Reassignment of CGA and CGG to tryptophan in Absconditabacteria

In support of this reassignment, we provide the following data files: an Excel spreadsheet with detailed information on all genomes belonging to the reassigned clade and close outgroups (including information on inferred CGA/CGG amino acid, CheckM genome completeness estimate from GTDB, presence/absence of specific tRNA genes, codon usage in aligned Pfam domains, and genomic GC content), full alignments of BUSCO genes featured in figures, and alignments of tRNA sequences from the reassigned clade and outgroups.

Results

Genetic code inference

Two clades in the Candidate Phyla Radiation– 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. We confirmed using Codetta that genomes annotated as Absconditabacteria and Gracilibacteria on NCBI were predicted to translate UGA as glycine by Codetta (see Shulgina & Eddy (2021), Table 1). Some of these genomes were additionally predicted to have reassignments of the canonical arginine codons CGA and/or CGG as tryptophan.

Species phylogenetic tree

Since many genomes derived from the uncultivated Candidate Phyla Radiation have only broad phylogenetic annotation in NCBI, we investigated the phylogenetic distribution of this reassignment on the Genome Taxonomy Database (GTDB). In Figure 1, we show a part of the GTDB phylogenetic tree that 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 not present in the genome or whether it missing due to an incomplete assembly, we marked the most complete genomes on the tree. We believe that the CheckM estimates of genome completeness (based on presence of conserved protein sets and are provided on GTDB for each genome) tend to underestimate the completeness of Candidate Phyla Radiation genomes due to lineage-specific gene loss. Therefore, we 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 1, we 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 1), all members of the Absconditabacteria and Gracilibacteria were inferred to use the known reassignment of UGA as glycine, while none of the outgroup genomes were predicted to have an amino acid translation of the stop codon UGA. Consistent with that prediction, all members of the UGA-reassigned clades do not have an release factor 2 (RF2) gene (which terminates translation at UAA and UGA codons) and most in-

stead contain 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, as they 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 tree (Figure 1), all 10 species were inferred to translate CGA as tryptophan and CGG as either tryptophan, uninferred, or arginine (in one species).

Figure 2 shows four example alignments of conserved single-copy bacterial genes across Absconditabacteria, Gracilibacteria, and outgroup species. The four aligned genes show that multiple CGA and CGG positions across all 10 Absconditabacteria genomes appear to be 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 CGAs and CGGs occur at positions broadly conserved for arginine.

The tRNAs present in Absconditabacteria genomes are consistent overall with the inferred codon translations (Figure 1). The $tRNA_{CCG}$ and $tRNA_{UCG}$ genes, which decode CGA and CGG codons, were predominantly tryptophan-type (classified by G73 discrminator 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.

One exception was the $tRNA_{CCG}$ from GCA_002414185.1 (Patescibacteria group bacterium UBA5124; #9 on tree), which is the only Absconditabacteria genome with an arginine translation inferred for CGG. The $tRNA_{CCG}$ could not be classified as either arginine- or tryptophan-type because it has a A73 discriminator base, which supports arginine 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 from across the Absconditabacteria, Gracilibacteria, and outgroup species (Figure 3), 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. We 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 and not arginine tRNAs.

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 1). At the extreme, Absconditabacteria genomes where CGG was either uninferred or inferred to code for arginine had the lowest CGG codon usage (<3 per 10,000 codons) resulting in fewer than 10 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 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. GC content-driven substitutions towards AT-rich arginine codons, such as AGA and AGG, 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 (typically GGN) (*1*).

Translation of CGA and CGG in Gracilibacteria

The Gracilibacteria genomes on the tree (Figure 1) split into two clades (dubbed here as "clade 1" and "clade 2") that differ in codon usage, genomic GC content, inferred CGA/CGG translation, 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 codon meaning are uninferred 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 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 2), CGA occurs at conserved arginine positions in members of Gracilibacteria clade 1 and clade 2; only one aligned CGA position in Gracilibacteria clade 1 is shown. 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 1). Clade 2 genomes all contain an arginine-type $tRNA_{UCG}$, which would translate both CGA and CGG codons are arginine. In contrast, none of the clade 1 genomes contain a $tRNA_{UCG}$. It is possible that we 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.

In Gracilibacteria clade 1, 6 of the 8 genomes contain a CGG-decoding $tRNA_{CCG}$. We could not confidently assign either arginine or tryptophan isotype to these tRNAs because they have an unusual D-loop sequence that makes it unclear which nucleotide is at position 20, contain the minor tryptophan identity element A1:U72 and a G73 discriminator which is compatible with either arginine or tryptophan isotype. We constructed a phylogenetic tree from an alignment of arginine and tryptophan tRNAs in Absconditabacteria, Gracilibacteria, and outgroups (Figure 3). In the consensus tree, the unusual $tRNA_{CCG}$ sequences from Gracilibacteria clade 1 form a clade within the cluster of tryptophan $tRNA_{CCA}$ from all groups and the reassigned $tRNA_{UCG}$ and $tRNA_{CCG}$ sequences from Absconditabacteria, and not with the arginine tRNAs from all groups. We 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 and 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 (while the Absconditabacteria $tRNA_{UCG}$ and $tRNA_{CCG}$ cluster with tryptophan 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 $tRNA_{CCG}$ is more similar in sequence to tryptophan tRNAs and not arginine tRNAs, and was most likely derived from a tryptophan tRNA; however, this does not necessarily mean that this tRNA is charged with tryptophan *in vivo*.

All together, the evidence paints a picture where Gracilibacteria clade 2 species decode CGA and CGG as arginine, while clade 1 genomes might treat CGA as an arginine codon and CGG as a tryptophan codon but many questions remain regarding 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 occurrence

of CGG in conserved tryptophan residues in conserved single-copy gene alignments in a few species 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 mulitple 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 non-cognate tRNAs, possibly by the arginine $tRNA_{GCG}$ or perhaps the unusual $tRNA_{CCG}$.

Methods

The methods used to generate the results can be found in Shulgina & Eddy (2021). Additionally, the following methods were used.

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 we additionally assessed genome completeness by tabulating the presence of a set of required tRNA genes found by tRNAscan-SE 2.0 (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. We excluded tRNAs that are involved in the decoding of the CGN-box which includes the reassigned codons.

tRNA phylogeny

Phylogenetic trees were inferred from tRNA alignments including tRNAs that decode the reassigned codon and a selection of tRNAs for the original and new amino acid from the reassigned

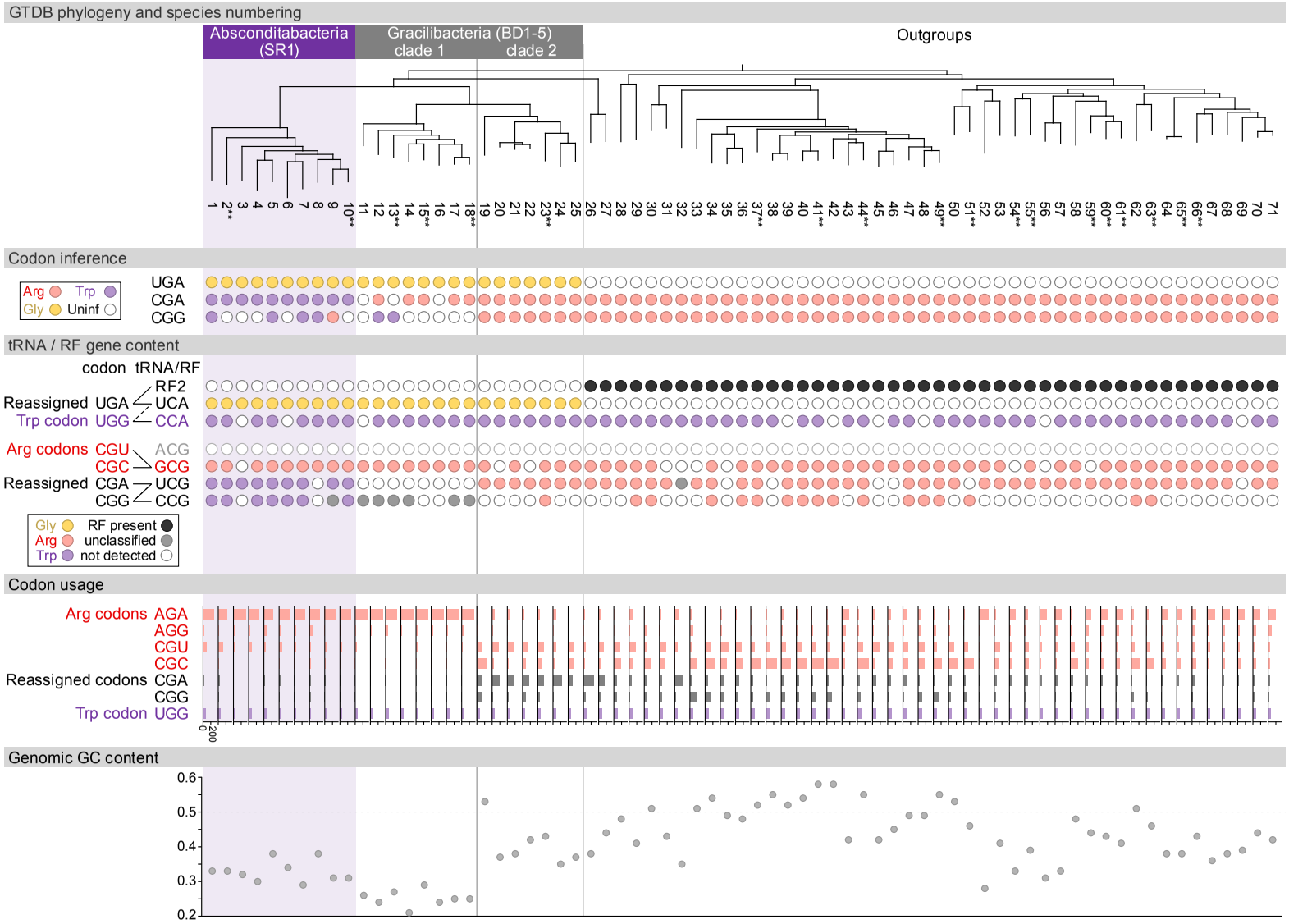
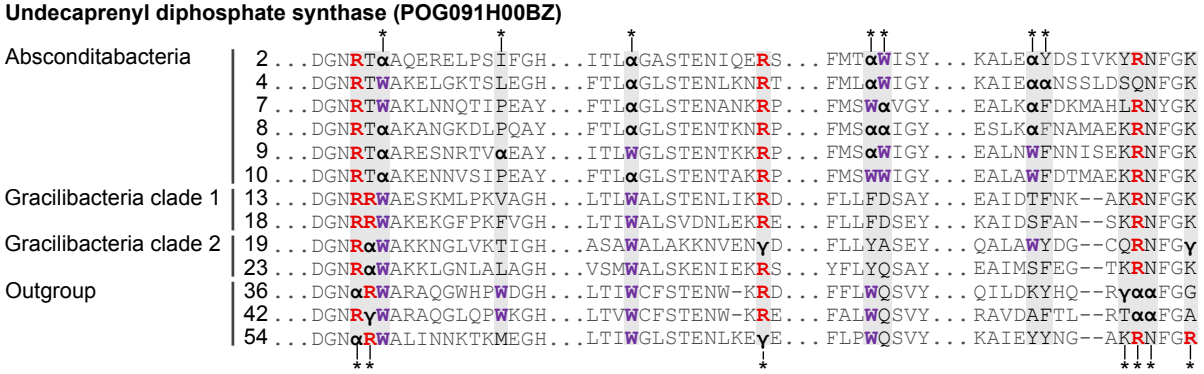

GTDB phylogeny and species numbering
Absconditabacteria (SR1)
Gracilibacteria (BD1-5) clade 1
clade 2
Outgroups
1 2** 3 4 5 6 7 8 9 10** 11 12 13** 14 15** 16 17 18** 19 20 21 22 23** 24 25 26 27 28 29 30 31 32 33 34 35 36 37** 38 39 40 41** 42 43 44** 45 46 47 48 49** 50 51** 52 53 54** 55** 56 57 58 59** 60** 61** 62 63** 64 65** 66** 67 68 69 70 71
Codon inference
Arg Trp
Gly Uninf
UGA
CGA
CGG
tRNA / RF gene content
codon tRNA/RF
RF2
Reassigned UGA UCA
Trp codon UGG CCA
Arg codons CGU ACG
CGC GCG
Reassigned CGA UCG
CGG CCG
Gly RF present
Arg unclassified
Trp not detected
Codon usage
Arg codons AGA
AGG
CGU
CGC
Reassigned codons CGA
CGG
Trp codon UGG
0
200
Genomic GC content
0.6
0.5
0.4
0.3
0.2

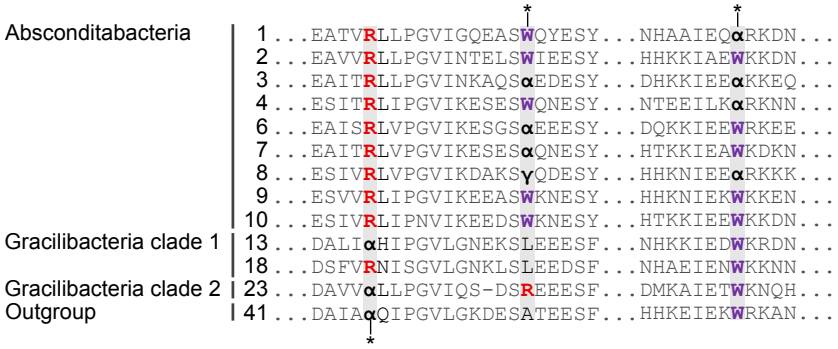
Figure 1: Phylogenetic tree from GTDB showing the Absconditabacteria, Gracilibacteria, and closest outgroup genomes (Peregrinibacteria and Peribacteria), each species indicated by a number which can be cross-referenced with the summary spreadsheet. Asterisks indicate genomes the most complete genomes, 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 the Candidate Phyla Radiation. 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.

clade and outgroups. Trees were inferred using a Bayesian approach implemented in MrBayes 3.2.7a (*2*) run for 40 million generations with sampling every 500 generations with default burn-in. We excluded columns corresponding to the anticodon and to the 3' half of stems (to remove the correlated information in basepaired stem regions). The remaining columns were partitioned into stem and loop regions and were modelled 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).

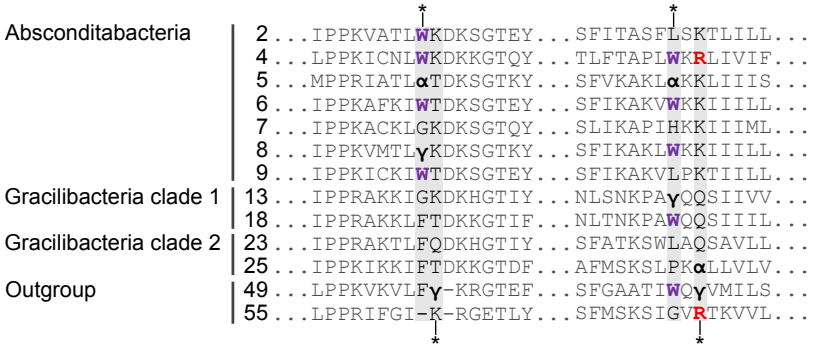
Undecaprenyl diphosphate synthase (POG091H00BZ)



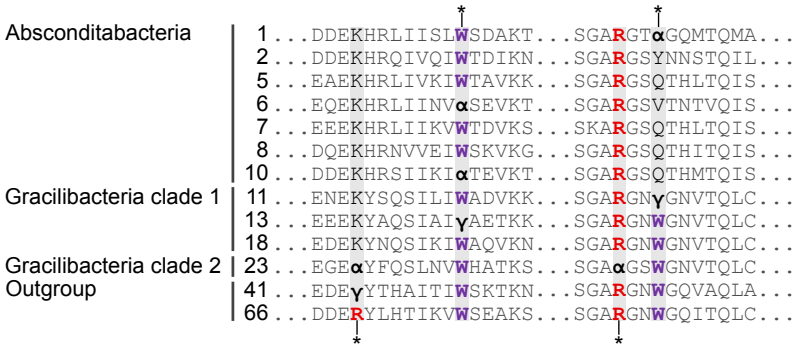
tRNA (guanine-N1)-methyltransferase (POG091H01WE)



Peptidase M50 (POG091H0131)



DNA-directed RNA polymerase subunit beta (POG091H02K5)



Reassigned codon symbols	
α	CGA
γ	CGG

Figure 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 nearby 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.

References

1. C. Rinke, *et al.*, *Nature* **499**, 431 (2013).

2. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (2003).

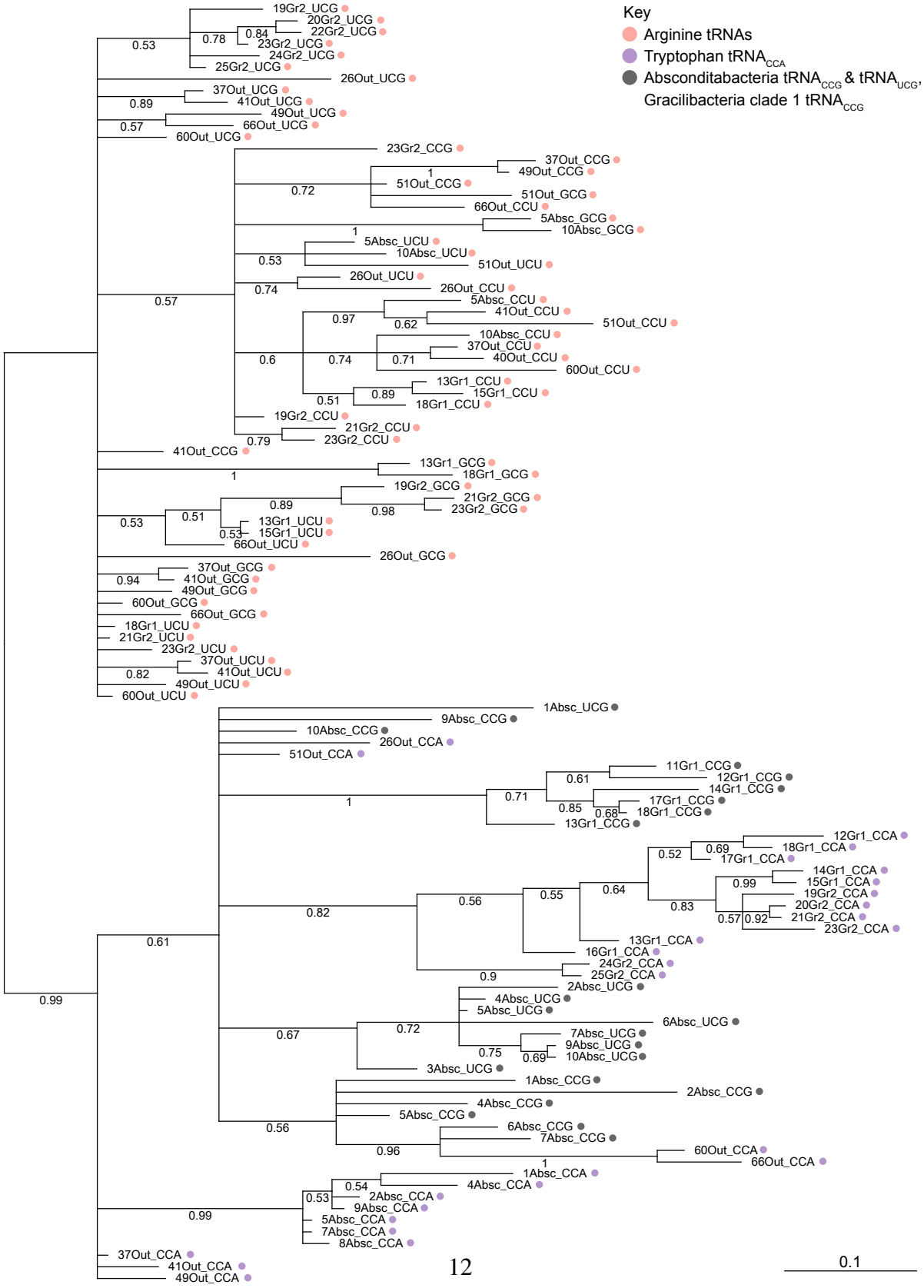

Key
Arginine tRNAs
Tryptophan tRNA$_{CCA}$
Absconditabacteria tRNA$_{CCG}$ & tRNA$_{UCG}$, Gracilibacteria clade 1 tRNA$_{CCG}$
19Gr2_UCG
20Gr2_UCG
22Gr2_UCG
23Gr2_UCG
24Gr2_UCG
25Gr2_UCG
26Out_UCG
37Out_UCG
41Out_UCG
49Out_UCG
66Out_UCG
60Out_UCG
23Gr2_CCG
37Out_CCG
49Out_CCG
51Out_CCG
51Out_GCG
66Out_CCU
5Absc_GCG
10Absc_GCG
5Absc_UCU
10Absc_UCU
51Out_UCU
26Out_UCU
26Out_CCU
5Absc_CCU
41Out_CCU
51Out_CCU
10Absc_CCU
37Out_CCU
40Out_CCU
60Out_CCU
13Gr1_CCU
15Gr1_CCU
18Gr1_CCU
19Gr2_CCU
21Gr2_CCU
23Gr2_CCU
41Out_CCG
13Gr1_GCG
18Gr1_GCG
19Gr2_GCG
21Gr2_GCG
23Gr2_GCG
13Gr1_UCU
15Gr1_UCU
66Out_UCU
26Out_GCG
37Out_GCG
41Out_GCG
49Out_GCG
60Out_GCG
66Out_GCG
18Gr1_UCU
21Gr2_UCU
23Gr2_UCU
37Out_UCU
41Out_UCU
49Out_UCU
60Out_UCU
1Absc_UCG
9Absc_CCG
10Absc_CCG
26Out_CCA
51Out_CCA
11Gr1_CCG
12Gr1_CCG
14Gr1_CCG
17Gr1_CCG
18Gr1_CCG
13Gr1_CCG
12Gr1_CCA
18Gr1_CCA
17Gr1_CCA
14Gr1_CCA
15Gr1_CCA
19Gr2_CCA
20Gr2_CCA
21Gr2_CCA
23Gr2_CCA
13Gr1_CCA
16Gr1_CCA
24Gr2_CCA
25Gr2_CCA
2Absc_UCG
4Absc_UCG
5Absc_UCG
6Absc_UCG
7Absc_UCG
9Absc_UCG
10Absc_UCG
3Absc_UCG
1Absc_CCG
2Absc_CCG
4Absc_CCG
5Absc_CCG
6Absc_CCG
7Absc_CCG
60Out_CCA
66Out_CCA
1Absc_CCA
4Absc_CCA
2Absc_CCA
9Absc_CCA
5Absc_CCA
7Absc_CCA
8Absc_CCA
37Out_CCA
41Out_CCA
49Out_CCA
0.1

Figure 3: 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.