Reassignment of CGG to tryptophan in a clade of Bacilli

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 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, 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 Bacilli clade using the predicted reassignment of AGG to methionine also belongs to the order o__RFN20, but to a different family.

Species phylogenetic tree

In Figure 1, 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) due to the tree topology.

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

Figure 2 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 3). 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 Codetta, 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 Codetta and does not have a tRNA_{CCG}, but is included in the reassigned clade because it is placed in the same clade as the reassigned species according to the GTDB tree. The entire reassigned clade plus the closest outgroup species s__UBA4855 sp002451465 (species #6 on tree) form the GTDB genus g__UBA4855, which is characterized by low genomic GC content and extremely low codon usage of CGG.

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

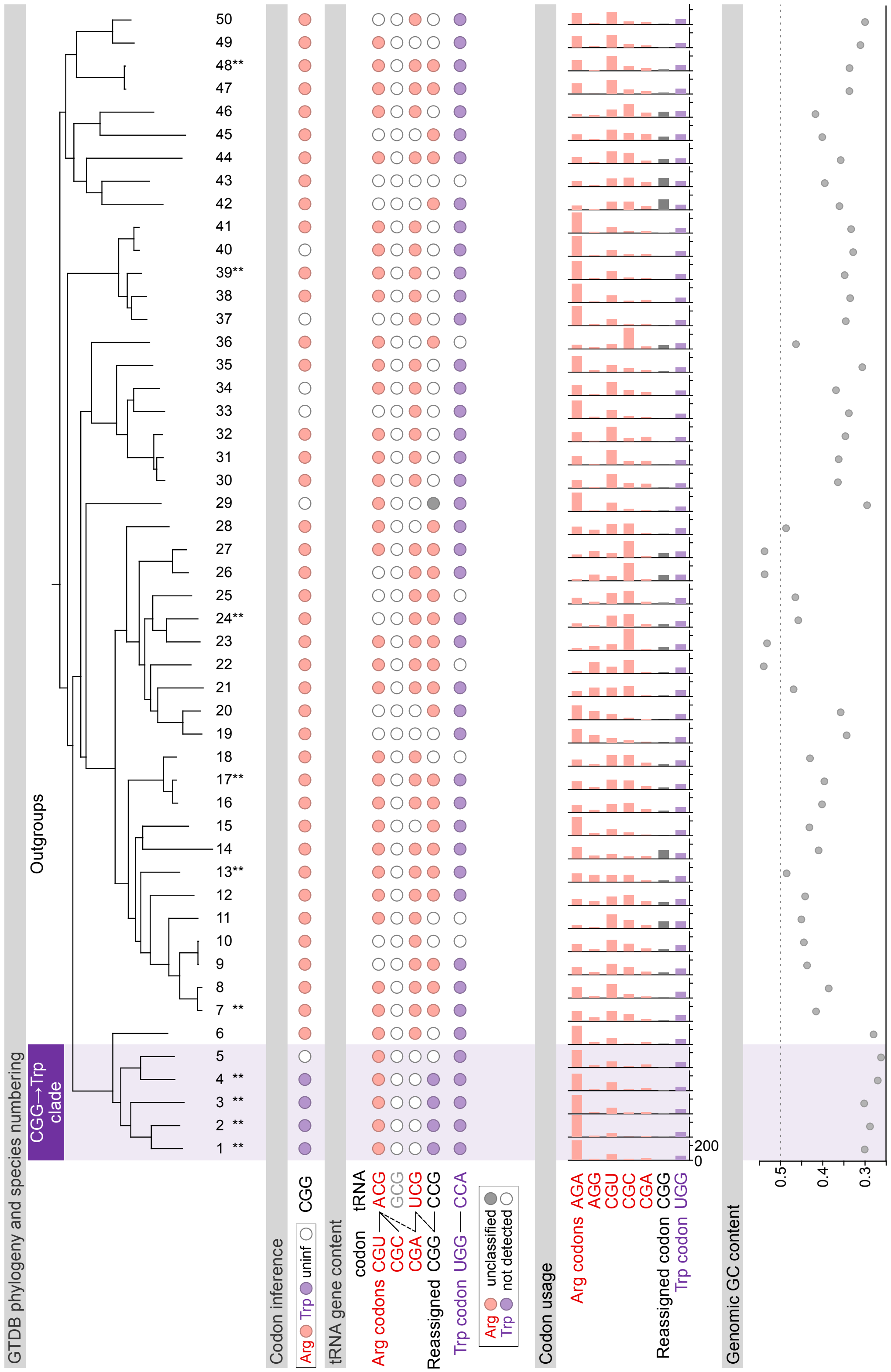
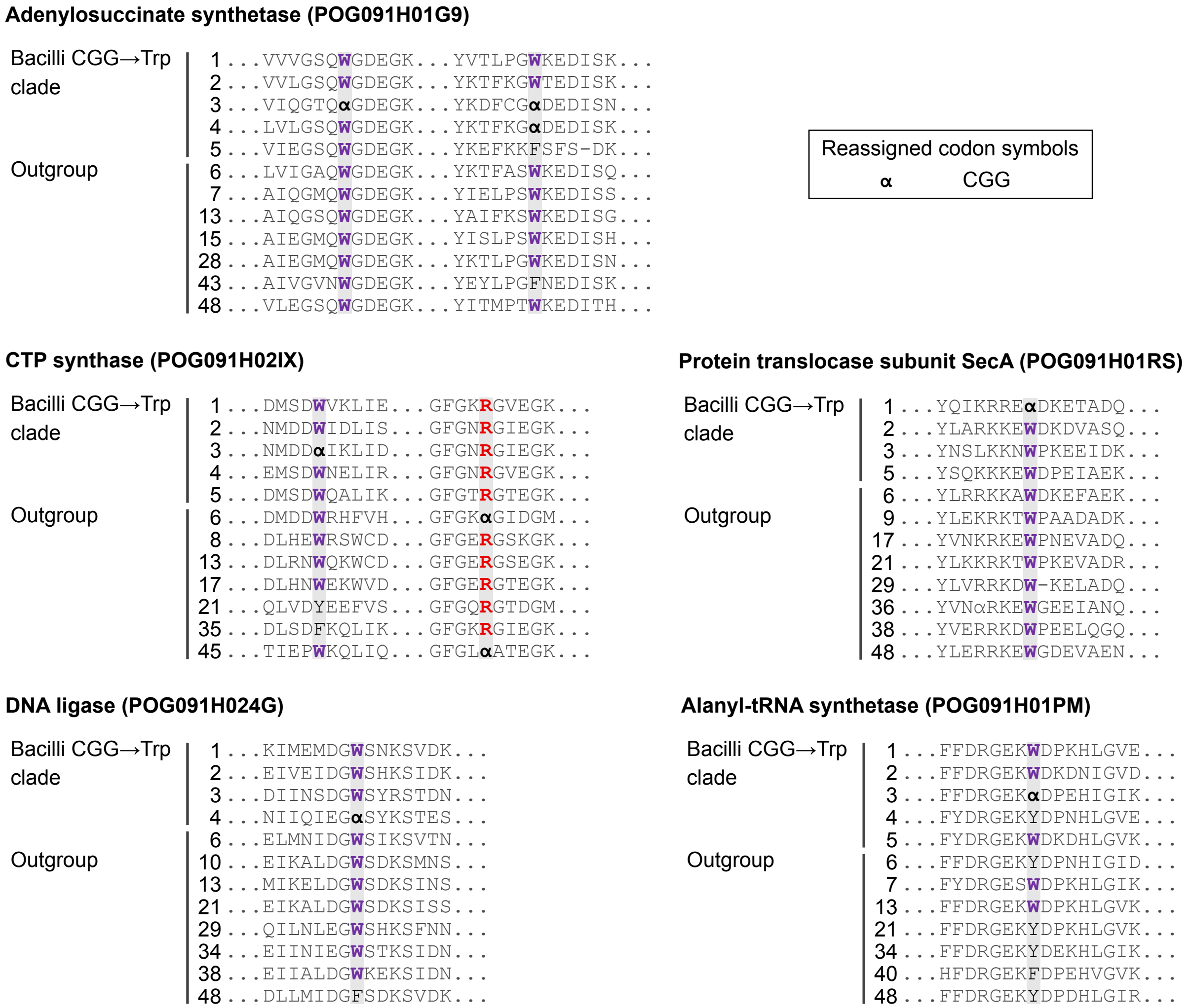

GTDB phylogeny and species numbering
Outgroups
CGG→Trp clade
1 **
2 **
3 **
4 **
5
6
7 **
8
9
10
11
12
13**
14
15
16
17**
18
19
20
21
22
23
24**
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39**
40
41
42
43
44
45
46
47
48**
49
50
Codon inference
Arg ● Trp ● uninf ○ CGG
tRNA gene content
codon tRNA
Arg codons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

Figure 1: 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 Codetta is indicated by colored circles (red: arginine, purple: tryptophan, white: uninferred). The presence of tRNA genes that recognize the CCA and CGN-codons is also indicated by filled circles, colored according to the predicted amino acid charging based on identity elements for tRNAs (see Methods). Anticodons in gray font are typically not found in Firmicutes. The lines connecting codons and tRNA anticodons represent the likely wobble decoding capabilities, with dashed lines representing weaker interactions. The anticodon ACG is presumed to be modified to ICG. The U34 of anticodon UCG is presumed to be modified in a way that restricts decoding to CGA and CGG, but could potentially recognize CGU and/or CGC depending on the true modification state. The remaining anticodons are not expected to be modified in a way that alters which codons are recognized. The codon usage for the reassigned codon CGG, the arginine codons AGA, AGG, CGU, CGC, and CGA, and the tryptophan codon CCA is the frequency per 10,000 codons aligned to Pfam domains. Genomic GC content is calculated over the entire genome.

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 (*1*) 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



Adenylosuccinate synthetase (POG091H01G9)

```
Bacilli CGG→Trp   1...VVVGSQWGDEGK...YVTLPGWKEDISK...
clade             2...VVLGSQWGDEGK...YKTFKGWTEDISK...
                  3...VIQGTQαGDEGK...YKDFCGαDEDISN...
                  4...LVLGSQWGDEGK...YKTFKGαDEDISK...
                  5...VIEGSQWGDEGK...YKEFKKFSFS-DK...
Outgroup          6...LVIGAQWGDEGK...YKTFASWKEDISQ...
                  7...AIQGMQWGDEGK...YIELPSWKEDISS...
                 13...AIQGSQWGDEGK...YAIFKSWKEDISG...
                 15...AIEGMQWGDEGK...YISLPSWKEDISH...
                 28...AIEGMQWGDEGK...YKTLPGWKEDISN...
                 43...AIVGVNWGDEGK...YEYLPGFNEDISK...
                 48...VLEGSQWGDEGK...YITMPTWKEDITH...
```

CTP synthase (POG091H02IX)

```
Bacilli CGG→Trp   1...DMSDWVKLIE...GFGKRGVEGK...
clade             2...NMDDWIDLIS...GFGNRGIEGK...
                  3...NMDDαIKLID...GFGNRGIEGK...
                  4...EMSDWNELIR...GFGNRGVEGK...
                  5...DMSDWQALIK...GFGTRGTEGK...
Outgroup          6...DMDDWRHFVH...GFGKαGIDGM...
                  8...DLHEWRSWCD...GFGERGSKGK...
                 13...DLRNWQKWCD...GFGERGSEGK...
                 17...DLHNWEKWVD...GFGERGTEGK...
                 21...QLVDYEEFVS...GFGQRGTDGM...
                 35...DLSDFKQLIK...GFGKRGIEGK...
                 45...TIEPWKQLIQ...GFGLαATEGK...
```

Protein translocase subunit SecA (POG091H01RS)

```
Bacilli CGG→Trp   1...YQIKRREαDKETADQ...
clade             2...YLARKKEWDKDVASQ...
                  3...YNSLKKNWPKEEIDK...
                  5...YSQKKKEWDPEIAEK...
                  6...YLRRKKAWDKEFAEK...
Outgroup          9...YLEKRKTWPAADADK...
                 17...YVNKRKEWPNEVADQ...
                 21...YLKKRKTWPKEVADR...
                 29...YLVRRKDW-KELADQ...
                 36...YVNαRKEWGEEIANQ...
                 38...YVERRKDWPEELQGQ...
                 48...YLERRKEWGDEVAEN...
```

DNA ligase (POG091H024G)

```
Bacilli CGG→Trp   1...KIMEMDGWSNKSVDK...
clade             2...EIVEIDGWSHKSIDK...
                  3...DIINSDGWSYRSTDN...
                  4...NIIQIEGαSYKSTES...
                  6...ELMNIDGWSIKSVTN...
Outgroup         10...EIKALDGWSDKSMNS...
                 13...MIKELDGWSDKSINS...
                 21...EIKALDGWSDKSISS...
                 29...QILNLEGWSHKSFNN...
                 34...EIINIEGWSTKSIDN...
                 38...EIIALDGWKEKSIDN...
                 48...DLLMIDGFSDKSVDK...
```

Alanyl-tRNA synthetase (POG091H01PM)

```
Bacilli CGG→Trp   1...FFDRGEKWDPKHLGVE...
clade             2...FFDRGEKWDKDNIGVD...
                  3...FFDRGEKαDPEHIGIK...
                  4...FYDRGEKYDPNHLGVE...
                  5...FYDRGEKWDKDHLGVK...
Outgroup          6...FFDRGEKYDPNHIGID...
                  7...FYDRGESWDPKHLGIK...
                 13...FFDRGEKWDPKHLGVK...
                 21...FFDRGEKYDPKHLGVK...
                 34...FFDRGEKYDEKHLGIK...
                 40...HFDRGEKFDPEHVGVK...
                 48...FFDRGEKYDPDHLGIR...
```

Figure 2: 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.

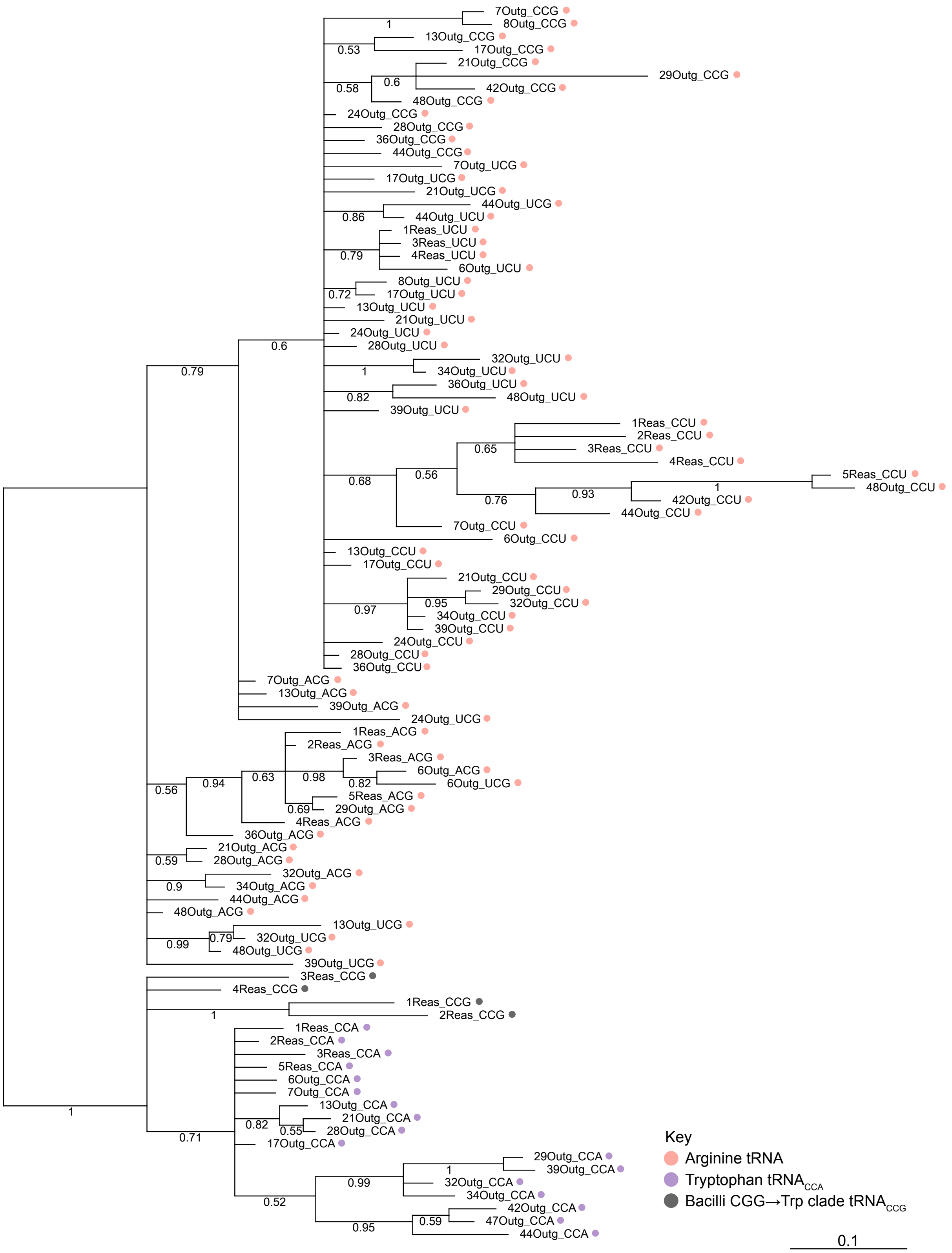

7Outg_CCG
8Outg_CCG
1
13Outg_CCG
0.53
17Outg_CCG
21Outg_CCG
29Outg_CCG
0.6
0.58
42Outg_CCG
48Outg_CCG
24Outg_CCG
28Outg_CCG
36Outg_CCG
44Outg_CCG
7Outg_UCG
17Outg_UCG
21Outg_UCG
44Outg_UCG
0.86
44Outg_UCU
1Reas_UCU
3Reas_UCU
0.79
4Reas_UCU
6Outg_UCU
8Outg_UCU
0.72
17Outg_UCU
13Outg_UCU
21Outg_UCU
24Outg_UCU
0.6
28Outg_UCU
32Outg_UCU
0.79
1
34Outg_UCU
36Outg_UCU
0.82
48Outg_UCU
39Outg_UCU
1Reas_CCU
2Reas_CCU
0.65
3Reas_CCU
4Reas_CCU
5Reas_CCU
0.68
0.56
1
48Outg_CCU
0.93
0.76
42Outg_CCU
44Outg_CCU
7Outg_CCU
6Outg_CCU
13Outg_CCU
17Outg_CCU
21Outg_CCU
29Outg_CCU
0.95
32Outg_CCU
0.97
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
0.56
0.94
0.63
0.98
0.82
6Outg_UCG
5Reas_ACG
0.69
29Outg_ACG
4Reas_ACG
36Outg_ACG
21Outg_ACG
0.59
28Outg_ACG
32Outg_ACG
0.9
34Outg_ACG
44Outg_ACG
48Outg_ACG
13Outg_UCG
0.79
32Outg_UCG
0.99
48Outg_UCG
39Outg_UCG
3Reas_CCG
4Reas_CCG
1Reas_CCG
1
2Reas_CCG
1Reas_CCA
2Reas_CCA
3Reas_CCA
5Reas_CCA
6Outg_CCA
1
7Outg_CCA
13Outg_CCA
21Outg_CCA
0.82
0.55
28Outg_CCA
0.71
17Outg_CCA
29Outg_CCA
1
39Outg_CCA
0.99
32Outg_CCA
0.52
34Outg_CCA
42Outg_CCA
0.59
47Outg_CCA
0.95
44Outg_CCA
Key
● Arginine tRNA
● Tryptophan $tRNA_{CCA}$
● Bacilli CGG→Trp clade $tRNA_{CCG}$
0.1

Figure 3: 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.

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).

References

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