Reassignment of CGG to glutamine in *Peptacetobacter*

In support of this 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 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

In our analysis by Codetta 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 *Peptacetobacter 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 (also known as *Peptacetobacter*; GTDB 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 1, 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 (Figure 1), 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 (***?***) 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).

Reassignment of CGG to glutamine in Clostridium_U

In the Clostridium_U clade on the GTDB tree (Figure 1), 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, 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 2 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 $tRNA_{CCG}$ 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 (***?***). The 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 Clostridium_U species and outgroups, the reassigned $tRNA_{CCG}$ branches outside of either the cluster of arginine or glutamine tRNAs (Figure 3). 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 Codetta, 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 Codetta 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 1).

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 1, 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 $\text{tRNA}^{\text{Arg}}_{\text{ACG}}$ 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.

Methods

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

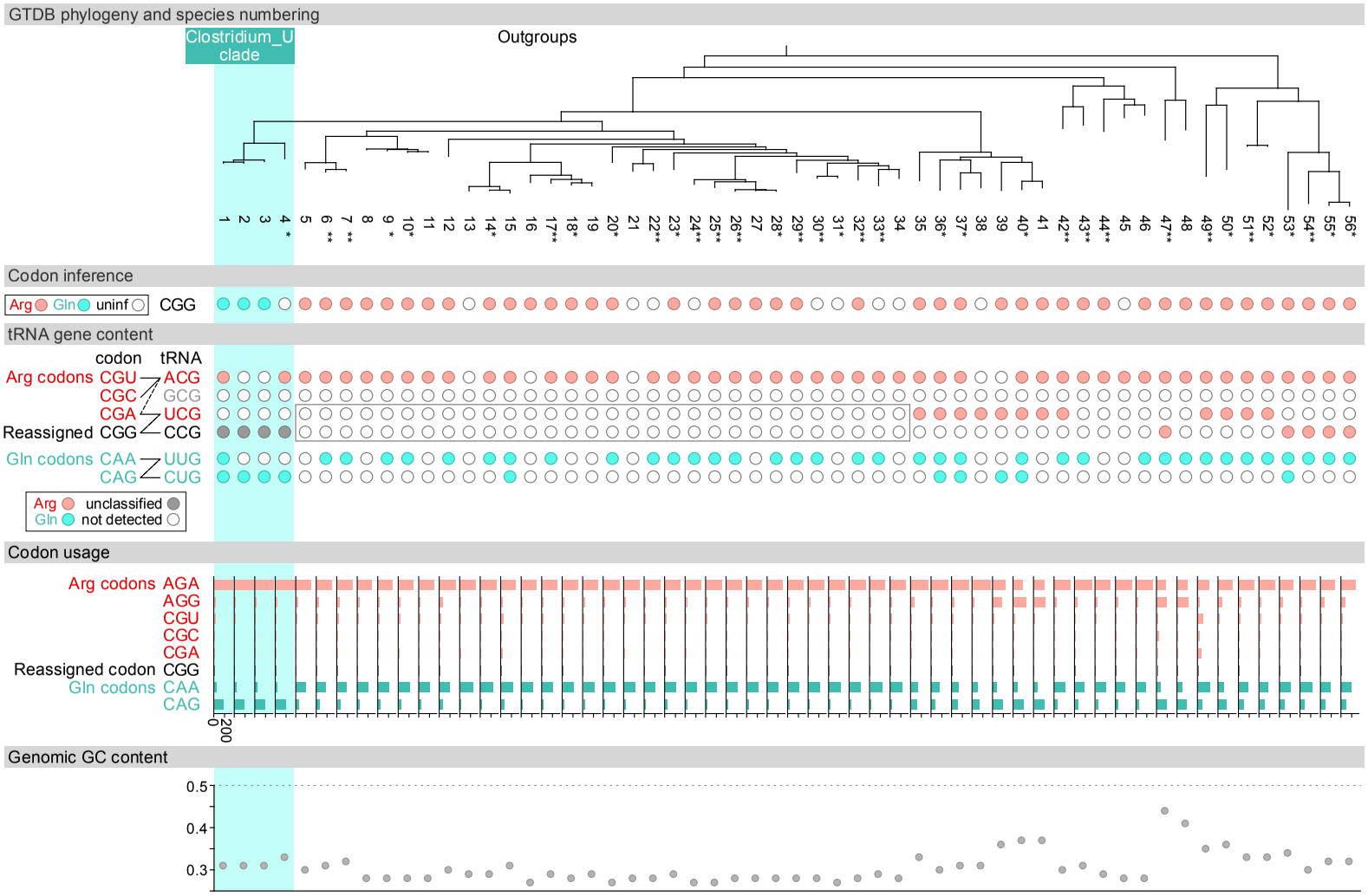
GTDB phylogeny and species numbering
Clostridium_U clade
Outgroups
1 2 3 4 * 5 6 ** 7 ** 8 9 * 10* 11 12 13 14* 15 16 17** 18* 19 20* 21 22** 23* 24** 25** 26** 27 28* 29** 30** 31* 32** 33** 34 35 36* 37* 38 39 40* 41 42** 43** 44** 45 46 47** 48 49** 50* 51** 52* 53* 54** 55* 56*
Codon inference
Arg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

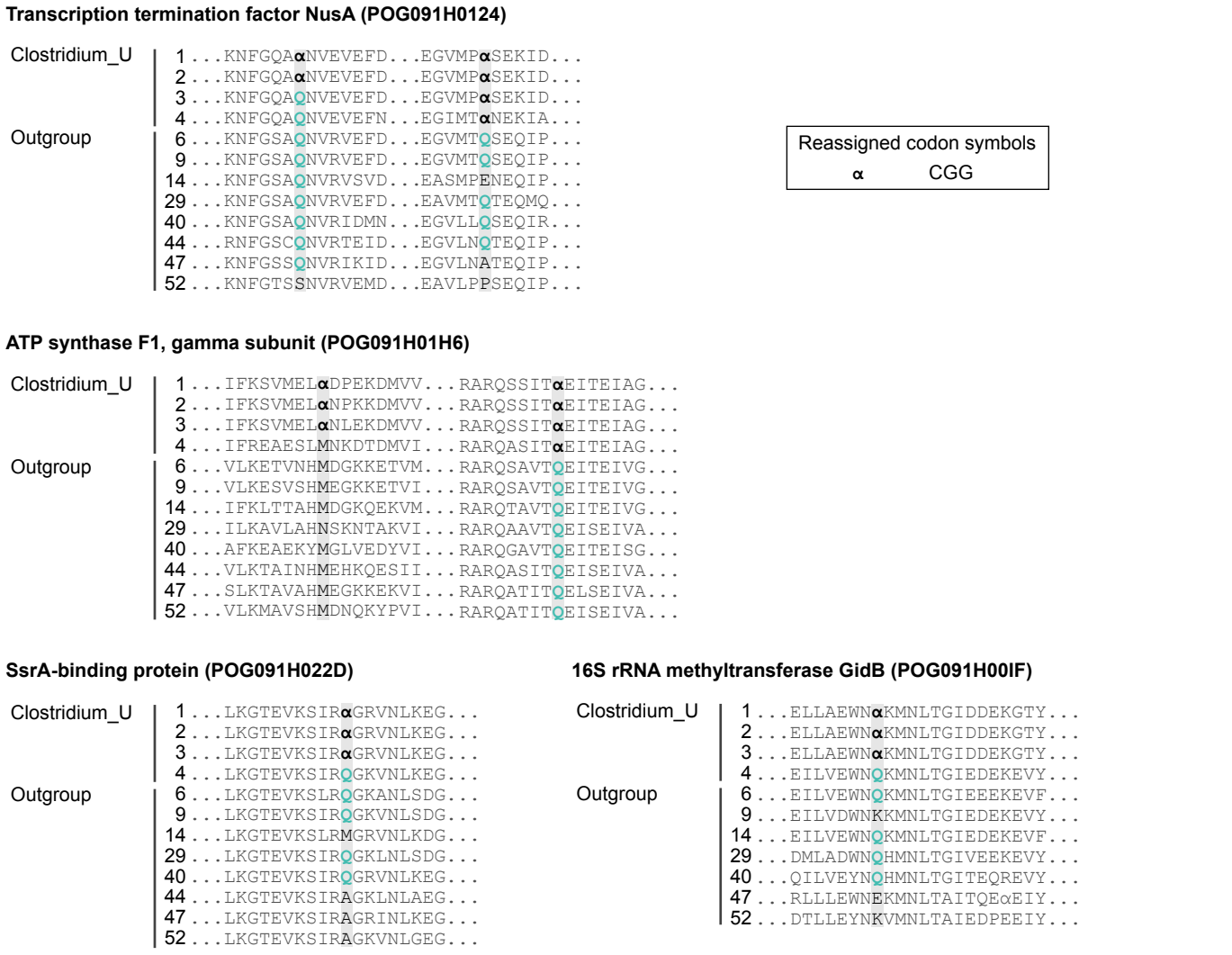

Figure 2: Multiple sequence alignments of transcription termination factor NusA (BUSCO POG091H0124), ATP synthase F1, gamma subunit (BUSCO POG091H01H6), SsrA-binding protein (BUSCO POG091H022D), and 16S rRNA methyltransferase GidB (BUSCO POG091H00IF) from the Clostridium_U clade and selected outgroup species. Alignment regions containing nearby CGG (α) positions are shown, with columns with CGG in Clostridium_U species sequences highlighted.

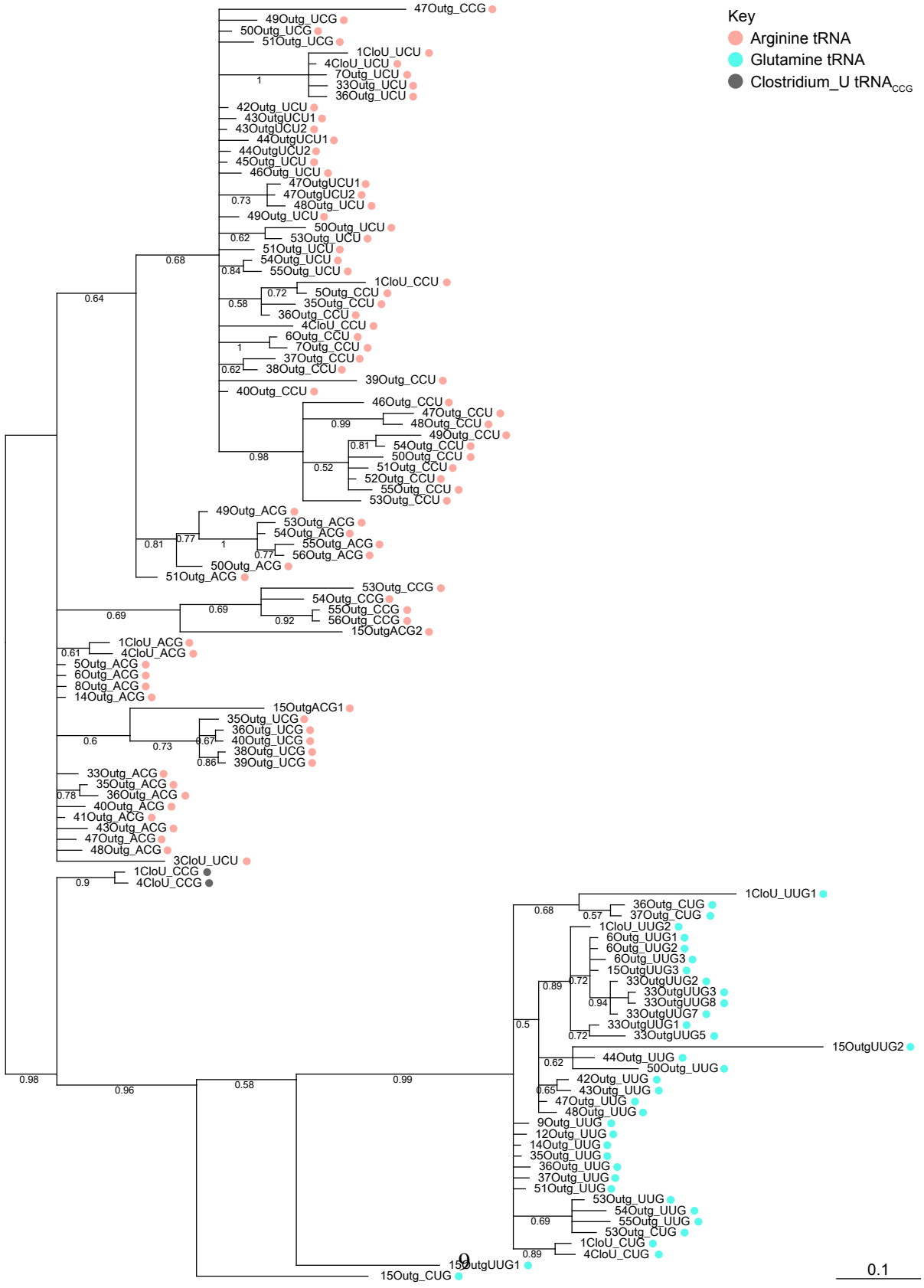
Key
Arginine tRNA
Glutamine tRNA
Clostridium_U tRNA$_{CCG}$
47Outg_CCG
49Outg_UCG
50Outg_UCG
51Outg_UCG
1CloU_UCU
4CloU_UCU
7Outg_UCU
33Outg_UCU
36Outg_UCU
42Outg_UCU
43OutgUCU1
43OutgUCU2
44OutgUCU1
44OutgUCU2
45Outg_UCU
46Outg_UCU
47OutgUCU1
47OutgUCU2
48Outg_UCU
49Outg_UCU
50Outg_UCU
53Outg_UCU
51Outg_UCU
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 3: 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.

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 clade and outgroups. Trees were inferred using a Bayesian approach implemented in MrBayes 3.2.7a 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 likelhood 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).

References

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

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