Reassignment of CGG to tryptophan in *Anaerococcus*

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

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 **??**). 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 Codetta, 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 **??**, gray box outline). On the tRNA phylogenetic tree, *Finegoldia* $tRNA_{CCG}$ sequences form a clade within other tryptophan tRNAs (Figure **??**). 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 Codetta 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.

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 1: (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.

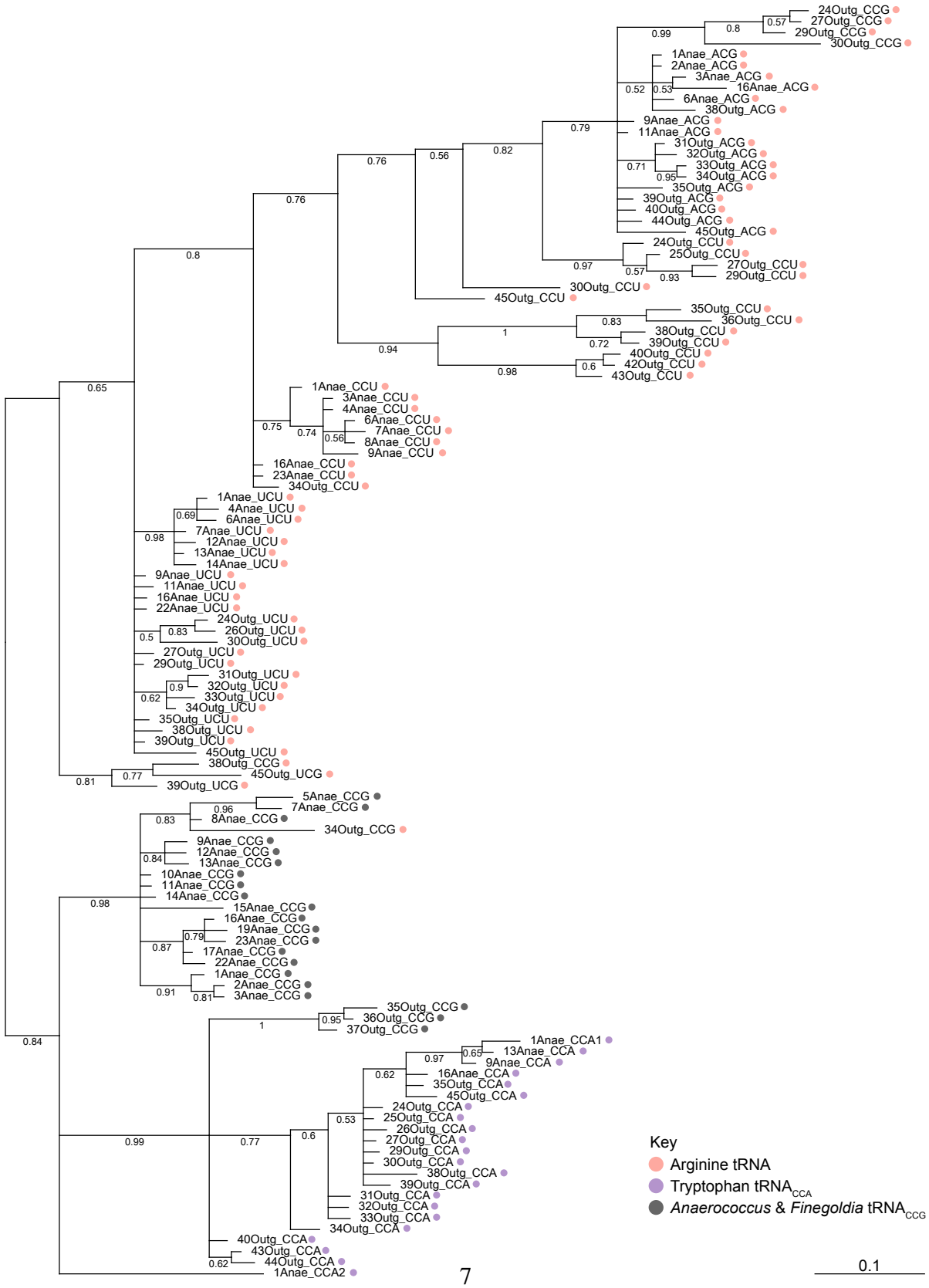
24Outg_CCG
27Outg_CCG
29Outg_CCG
30Outg_CCG
1Anae_ACG
2Anae_ACG
3Anae_ACG
16Anae_ACG
6Anae_ACG
38Outg_ACG
9Anae_ACG
11Anae_ACG
31Outg_ACG
32Outg_ACG
33Outg_ACG
34Outg_ACG
35Outg_ACG
39Outg_ACG
40Outg_ACG
44Outg_ACG
45Outg_ACG
24Outg_CCU
25Outg_CCU
27Outg_CCU
29Outg_CCU
30Outg_CCU
45Outg_CCU
35Outg_CCU
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 2: 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.

Methods

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

Genome completeness estimate by tRNA presence

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

tRNA phylogeny

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