Additional results for CGA / CGG reassignments

For each CGA and/or CGG reassignment, we provide the following data files: an Excel spreadsheet with detailed information on all genomes belonging to the reassigned clades 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, alignments of tRNA sequences from the reassigned clade and outgroups, and a consensus phylogenetic tree of the tRNA sequences.

Reassignment of CGA and CGG to tryptophan in Absconditabacteria

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 that genomes annotated as Absconditabacteria and Gracilibacteria on NCBI were predicted to translate UGA as glycine by our genetic code inference method (paper 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 supports 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, 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 positions 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 positions. 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 $tRNA_{CCG}$.

Reassignment of CGG to glutamine in a clade of Clostridia

Genetic code inference

In our analysis of all sequenced bacterial genomes, we noticed that 6 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 *Clostridium hiranonis*. Four of these genomes are included in the Genome Taxonomy Database (GTDB) phylogeny, where they are assigned to three species clusters within the GTDB genus Clostridium_U (family Peptostreptococcaceae, order Peptostreptococcales, class Clostridia), which additionally includes *Peptacetobacter hominis* (our program did not infer an amino acid translation for CGG in this species).

Species phylogenetic tree

In Figure 4, we show a part of the GTDB phylogenetic tree that includes all members of the GTDB genus Clostridium_U (4 species clusters) and outgroup species including the remaining species in the family Peptostreptococcaceae and the families Peptoclostridiaceae and Filifactoraceae (52 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 genome, we marked the most complete genomes on the tree

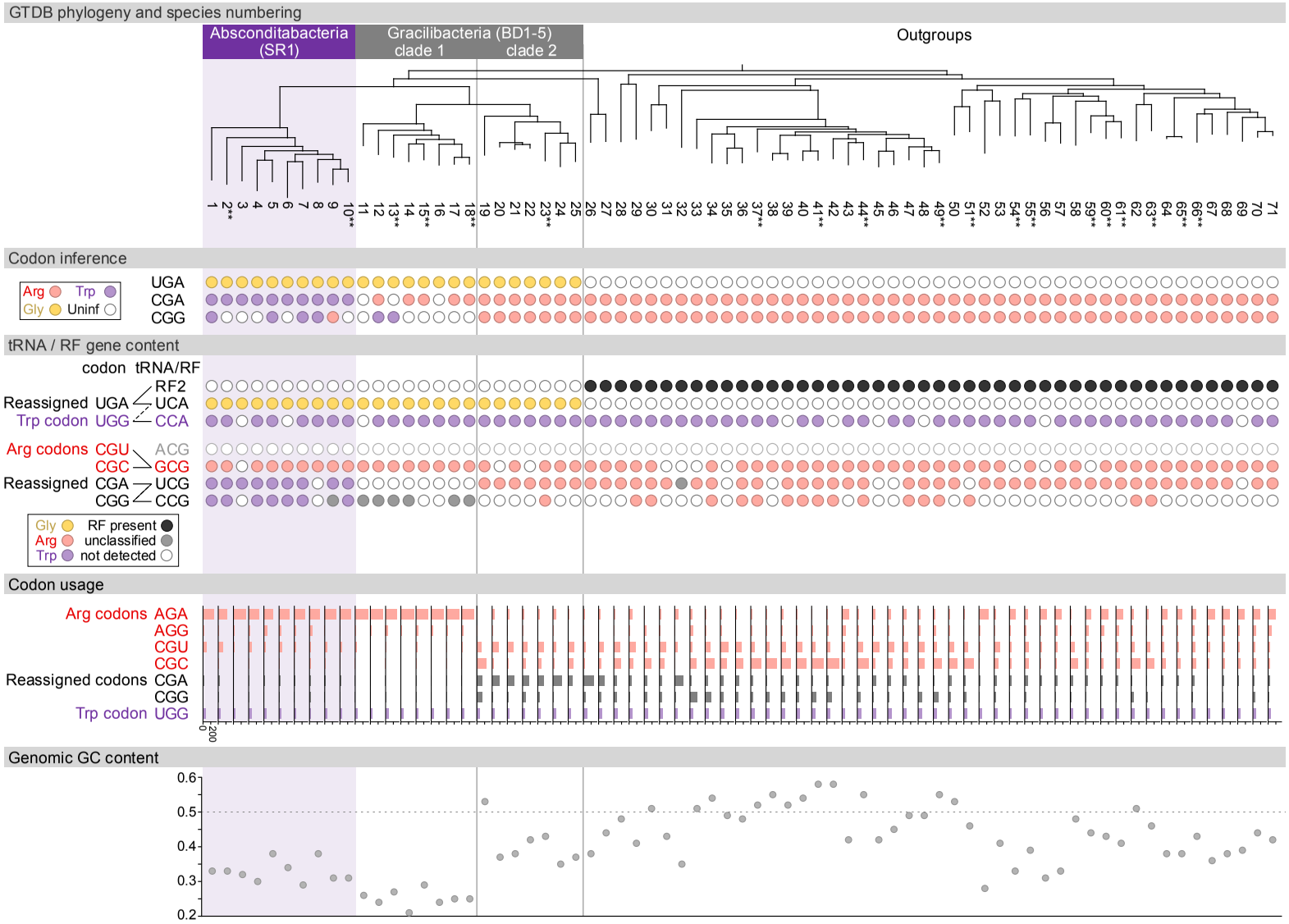
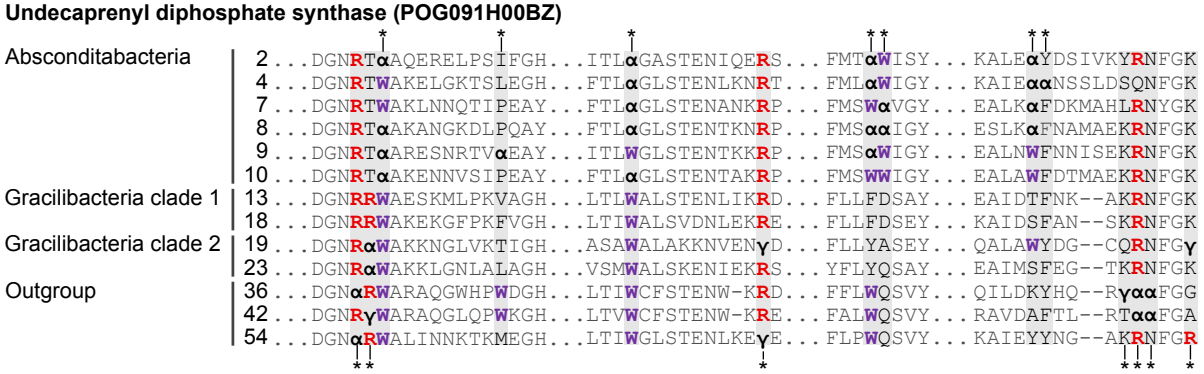

GTDB phylogeny and species numbering
Absconditabacteria (SR1)
Gracilibacteria (BD1-5) clade 1
clade 2
Outgroups
1 2** 3 4 5 6 7 8 9 10** 11 12 13** 14 15** 16 17 18** 19 20 21 22 23** 24 25 26 27 28 29 30 31 32 33 34 35 36 37** 38 39 40 41** 42 43 44** 45 46 47 48 49** 50 51** 52 53 54** 55** 56 57 58 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

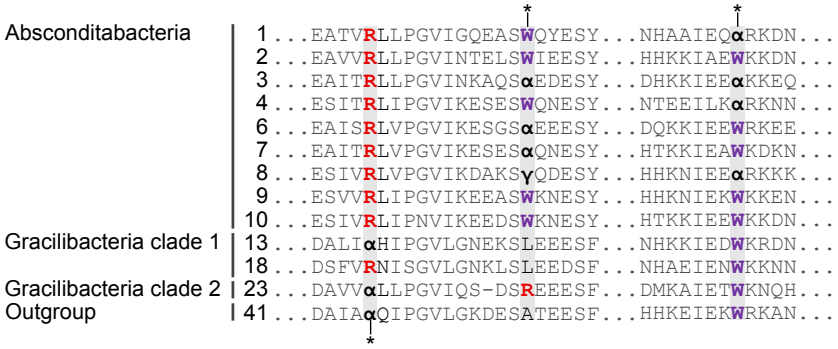
Figure 1: Phylogenetic tree from GTDB showing the Absconditabacteria, Gracilibacteria, and closest outgroup genomes (Peregrinibacteria and Peribacteria), each species indicated by a number which can be cross-referenced with the summary spreadsheet. Asterisks indicate genomes the most complete genomes, 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 our method 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 positions. Genomic GC content is calculated over the entire genome.

(Figure 4), 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). We 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. We attribute this due to the fact that tRNA genes are known to be highly clustered in Firmicutes (*2*) 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).

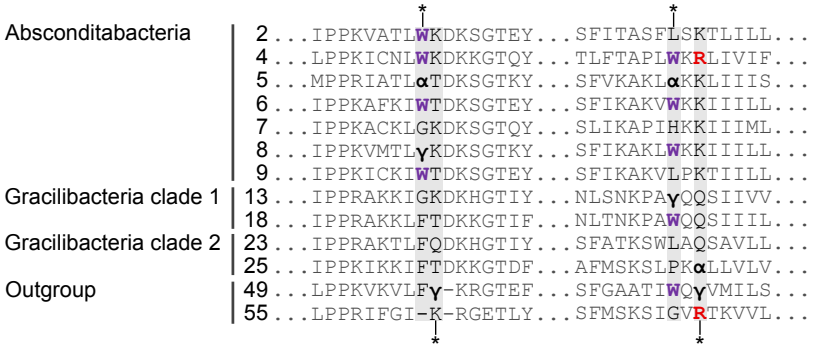
Undecaprenyl diphosphate synthase (POG091H00BZ)



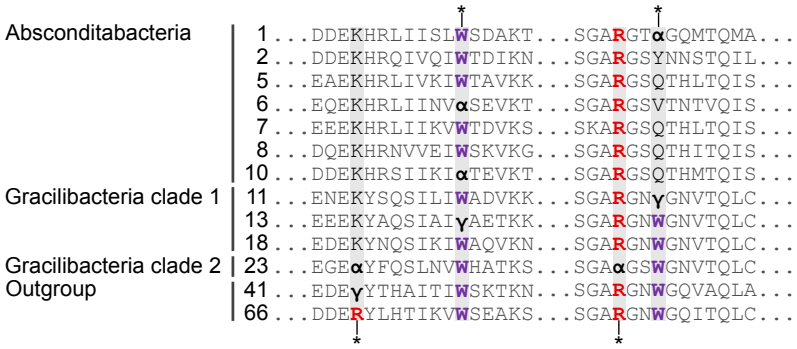
tRNA (guanine-N1)-methyltransferase (POG091H01WE)



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.

Reassignment of CGG to glutamine in Clostridium_U

In the Clostridium_U clade on the GTDB tree (Figure 4), 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 *Clostridium hiranonis* (species #1 on tree) was inferred to translate CGG as arginine; however, we 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 positions).

Figure 5 shows four example alignments of conserved single-copy bacterial genes across Clostridium_U and outgroup species. The four aligned genes contain several positions encoded by CGG in Clostridium_U genomes and these residues are primarily conserved for glutamine in other species. The occurence of CGG at conserved glutamine positions supports CGG being used as a glutamine codon in the reassigned clade.

In all Clostridium_U genomes, CGG is decoded by a $tRNA_{CCG}$. We 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 weak 1:72 base pair in the acceptor stem. Therefore, we could not determine which aminoacyl-tRNA synthetase recognizes the $tRNA_{CCG}$ from the tRNA sequence alone. It is possible that the glutamyl-tRNA synthetase (GlnRS) still recognizes the

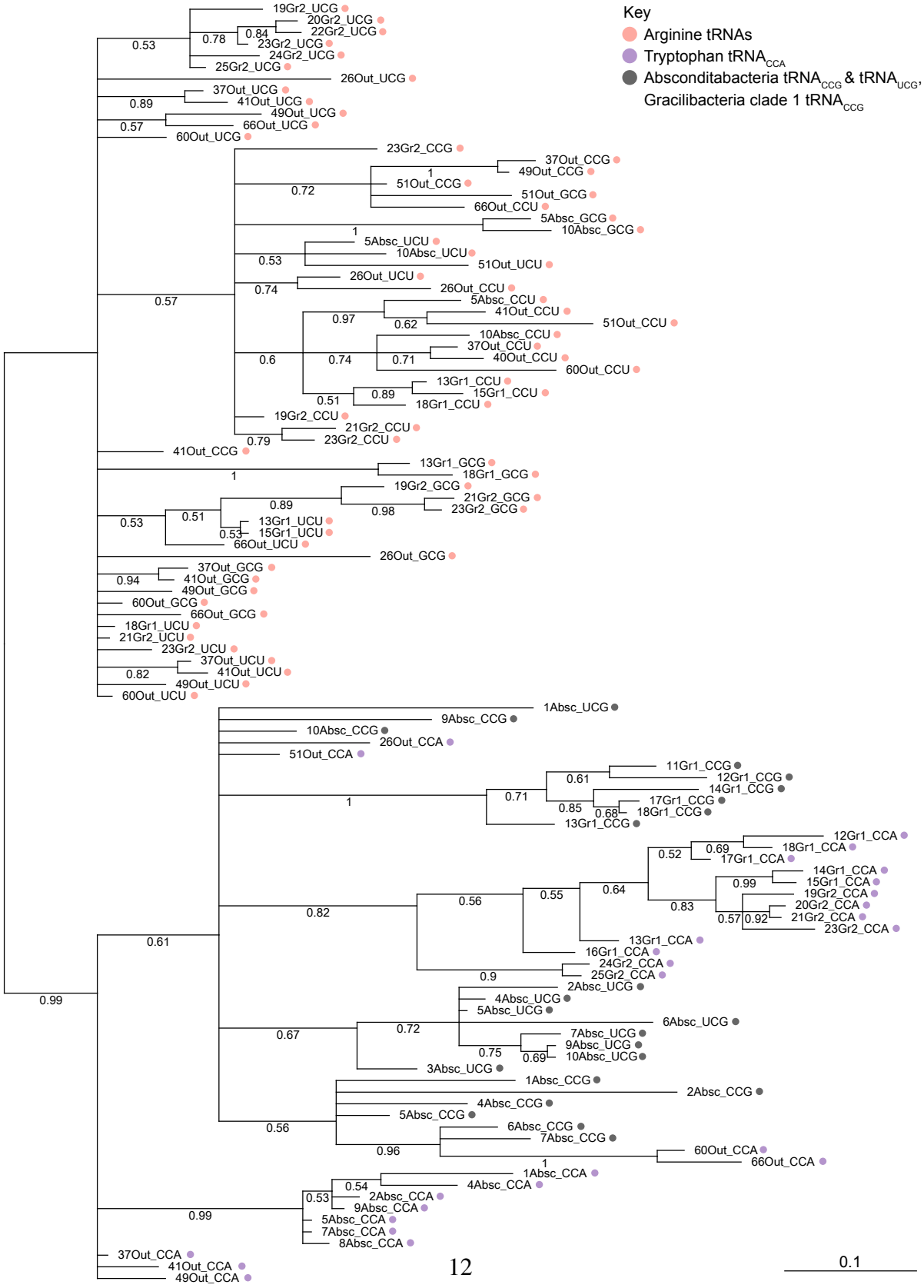

Key
Arginine tRNAs
Tryptophan tRNA$_{CCA}$
Absconditabacteria tRNA$_{CCG}$ & tRNA$_{UCG}$, Gracilibacteria clade 1 tRNA$_{CCG}$
19Gr2_UCG
20Gr2_UCG
22Gr2_UCG
23Gr2_UCG
24Gr2_UCG
25Gr2_UCG
26Out_UCG
37Out_UCG
41Out_UCG
49Out_UCG
66Out_UCG
60Out_UCG
23Gr2_CCG
37Out_CCG
49Out_CCG
51Out_CCG
51Out_GCG
66Out_CCU
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.

$tRNA_{CCG}$ despite missing many of the known identity elements because the relaxed specificity of GlnRS towards non-cognate tRNAs. 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 (*3*). The charging of $tRNA_{CCG}$ could be experimentally confirmed in the future for culturable species such as *Clostridium hiranonis*.

In a tRNA phylogeny built from arginine and glutamine tRNA sequences from Clostridium_U species and outgroups, the reassigned $tRNA_{CCG}$ branches outside of either the cluster of arginine or glutamine tRNAs (Figure 6). We compared the likelihood of two phylogenetic models, one where the Clostridium_U $tRNA_{CCG}$ 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.

In some Clostridium_U 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. This may be due to incomplete genome assembly because some members of Clostridium_U do have these tRNA genes, and all of the four Clostridium_U species are missing at least a few other required tRNA genes.

CGG is a rare codon in all members of Clostridium_U, with usage of 4-6 per 10,000 codons in aligned Pfam domains. The low genomic GC content in Clostridium_U 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 of CGG, this 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 Clostridium_U species by the genetic code inference method, and is supported by examining 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 examination of the tRNA sequence. *Peptacetobacter hominis* (GCA_006861675.1, species #4 on tree) did not have an inferred CGG meaning by the genetic code inference program despite 145 Pfam positions aligned to CGG; however, the top amino acid model for CGG is glutamine (model probability of 0.984, below the 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.

CGG decoding in the outgroup species

In Clostridium_U, 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).

In the sister group to Clostridium_U within the Peptostreptococcaceae (species #5-34 on tree, genera *Clostridioides*, *Asaccharospora*, *Intestinibacter*, *Terrisporobacter*, *Paeniclostridium*, *Paraclostridium*, and *Romboutsia*), we could not find any tRNA capable of decoding CGG (Figure 4, gray box outline). We wanted to confirm that this is not due to incomplete genome assembly, so we analyzed the tRNA gene content of all 2,114 genome assemblies belonging to the 29 species in this clade (predominantly consisting of 2,014 *Clostridiodes 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.

If a tRNA to decode CGG is truly missing in this clade, then the few CGGs in coding regions might be decoded inefficiently by the $tRNA^{Arg}_{ACG}$ or by a non-cognate tRNA. The lack of a tRNA to decode CGG in some species, in conjunction with low GC content, may explain the extremely low usage of CGG in Peptostreptococcaceae genomes.

Reassignment of CGG to tryptophan in a clade of Bacilli

Genetic code inference

In our analysis of all sequenced bacterial genomes, 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 Genome Taxonomy Database (GTDB) phylogeny, where they belong to 4 species clusters that form a monophyletic clade within the GTDB genus g__UBA4855 (family f__CAG-826, order o__RFN20, class c__Bacilli). The clade using the predicted reassignment of AGG to methionine also belongs to the order o__RFN20, but to a different family.

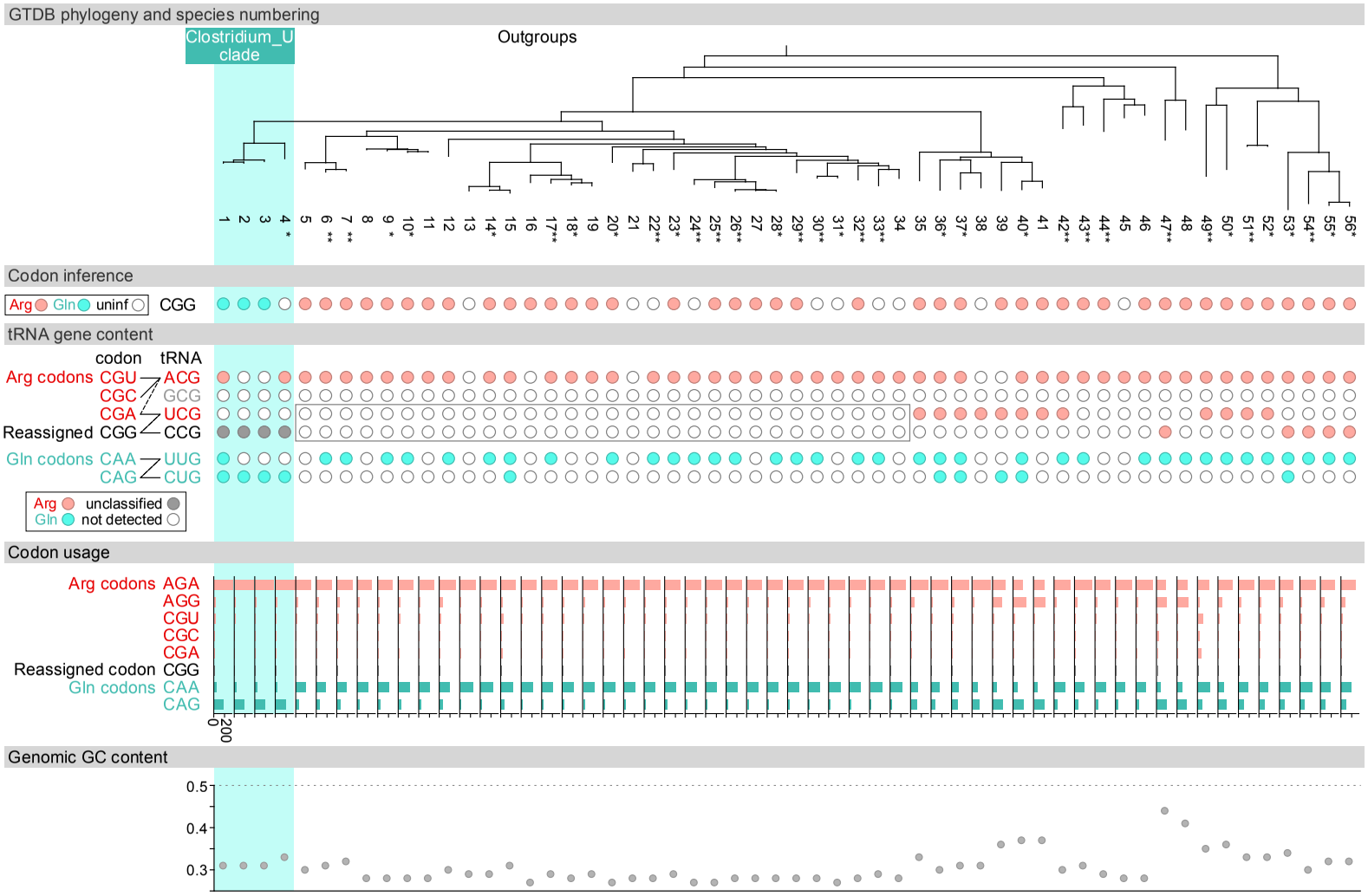

GTDB phylogeny and species numbering
Clostridium_U clade
Outgroups
1 2 3 4 * 5 6 ** 7 ** 8 9 * 10* 11 12 13 14* 15 16 17** 18* 19 20* 21 22** 23* 24** 25** 26** 27 28* 29** 30** 31* 32** 33** 34 35 36* 37* 38 39 40* 41 42** 43** 44** 45 46 47** 48 49** 50* 51** 52* 53* 54** 55* 56*
Codon inference
Arg Gln uninf
CGG
tRNA gene content
codon tRNA
Arg codons CGU ACG
CGC GCG
CGA UCG
Reassigned CGG CCG
Gln codons CAA UUG
CAG CUG
Arg unclassified
Gln not detected
Codon usage
Arg codons AGA
AGG
CGU
CGC
CGA
Reassigned codon CGG
Gln codons CAA
CAG
0
200
Genomic GC content
0.5
0.4
0.3

Figure 4: Phylogenetic tree from GTDB showing the Clostridium_U clade and closest outgroup genomes, each species indicated by a number which can be cross-referenced with the summary spreadsheet. 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 our method 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 the 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 positions. Genomic GC content is calculated over the entire genome.

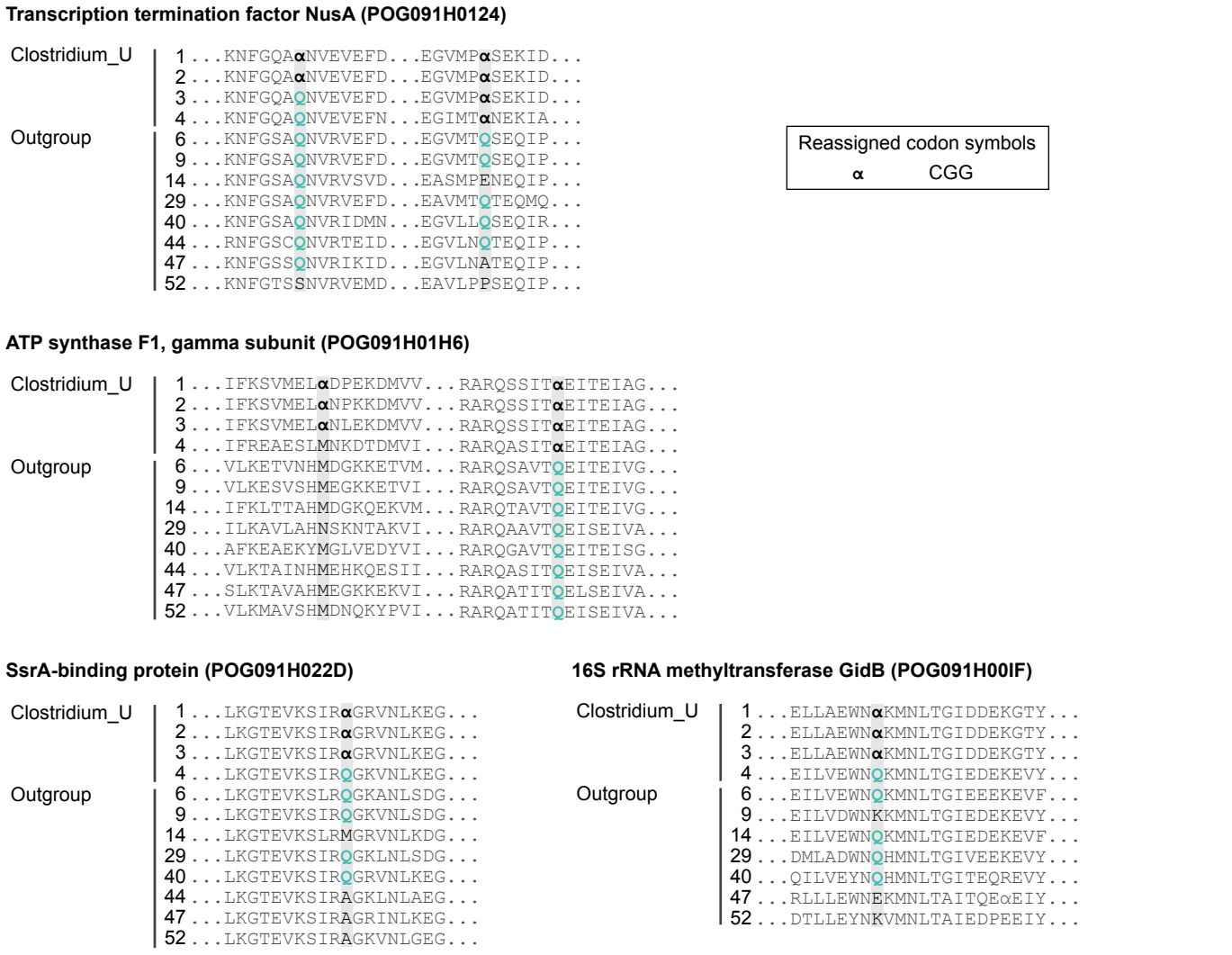


Figure 5: Multiple sequence alignments of transcription termination factor NusA (BUSCO POG091H0124), ATP synthase F1, gamma subunit (BUSCO POG091H01H6), SsrA-binding protein (BUSCO POG091H022D), and 16S rRNA methyltransferase GidB (BUSCO POG091H00IF) from the Clostridium_U clade and selected outgroup species. Alignment regions containing nearby CGG (α) positions are shown, with columns with CGG in Clostridium_U species sequences highlighted.

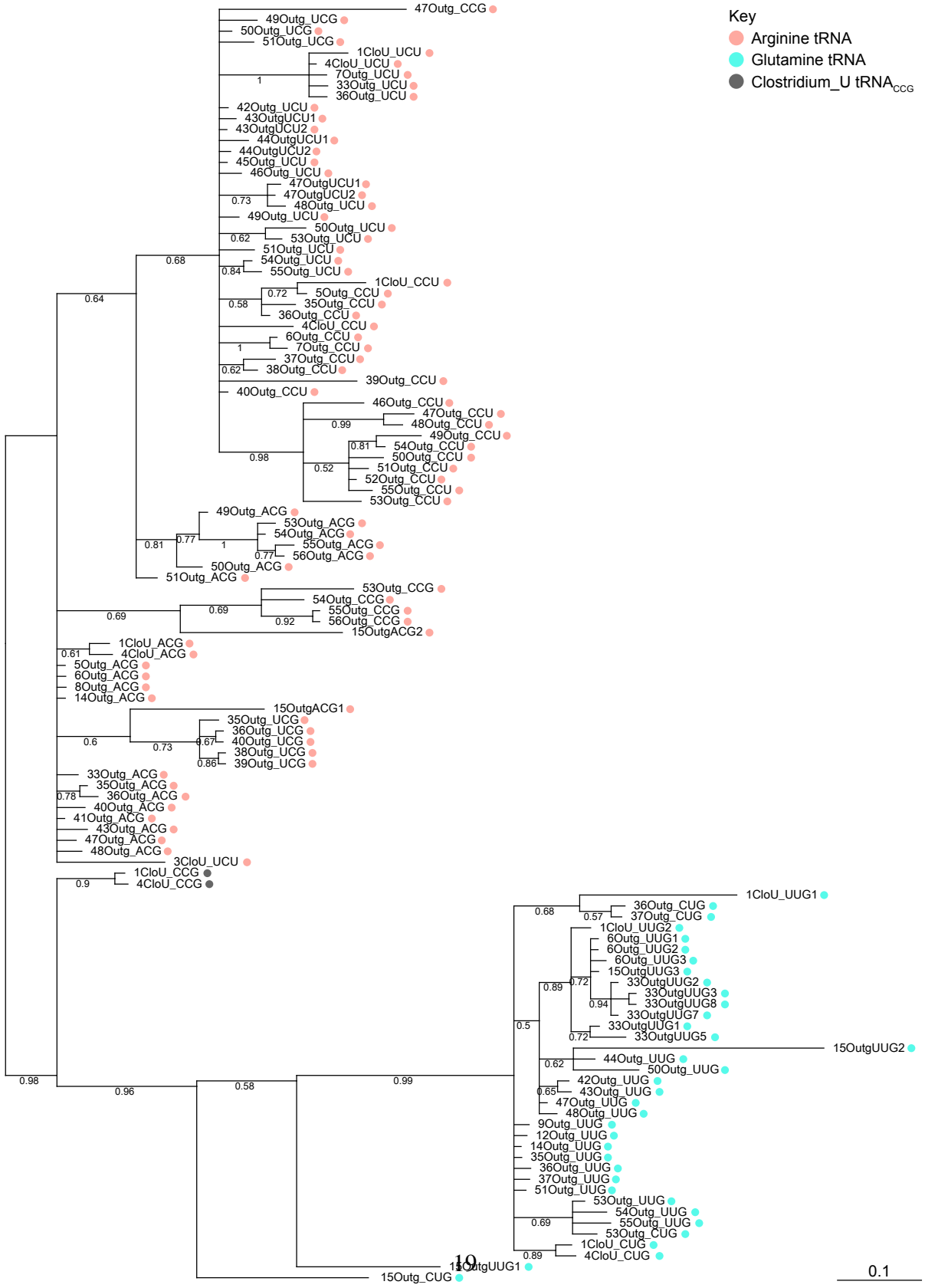

Key
Arginine tRNA
Glutamine tRNA
Clostridium_U tRNA$_{CCG}$
47Outg_CCG
49Outg_UCG
50Outg_UCG
51Outg_UCG
1CloU_UCU
4CloU_UCU
7Outg_UCU
33Outg_UCU
36Outg_UCU
42Outg_UCU
43OutgUCU1
43OutgUCU2
44OutgUCU1
44OutgUCU2
45Outg_UCU
46Outg_UCU
47OutgUCU1
47OutgUCU2
48Outg_UCU
49Outg_UCU
50Outg_UCU
53Outg_UCU
51Outg_UCU
54Outg_UCU
55Outg_UCU
1CloU_CCU
5Outg_CCU
35Outg_CCU
36Outg_CCU
4CloU_CCU
6Outg_CCU
7Outg_CCU
37Outg_CCU
38Outg_CCU
39Outg_CCU
40Outg_CCU
46Outg_CCU
47Outg_CCU
48Outg_CCU
49Outg_CCU
54Outg_CCU
50Outg_CCU
51Outg_CCU
52Outg_CCU
55Outg_CCU
53Outg_CCU
49Outg_ACG
53Outg_ACG
54Outg_ACG
55Outg_ACG
56Outg_ACG
50Outg_ACG
51Outg_ACG
53Outg_CCG
54Outg_CCG
55Outg_CCG
56Outg_CCG
15OutgACG2
1CloU_ACG
4CloU_ACG
5Outg_ACG
6Outg_ACG
8Outg_ACG
14Outg_ACG
15OutgACG1
35Outg_UCG
36Outg_UCG
40Outg_UCG
38Outg_UCG
39Outg_UCG
33Outg_ACG
35Outg_ACG
36Outg_ACG
40Outg_ACG
41Outg_ACG
43Outg_ACG
47Outg_ACG
48Outg_ACG
3CloU_UCU
1CloU_CCG
4CloU_CCG
1CloU_UUG1
36Outg_CUG
37Outg_CUG
1CloU_UUG2
6Outg_UUG1
6Outg_UUG2
6Outg_UUG3
15OutgUUG3
33OutgUUG2
33OutgUUG3
33OutgUUG8
33OutgUUG7
33OutgUUG1
33OutgUUG5
15OutgUUG2
44Outg_UUG
50Outg_UUG
42Outg_UUG
43Outg_UUG
47Outg_UUG
48Outg_UUG
9Outg_UUG
12Outg_UUG
14Outg_UUG
35Outg_UUG
36Outg_UUG
37Outg_UUG
51Outg_UUG
53Outg_UUG
54Outg_UUG
55Outg_UUG
53Outg_CUG
1CloU_CUG
4CloU_CUG
15OutgUUG1
15Outg_CUG
0.1

Figure 6: An unrooted phylogenetic tree of arginine, glutamine, and reassigned tRNA sequences from Clostridium_U 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 Clostridium_U 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.

Species phylogenetic tree

In Figure 7, we 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). We designated the reassigned clade as a 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).

In order to help assess whether a tRNA gene is used by members of a clade or whether it missing due to an incomplete genome, we marked the most complete genomes on the tree (Figure 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.

Reassignment of CGG to tryptophan in the reassigned clade

In the reassigned clade (Figure 7, species #1-5), four species were inferred to translate CGG as tryptophan and one species did not have an inferred meaning for CGG (species s__UBA4855 sp002438785) due to one or two aligned Pfam positions 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 positions).

Figure 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 we 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. 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. 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 an incomplete genome assembly, tRNA gene detection failure, or a bona fide inability to translate CGG.

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 9). We 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.

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). 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. 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). If low genomic GC content favored substitutions away from CGG prior to reassignment of CGG, this 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 tryptophan positions in proteins has led to CGG being inferred as a tryptophan codon in reassigned species by the genetic code inference method, and is supported by examining 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 the genetic code inference program 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 GTDB genus g__UBA4855, which is characterized by low genomic GC content and extremely low codon usage of CGG.

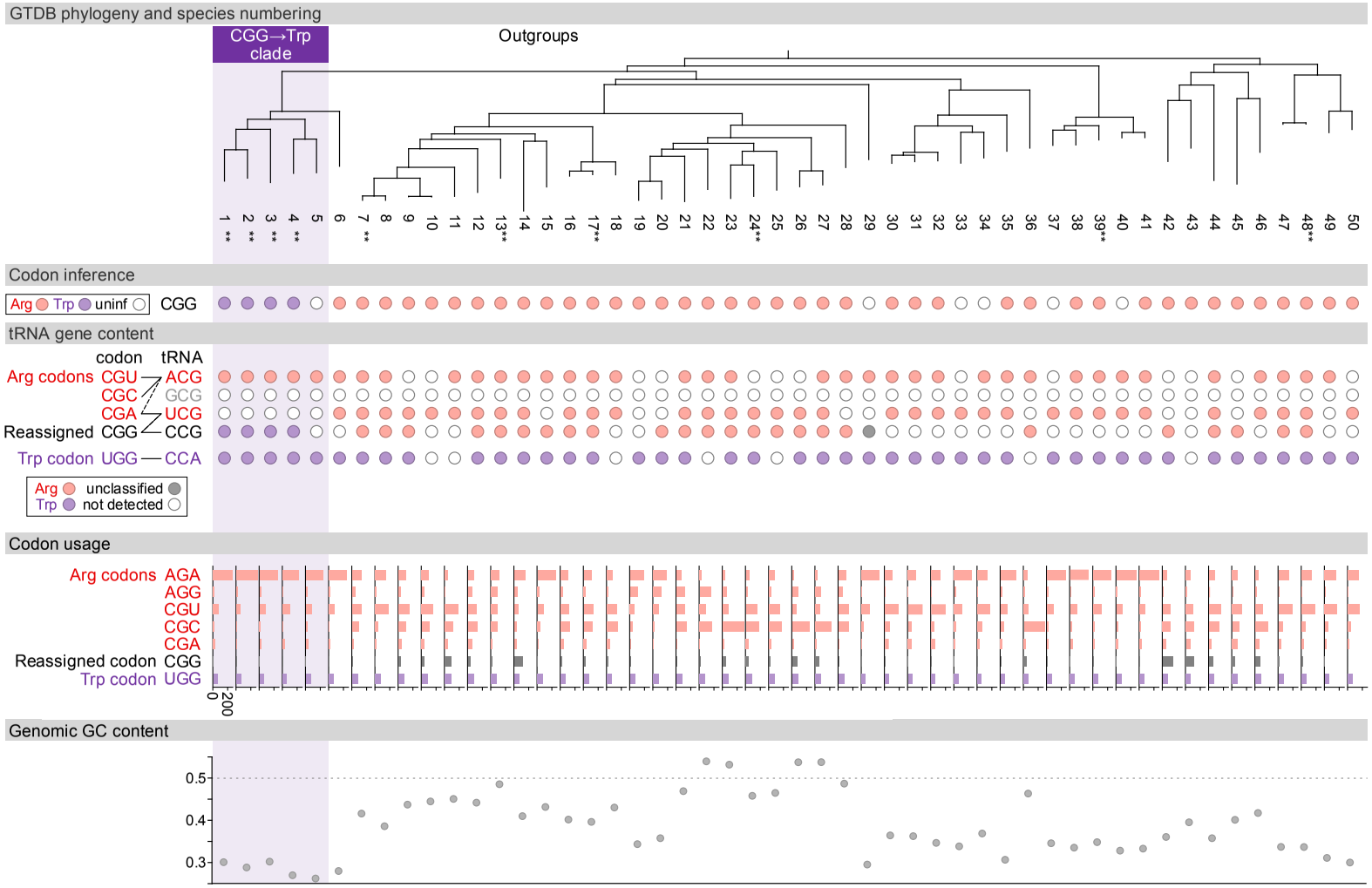

GTDB phylogeny and species numbering
CGG→Trp clade
Outgroups
1 ** 2 ** 3 ** 4 ** 5 6 7 ** 8 9 10 11 12 13** 14 15 16 17** 18 19 20 21 22 23 24** 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39** 40 41 42 43 44 45 46 47 48** 49 50
Codon inference
Arg Trp uninf
CGG
tRNA gene content
codon tRNA
Arg codons CGU ACG
CGC GCG
CGA UCG
Reassigned CGG CCG
Trp codon UGG CCA
Arg unclassified
Trp not detected
Codon usage
Arg codons AGA
AGG
CGU
CGC
CGA
Reassigned codon CGG
Trp codon UGG
0 200
Genomic GC content
0.5
0.4
0.3

Figure 7: Phylogenetic tree from GTDB showing the Bacilli CGG→Trp clade and closest outgroup genomes, each species indicated by a number which can be cross-referenced with the summary spreadsheet. 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 our method 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 the 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 positions. Genomic GC content is calculated over the entire genome.

Reassignment of CGG to tryptophan in *Anaerococcus*

Genetic code inference

In our analysis of all sequenced bacterial genomes, 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 Genome Taxonomy Database (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

In Figure 10A, we 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 assess whether a tRNA gene is used by members of a clade

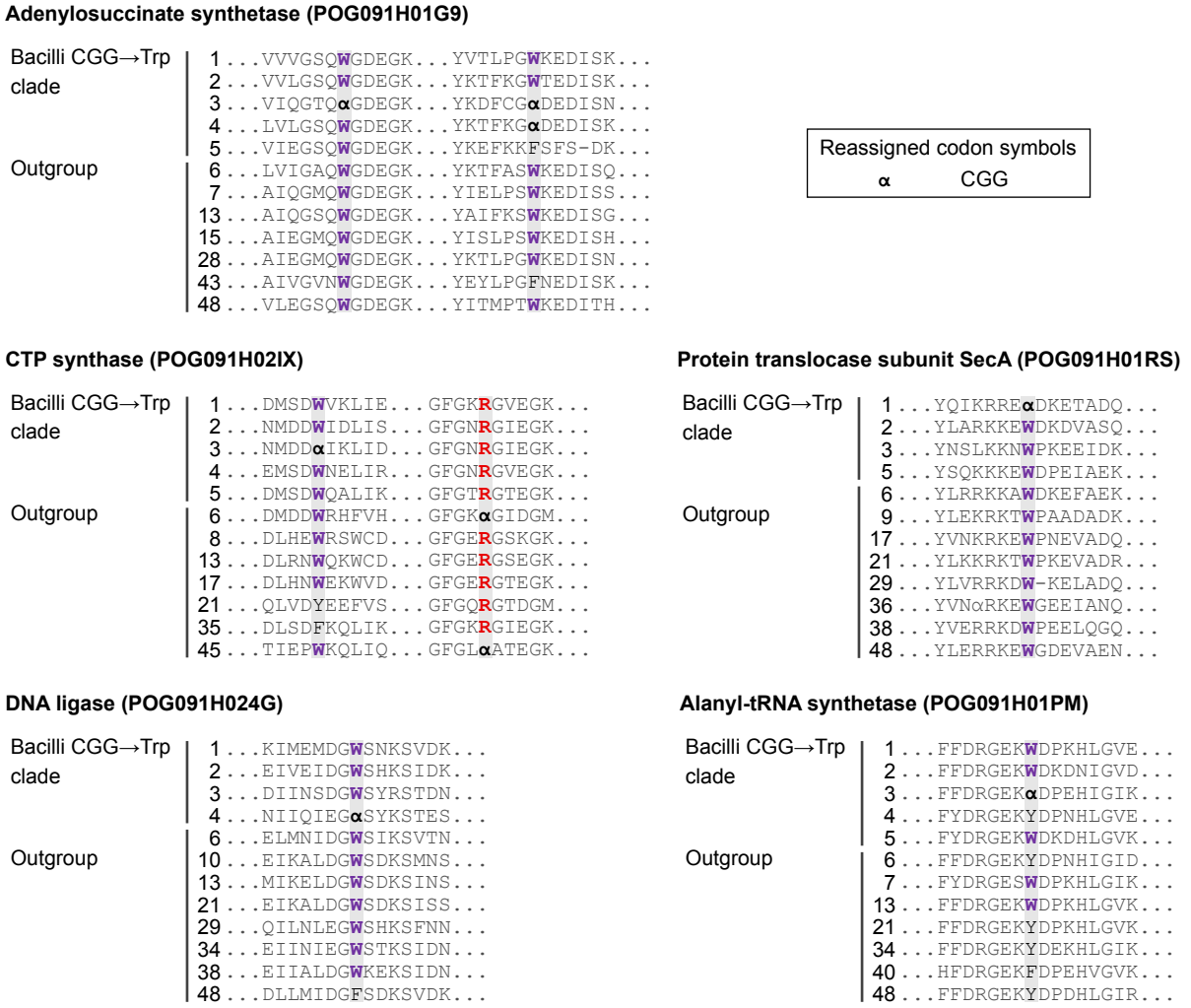


Figure 8: Multiple sequence alignments of adenylosuccinate synthetase (BUSCO POG091H01G9), CTP synthase (BUSCO POG091H02IX), protein translocase subunit SecA (BUSCO POG091H01RS), DNA ligase (BUSCO POG091H024G), and alanyl-tRNA synthetase (BUSCO POG091H01PM) from the Bacilli CGG→Trp clade and selected outgroup species. Alignment regions containing CGG (α) at conserved positions are shown, with columns with CGG in Bacilli CGG→Trp clade and the closest outgroup (species #6) sequences highlighted.

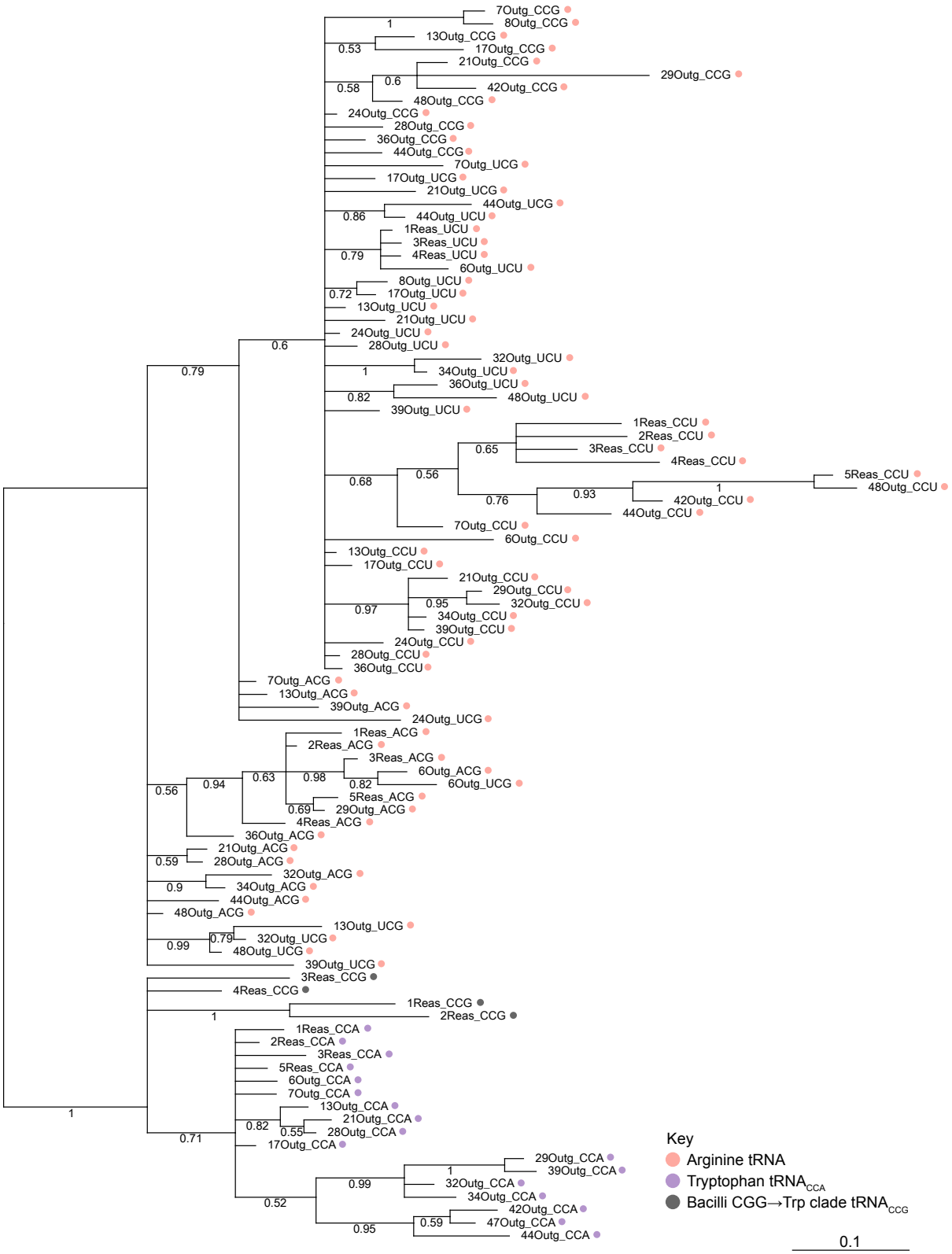

7Outg_CCG
8Outg_CCG
13Outg_CCG
17Outg_CCG
21Outg_CCG
29Outg_CCG
42Outg_CCG
48Outg_CCG
24Outg_CCG
28Outg_CCG
36Outg_CCG
44Outg_CCG
7Outg_UCG
17Outg_UCG
21Outg_UCG
44Outg_UCG
44Outg_UCU
1Reas_UCU
3Reas_UCU
4Reas_UCU
6Outg_UCU
8Outg_UCU
17Outg_UCU
13Outg_UCU
21Outg_UCU
24Outg_UCU
28Outg_UCU
32Outg_UCU
34Outg_UCU
36Outg_UCU
48Outg_UCU
39Outg_UCU
1Reas_CCU
2Reas_CCU
3Reas_CCU
4Reas_CCU
5Reas_CCU
48Outg_CCU
42Outg_CCU
44Outg_CCU
7Outg_CCU
6Outg_CCU
13Outg_CCU
17Outg_CCU
21Outg_CCU
29Outg_CCU
32Outg_CCU
34Outg_CCU
39Outg_CCU
24Outg_CCU
28Outg_CCU
36Outg_CCU
7Outg_ACG
13Outg_ACG
39Outg_ACG
24Outg_UCG
1Reas_ACG
2Reas_ACG
3Reas_ACG
6Outg_ACG
6Outg_UCG
5Reas_ACG
29Outg_ACG
4Reas_ACG
36Outg_ACG
21Outg_ACG
28Outg_ACG
32Outg_ACG
34Outg_ACG
44Outg_ACG
48Outg_ACG
13Outg_UCG
32Outg_UCG
48Outg_UCG
39Outg_UCG
3Reas_CCG
4Reas_CCG
1Reas_CCG
2Reas_CCG
1Reas_CCA
2Reas_CCA
3Reas_CCA
5Reas_CCA
6Outg_CCA
7Outg_CCA
13Outg_CCA
21Outg_CCA
28Outg_CCA
17Outg_CCA
29Outg_CCA
39Outg_CCA
32Outg_CCA
34Outg_CCA
42Outg_CCA
47Outg_CCA
44Outg_CCA
Key
Arginine tRNA
Tryptophan tRNA$_{CCA}$
Bacilli CGG→Trp clade tRNA$_{CCG}$
0.1

Figure 9: An unrooted phylogenetic tree of arginine, tryptophan, and reassigned tRNA sequences from Bacilli 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 Bacilli 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.

or whether it missing due to an incomplete genome, we marked the most complete genomes on the tree (Figure 10A), 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.

Reassignment of CGG to tryptophan in the reassigned clade

In *Anaerococcus* (Figure 10A, species #1-23), two species were inferred to translate CGG as tryptophan, two species were inferred to translate CGG was arginine, and the remaining 19 species did not have an inferred meaning for CGG due to fewer than 18 aligned Pfam positions at CGG codons. We refer to all of *Anaerococcus* as the reassigned clade based on the presence of the tryptophan-type $tRNA_{CCG}$ (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 positions).

We 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 10B).

In *Anaerococcus*, all 23 species contain 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. 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*, we built a tRNA phylogenetic tree from sequences of arginine and tryptophan tRNAs from *Anaerococcus* and outgroup species (Figure 11). In this phylogenetic tree, the *Anaerococcus* $tRNA_{CCG}$ sequences form a clade 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. We 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, favoring the tryptophan model. This indicates that the *Anaerococcus* $tRNA_{CCG}$ genes are moer similar in sequences to $tRNA^{Trp}_{CCA}$ than 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 the genetic code inference method, while most species did not have an inferred amino acid for CGG and two species were predicted to translate CGG as arginine. 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 positions to make a confident amino acid prediction. Despite the sparse distribution of CGG translation as tryptophan, all *Anaerococcus* species encode a $tRNA_{CCG}$ gene that is consistent with a tryptophan isotype.

CGG translation in the outgroup

The outgroup consists of all remaining 22 species in the family f__Helcococcaceae. All species were predicted to translate CGG as either arginine or no amino acid meaning was predicted. 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, we 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 11, gray box outline). On the tRNA phylogenetic tree, *Finegoldia* $tRNA_{CCG}$ sequences form a clade within other tryptophan tRNAs (Figure 11). We compared the likelihood of two phylogenetic models, one where the *Finegoldia* $tRNA_{CCG}$ sequences are constrainted 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, indicated 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 positions for the genetic code inference method to predict an amino acid inference, and no CGGs could be found in well-aligned regions of BUSCO alignments. It is possible that this represents an early stage in a reassignment of CGG to tryptophan. It is unclear whether the *Finegoldia* and *Anaerococcus* $tRNA_{CCG}$ genes are related or independently evolved.

References

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

2. T. T. T. Tran, H. Belahbib, V. Bonnefoy, E. Talla, *Genome Biology and Evolution* **8**, 282 (2015).

3. J. Normanly, L. G. Kleina, J. M. Masson, J. Abelson, J. H. Miller, *Journal of Molecular Biology* **213**, 719 (1990).

B **DNA ligase (POG091H024G)**

Anaerococcus	1	...KFEAEEYTTTLREVV**W**NVGRSGKVTPSAILDP...
	3	...KFEAEEYTTTLRKVV**W**NVGRTGKVTPSAILDP...
	4	...KYEAEEFTTTLKEVV**W**NVGRTGKVTPSAILEP...
	5	...KFEPEEFTTKLIDVV**W**NVGRTGKVTPSAILEP...
	11	...KFEAEEYTTTLLDVV**W**NVGRTGKVTPSAILEP...
	14	...KFEAEEYTTTLLDVV**α**NVGRTGKVTPSALLEP...
	15	...KFEAEEYTTTLLDVV**W**NVGRTGKVTPSAILEP...
	21	...KYDPEEYTTKLLDVV**W**NVGRTGKVTPSAILEP...
Outgroup	27	...KFEAEEVTTILαSVE**W**NVGαTGKVTPIALLDP...
	30	...KFEAEEVTTILQAVE**W**NVGRTGKVTPTAQLDP...
	39	...KFEPEEVTTILKEVI**W**NVGRTGKVTPTAILEP...
	41	...KFEAEEYSTILKEVV**W**NVGRTGKVTPTAILEP...

Reassigned codon symbols	
α	CGG

Figure 10: (A) Phylogenetic tree from GTDB showing *Anaerococcus* and closest outgroup genomes, each species indicated by a number which can be cross-referenced with the summary spreadsheet. 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 our method 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 the 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 positions. 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 (α) at conserved positions is shown, with a column with CGG in a single *Anaerococcus* species highlighted.

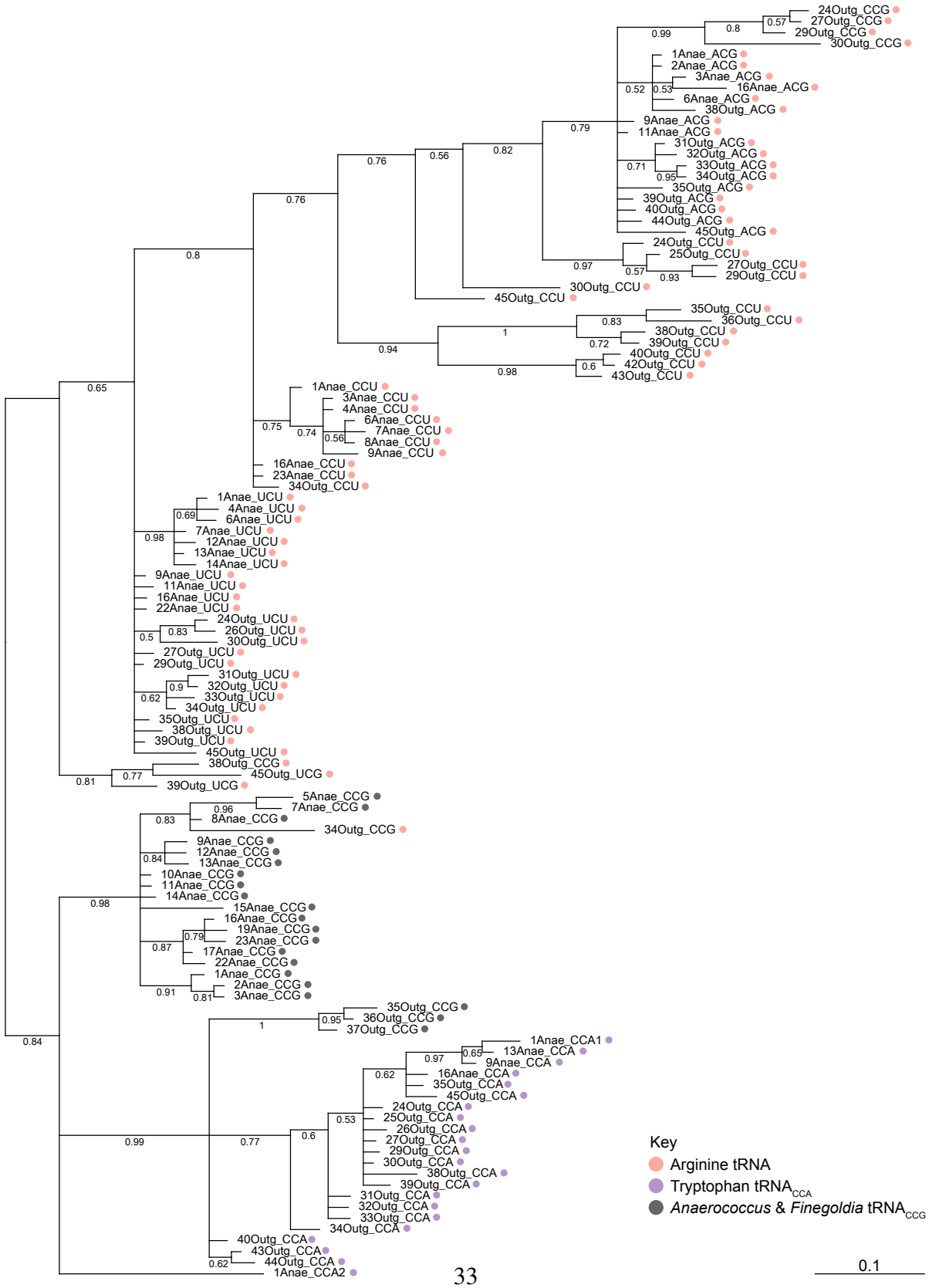

24Outg_CCG
27Outg_CCG
29Outg_CCG
30Outg_CCG
1Anae_ACG
2Anae_ACG
3Anae_ACG
16Anae_ACG
6Anae_ACG
38Outg_ACG
9Anae_ACG
11Anae_ACG
31Outg_ACG
32Outg_ACG
33Outg_ACG
34Outg_ACG
35Outg_ACG
39Outg_ACG
40Outg_ACG
44Outg_ACG
45Outg_ACG
24Outg_CCU
25Outg_CCU
27Outg_CCU
29Outg_CCU
30Outg_CCU
45Outg_CCU
35Outg_CCU
36Outg_CCU
38Outg_CCU
39Outg_CCU
40Outg_CCU
42Outg_CCU
43Outg_CCU
1Anae_CCU
3Anae_CCU
4Anae_CCU
6Anae_CCU
7Anae_CCU
8Anae_CCU
9Anae_CCU
16Anae_CCU
23Anae_CCU
34Outg_CCU
1Anae_UCU
4Anae_UCU
6Anae_UCU
7Anae_UCU
12Anae_UCU
13Anae_UCU
14Anae_UCU
9Anae_UCU
11Anae_UCU
16Anae_UCU
22Anae_UCU
24Outg_UCU
26Outg_UCU
30Outg_UCU
27Outg_UCU
29Outg_UCU
31Outg_UCU
32Outg_UCU
33Outg_UCU
34Outg_UCU
35Outg_UCU
38Outg_UCU
39Outg_UCU
45Outg_UCU
38Outg_CCG
45Outg_UCG
39Outg_UCG
5Anae_CCG
7Anae_CCG
8Anae_CCG
34Outg_CCG
9Anae_CCG
12Anae_CCG
13Anae_CCG
10Anae_CCG
11Anae_CCG
14Anae_CCG
15Anae_CCG
16Anae_CCG
19Anae_CCG
23Anae_CCG
17Anae_CCG
22Anae_CCG
1Anae_CCG
2Anae_CCG
3Anae_CCG
35Outg_CCG
36Outg_CCG
37Outg_CCG
1Anae_CCA1
13Anae_CCA
9Anae_CCA
16Anae_CCA
35Outg_CCA
45Outg_CCA
24Outg_CCA
25Outg_CCA
26Outg_CCA
27Outg_CCA
29Outg_CCA
30Outg_CCA
38Outg_CCA
39Outg_CCA
31Outg_CCA
32Outg_CCA
33Outg_CCA
34Outg_CCA
40Outg_CCA
43Outg_CCA
44Outg_CCA
1Anae_CCA2
Key
Arginine tRNA
Tryptophan tRNA$_{CCA}$
Anaerococcus & *Finegoldia* tRNA$_{CCG}$
0.1

Figure 11: 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.