20**, 4846–4853.

Lovett, P. S., Ambulos, N. P., Jr, Mulbry, W., Noguchi, N. and Rogers, E. J. (1991). UGA can be decoded as tryptophan at low efficiency in *Bacillus subtilis*. *Journal of Bacteriology*, **173**, 1810–1812.

Lozupone, C. A., Knight, R. D. and Landweber, L. F. (2001). The molecular basis of nuclear genetic code change in ciliates. *Current Biology*, **11**, 65–74.

Ma, N. J., Hemez, C. F., Barber, K. W., Rinehart, J. and Isaacs, F. J. (2018). Organisms with alternative genetic codes resolve unassigned codons via mistranslation and ribosomal rescue. *eLife*, **7**.

Mandal, D., Köhrer, C., Su, D., Russell, S. P., Krivos, K., Castleberry, C. M., Blum, P., Limbach, P. A., Söll, D. and RajBhandary, U. L. (2010). Agmatidine, a modified cytidine in the anticodon of archaeal $tRNA^{Ile}$, base pairs with adenosine but not with guanosine. *PNAS*, **107**, 2872–2877.

Martin, R. P., Sibler, A. P., Gehrke, C. W., Kuo, K., Edmonds, C. G., McCloskey, J. A. and Dirheimer, G. (1990). 5-[[(carboxymethyl)amino]methyl]uridine is found in the anticodon of yeast mitochondrial tRNAs recognizing two-codon families ending in a purine. *Biochemistry*, **29**, 956–959.

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

Masuda, I., Matsuzaki, M. and Kita, K. (2010). Extensive frameshift at all AGG and CCC codons in the mitochondrial cytochrome c oxidase subunit 1 gene of *Perkinsus marinus* (Alveolata; Dinoflagellata). *Nucleic Acids Research*, **38**, 6186–6194.

Matsugi, J., Murao, K. and Ishikura, H. (1998). Effect of *B. subtilis* tRNA(Trp) on readthrough rate at an opal UGA codon. *Journal of Biochemistry*, **123**, 853–858.

Matsuyama, S., Ueda, T., Crain, P. F., McCloskey, J. A. and Watanabe, K. (1998). A novel wobble rule found in starfish mitochondria. presence of 7-methylguanosine at the anticodon wobble position expands decoding capability of tRNA. *Journal of Biological Chemistry*, **273**, 3363–3368.

McCutcheon, J. P., McDonald, B. R. and Moran, N. A. (2009). Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genetics*, **5**, e1000565.

— and Moran, N. A. (2010). Functional convergence in reduced genomes of bacterial symbionts spanning 200 my of evolution. *Genome Biology and Evolution*, **2**, 708–718.

— and — (2011). Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, **10**, 13–26.

Meinnel, T., Mechulam, Y., Lazennec, C., Blanquet, S. and Fayat, G. (1993). Critical role of the acceptor stem of tRNAs Met in their aminoacylation by *Escherichia coli* methionyl-tRNA synthetase. *Journal of Molecular Biology*, **229**, 26–36.

Meyer, F., Schmidt, H. J., Plümper, E., Hasilik, A., Mersmann, G., Meyer, H. E., Engström, A. and Heckmann, K. (1991). UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*. *PNAS*, **88**, 3758–3761.

Miller, J. H., Coulondre, C. and Farabaugh, P. J. (1978). Correlation of nonsense sites in the laci gene with specific codons in the nucleotide sequence. *Nature*, **274**, 770–775.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D. and Alex, B. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, **49**, D412–D419.

Mühlhausen, S., Findeisen, P., Plessmann, U., Urlaub, H. and Kollmar, M. (2016). A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. *Genome Research*, **26**, 945–955.

— and Kollmar, M. (2014a). Molecular phylogeny of sequenced Saccharomycetes reveals polyphyly of the alternative yeast codon usage. *Genome Biology and Evolution*, **6**, 3222–3237.

— and Kollmar, M. (2014b). Predicting the fungal CUG codon translation with bagheera. *BMC Genomics*, **15**, 411.

—, Schmitt, H. D., Pan, K.-T., Plessmann, U., Urlaub, H., Hurst, L. D. and Kollmar, M. (2018). Endogenous stochastic decoding of the CUG codon by competing Ser- and Leu-tRNAs in *Ascoidea asiatica*. *Current Biology*, **28**, 2046–2057.e5.

Mukai, T., Hayashi, A., Iraha, F., Sato, A., Ohtake, K., Yokoyama, S. and Sakamoto, K. (2010). Codon reassignment in the *Escherichia coli* genetic code. *Nucleic Acids Research*, **38**, 8188–8195.

—, Yamaguchi, A., Ohtake, K., Takahashi, M., Hayashi, A., Iraha, F., Kira, S., Yanagisawa, T., Yokoyama, S., Hoshi, H., Kobayashi, T. and Sakamoto, K. (2015). Reassignment of a rare sense codon to a non-canonical amino acid in *Escherichia coli*. *Nucleic Acids Research*, **43**, 8111–8122.

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

Muto, A. and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *PNAS*, **84**, 166–169.

Nakanishi, K., Ogiso, Y., Nakama, T., Fukai, S. and Nureki, O. (2005). Structural basis for anticodon recognition by methionyl-tRNA synthetase. *Nature Structural and Molecular Biology*, **12**, 931–932.

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Nedelcu, A. M., Lee, R. W., Lemieux, C., Gray, M. W. and Burger, G. (2000). The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. *Genome Research*, **10**, 819–831.

Nishimura, Y., Shiratori, T., Ishida, K.-I., Hashimoto, T., Ohkuma, M. and Inagaki, Y. (2019). Horizontally-acquired genetic elements in the mitochondrial genome of a centrohelid *Marophrys* sp. SRT127. *Scientific Reports*, **9**, 4850.

Normanly, J., Kleina, L. G., Masson, J. M., Abelson, J. and Miller, J. H. (1990). Construction of *Escherichia coli* amber suppressor tRNA genes. III. Determination of tRNA specificity. *Journal of Molecular Biology*, **213**, 719–726.

Noutahi, E., Calderon, V., Blanchette, M., El-Mabrouk, N. and Lang, B. F. (2019). Rapid genetic code evolution in green algal mitochondrial genomes. *Molecular Biology and Evolution*, **36**, 766–783.

—, —, —, Lang, B. F. and El-Mabrouk, N. (2017). CoreTracker: Accurate codon reassignment prediction, applied to mitochondrial genomes. *Bioinformatics*, **33**, 3331–3339.

Nyerges, A., Vinke, S., Flynn, R., Owen, S. V., Rand, E. A., Budnik, B., Keen, E., Narasimhan, K., Marchand, J. A., Baas-Thomas, M., Liu, M., Chen, K., Chiappino-Pepe, A., Hu, F., Baym, M. and Church, G. M. (2022). Swapped genetic code blocks viral infections and gene transfer. *bioRxiv*, p. 2022.07.08.499367.

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

Ogle, J. M., Brodersen, D. E., Clemons, W. M., Jr, Tarry, M. J., Carter, A. P. and Ramakrishnan, V. (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, **292**, 897–902.

Ohama, T., Muto, A. and Osawa, S. (1990a). Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content. *Nucleic Acids Research*, **18**, 1565–1569.

—, Osawa, S., Watanabe, K. and Jukes, T. H. (1990b). Evolution of the mitochondrial genetic code. IV. AAA as an asparagine codon in some animal mitochondria. *Journal of Molecular Evolution*, **30**, 329–332.

Osawa, S. and Jukes, T. H. (1989). Codon reassignment (codon capture) in evolution. *Journal of Molecular Evolution*, **28**, 271–278.

—, Ohama, T., Yamao, F., Muto, A., Jukes, T. H., Ozeki, H. and Umesono, K. (1988). Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. *PNAS*, **85**, 1124–1128.

Oudot-Le Secq, M.-P. and Green, B. R. (2011). Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Gene*, **476**, 20–26.

Pallanck, L., Pak, M. and Schulman, L. H. (1994). trna discrimination in aminoacylation. In *tRNA: Structure, Biosynthesis, and Function*, ASM Press, pp. 251–292.

Pánek, T., Žihala, D., Sokol, M., Derelle, R., Klimeš, V., Hradilová, M., Zadrobílková, E., Susko, E., Roger, A. J., Čepička, I. and Eliáš, M. (2017). Nuclear genetic codes with a different meaning of the UAG and the UAA codon. *BMC Biol.*, **15**, 8.

Paquin, B., Laforest, M. J., Forget, L., Roewer, I., Wang, Z., Longcore, J. and Lang, B. F. (1997). The fungal mitochondrial genome project: Evolution of fungal mitochondrial genomes and their gene expression. *Current Genetics*, **31**, 380–395.

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J. and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, **38**, 1079–1086.

—, Imelfort, M., Skennerton, C. T., Hugenholtz, P. and Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, **25**, 1043–1055.

—, Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P. and Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, **2**, 1533–1542.

Perona, J. J. and Hadd, A. (2012). Structural diversity and protein engineering of the aminoacyl-tRNA synthetases. *Biochemistry*, **51**, 8705–8729.

—, Swanson, R. N., Rould, M. A., Steitz, T. A. and Söll, D. (1989). Structural basis for misaminoacylation by mutant *E. coli* glutaminyl-tRNA synthetase enzymes. *Science*, **246**, 1152–1154.

Perret, V., Garcia, A., Grosjean, H., Ebel, J. P., Florentz, C. and Giegé, R. (1990). Relaxation of a transfer RNA specificity by removal of modified nucleotides. *Nature*, **344**, 787–789.

Perseke, M., Hetmank, J., Bernt, M., Stadler, P. F., Schlegel, M. and Bernhard, D. (2011). The enigmatic mitochondrial genome of *Rhabdopleura compacta* (Pterobranchia) reveals insights into selection of an efficient tRNA system and supports monophyly of Ambulacraria. *BMC Evolutionary Biology*, **11**, 134.

Petropoulos, A. D., McDonald, M. E., Green, R. and Zaher, H. S. (2014). Distinct roles for release factor 1 and release factor 2 in translational quality control. *Journal of Biological Chemistry*, **289**, 17589–17596.

Prat, L., Heinemann, I. U., Aerni, H. R., Rinehart, J., O'Donoghue, P. and Söll, D. (2012). Carbon source-dependent expansion of the genetic code in bacteria. *PNAS*, **109**, 21070–21075.

Rafels-Ybern, À., Torres, A. G., Camacho, N., Herencia-Ropero, A., Roura Frigolé, H., Wulff, T. F., Raboteg, M., Bordons, A., Grau-Bove, X., Ruiz-Trillo, I. and Ribas de Pouplana, L. (2018). The expansion of inosine at the wobble position of tRNAs, and its role in the evolution of proteomes. *Molecular Biology and Evolution*, **36**, 650–662.

Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Göker, M., Salamov, A. A., Wisecaver, J. H., Long, T. M., Calvey, C. H., Aerts, A. L., Barry, K. W., Choi, C., Clum, A., Coughlan, A. Y., Deshpande, S., Douglass, A. P., Hanson, S. J., Klenk, H.-P., LaButti, K. M., Lapidus, A., Lindquist, E. A., Lipzen, A. M., Meier-Kolthoff, J. P., Ohm, R. A., Otillar, R. P., Pangilinan, J. L., Peng, Y., Rokas, A., Rosa, C. A., Scheuner, C., Sibirny, A. A., Slot, J. C., Stielow, J. B., Sun, H., Kurtzman, C. P., Blackwell, M., Grigoriev, I. V. and Jeffries, T. W. (2016). Comparative genomics of biotechnologically important yeasts. *PNAS*, **113**, 9882–9887.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A. E., Malfatti, S., Swan, B. K., Gies, E. a., Dodsworth, J. a., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. a., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P. and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.

Roberts, J. W. and Carbon, J. (1974). Molecular mechanism for missense suppression in *E. coli*. *Nature*, **250**, 412–414.

Rodnina, M. V. (2018). Translation in prokaryotes. *Cold Spring Harbor Perspectives in Biology*, **10** (9), a032664.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rould, M. A., Perona, J. J., Söll, D. and Steitz, T. A. (1989). Structure of *E. coli* glutaminyl-tRNA synthetase complexed with $tRNA^{Gln}$ and ATP at 2.8 å resolution. *Science*, **246**, 1135–1142.

Roy, B., Leszyk, J. D., Mangus, D. A. and Jacobson, A. (2015). Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. *PNAS*, **112**, 3038–3043.

Sakamoto, K., Kawai, G., Niimi, T., Satoh, T., Sekine, M., Yamaizumi, Z., Nishimura, S., Miyazawa, T. and Yokoyama, S. (1993). A modified uridine in the first position of the anticodon of a minor species of arginine tRNA, the argu gene product, from *Escherichia coli*. *European Journal of Biochemistry*, **216**, 369–375.

Sakurai, M., Ohtsuki, T., Suzuki, T. and Watanabe, K. (2005). Unusual usage of wobble modifications in mitochondrial tRNAs of the nematode *Ascaris suum*. *FEBS Letters*, **579**, 2767–2772.

Salem, H., Bauer, E., Kirsch, R., Berasategui, A., Cripps, M., Weiss, B., Koga, R., Fukumori, K., Vogel, H., Fukatsu, T. and Kaltenpoth, M. (2017). Drastic genome reduction in an herbivore's pectinolytic symbiont. *Cell*, **171**, 1520–1531.e13.

Sánchez-Silva, R., Villalobo, E., Morin, L. and Torres, A. (2003). A new noncanonical nuclear genetic code: Translation of UAA into glutamate. *Current Biology*, **13**, 442–447.

Santos, M. A., Cheesman, C., Costa, V., Moradas-Ferreira, P. and Tuite, M. F. (1999). Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida* spp. *Molecular Microbiology*, **31**, 937–947.

Sato, Y., Hirayama, M., Morimoto, K., Yamamoto, N., Okuyama, S. and Hori, K. (2011). High mannose-binding lectin with preference for the cluster of α1–2-mannose from the green alga *Boodlea coacta* is a potent entry inhibitor of HIV-1 and influenza viruses. *Journal of Biological Chemistry*, **286**, 19446–19458.

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

Schmidt, C. A. and Matera, A. G. (2020). tRNA introns: Presence, processing, and purpose. *WIREs RNA*, **11**, e1583.

Schneider, S. U., Leible, M. B. and Yang, X. P. (1989). Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage. *Molecular Genetics and Genomics*, **218**, 445–452.

Schulman, L. H. and Pelka, H. (1983). Anticodon loop size and sequence requirements for recognition of formylmethionine tRNA by methionyl-tRNA synthetase. *PNAS*, **80**, 6755–6759.

— and — (1988). Anticodon switching changes the identity of methionine and valine transfer RNAs. *Science*, **242**, 765–768.

Schultz, D. W. and Yarus, M. (1994). Transfer RNA mutation and the malleability of the genetic code. *Journal of Molecular Biology*, **235**, 1377–1380.

— and — (1996). On malleability in the genetic code. *Journal of Molecular Evolution*, **42**, 597–601.

Seah, B. K. B., Singh, A. and Swart, E. C. (2022). Karyorelict ciliates use an ambiguous genetic code with context-dependent stop/sense codons. *Peer Community Journal*, **2**, e42.

Seilhamer, J. J. and Cummings, D. J. (1982). Altered genetic code in *Paramecium* mitochondria: Possible evolutionary trends. *Molecular and General Genetics*, **187**, 236–239.

Seit-Nebi, A., Frolova, L. and Kisselev, L. (2002). Conversion of omnipotent translation termination factor eRF1 into ciliate-like UGA-only unipotent eRF1. *EMBO Reports*, **3**, 881–886.

Sekimoto, S., Beakes, G. W., Gachon, C. M. M., Muller, D. G., Kupper, F. C. and Honda, D. (2008). The development, ultrastructural cytology, and molecular phylogeny of the basal oomycete *Eurychasma dicksonii*, infecting the filamentous phaeophyte algae *Ectocarpus siliculosus* and *Pylaiella littoralis*. *Protist*, **159**, 299–318.

Sengupta, S. and Higgs, P. G. (2005). A unified model of codon reassignment in alternative genetic codes. *Genetics*, **170**, 831–840.

—, Yang, X. and Higgs, P. G. (2007). The mechanisms of codon reassignments in mitochondrial genetic codes. *Journal of Molecular Evolution*, **64**, 662–688.

Shackelton, L. A. and Holmes, E. C. (2008). The role of alternative genetic codes in viral evolution and emergence. *Journal of Theoretical Biology*, **254**, 128–134.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A. B., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R.-I., Čadež, N., Libkind, D., Rosa, C. A., DeVirgilio, J., Hulfachor, A. B., Groenewald, M., Kurtzman, C. P., Hittinger, C. T. and Rokas, A. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell*, **175**, 1533–1545.e20.

Sherman, J., Rogers, M. and Soll, D. (1994). Recognition in the glutamine trna system: From structure to function. In *tRNA: Structure, Biosynthesis, and Function*, ASM Press, pp. 395–409.

Sherman, J. M., Rogers, M. J. and Söll, D. (1992). Competition of aminoacyl-tRNA synthetases for tRNA ensures the accuracy of aminoacylation. *Nucleic Acids Research*, **20**, 1547–1552.

Shiba, K., Schimmel, P., Motegi, H. and Noda, T. (1994). Human glycyl-tRNA synthetase. wide divergence of primary structure from bacterial counterpart and species-specific aminoacylation. *Journal of Biological Chemistry*, **269**, 30049–30055.

Shulgina, Y. and Eddy, S. R. (2021). A computational screen for alternative genetic codes in over 250,000 genomes. *eLife*, **10**, e71402.

Sibler, A. P., Dirheimer, G. and Martin, R. P. (1986). Codon reading patterns in *Saccharomyces cerevisiae* mitochondria based on sequences of mitochondrial tRNAs. *FEBS Letters*, **194**, 131–138.

Simon, C. and Daniel, R. (2011). Metagenomic analyses: Past and future trends. *Applied and Environmental Microbiology*, **77**, 1153–1161.

Soma, A., Kumagai, R., Nishikawa, K. and Himeno, H. (1996). The anticodon loop is a major identity determinant of *Saccharomyces cerevisiae* tRNALeu. *Journal of Molecular Biology*, **263**, 707–714.

Spanjaard, R. A., Chen, K., Walker, J. R. and van Duin, J. (1990). Frameshift suppression at tandem AGA and AGG codons by cloned tRNA genes: Assigning a codon to *argU* tRNA and T4 $tRNA^{Arg}$. *Nucleic Acids Research*, **18**, 5031–5036.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. and Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, **26**, 148–153.

Spruyt, N., Delarbre, C., Gachelin, G. and Laudet, V. (1998). Complete sequence of the amphioxus *Branchiostoma lanceolatum* mitochondrial genome: Relations to vertebrates. *Nucleic Acids Research*, **26**, 3279–3285.

Srinivasan, G., James, C. M. and Krzycki, J. A. (2002). Pyrrolysine encoded by UAG in Archaea: Charging of a UAG-decoding specialized tRNA. *Science*, **296**, 1459–1462.

Steenwyk, J. L., Opulente, D. A., Kominek, J., Shen, X.-X., Zhou, X., Labella, A. L., Bradley, N. P., Eichman, B. F., Čadež, N., Libkind, D., DeVirgilio, J., Hulfachor, A. B., Kurtzman, C. P., Hittinger, C. T. and Rokas, A. (2019). Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. *PLoS Biology*, **17**, e3000255.

Stehlin, C., Burke, B., Yang, F., Liu, H., Shiba, K. and Musier-Forsyth, K. (1998). Species-specific differences in the operational RNA code for aminoacylation of $tRNA^{Pro}$. *Biochemistry*, **37**, 8605–8613.

Sueoka, N. (1961). Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *PNAS*, **47**, 1141–1149.

— (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *PNAS*, **48**, 582–592.

Sun, J., Evans, P. N., Gagen, E. J., Woodcroft, B. J., Hedlund, B. P., Woyke, T., Hugenholtz, P. and Rinke, C. (2021). Recoding of stop codons expands the metabolic potential of two novel asgardarchaeota lineages. *ISME Communications*, **1**, 1–14.

Sunagawa, S., Karsenti, E., Bowler, C. and Bork, P. (2015). Computational eco-systems biology in Tara Oceans: Translating data into knowledge. *Molecular Systems Biology*, **11**, 809.

Suzuki, T., Miyauchi, K., Suzuki, T., Yokobori, S.-I., Shigi, N., Kondow, A., Takeuchi, N., Yamagishi, A. and Watanabe, K. (2011). Taurine-containing uridine modifications in tRNA anticodons are required to decipher non-universal genetic codes in ascidian mitochondria. *Journal of Biological Chemistry*, **286** (41), 35494–35498.

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

Swart, E. C., Bracht, J. R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J. S., Goldman, A. D., Nowacki, M., Schotanus, K., Jung, S., Fulton, R. S., Ly, A., McGrath, S., Haub, K., Wiggins, J. L., Storton, D., Matese, J. C., Parsons, L., Chang, W.-J., Bowen, M. S., Stover, N. A., Jones, T. A., Eddy, S. R., Herrick, G. A., Doak, T. G., Wilson, R. K., Mardis, E. R. and Landweber, L. F. (2013). The *Oxytricha trifallax* macronuclear genome: A complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biology*, **11**, e1001473.

—, Serra, V., Petroni, G. and Nowacki, M. (2016). Genetic codes with no dedicated stop codon: Context-dependent translation termination. *Cell*, **166**, 691–702.

Sylvers, L. A., Rogers, K. C., Shimizu, M., Ohtsuka, E. and Söll, D. (1993). A 2-thiouridine derivative in $tRNA^{Glu}$ is a positive determinant for aminoacylation by *Escherichia coli* glutamyl-tRNA synthetase. *Biochemistry*, **32**, 3836–3841.

Szweykowska-Kulinska, Z., Senger, B., Keith, G., Fasiolo, F. and Grosjean, H. (1994). Intron-dependent formation of pseudouridines in the anticodon of *Saccharomyces cerevisiae* minor tRNA(Ile). *The EMBO Journal*, **13**, 4636–4644.

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H. and Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature*, **387**, 700–702.

Takemoto, C., Spremulli, L. L., Benkowski, L. A., Ueda, T., Yokogawa, T. and Watanabe, K. (2009). Unconventional decoding of the AUA codon as methionine by mitochondrial tRNA met with the anticodon f^5CAU as revealed with a mitochondrial *in vitro* translation system. *Nucleic Acids Research*, **37**, 1616–1627.

Taniguchi, T., Miyauchi, K., Nakane, D., Miyata, M., Muto, A., Nishimura, S. and Suzuki, T. (2013). Decoding system for the AUA codon by $tRNA^{Ile}$ with the UAU anticodon in *Mycoplasma mobile*. *Nucleic Acids Research*, **41**, 2621–2631.

Taylor, D. J., Ballinger, M. J., Bowman, S. M. and Bruenn, J. A. (2013). Virus-host co-evolution under a modified nuclear genetic code. *PeerJ*, p. e50.

Telford, M. J., Herniou, E. A., Russell, R. B. and Littlewood, D. T. (2000). Changes in mitochondrial genetic codes as phylogenetic characters: Two examples from the flatworms. *PNAS*, **97**, 11359–11364.

Temperley, R., Richter, R., Dennerlein, S., Lightowlers, R. N. and Chrzanowska-Lightowlers, Z. M. (2010). Hungry codons promote frameshifting in human mitochondrial ribosomes. *Science*, **327**, 301.

Tomita, K., Ueda, T., Ishiwa, S., Crain, P. F., McCloskey, J. A. and Watanabe, K. (1999a). Codon reading patterns in *Drosophila melanogaster* mitochondria based on their tRNA sequences: A unique wobble rule in animal mitochondria. *Nucleic Acids Research*, **27**, 4291–4297.

—, — and Watanabe, K. (1998). 7-methylguanosine at the anticodon wobble position of squid mitochondrial $tRNA^{Ser}$ GCU: Molecular basis for assignment of AGA/AGG codons as serine in invertebrate mitochondria. *Biochimica et Biophysica Acta*, **1399**, 78–82.

—, — and — (1999b). The presence of pseudouridine in the anticodon alters the genetic code: A possible mechanism for assignment of the AAA lysine codon as asparagine in echinoderm mitochondria. *Nucleic Acids Research*, **27**, 1683–1689.

Tran, T. T. T., Belahbib, H., Bonnefoy, V. and Talla, E. (2015). A comprehensive tRNA genomic survey unravels the evolutionary history of tRNA arrays in prokaryotes. *Genome Biology and Evolution*, **8**, 282–295.

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.

Turmel, M., Lemieux, C., Burger, G., Lang, B. F., Otis, C., Plante, I. and Gray, M. W. (1999). The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*. Two radically different evolutionary patterns within green algae. *Plant Cell*, **11**, 1717–1730.

—, Otis, C. and Lemieux, C. (2010). A deviant genetic code in the reduced mitochondrial genome of the picoplanktonic green alga *Pycnococcus provasolii*. *Journal of Molecular Evolution*, **70**, 203–214.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

Uhlenbeck, O. C. and Schrader, J. M. (2018). Evolutionary tuning impacts the design of bacterial tRNAs for the incorporation of unnatural amino acids by ribosomes. *Current Opinion in Chemical Biology*, **46**, 138–145.

UniProt Consortium (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, **49** (D1), D480–D489.

Vargas-Rodriguez, O., Badran, A. H., Hoffman, K. S., Chen, M., Crnković, A., Ding, Y., Krieger, J. R., Westhof, E., Söll, D. and Melnikov, S. (2021). Bacterial translation machinery for deliberate mistranslation of the genetic code. *PNAS*, **118**.

Varshney, U., Lee, C. P. and RajBhandary, U. L. (1991). Direct analysis of aminoacylation levels of tRNAs in vivo. application to studying recognition of *Escherichia coli* initiator tRNA mutants by glutaminyl-tRNA synthetase. *Journal of Biological Chemistry*, **266**, 24712–24718.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas,

M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H. and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Volokhov, D. V., Neverov, A. A., George, J., Kong, H., Liu, S. X., Anderson, C., Davidson, M. K. and Chizhikov, V. (2007). Genetic analysis of housekeeping genes of members of the genus *Acholeplasma*: Phylogeny and complementary molecular markers to the 16S rRNA gene. *Molecular Phylogenetics and Evolution*, **44**, 699–710.

Watanabe, K. and Yokobori, S.-I. (2011). tRNA modification and genetic code variations in animal mitochondria. *Journal of Nucleic Acids*, **2011**, 623095.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V. and Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, **35**, 543–548.

Weixlbaumer, A., Jin, H., Neubauer, C., Voorhees, R. M., Petry, S., Kelley, A. C. and Ramakrishnan, V. (2008). Insights into translational termination from the structure of RF2 bound to the ribosome. *Science*, **322**, 953–956.

Wideman, J. G., Monier, A., Rodríguez-Martínez, R., Leonard, G., Cook, E., Poirier, C., Maguire, F., Milner, D. S., Irwin, N. A. T., Moore, K., Santoro, A. E., Keeling, P. J., Worden, A. Z. and Richards, T. A. (2020). Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. *Nature Microbiology*, **5**, 154–165.

Wolff, P., Villette, C., Zumsteg, J., Heintz, D., Antoine, L., Chane-Woon-Ming, B., Droogmans, L., Grosjean, H. and Westhof, E. (2020). Comparative patterns of modified nucleotides in individual tRNA species from a mesophilic and two thermophilic archaea. *RNA*, **26**, 1957–1975.

Wolstenholme, D. R. (1992). Animal mitochondrial DNA: Structure and evolution. *International Review of Cytology*, **141**, 173–216.

Wong, L. E., Li, Y., Pillay, S., Frolova, L. and Pervushin, K. (2012). Selectivity of stop codon recognition in translation termination is modulated by multiple conformations of GTS loop in eRF1. *Nucleic Acids Research*, **40**, 5751–5765.

Yamao, F., Iwagami, S., Azumi, Y., Muto, A., Osawa, S., Fujita, N. and Ishihama, A. (1988). Evolutionary dynamics of tryptophan tRNAs in *Mycoplasma capricolum*. *Molecular and General Genetics*, **212**, 364–369.

—, Muto, A., Kawauchi, Y., Iwami, M., Iwagami, S., Azumi, Y. and Osawa, S. (1985). UGA is read as tryptophan in *Mycoplasma capricolum*. *PNAS*, **82**, 2306–2309.

Yan, Y., Maurer-Alcalá, X. X., Knight, R., Kosakovsky Pond, S. L. and Katz, L. A. (2019). Single-cell transcriptomics reveal a correlation between genome architecture and gene family evolution in ciliates. *MBio*, **10**, e02524–19.

Yokobori, S., Suzuki, T. and Watanabe, K. (2001). Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. *Journal of Molecular Evolution*, **53**, 314–326.

Yokobori, S.-I., Kitamura, A., Grosjean, H. and Bessho, Y. (2013). Life without tRNAArg–adenosine deaminase TadA: Evolutionary consequences of decoding the four CGN codons as arginine in mycoplasmas and other mollicutes. *Nucleic Acids Research*, **41**, 6531–6543.

Yokogawa, T., Suzuki, T., Ueda, T., Mori, M., Ohama, T., Kuchino, Y., Yoshinari, S., Motoki, I., Nishikawa, K. and Osawa, S. (1992). Serine tRNA complementary to the nonuniversal serine codon CUG in *Candida cylindracea*: Evolutionary implications. *PNAS*, **89**, 7408–7411.

Yokoyama, S., Muramatsu, T., Horie, N., Nishikawa, K., Matsuda, A., Ueda, T., Yamaizumi, Z., Kuchino, Y., Nishimura, S. and Miyazawa, T. (1989). A novel lysine-substituted nucleoside in the first position of the anticodon of minor isoleucine tRNA from *Escherichia coli*. *Pure and Applied Chemistry*, **61**, 573–576.

Young, M., Artsatbanov, V., Beller, H. R., Chandra, G., Chater, K. F., Dover, L. G., Goh, E.-B., Kahan, T., Kaprelyants, A. S., Kyrpides, N., Lapidus, A., Lowry, S. R., Lykidis, A., Mahillon, J., Markowitz, V., Mavromatis, K., Mukamolova, G. V., Oren, A., Rokem, J. S., Smith, M. C. M., Young, D. I. and Greenblatt, C. L. (2010). Genome sequence of the Fleming strain of *Micrococcus luteus*, a simple free-living actinobacterium. *Journal of Bacteriology*, **192**, 841–860.

Záhonová, K., Kostygov, A. Y., Ševčíková, T., Yurchenko, V. and Eliáš, M. (2016). An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Current Biology*, **26**, 2364–2369.

—, Lax, G., Sinha, S. D., Leonard, G., Richards, T. A., Lukeš, J. and Wideman, J. G. (2021). Single-cell genomics unveils a canonical origin of the diverse mitochondrial genomes of euglenozoans. *BMC Biology*, **19**, 103.

Zhang, H., Gong, X., Zhao, Q., Mukai, T., Vargas-Rodriguez, O., Zhang, H., Zhang, Y., Wassel, P., Amikura, K., Maupin-Furlow, J., Ren, Y., Xu, X., Wolf, Y. I., Makarova, K. S., Koonin, E. V., Shen, Y., Söll, D. and Fu, X. (2022). The tRNA discriminator base defines the mutual orthogonality of two distinct pyrrolysyl-tRNA synthetase/tRNAPyl pairs in the same organism. *Nucleic Acids Research*, **50**, 4601–4615.

Zhang, Y., Baranov, P. V., Atkins, J. F. and Gladyshev, V. N. (2005). Pyrrolysine and selenocysteine use dissimilar decoding strategies. *Journal of Biological Chemistry*, **280**, 20740–20751.

Žihala, D. and Eliáš, M. (2019). Evolution and unprecedented variants of the mitochondrial genetic code in a lineage of green algae. *Genome Biology and Evolution*, **11**, 2992–3007.

—, Salamonová, J. and Eliáš, M. (2020). Evolution of the genetic code in the mitochondria of Labyrinthulea (Stramenopiles). *Mol. Phylogenet. Evol.*, **152**, 106908.

Appendix A

Appendix to Chapter 1

All known alternative genetic codes

Phylum	Phylogenetic extent	Reassignment	MSA	tRNA	Activity	Direct	References
Firmicutes	Entomoplasmatales & Mycoplasmatales	UGA Stop→Trp	✓	✓	✓	✓	[1,2,3,4]
Firmicutes	*Enterosoma* spp	AGG Arg→Met	✓	✓			[5]
Firmicutes	*Peptacetobacter* spp	CGG Arg→Gln	✓	✓			[5]
Firmicutes	*Onthovivens* spp	CGG Arg→Trp	✓	✓			[5]
Firmicutes	*Anaerococcus*	CGG Arg→Trp	✓	✓			[5]
Proteobacteria	*Hodgkinia cicadicola*	UGA Stop→Trp	✓	✓		✓	[6]
Proteobacteria	*Nasuia deltocephalinicola*	UGA Stop→Trp	✓	✓			[7]
Proteobacteria	*Zinderia insecticola*	UGA Stop→Trp	✓				[8]
Proteobacteria	*Stammera capleta*	UGA Stop→Trp	✓				[9]
CPR	Gracilibacteria	UGA Stop→Gly	✓	✓	✓	✓	[10,11]
CPR	Absconditabacteria	UGA Stop→Gly	✓	✓	✓		[12]
		CGR Arg→Trp	✓	✓			[5]

Table A.1: All bacterial clades known to use alternative genetic codes with the degree of support for the codon reassignment and references. Categories for support are: MSA– multiple sequence alignment or profile alignment showing codon at conserved positions for the new meaning; tRNA– presence of tRNA / release factor gene consistent with the new translation; Activity– experiments showing tRNA / aaRS activity *in vitro* or in a heterologous system; and Direct– direct protein evidence, either protein sequencing or mass spectrometry of proteins from the reassigned organism. [1] Yamao *et al.* (1985), [2] Ivanova *et al.* (2014), [3] Citti *et al.* (1992), [4] Futo *et al.* (1995), [5] Shulgina and Eddy (2021), [6] McCutcheon *et al.* (2009), [7] Bennett and Moran (2013), [8] McCutcheon and Moran (2010), [9] Salem *et al.* (2017), [10] Rinke *et al.* (2013), [11] Hanke *et al.* (2014), [12] Campbell *et al.* (2013)

Phylum	Phylogenetic extent	Reassignment	UAGs	tRNA	PylRS	Activity	Direct	Ref
Halobacteriota	Methanosarcinaceae	UAG Stop→Pyl	✓	✓	✓	✓	✓	[1,2,3]
Halobacteriota	Methermicoccaceae	UAG Stop→Pyl/Stop		✓	✓			[3,4]
Halobacteriota	Methanonatronarchaeia	UAG Stop→Pyl	✓	✓	✓	✓		[3,5]
Thermoplasmatota	Methanomassiliicoccales	UAG Stop→Pyl	✓	✓	✓	✓		[3,5,6]
Thermoproteota	Methanomethylicia	UAG Stop→Pyl			✓			[3]
Thermoproteota	Bathyarchaeota JdFR-11	UAG Stop→Pyl	✓	✓	✓			[3,7]
Thermoproteota	Korarchaeota KS3-KO24	UAG Stop→Pyl		✓	✓			[3]
Thermoproteota	Nitrososphaeria archaeon	UAG Stop→Pyl			✓			[3]
Asgardarchaeota	Borrarchaeaceae	UAG Stop→Pyl/Stop	✓	✓	✓			[3,7]
Hydrothermarchaeota	*Hydrothermarchaeum*	UAG Stop→Pyl		✓	✓			[3,7]
(Stygia)	Persephonarchaea MSBL1	UAG Stop→Pyl/Stop	✓	✓	✓			[4]

Table A.2: All archaeal clades known to use alternative genetic codes, which are restricted to reassignments of UAG to pyrrolysine. For each clade, I indicate the degree of support for the reassignment and references. Some UAG codons are marked as both pyrrolysine and stop due to high UAG usage at the end of reading frames. Categories for support are: UAGs– in-frame UAG codons at known pyrrolysine positions; tRNA– presence of tRNA / release factor gene consistent with the new translation; PylRS– presence of pyrrolysine-tRNA synthetase gene; Activity– experiments showing tRNA / PylRS activity *in vitro* or in a heterologous system; and Direct– direct protein evidence, either protein sequencing or mass spectrometry of proteins from the reassigned organism. [1] James *et al.* (2001), [2] Hao *et al.* (2002), [3] Guo *et al.* (2022), [4] Guan *et al.* (2017), [5] Zhang *et al.* (2022), [6] Borrel *et al.* (2014), [7] Sun *et al.* (2021)

Table A.3: All eukaryotic clades known to use alternative genetic codes with the degree of support for the codon reassignment and references. Categories for support are: MSA– multiple sequence alignment or profile alignment showing codon at conserved positions for the new meaning; tRNA– presence of tRNA / release factor gene consistent with the new translation; Activity– experiments showing tRNA / aaRS activity *in vitro* or in a heterologous system; and Direct– direct protein evidence, either protein sequencing or mass spectrometry of proteins from the reassigned organism. [1] Kawaguchi *et al.* (1989), [2] Yokogawa *et al.* (1992), [3] Mühlhausen *et al.* (2016), [4] Mühlhausen *et al.* (2018), [5] Krassowski *et al.* (2018), [6] Shulgina and Eddy (2021), [7] Karpov *et al.* (2013), [8] Záhonová *et al.* (2016), [9] Keeling and Leander (2003), [10] De Koning *et al.* (2008), [11] Keeling and Doolittle (1996), [12] Keeling and Doolittle (1997), [13] Pánek *et al.* (2017), [14] Schneider *et al.* (1989), [15] Cocquyt *et al.* (2010), [16] Sato *et al.* (2011), [17] Li *et al.* (2021), [18] Bachvaroff (2019), [19] Caron and Meyer (1985), [20] Horowitz and Gorovsky (1985), [21] Hanyu *et al.* (1986), [22] Yan *et al.* (2019), [23] Helftenbein (1985), [24] Swart *et al.* (2013), [25] Seah *et al.* (2022), [26] Swart *et al.* (2016), [27] Meyer *et al.* (1991), [28] Turanov *et al.* (2009), [29] Heaphy *et al.* (2016), [30] Sánchez-Silva *et al.* (2003), [31] Hatanaka *et al.* (2007), [32] Lozupone *et al.* (2001)

Group	Phylogenetic extent	Reassignment	MSA	tRNA	Activity	Direct	References
Ascomycota	CUG-Ser clade	CUG Leu→Ser	✓	✓	✓	✓	[1,2,3]
Ascomycota	*Ascoidea*, *Saccharomycopsis*	CUG Leu→Ser/Leu	✓	✓		✓	[4,5,6]
Ascomycota	CUG-Ala clade	CUG Leu→Ala	✓	✓		✓	[3,5]
Aphelida	*Amoeboaphelidium protococcarum*	UAR Stop→Gln	✓				[7]
Trypanosomes	*Blastocrithidia*	UAR Stop→Stop/Glu	✓				[8]
		UGA Stop→Trp	✓				
Oxymonads	*Streblomastix*	UAR Stop→Gln	✓				[9,10]
Diplomonads	Hexamitinae	UAR Stop→Gln	✓	✓	✓		[11,12]
Caviomonads	*Iotanema spirale*	UAG Stop→Gln	✓				[13]
Ulvophyceae	Dasycladales, Cladophorales, *Blastophysa*, *Trentepohlia*	UAR Stop→Gln	✓			✓	[14,15,16]
Ulvophyceae	*Scotinosphaera*	UAG Stop→Ser	✓				[17]
Rhizaria	Rhizaria sp. exLh	UAG Stop→Leu	✓				[13]
Dinophyceae	*Amoebophrya* sp. ex *Karlodinium veneficum*	UAR Stop→Gln	✓	✓			[18]
		UGA Stop→Stop/Trp	✓				
Ciliophora	Oligohymenophorea	UAR Stop→Gln	✓	✓	✓	✓	[19,20,21]
Ciliophora	Spirotrichea	UAR Stop→Gln	✓	✓			[22,23,24]
Ciliophora	Karyorelictea	UAR Stop→Gln	✓				[25,26,22]
		UGA Stop→Stop/Trp	✓				
Ciliophora	Euplotidae	UGA Stop→Cys	✓	✓		✓	[27,28,22]
Ciliophora	Mesodiniidae	UAR Stop→Tyr	✓				[29,22]
Ciliophora	*Blepharisma*	UGA Stop→Trp	✓				[22,26,29]
Ciliophora	*Condylostoma magnum*	UAR Stop→Stop/Gln	✓	✓			[26,29,22]
		UGA Stop→Stop/Trp	✓			✓	
Ciliophora	Peritrichia	UAR Stop→Glu	✓	✓			[30,26,29]
Ciliophora	*Cryptocaryon irritans*	UAR Stop→Gln	✓				[31,26,29]
Ciliophora	*Tiarina fusa*	UAR Stop→Gln	✓				[26,29]
Ciliophora	*Aristerostoma* sp	UAR Stop→Gln	✓				[26,29]
Ciliophora	Nassulidae, *Zosterograptus*	UAR Stop→Gln	✓				[32]

Table A.3 (Continued)

Table A.4: All animal clades known to use a mitochondrial alternative genetic code. For each clade, I indicate the degree of support for the reassignment and references. Categories for support are: MSA– multiple sequence alignment or profile alignment showing codon at conserved positions for the new meaning; tRNA– presence of tRNA / release factor gene consistent with the new translation. [1] Gray *et al.* (1998), [2] Beagley *et al.* (1998), [3] Burger *et al.* (2003), [4] Sengupta *et al.* (2007), [5] Lavrov *et al.* (2013), [6] (Barrell *et al.*, 1979), [7] Fox (1987), [8] Boore *et al.* (1999), [9] Noutahi *et al.* (2017), [10] Spruyt *et al.* (1998), [11] Durrheim *et al.* (1993), [12] Himeno *et al.* (1987), [13] Perseke *et al.* (2011), [14] Li *et al.* (2019), [15] Clary and Wolstenholme (1985), [16] Abascal *et al.* (2006a), [17] Garey and Wolstenholme (1989), [18] Ohama *et al.* (1990b), [19] Telford *et al.* (2000), [20] Jacob *et al.* (2009)

	Phylum	Phylogenetic extent	Reassignment	MSA	tRNA	References
		all Metazoa, Choanozoa	UGA Stop→Trp	✓	✓	[1,2,3,4]
Deuterostomes	Porifera	*Clathrina clathrus*	UGA Stop→Trp	✓	✓	[5]
			UAG Stop→Tyr	✓	✓	
			CGN Arg→Gly	✓	✓	
	Chordata	Vertebrata	UGA Stop→Trp	✓	✓	[6,7]
			AUA Ile→Met	✓	✓	
			AGR Arg→0	✓		
	Chordata	*Brachiostoma*, *Epigonichthys*	UGA Stop→Trp	✓	✓	[8,9]
			AUA Ile→Met	✓	✓	
			AGR Arg→Ser	✓		
	Chordata	*Branchiostoma lanceolatum*	UGA Stop→Trp	✓	✓	[10]
			AUA Ile→Met	✓	✓	
			AGR Arg→Gly	✓	✓	
	Chordata	Tunicata	UGA Stop→Trp	✓	✓	[11,9]
			AUA Ile→Met	✓	✓	
			AGR Arg→Gly	✓	✓	
	Echinodermata	Echinodermata	UGA Stop→Trp	✓	✓	[12,9]
			AGR Arg→Ser	✓		
			AAA Lys→Asn	✓	✓	
	Hemichordata	Rhabdopleurida	UGA Stop→Trp	✓	✓	[13,14]
			AGA Arg→Ser	✓		
			AGG Arg→Lys	✓		
	Hemichordata	Cephalodiscida	UAA Stop→Tyr	✓		[14]
			UGA Stop→Trp	✓	✓	
			AGA Arg→Ser	✓		
			AGG Arg→Lys	✓		
	Hemichordata	Enteropneusta	UGA Stop→Trp	✓	✓	[13,14]
			AGR Arg→Ser	✓		
Protostomes	all Protostomes (“invertebrate code”)		UGA Stop→Trp	✓	✓	[15,7,9,16]
			AUA Ile→Met	✓		
			AGR Arg→Ser	✓		
	Arthropoda	various	UGA Stop→Trp	✓	✓	[9,16]
			AUA Ile→Met	✓		
			AGA Arg→Ser	✓		
			AGG Arg→Lys	✓		
	Platyhelminthes	Platyhelminthes	UGA Stop→Trp	✓	✓	[17,18,19]
			AGR Arg→Ser	✓	✓	
			AAA Lys→Asn	✓		
	Nematoda	*Radopholus*	UAA Stop→Tyr	✓		[20]
			UGA Stop→Trp	✓	✓	
			AUA Ile→Met	✓		
			AGR Arg→Ser	✓	✓	

Table A.4 (Continued)

	Group	Phylogenetic extent	Reassignment	MSA	tRNA	References
Fungi	Ascomycota	Saccharomyceta	UGA Stop→Trp	✓	✓	[1,2]
	Ascomycota	Saccharomycetaceae	UGA Stop→Trp	✓	✓	[3,4,5]
			AUA Ile→Met	✓		
			CUN Leu→Thr	✓	✓	
	Ascomycota	*Eremothecium*	UGA Stop→Trp	✓	✓	[4]
			AUA Ile→Met	✓		
			CUN Leu→Ala	✓	✓	
	Basidiomycota	Basidiomycota	UGA Stop→Trp	✓		[6,7]
	Chytridiomycota	Chytridiomycota	UAG Stop→Leu	✓	✓	[8]
Others	Haptophyta	Prymnesiophyceae	UGA Stop→Trp	✓	✓	[9,7]
	Centroplasthelida	*Marophrys* sp. SRT127	UGA Stop→Trp	✓		[10]
	Euglenozoa	Euglenozoa	UGA Stop→Trp	✓	✓	[11,12]
	Discosea	Centramoebida	UGA Stop→Trp	✓		[13]
	?	*Ancoracysta twista*	UGA Stop→Trp	✓	✓	[14]

Table A.5: All fungal and miscellaneous microbial eukaryotic clades known to use a mitochondrial alternative genetic code. For each clade, I indicate the degree of support for the reassignment and references. Categories for support are: MSA– multiple sequence alignment or profile alignment showing codon at conserved positions for the new meaning; tRNA– presence of tRNA / release factor gene consistent with the new translation. [1] Heckman *et al.* (1980), [2] Fox (1987), [3] Li and Tzagoloff (1979), [4] Ling *et al.* (2014), [5] Noutahi *et al.* (2017), [6] Paquin *et al.* (1997), [7] Sengupta *et al.* (2007), [8] Laforest *et al.* (1997), [9] Hayashi-Ishimaru *et al.* (1997), [10] Nishimura *et al.* (2019), [11] Benne *et al.* (1983), [12] Záhonová *et al.* (2021), [13] Burger *et al.* (1995), [14] Janouškovec *et al.* (2017)

	Phylogenetic extent	Reassignment	MSA	tRNA	References
	Rhodophyta except Cyanidiales	UGA Stop→Trp	✓	✓	[1,2]
Chlorophyta	*Pycnococcus provasolii*	UGA Stop→Trp	✓	✓	[3,4]
		AUA Ile→Met	✓	✓	
		UUR Leu→Stop	✓		
	Pedinomonas minor	UGA Stop→Trp	✓	✓	[5, 4]
	Sphaeropleales	UCR Ser→Stop	✓		[4,6]
	Atractomorpha echinata	CGG Arg→Leu	✓		[6]
	Bracteacoccus	UCR Ser→Stop	✓		[4 6]
		AGG Arg→Ala	✓	✓	
	Chromochloris zofingiensis	UCR Ser→Stop	✓		[4,6]
		AGG Arg→Met	✓	✓	
		CGG Arg→Leu	✓	✓	
	Rotundella rotunda	UCR Ser→Stop	✓		[6]
		AGR Arg→Ala	✓		
	Scenedesmaceae	UAG Stop→Leu	✓	✓	[7,8,4,6]
		UCR Ser→Stop	✓		
		AGG Arg→Ala	✓	✓	
	Coellastrella, *Hariotina*	UAG Stop→Leu	✓	✓	[6]
		UCR Ser→Stop	✓		
		AGR Arg→Ala	✓	✓	
	Chlorotetraedron incus	UCR Ser→Stop	✓		[4,6]
		AGG Arg→Ala	✓	✓	
	Neochloris aquatica	UAG Stop→Ala	✓	✓	[4,6]
		UCR Ser→Stop	✓		
		AGR Arg→Ala	✓	✓	
	Hydrodictyaceae	UAG Stop→Ala	✓	✓	[4,6,7]
		UCR Ser→Stop	✓		

Table A.6: All red and green algae known to use a mitochondrial alternative genetic code. For each clade, I indicate the degree of support for the reassignment and references. Categories for support are: MSA– multiple sequence alignment or profile alignment showing codon at conserved positions for the new meaning; tRNA– presence of tRNA / release factor gene consistent with the new translation. [1] Boyen *et al.* (1994), [2] Boyen *et al.* (2001), [3] Turmel *et al.* (2010), [4] Noutahi *et al.* (2019), [5] Turmel *et al.* (1999), [6] Žihala and Eliáš (2019), [7] Hayashi-Ishimaru *et al.* (1996), [8] Nedelcu *et al.* (2000)

	Broad group	Phylogenetic extent	Reassignment	MSA	tRNA	References
Stramenopiles	Bacillariophyta	Thalassiosirales	UGA Stop→Trp	✓	✓	[1,2]
	PX clade	Xanthophyceae	AUA Ile→Met	✓		[3]
	Opalozoa	*Cafeteria roenbergensis*, Stramenopile MAST12	UGA Stop→Trp	✓	✓	[4,5]
	All Labyrinthulea		AUA Ile→Met	✓		[5,6]
			UUA Leu→Stop	✓		
	Labyrinthulea	Thraustochytrid S2, S4	AUA Ile→Met	✓		[5,6]
			UUA Leu→Stop	✓		
			UAR Stop→Tyr	✓	✓	
			AGR Arg→Stop	✓		
	Labyrinthulea	*Aplanochytrium*	AUA Ile→Met	✓		[6]
			UUA Leu→Stop	✓		
			UAG Stop→Tyr	✓	✓	
	Eogyrea	Stramenopile MAST8b	AUA Ile→Met	✓		[6]
			AGR Arg→Ser	✓	✓	
Alveolata	Oomycota	*Eurychasma dicksonii*	UGA Stop→Trp	✓		[7]
	Perkinsozoa	*Perkinsus marinus*	UGA Stop→Trp	✓	✓	[8,9]
	Ciliophora	Ciliophora	UGA Stop→Trp	✓	✓	[10,11]
	Rhizaria	Cercozoa	UGA Stop→Trp	✓	✓	[12,5]

Table A.7: All stramenopile, alveolate, and rhizarian (SAR) clades known to use a mitochondrial alternative genetic code. For each clade, I indicate the degree of support for the reassignment and references. Categories for support are: MSA– multiple sequence alignment or profile alignment showing codon at conserved positions for the new meaning; tRNA– presence of tRNA / release factor gene consistent with the new translation. [1] Ehara *et al.* (2000), [2] Oudot-Le Secq and Green (2011), [3] Ehara *et al.* (1997), [4] Gray *et al.* (1998), [5] Wideman *et al.* (2020), [6] Žihala *et al.* (2020), [7] Sekimoto *et al.* (2008), [8] Masuda *et al.* (2010), [9] Gornik *et al.* (2022), [10] Seilhamer and Cummings (1982), [11] Burger *et al.* (2000), [12] Inagaki *et al.* (1998)

Appendix B

Appendix to Chapter 2

Supplementary Tables and Figures

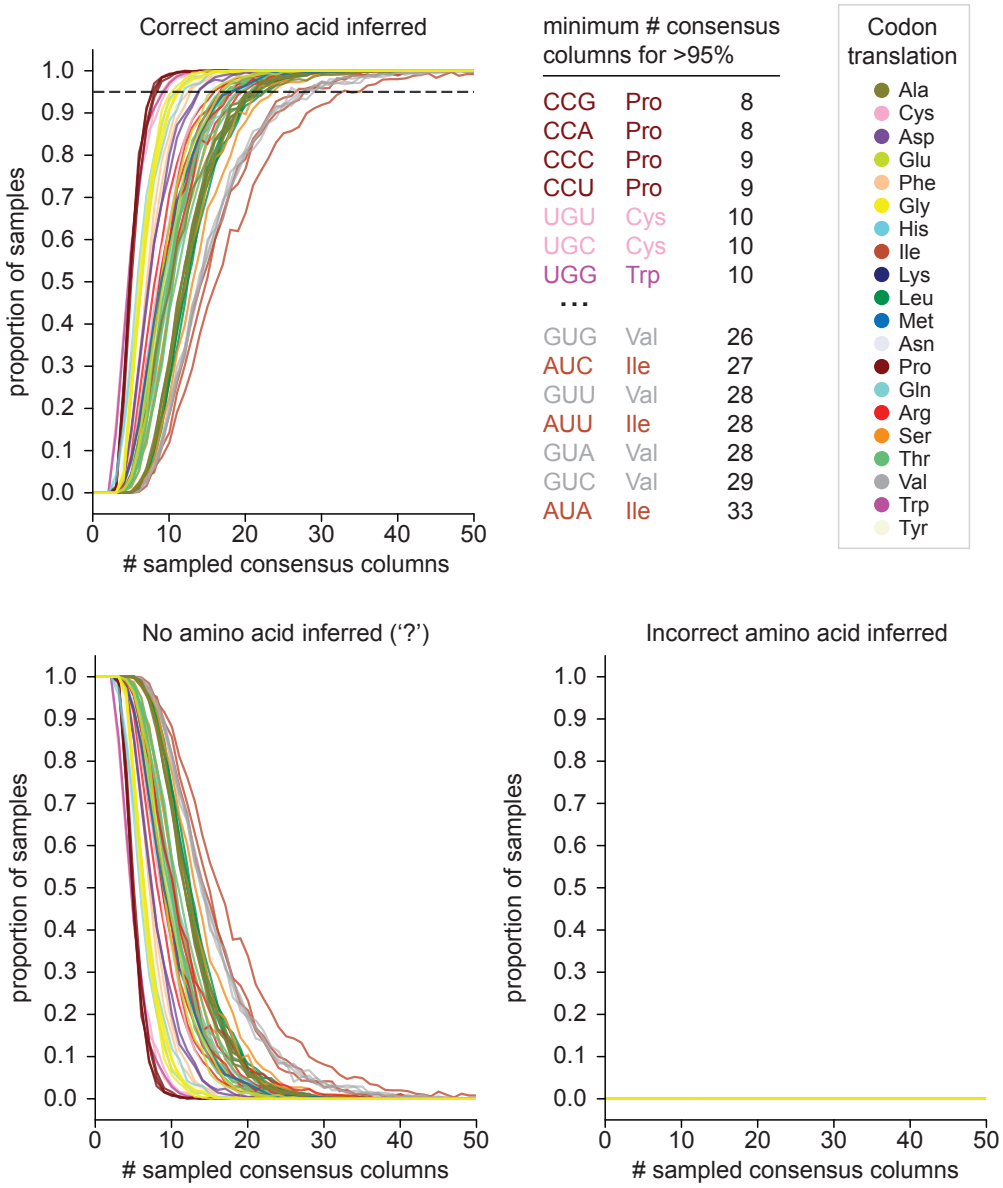

Figure B.1: Codetta error rate and power estimated by subsampling Pfam consensus columns aligned to an *E. coli* genome. Plots show the proportion of consensus column samples that resulted in the correct inference (power), incorrect inference (error rate), and no amino acid inference ('?' inference) under a decoding threshold of 0.9999 for the 61 sense codons, color-coded by the correct amino acid translation. The dashed line indicates 95% of samples with the correct amino acid inferred. The codons that required the fewest and greatest number of consensus columns for >95% correct inference are listed.

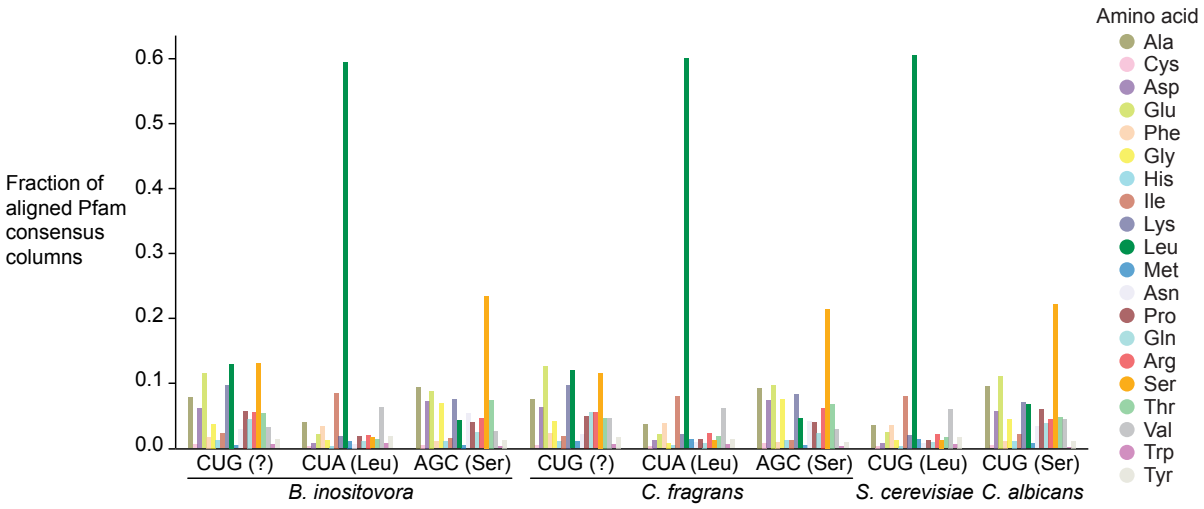

Figure B.2: Distribution of the highest probability amino acid for all aligned Pfam consensus columns to CUG, CUA (rare leucine codon), and AGC (rare serine codon) in *B. inositovora* and *C. fragrans* and to CUG in *S. cerevisiae* and *C. albicans*. Codetta codon inference is labelled in parentheses.

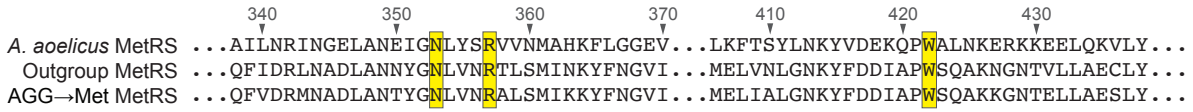

Figure B.3: A multiple sequence alignment of the methionyl-tRNA synthetase (MetRS) sequence from *Aquifex aeolicus* and the predicted MetRS sequences from the outgroup species #6 (GCA_900316035.1) and *Enterosoma* AGG→Met clade species #2 (GCA_002404995.1). Residues are numbered in reference to the *A. aeolicus* sequence. Residues in contact with the tRNA anticodon in the *A. aeolicus* crystal structure (PDB 2CSX/2CT8, Nakanishi *et al.* (2005)) are highlighted in yellow.

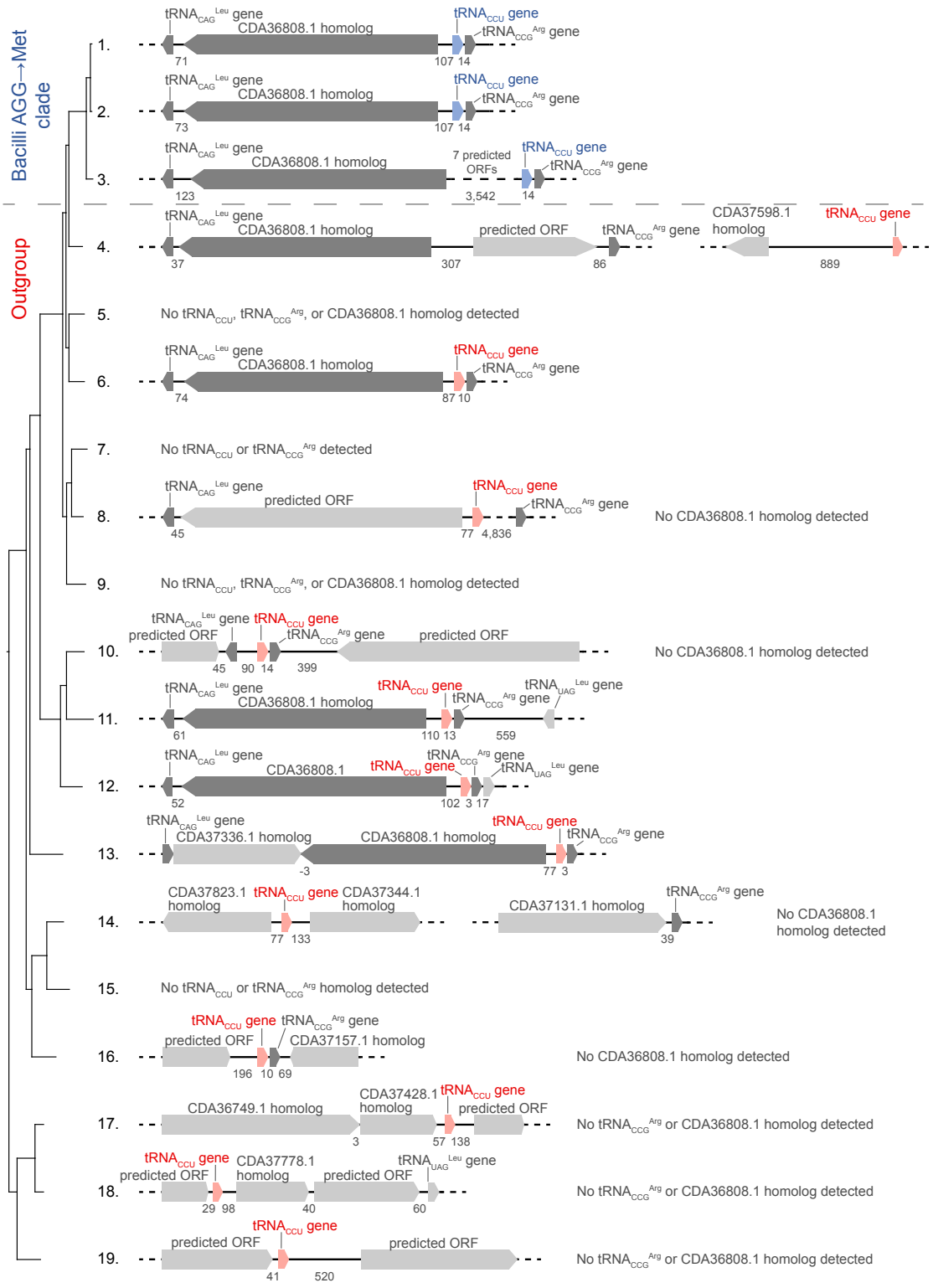

Figure B.4: A visualization of the local genomic context surrounding the tRNA$_{CCU}$ gene, tRNA$_{CCG}^{Arg}$ gene, and the CDA36808.1 homolog, if detected, in the *Enterosoma* AGG→Met clade and outgroup species. The species tree and numberings are the same as Figure 2.3 and more information for each species can be found in the source data for Figure 2.3.

Appendix C

Appendix to Chapter 3

Comparison of Codetta and FACIL codon predictions

For both the yeast genome test set and the mitochondrial genome test set, each species has a ground truth annotation of the expected translation for each codon which we compare to the predictions from Codetta and FACIL (Supplementary Files). If the translation predicted by Codetta or FACIL is consistent with the ground truth annotation, then that prediction is considered to be correct; if it differs from the ground truth, it is considered to be incorrect; and if the codon is uninferred, it is not counted as either (see Table C.1). For annotated stop codons, only amino acid predictions are considered to be incorrect, since FACIL explicitly predicts stop codons while Codetta is expected to leave stop codons uninferred. Furthermore, any codon prediction that differs from the standard genetic code is considered to be a "predicted codon reassignment", which could be correct (a known, annotated codon reassignment) or incorrect (not consistent with the ground truth annotation) (see Table C.2).

We first compared the predictions against the annotated genetic code and calculated overall rates of correct, incorrect, and uninferred codon predictions for both Codetta and FACIL

(Table C.1). In the yeast genome comparison (Table C.1a), Codetta predicted the correct amino acid for 30,357 out of 30,366 sense codons (including reassigned CUG codons). Nine sense codons were left uninferred by Codetta, which include six cases of likely ambiguous translation of CUG and three cases of rare codon usage (Shulgina and Eddy, 2021). No sense codons were incorrectly predicted by Codetta. In contrast, FACIL selected the incorrect amino acid for 53 sense codons (mostly AUG codons) and left 563 sense codons uninferred, including 109 reassigned CUG codons. Both methods had similar rates of incorrect stop codon predictions in both the yeast genome test set and mitochondrial genome test set.

In the mitochondrial comparison (Table C.1b), FACIL had slightly more correctly predicted sense codons (4,283 (84.7%) for FACIL and 3,908 (77.2%) for Codetta); for reassigned sense codons, the number of correctly predicted codons was more even (170 for FACIL and 190 for Codetta). FACIL again had more incorrect sense codon predictions (78, 1.5%) compared to three for Codetta. The errors made by FACIL were distributed across reassigned sense codons, AUG codons, and other sense codons. The three unexpected sense codon predictions made by Codetta may be real codon reassignments, pending further support. Multiple sequence alignments of mitochondrial proteins show that the three potentially reassigned codons occur at conserved positions for the predicted amino acid (Figures C.2-C.4).

In the main text, we compare positive predictive values (PPV) for Codetta and FACIL on the two test sets by taking the predicted codon reassignments (codon predictions that differ from the standard genetic code) and evaluating how many of the predictions are correct (consistent with ground truth; known codon reassignments) or incorrect. Table C.2 shows the Codetta and FACIL predictions relative to the standard genetic code and a breakdown of the predicted codon reassignments (used to make Figure 3.1).

Supplementary Tables and Figures

(a) Yeast test set, predictions relative to annotated genetic code

			Codetta (Pfam)			FACIL (Pfam)		
			Correct	Incorrect	Uninf.	Correct	Incorrect	Uninf.
Sense codons	CUG	Leu (std)	332	0	0	332	0	0
		Ser	156	0	0	54	0	102
		Ala	6	0	0	5	0	1
		Ser/Leu	6 S	0	6	5 L, 1 S	0	6
	AUG		506	0	0	0	52	454
	All other sense codons		29,851	0	3	29,853	1	0
	Stop codons		-	6	1,512	1,510	7	1

(b) Mitochondrial test set, predictions relative to annotated genetic code

		Codetta (mito proteins)			FACIL (Pfam)		
		Correct	Incorrect	Uninf.	Correct	Incorrect	Uninf.
Sense	Sense/stop-to-sense	190	0	67	170	11	76
	AUG codons	78	0	4	29	12	41
	All other sense codons	3,640	3	1,077	4,084	55	581
Stop	Sense-to-stop reassign.	-	0	24	19	0	5
	Unchanged stops	-	0	165	119	0	46

Table C.1: Per-codon comparison of Codetta and FACIL relative to the annotated genetic code on (a) 506 yeast genomes (32,384 codon predictions) and (b) 82 mitochondrial genomes (5,248 codon predictions). We show totals separately for sense codons and stop codons and further split reassigned codons and AUG (frequent FACIL error). For stop codons, only amino acid predictions are considered incorrect.

(a) Codetta, yeast test set

Codetta	Relative to std code			Codon reassign.		
	?	Same	Codon reassign.	Known	Error	Codon reassignment breakdown
Standard sense	9	30,689	168	168	0	K: 156 CUG-Ser 6 CUG-Ser/Leu 6 CUG-Ala
Standard stop	1,512	-	6	0	6	E: 6 stop-to-sense error

(b) FACIL, yeast test set

FACIL	Relative to std code			Codon reassign.		
	?	Same	Codon reassign.	Known	Error	Codon reassignment breakdown
Standard sense	563	30,190	113	60	53	K: 54 CUG-Ser 1 CUG-Ser/Leu 5 CUG-Ala E: 52 AUG error 1 non-AUG sense error
Standard stop	1	1,510	7	0	7	E: 7 stop-to-sense error

(c) Codetta, mitochondrial test set

Codetta	Relative to std code			Codon reassign.		
	?	Same	Codon reassign.	Known	Error	Codon reassignment breakdown
Standard sense	1,169	3,718	115	112	3	K: 112 known sense-to-sense E: 3 non-AUG sense error
Standard stop	168	-	78	78	0	K: 78 known stop-to-sense

(d) FACIL, mitochondrial test set

FACIL	Relative to std code			Codon reassign.		
	?	Same	Codon reassign.	Known	Error	Codon reassignment breakdown
Standard sense	702	4,113	187	109	78	K: 90 known sense-to-sense 19 known sense-to-stop E: 12 AUG error 58 non-AUG sense error 8 sense-to-stop error
Standard stop	47	119	80	80	0	K: 80 known stop-to-sense

Table C.2: Breakdown of codon inferences by Codetta and FACIL relative to the standard genetic code on the yeast genomes test set (a,b) and the mitochondrial genomes test set (c,d). Comparing relative to the standard genetic code allows us to evaluate how many of the predicted codon reassignments are known vs prediction errors, and to calculate positive predictive value (PPV). The codon reassignment breakdowns are plotted in Figure 3.1.

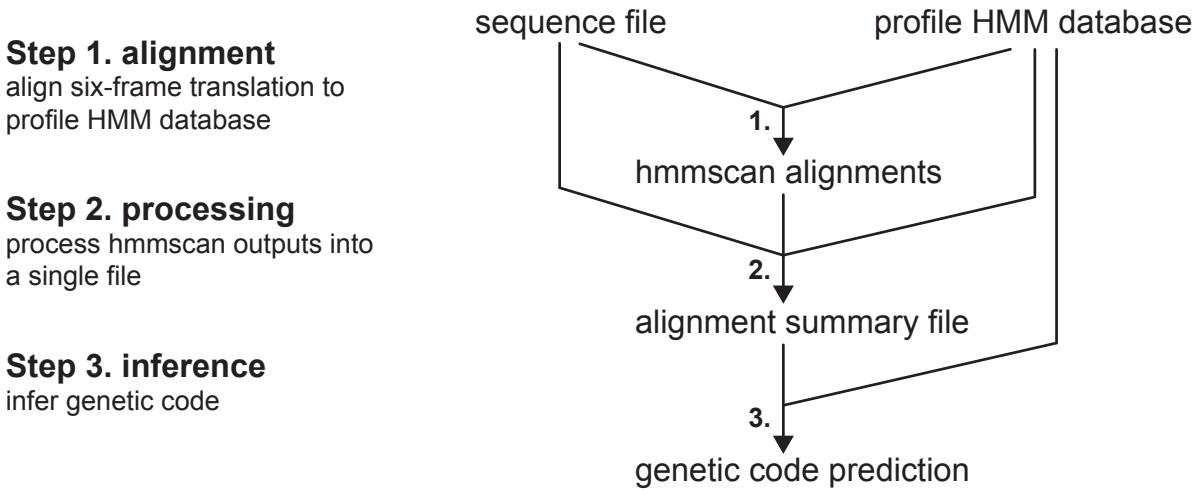

Figure C.1: Schematic visualization of the main steps of Codetta, including the input and output files for each step.

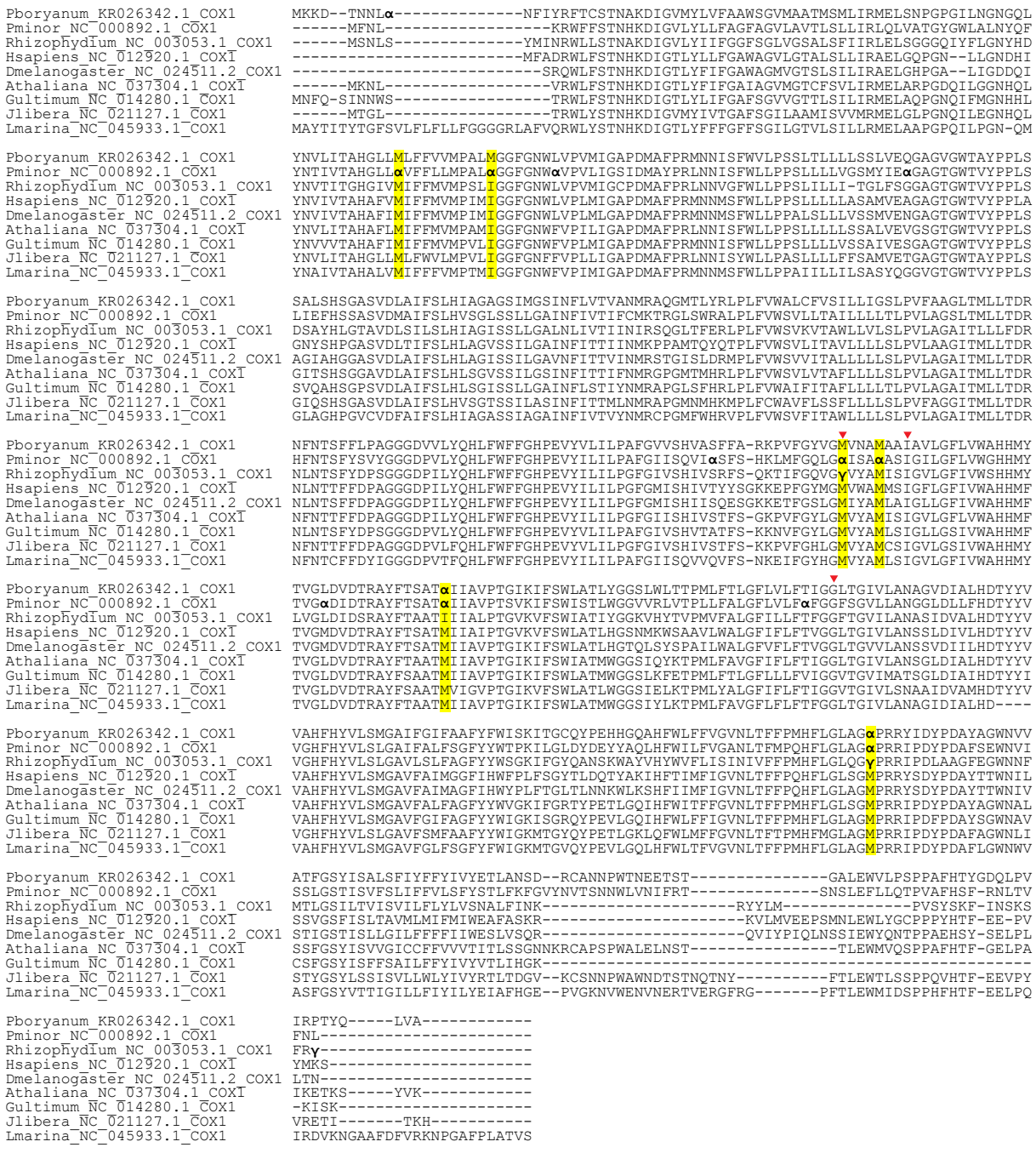

Figure C.2: Alignment of cytochrome c oxidase subunit 1 (COX1) predicted protein sequences from the three mitochondrial genomes with an unexpected Codetta inference and six reference species. In *Pseudopediastrum boryanum* (KR026342.1) and *Pedinomonas minor* (NC_000892.1), AUA codons are represented by alphas (α). In *Rhizophydium* sp. 136 (NC_003053.1), AAG codons are represented by gammas (γ). Yellow highlighted columns contain the codons under question and are conserved for the same amino acid in all six reference species. Red triangles indicate residues located within the active site region of COX1, associated with a human disease, and reported to have a biochemical effect when mutated (from Lloyd and McGeehan (2013)).

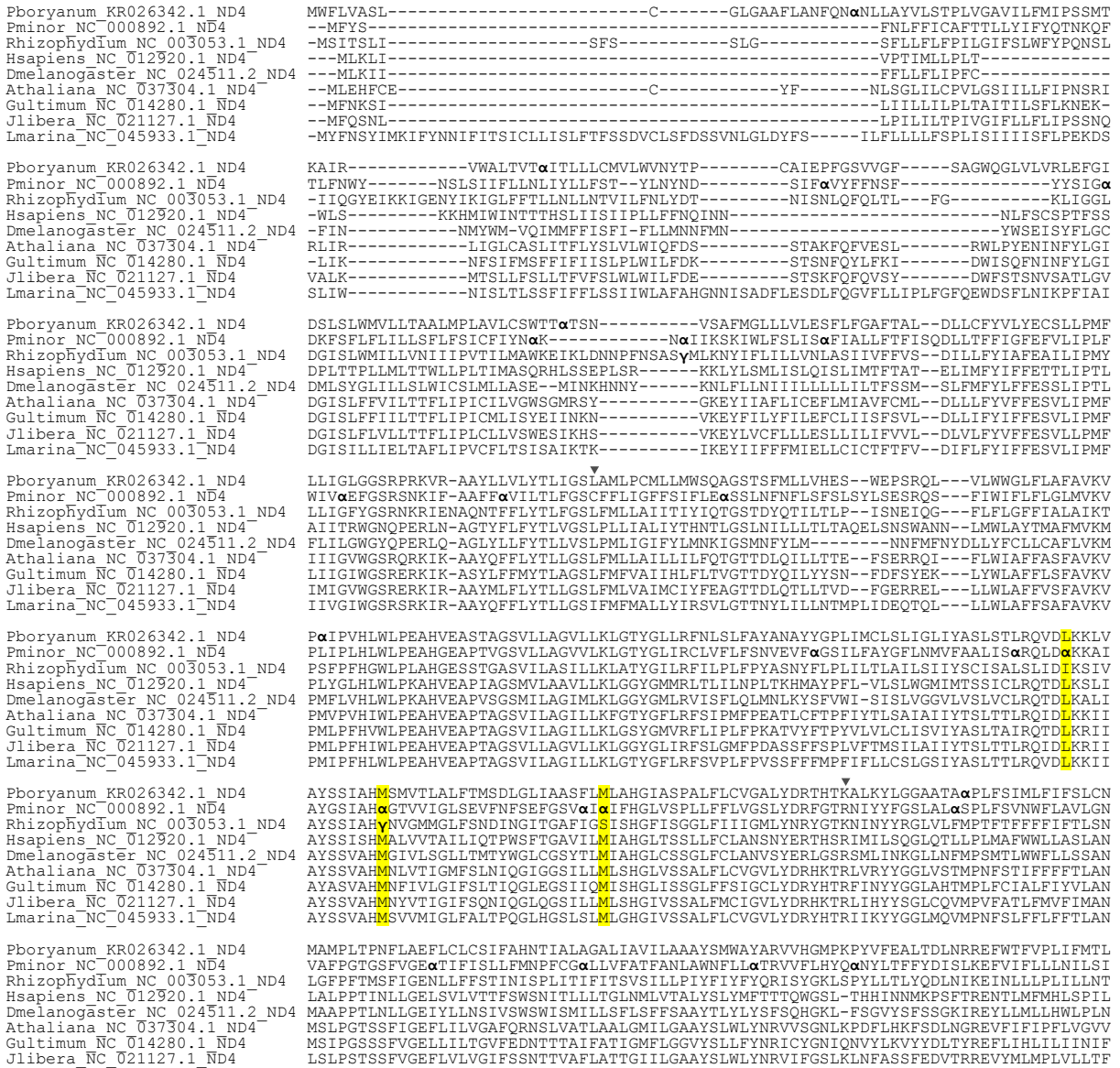

Figure C.3: Alignment of NADH dehydrogenase subunit 4 (ND4) predicted protein sequences from the three mitochondrial genomes with an unexpected Codetta inference and six reference species. In *Pseudopediastrum boryanum* (KR026342.1) and *Pedinomonas minor* (NC_000892.1), AUA codons are represented by alphas (α). In *Rhizophydium* sp. 136 (NC_003053.1), AAG codons are represented by gammas (γ). Yellow highlighted columns contain the codons under question and are conserved for the same amino acid in all six reference species. Triangles indicate positions associated with complex I diseases when mutated (from Bridges *et al.* (2011)).

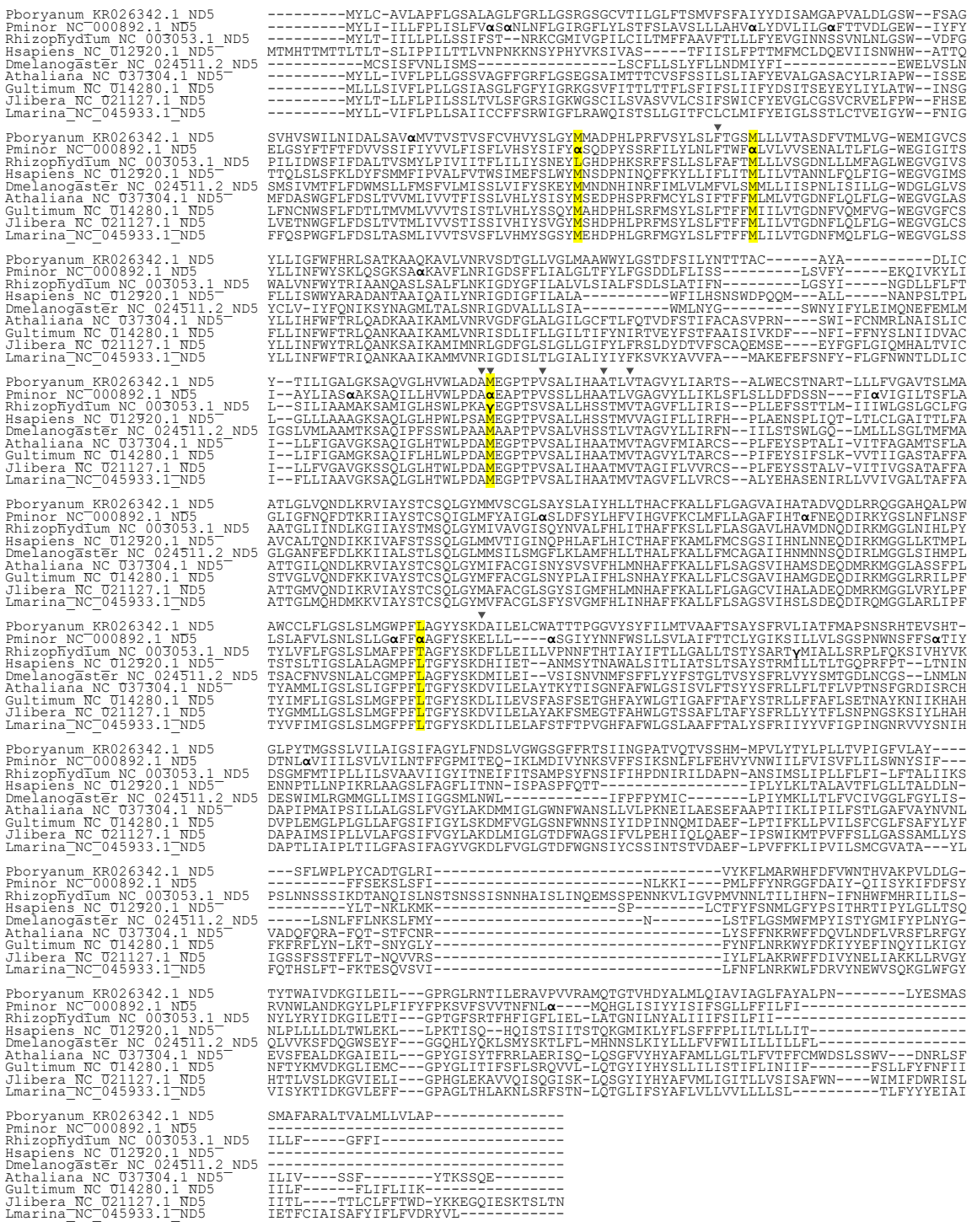

Figure C.4: Alignment of NADH dehydrogenase subunit 5 (ND5) predicted protein sequences from the three mitochondrial genomes with an unexpected Codetta inference and six reference species. In *Pseudopediastrum boryanum* (KR026342.1) and *Pedinomonas minor* (NC_000892.1), AUA codons are represented by alphas (α). In *Rhizophydium* sp. 136 (NC_003053.1), AAG codons are represented by gammas (γ). Yellow highlighted columns contain the codons under question and are conserved for the same amino acid in all six reference species. Triangles indicate positions associated with complex I diseases when mutated (from Bridges *et al.* (2011)).