Conservation and evolution of the programmed ribosomal frameshift in *prfB* across the bacterial domain

Authors: Cassidy R. Prince[1], Isabella N. Lin[1], Heather A. Feaga[1]*

Affiliations:

[1] Department of Microbiology, Cornell University, Ithaca, NY 14853

* Corresponding author haf54@cornell.edu

Keywords: translation termination; RF-2; *prfB*; programmed ribosomal frameshifting; ribosome; evolution; Actinobacteria

Abstract

When the ribosome reaches a stop codon, translation is terminated by a release factor. Bacteria encode two release factors, RF1 and RF2. In many bacteria, the gene encoding RF2 (*prfB)* contains an in-frame premature stop codon near the beginning of the open reading frame. A programmed ribosomal frameshift is therefore required to translate full-length RF2. While the molecular mechanism of the programmed ribosomal frameshift has been extensively characterized in *Escherichia coli*, bioinformatic analysis of the evolution and conservation of this motif has been limited to few genomes. By analyzing >12,000 bacterial genomes, we sought to thoroughly characterize the conserved frameshifting elements within the programmed frameshifting motif and identify genomic features of phyla that have lost the motif altogether. We find that the programmed ribosomal frameshift in *prfB* was likely present in the last common ancestor of bacteria and that the motif elements are almost completely conserved, including the identity of the internal stop codon. We find that loss of the programmed frameshift motif is highly correlated with RF2-specific stop codon usage, suggesting that stop codon usage has shaped the conservation of this regulatory mechanism. In support of this model, the programmed frameshift in *prfB* is entirely absent in Actinobacteriota, which have particularly high RF2 specific stop codon usage. Finally, we show that a model member of Actinobacteriota fails to produce full-length RF2 when provided with an allele of *prfB* that contains the programmed frameshifting motif. Altogether, our work provides a thorough characterization of RF2 regulation across the bacterial domain.

Introduction

Ribosomes translate mRNA into protein by iteratively decoding one codon at a time. Reading frame maintenance is essential for translation fidelity. Frameshifting alters the identity of downstream amino acids and typically leads to premature termination since the ribosome quickly encounters a stop codon when it deviates from an open reading frame (1). In rare cases, ribosomal frameshifting is required to make a full-length protein (2–5). These types of frameshifts are referred to as programmed ribosomal frameshifts (6–9)

A textbook example of programmed ribosomal frameshifting in bacteria is found in *prfB*, the gene that encodes Release Factor 2 (RF2) (10, 11). RF2 terminates translation at UGA and UAA stop codons (12). In *Escherichia coli prfB*, there is an in-frame UGA codon at position 26 of the open reading frame (10) (Fig. 1). A purine rich, internal Shine-Dalgarno (SD)-like sequence is located 6 nucleotides upstream of this stop codon and a CTT leucine codon immediately precedes the stop codon (13). The UGA stop codon is followed by a C nucleotide, making it a weak stop codon (14–16). *In vitro* and *in vivo* studies using *E. coli* show that when concentrations of RF2 are high, RF2 terminates translation at the premature stop codon, creating a truncated peptide (11, 17). When RF2 concentrations are low, the ribosome pauses at the stop codon, and the SD-like sequence in the mRNA base-pairs with the anti-SD at the 3' end of the 16S rRNA, displacing the E-site tRNA (13, 17, 18). The leucine tRNA in the P site then slips into the +1 frame, causing the ribosome to bypass the internal stop codon and continue translation to produce full-length RF2 (11, 13). Since RF2 levels directly regulate the translation of the *prfB* transcript, this frameshifting mechanism autoregulates RF2 levels.

The first bioinformatic analysis of the conservation of the frameshifting motif was performed by Baranov and colleagues in 2002, with 87 bacterial *prfB* sequences (19). This foundational work determined that 70% of these *prfB* sequences required the programmed ribosomal frameshift to produce RF2, and that in all but one sequence, the in-frame stop codon interrupting the *prfB* open reading frame was UGA. The slippery sequence (CTT) and the C

nucleotide immediately following the stop codon were all highly conserved (20). Follow-up on a larger set of 259 genomes uncovered the regulatory frameshift in 87% of these genomes, indicating that the prevalence of this frameshift was even more widespread than previously determined (21).

Tens of thousands of bacterial genomes are now available, facilitating more thorough analyses of the conservation of the *prfB* programmed frameshift as an auto-regulatory mechanism. Our analyses are also aided by recent developments in bacterial phylogeny, which better resolve the bacterial species tree. Using a diverse and comprehensive genome set, we sought to answer remaining questions about the evolution and conservation of the programmed frameshift autoregulatory mechanism in *prfB*. For example, was the programmed frameshift present in the ancestral *prfB*, and what are the genome characteristics that are associated with absence of the programmed frameshift?

We surveyed 12,751 bacterial genomes across 21 phyla to determine the prevalence and conserved features of programmed ribosomal frameshifting in *prfB*. We find that most bacterial genomes encode a *prfB* that requires a ribosomal frameshift to produce full-length RF2 and that the programmed frameshift was most likely present in the common ancestor of bacteria. In our dataset, the frameshift sequence motif is extremely well-conserved, and most strikingly, the identity of the premature stop codon is nearly always an RF2-specific UGA stop codon. We did not find any instance of the RF1-specific stop codon within the motif, indicating that the autoregulatory role of the programmed ribosomal frameshift in *prfB* is conserved across the bacterial domain. Next, we examined the species that lack the autoregulatory frameshift motif and found that they have significantly higher RF2 stop codon usage, which may explain why RF2 is not autoregulated by this mechanism in these organisms. Consistent with this model, the Actinobacterium *Mycobacterium smegmatis*, which has high RF2 stop codon usage, exhibits low frameshifting efficiency at the motif. Cumulatively, our results support the autoregulatory function of the *prfB* frameshift across the bacterial domain, suggest that this autoregulatory mechanism

was present in ancestral *prfB*, and that this autoregulation was lost in phyla that exhibit high RF2-specific stop codon usage.

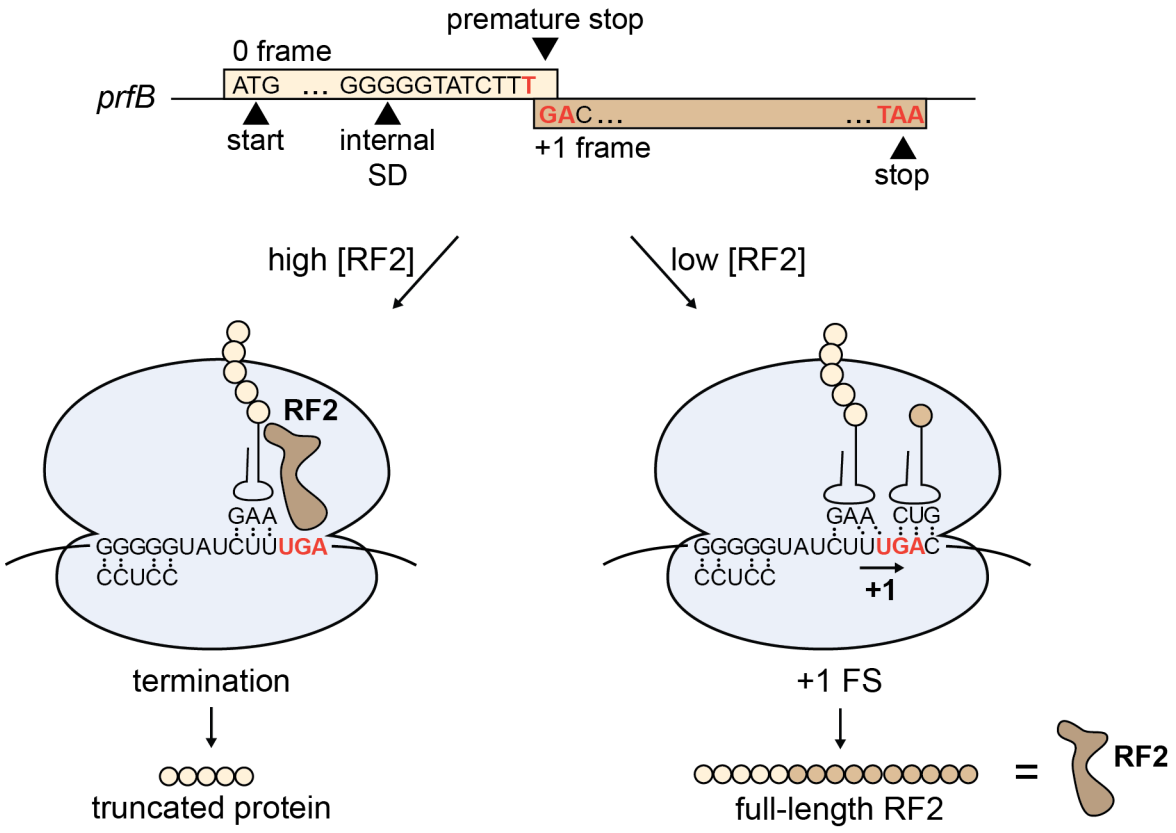


Figure 1. Schematic of *prfB* programmed ribosomal frameshift mechanism. (Top) DNA sequence elements of the *prfB* gene sequence in *E. coli*. (Bottom) The ribosome pauses at the internal SD-like sequence GGGGG, which base-pairs with the anti-SD sequence on the small ribosomal subunit. If RF2 levels are high, RF2 will enter the A-site and terminate translation at the premature UGA stop codon. If RF2 levels are low, the ribosome will slip into the +1 reading frame, causing a GAC codon to be positioned in the A-site instead of the premature stop codon. Therefore, ribosomal frameshifting results in production of full-length RF2.

Results

Sequence elements of the programmed frameshift within *prfB* are hyper-conserved

To determine the prevalence of the *prfB* programmed frameshift motif, we analyzed the *prfB* sequences of 12,751 representative bacterial species genomes from the NCBI RefSeq database as annotated by NCBI (22). We identified the *prfB* sequence in each of these genomes and determined whether they contained a premature, in-frame stop codon within the *prfB* reading frame. Of the 12,751 genomes surveyed, 8160 (64%) contain a premature stop codon in *prfB*, indicating that these organisms require a programmed ribosomal frameshift to produce functional RF2.

We aligned the *prfB* sequences to generate a nucleotide sequence logo of the programmed frameshift motif (Fig. 2A). Within this motif, the purine-rich internal Shine-Dalgarno-like sequence is hyper-conserved, with the consensus sequence approaching “AGGGGG” (Fig. 2A). An even more highly conserved slippery sequence of “CTTT” occurs 3 nucleotides downstream of the internal Shine-Dalgarno-like sequence. The last T of the slippery sequence is the first T of the premature stop codon. The cytosine following the premature stop codon is also highly conserved, likely because “TGAC” causes poor termination efficiency and would permit more frequent frameshifting (14). This canonical “CTTTGA” sequence is present in nearly every *prfB* frameshift motif. Exceptions included 21 genomes with a long poly-thymine tract in the slippery sequence, in which $tRNA_{Phe}$ would decode the codon before the stop codon instead of $tRNA_{Leu}$ (i.e. “TTTTTGA”). These genomes belong to predominantly low-GC organisms (mean GC of 34.6%) in Aquificota and Gammaproteobacteria. In these genomes, the *prfB* sequence retains the internal SD-like sequence as well as the spacing between the SD-like sequence and the slippery sequence, and the identity of the premature stop codon is TGA.

We also aligned the *prfB* sequences of species that do not contain a premature stop codon to determine whether these sequences retained any elements of the programmed frameshift motif. In these *prfB* sequences the internal SD-like sequence and slippery sequence are not found

in the analogous region of the sequence, suggesting that these elements are lost in organisms that encode a fully in-frame RF2 (Fig. 2A). This observation is expected since these elements would still promote ribosomal frameshifting, but without the presence of the autoregulatory TGA stop codon.

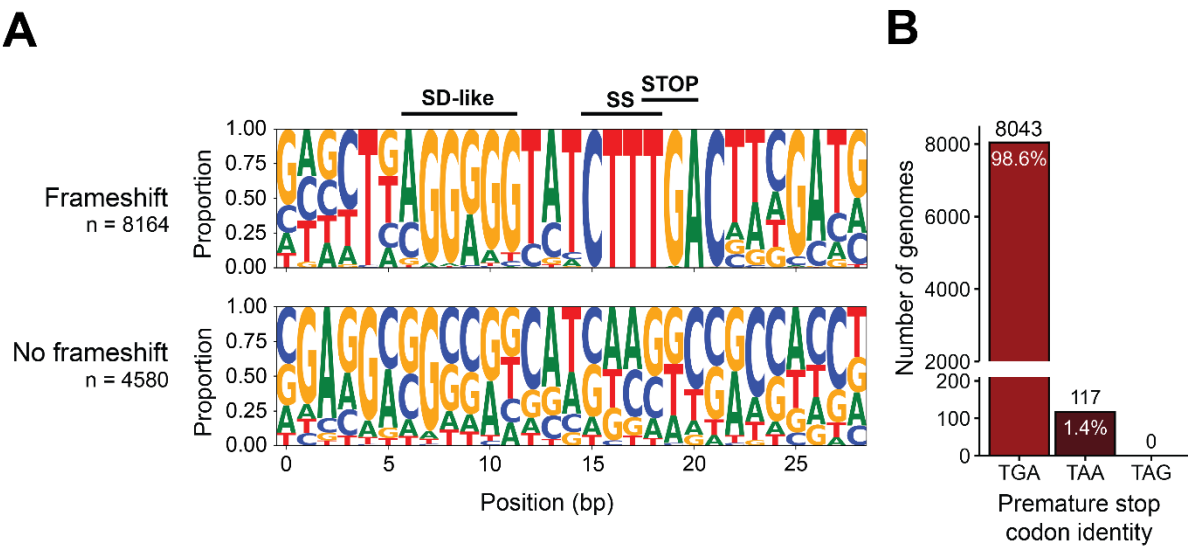


Figure 2. The premature stop codon in the *prfB* frameshifting motif is recognizable by RF2 in 100% of genomes surveyed. (A) Sequence logos of *prfB* frameshifting motif regions for genomes with and without a premature stop codon in *prfB*. The "No frameshift" logo captures the region in which the frameshifting motif is expected based on a full alignment of all *prfB* sequences. Labels above the black lines correspond to the following frameshifting motif components: SD-like, Shine-Dalgarno-like sequence; SS, slippery sequence; STOP, premature stop codon. (B) Identity of the premature stop codon within the *prfB* frameshifting motif.

The RF1-specific TAG stop codon is not detected as the premature stop codon in the *prfB* programmed frameshift motif

Bacteria terminate translation using one of three stop codons: TAA, TAG, or TGA. TAA is recognized by either RF1 or RF2, whereas TAG is RF1-specific and TGA is RF2-specific (23). To assess the conservation of the RF2-specific stop codon within the programmed frameshift motif, we determined the identity of the premature in-frame stop codon in the 8160 genomes that contain the motif. We found that 98.6% of genomes with the motif contain the RF2-specific TGA stop

157 codon as the premature stop codon (“CTTTGA”) (Fig. 2B). 1.4% of genomes contain a TAA
158 premature stop codon (“CTTTAA”). Since RF2 terminates translation at TAA as well as TGA, RF2
159 expression in these taxa would also be sensitive to the levels of RF2 in the cell. Genomes
160 encoding TAA as the premature stop codon are observed randomly amongst phyla and retained
161 among strains of a species (Fig. S1). These findings suggest that the TAA codon is poorly
162 tolerated since it does not become fixed within particular clades and that the premature TGA is
163 preferred for RF2 autoregulation (Fig. 2B). Most strikingly, none of the 8160 genomes encode an
164 in-frame RF1-specific TAG stop codon. The ubiquity of TGA as the premature stop codon in the
165 motif and the absence of TAG suggests that the universal purpose of the programmed frameshift
166 motif is indeed RF2 autoregulation.
167

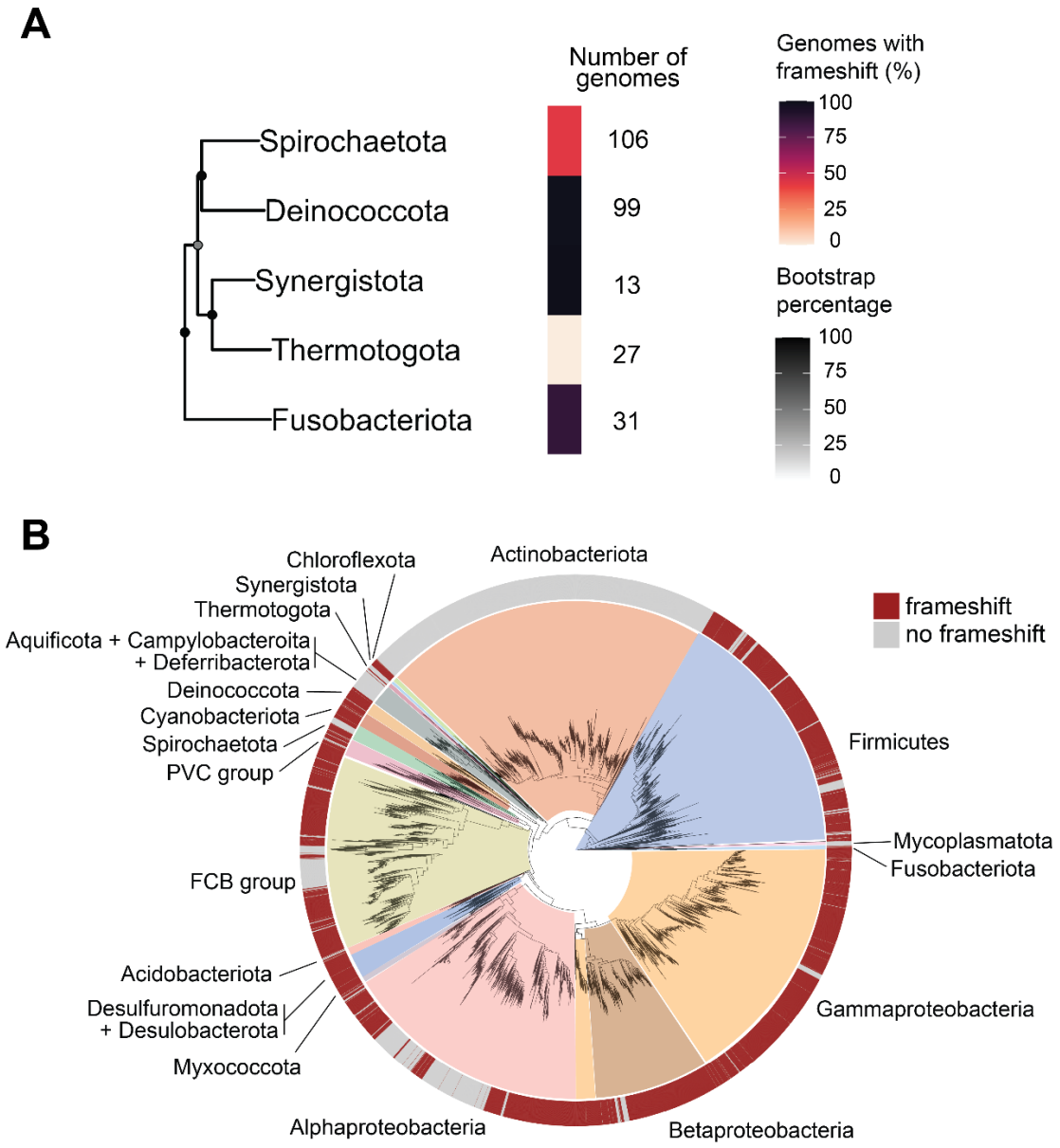
168 **The programmed frameshift *prfB* was likely present in the last common ancestor of**
169 **bacteria**

170 Bioinformatic analyses of the programmed frameshift motif in *prfB* sequences by Baranov and
171 colleagues found that the motif is absent in *Aquifex aeolicus* and *Thermotoga maritime.* At the
172 time, these taxa were thought to be the deepest branching taxa and most closely related to the
173 ancestor of eubacteria. Therefore, absence of the programmed frameshift in the *prfB* of *A.*
174 *aeolicus* and *T. maritime* suggested it may have been absent in the ancestral *prfB*. Recent efforts
175 to root the bacterial tree position the root in the neighborhood of Fusobacteriota (24). This
176 neighborhood also includes the Spirochaeotota, and a clade comprising Deinococcota,
177 Synergistota, and Thermotogota (DST clade). Consistent with Baranov and colleagues, we did
178 not detect the programmed frameshift in any of the 27 Thermotogota genomes we surveyed (Fig.
179 3A). However, we detected the programmed frameshift in >85% of *prfB* sequences from most
180 other lineages in this neighborhood. The programmed frameshift is present in 87% of *prfB*
181 sequences from Fusobacteriota (n = 31), 100% of Synergistota (n = 13), 99% of Deinococcota (n
182 = 99), and 42% of Spirochaetota (n = 106) (Fig. 3A). The presence of the programmed frameshift

183 motif in taxa closest to the root of the bacterial tree suggests that the ancestral *prfB* contained the

184 programmed frameshift.

185



186

187 **Figure 3. The *prfB* frameshifting motif is broadly distributed across bacterial phyla.** (A)
188 Percent of genomes containing the *prfB* frameshifting motif within phyla most closely related to
189 the last bacterial common ancestor. A subtree of the large 16S tree was created using a random
190 single representative genome for each phylum. The number of analyzed genomes per phylum is
191 located to the right of each bar. Darkened circles indicate bootstrap values for each node. (B) A
192 16S maximum-likelihood phylogenetic tree showing distribution of genomes that encode the
193 programmed frameshift within *prfB*. Phyla with more than 10 available and high-quality reference
194 genomes are shown.

195

196 Next, we determined the distribution of the *prfB* programmed frameshift throughout the
197 bacterial domain (Fig. 3B, Fig. S2). Only 3 phyla with >10 available genomes completely lacked
198 the frameshift motif: Actinobacteriota, Mycoplasmatota, and Thermotogota. In 14 out of 19 phyla
199 with >10 available genomes, >50% of genomes contain the premature stop and frameshifting
200 motif (Fig. S2), demonstrating strong conservation of the programmed frameshift. For phyla where
201 most genomes contain the motif, it is likely and parsimonious that the common ancestor of the
202 phylum contained the motif, and that the motif was lost in recent lineages. Altogether, our results
203 suggest that the programmed ribosomal frameshift in *prfB* is widespread throughout bacteria,
204 including in deeply rooted branches of the bacterial domain.
205

206 **The *prfB* programmed frameshift motif is absent in Actinobacteriota and Mycoplasmatota**

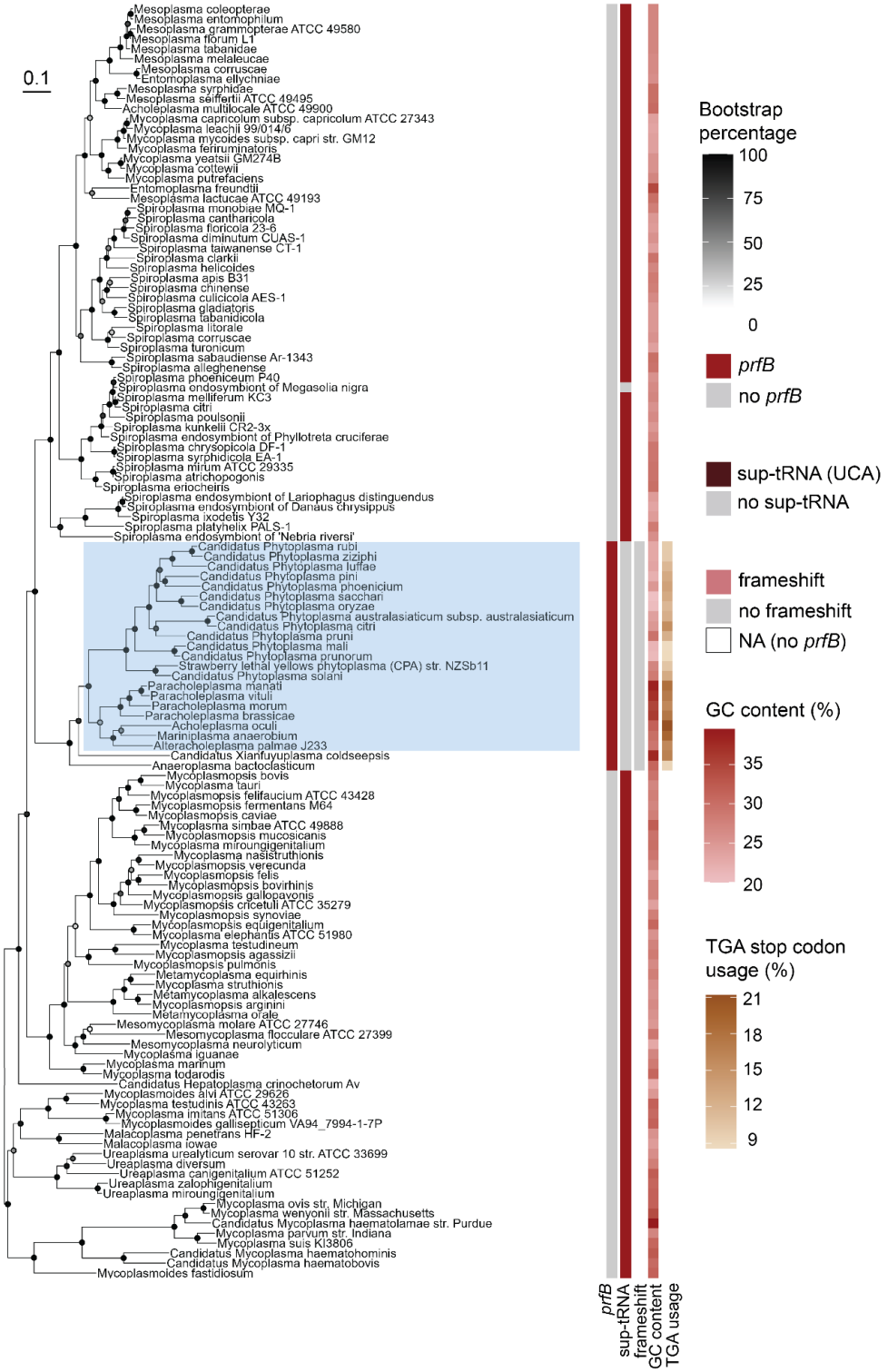
207 It was previously hypothesized that the lineages of Mycobacteria and Streptomyces lost the
208 programmed frameshift in *prfB* based on its absence in two species – *Mycobacterium tuberculosis*
209 and *Streptomyces coelicolor* (25). Expanding upon this, we surveyed the *prfB* sequences of the
210 Actinobacteriota phylum that includes these lineages and did not detect the programmed
211 frameshift in any of these genomes (n = 2658) (Fig. 3B), supporting a loss event in the common
212 Actinobacterial ancestor.

213 Our analysis also indicates that the *prfB* programmed frameshift was likely absent in the
214 common ancestor of the Mycoplasmatota phylum. Mycoplasmatota species are most closely
215 related to Firmicutes (26), but have significantly reduced genome sizes, low GC content (27), and
216 particularly low TGA stop codon usage (Fig. 4). Many species within Mycoplasmatota lack the
217 *prfB* gene completely (28, 29) and utilize nearly zero TGA stop codons (12). Instead, TGA codons
218 are decoded as tryptophan by suppressor tRNAs (27, 29–31). We found that all genomes lacking
219 *prfB* encoded a suppressor tRNA for the TGA stop codon (Fig. 4). Only 23 out of 128
220 Mycoplasmatota genomes retained *prfB,* and none of these sequences contain the programmed

221 frameshift motif (Fig. 4). Species that retain *prfB* are predominantly found within the

222 Acholeplasmataceae family (Fig. 4), including important *Phytoplasma* plant pathogens (32).

223



224

Figure 4. Most Mycoplasmatota species lack prfB and suppress TGA stop codons. Core SNP maximum-likelihood tree of 128 Mycoplasmatota species. From left to right, columns represent *prfB* gene presence, TGA stop codon suppressor tRNA presence, *prfB* frameshift motif presence, GC content, and TGA stop codon usage. The Acholeplasmataceae family is highlighted in blue. Darkened circles indicate bootstrap values for each node.

Both free-living and obligate intracellular species retain the *prfB* programmed frameshift motif

Alphaproteobacteria are unique in that about 50% of genomes in this phylum encode the programmed ribosomal frameshift within *prfB* (Fig. S2). Since this phylum comprises both free-living and obligate intracellular bacteria, we questioned whether intracellular species would be less likely to autoregulate RF2 levels with the programmed ribosomal frameshift. We observed that Alphaproteobacterial genomes that do not use the frameshift motif in *prfB* fall predominantly in two orders: Rhodobacterales (n = 486 genomes) and Sphingomonadales (n = 338 genomes). Within Rhodobacterales, only two genomes appear to utilize the programmed frameshift motif. These were the only genomes of their respective genera, so additional genomes for the species could not be consulted to rule out sequence inaccuracy. Within Sphingomonadales, only 15 genomes contained the frameshifting motif, and these genomes were in the family Sphingosinicellaceae. Members of both Rhodobacterales and Sphingomonadales tend to be free-living and are found in soils, water, and sediments (33). Other free-living orders of the Alphaproteobacteria such as Caulobacterales and Rhodospirillales predominantly encode the programmed frameshift motif in *prfB*. Meanwhile, the obligate intracellular Alphaproteobacterial genera including *Rickettsia* and *Wolbachia* maintain the *prfB* gene and its frameshifting motif. Therefore, we conclude that the programmed frameshift motif is conserved in both free-living and in obligate intracellular species within the Alphaproteobacteria.

Genomes that lack the programmed frameshifting motif in *prfB* have higher GC content and more TGA stop codon usage

254 We next explored genome characteristics of organisms that do not utilize the programmed
255 frameshift in *prfB*. We found that genomes that lack the programmed frameshift motif have
256 significantly higher GC content than genomes with the motif (62% average GC, $p < 2.2e-16$) (Fig.
257 5A). Our finding remains significant even with the removal of the well-represented high-GC
258 Actinobacterial genomes ($p < 1.1e-07$) (Fig. 5B). GC content positively correlates with RF2-
259 specific TGA codon usage (12) (Fig. S3). Therefore, we hypothesized that organisms lacking RF2
260 autoregulation would also encode more RF2-specific stop codons. To investigate this, we
261 compared terminal stop codon usage between genomes with and without the programmed
262 frameshift for a random subset of 1000 genomes. Genomes that lost the *prfB* frameshift motif
263 have significantly higher RF2-specific TGA terminal stop codon usage than genomes that retained
264 the motif ($p < 2.2e-16$) (Fig. 5A). Again, our findings are significant even when Actinobacterial
265 genomes are excluded ($p < 2.2$ e-06) (Fig. 5B). Therefore, a higher demand for RF2 due to
266 increased RF2-specific TGA stop codon usage may help explain the loss of the RF2
267 autoregulation in some species.
268
269

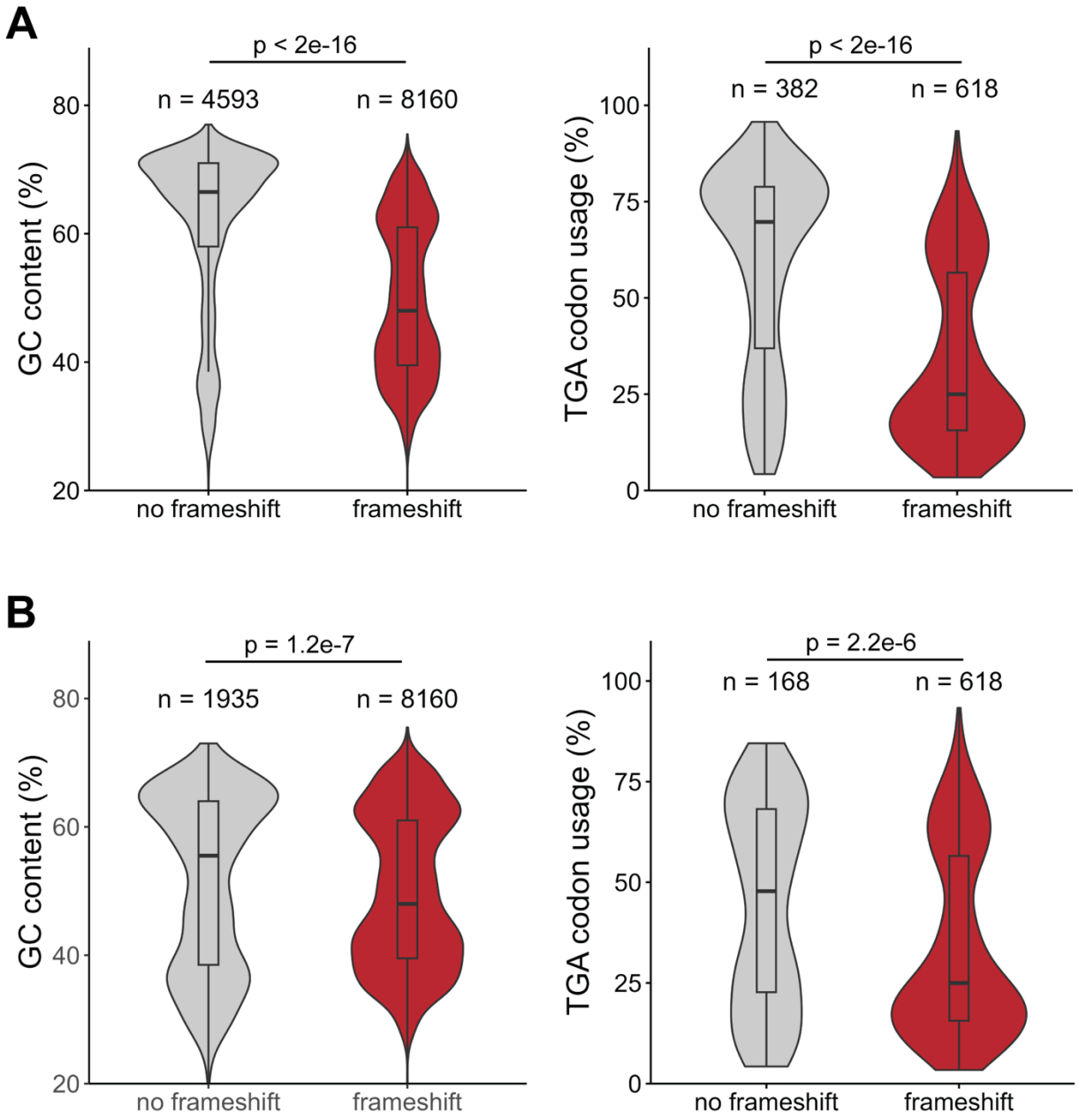


Figure 5. Genomes that encode the programmed frameshift motif within *prfB* have significantly higher GC content and TGA stop codon usage. (A) GC content and TGA stop codon usage of a subset of 1000 random genomes separated by genomes with and without the *prfB* frameshift motif. (B) Distributions of GC content and TGA stop codon usage of the random subset without Actinobacterial genomes are also shown. P-values indicate the results of a Welch two-sample *t*-test.

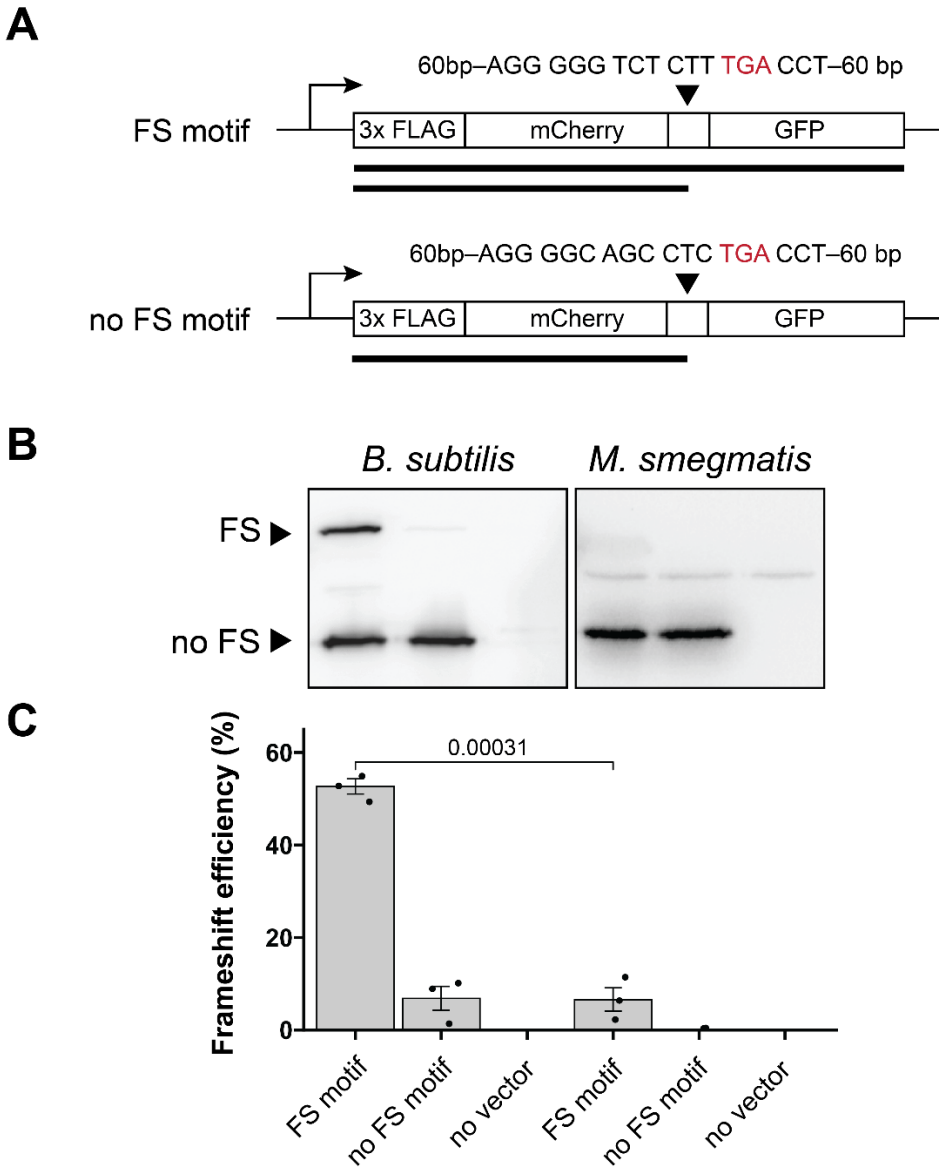
278 **Ribosomal frameshifting is inefficient at the *prfB* programmed frameshift motif in the**
279 **Actinobacterium *Mycobacterium smegmatis***

280 No surveyed Actinobacterial genomes (n = 2658) contain the *prfB* frameshift. Therefore, we
281 hypothesized that Actinobacteriota may exhibit poor frameshift efficiency at the motif. To assay
282 frameshifting efficiency, we compared the frameshifting efficiencies of an organism that natively
283 contains the frameshift motif, *B. subtilis*, and an Actinobacterium that lacks the motif,
284 *Mycobacterium smegmatis*. We designed analogous constructs for *B. subtilis* and *M. smegmatis*
285 that contain two fused protein sequences (encoding mCherry and GFP) separated by the *prfB*
286 frameshift motif from *B. subtilis* (denoted as 'FS motif'). We included ~60 bp upstream and
287 downstream of the frameshifting motif to ensure that the context surrounding the frameshift motif
288 is also maintained. In *M. smegmatis*, the fluorescent proteins and the 60 bp regions upstream and
289 downstream of the *prfB* motif were codon optimized for *M. smegmatis* to avoid ribosomes stalling
290 at rare codons. As a control for production of the truncated protein we used a construct identical
291 to the *prfB* FS motif construct including the premature TGA stop codon but lacking the
292 frameshifting motif elements (denoted as 'no FS motif'). Construct schematics are depicted in Fig.
293 6A.

294 Upon expressing the previously described constructs *in vivo*, we observed that ribosomal
295 frameshifting occurs at the *prfB* motif with an efficiency of 52.7% ± 2.8% in *B. subtilis* (Fig. 6B and
296 6C). In contrast, *M. smegmatis* exhibits ribosomal frameshifting at the motif with an efficiency of
297 6.6% ± 4.4% (Fig. 6B and 6C). As expected, neither organism exhibits strong ribosomal
298 frameshifting at the *prfB* variant that lacks the frameshifting motif elements (Fig. 6B). The
299 abnormally low frameshift efficiency in *M. smegmatis* at the *prfB* motif supports the idea that
300 members of the Actinobacteriota have lost this form of RF2 autoregulation because they are
301 unable to efficiently utilize the programmed frameshifting motif.

302

303



304

305 **Figure 6. *Mycobacterium smegmatis* cannot efficiently frameshift at the canonical *prfB***
306 **frameshifting motif.** (A) Schematics for analogous prfB frameshifting reporters in *B. subtilis* and
307 *M. smegmatis*. (B) Western blots showing both frameshifted and termination products resulting
308 from expression of the pictured constructs. Blots are representative of three biological replicates.
309 (C) Quantification of western blots. Frameshifting efficiency = frameshifted protein
310 produced/(frameshifted + non-frameshifted protein produced). P-values indicate the results of a
311 Welch two-sample *t*-test.
312

313

314

Discussion

The autoregulation of *prfB* expression by RF2 levels is a textbook example of programmed ribosomal frameshifting, and the molecular mechanism underlying the programmed frameshift has been precisely characterized in *E. coli* (11, 13). Using >12,000 bacterial genomes and referencing a recently updated bacterial phylogenetic tree (24), we present evidence that this regulatory frameshift motif was likely present in the ancestral *prfB*. We find that loss of the programmed frameshift highly correlates with RF2 specific stop codon usage. This finding is consistent with the expectation that the requirement for higher RF2 levels will select for inactivation of an autoregulatory mechanism that reduces RF2 expression.

Our analysis of the programmed frameshifting elements in *prfB* is consistent with previous reports analyzing fewer than 100 genomes (19, 21, 25) and is in perfect agreement with the highly characterized mechanistic features of the ribosomal frameshift. In particular, the SD-like sequence, the spacing between the SD-like sequence and the premature stop codon, the slippery sequence preceding the stop codon, and the C nucleotide immediately following the stop codon are almost universally conserved. Most striking, the identity of the premature stop codon is TGA in >99% of surveyed genomes and is TAA in all remaining genomes surveyed. There is evidence that TAA may be primarily recognized by RF2 in most bacterial taxa (34–38), with the exception of *E. coli* K-12 strains which encode a less efficient RF2 (36). Therefore, even *prfB* genes containing a premature TAA stop codon would likely be subject to regulation by RF2 levels. The precise conservation of each of the frameshifting elements in taxa throughout the bacterial domain suggests that nearly all bacteria that autoregulate RF2 levels do so by the same molecular mechanism.

Previous work postulated that the ancestral *prfB* lacked the programmed ribosomal frameshift. This hypothesis was informed by an understanding of the bacterial tree as being rooted closest to Thermotogota and the absence of the programmed frameshift in *prfB* sequences from this phylum. We similarly did not detect the ribosomal frameshift in *prfB* sequences in

341 Thermotogota (Fig. 3A). However, the overwhelming majority of *prfB* sequences of all phyla in
342 proximity to Thermotogota do indeed utilize the programmed frameshift (Fig. 3A). Because current
343 models place the root of the bacterial tree in the neighborhood of Fusobacteria, and the majority
344 of taxa in this neighborhood, except Thermotogota, use the programmed frameshift in *prfB*, it is
345 likely and parsimonious that the ancestral *prfB* was regulated by a programmed ribosomal
346 frameshift.

347 Having determined that the ancestral *prfB* likely encoded the programmed frameshift, we
348 next examined the genome characteristics of taxa that lost the frameshift motif. We found that
349 bacteria that have lost the programmed frameshift in *prfB* had significantly higher RF2-specific
350 stop codon usage ($p < 2.2e^{-16}$) (Fig. 5A). Even when we excluded Actinobacteriota, which make
351 up a large proportion of genomes that do not encode the motif, this finding was still highly
352 significant ($p = 2.2e^{-06}$) (Fig. 5B). Bacterial release factor concentrations correlate with cognate
353 stop codon usage (12, 39). The direction of causality is unknown, but a recent *in silico* study
354 proposed that release factor concentrations adapted to stop codon usage (40). High GC content
355 strongly correlates with high RF2-specific TGA stop codon usage but not with TAG stop codon
356 usage (Fig S3) (12). Therefore, a likely explanation for RF2 autoregulation loss is that high GC
357 content and high RF2-specific stop codon usage increased the demand for RF2, and RF2
358 autoregulation was subsequently lost to satisfy this demand.

359 Across the bacterial tree of life, the most notable loss event occurred in the Actinobacterial
360 ancestor. None of the 2658 Actinobacterial genomes in our survey encoded the programmed
361 frameshift in *prfB*. Therefore, we hypothesized that Actinobacteriota may exhibit poor ribosomal
362 frameshifting efficiency at this motif. In support of this hypothesis, when we provided *M.*
363 *smegmatis* with a reporter encoding the *prfB* frameshift motif, the frameshift efficiency was ~6%.
364 We cannot conclude from this result whether *M. smegmatis* ribosomes are less able to frameshift
365 at this frameshift motif due to structural differences, or whether RF2 levels and activity in *M.*
366 *smegmatis* are high enough that termination at the UGA stop codon is favored over frameshifting.

367 In either case, these results suggest that encoding the frameshift motif in *prfB* would be
368 detrimental in *M. smegmatis* because it may result in decreased RF2 levels. Interestingly, the only
369 other organism reported to have similarly low frameshifting efficiency at the *prfB* motif is
370 *Flavobacterium johnsoniae,* which rarely uses Shine-Dalgarno sequences during translation
371 initiation (41, 42). Surprisingly, *F. johnsoniae* retains the *prfB* frameshift motif despite its low
372 frameshifting efficiency. The low efficiency may be tolerated because the *F. johnsoniae* genome
373 has an unusually low proportion of RF2-specific stop codons (7% of all stop codons) (41), and
374 therefore a low demand for RF2.

375 Although RF2 specific stop codon usage is highly correlated with loss of the motif, there
376 are notable exceptions. In particular, Mycoplasmatota have low genomic GC content and low TGA
377 stop codon usage, but Mycoplasmatota genomes that encode *prfB* all lack the frameshifting motif
378 (Fig. 4). The ancestral Mycoplasmatota underwent extreme genome reduction, and therefore,
379 many species in the phylum have lost *prfB,* instead decoding UGA as tryptophan (29, 43). It is
380 possible that in lineages that retain *prfB*, RF2 may be actively decaying and therefore lowly
381 expressed, even without autoregulation. Therefore, RF2 autoregulation is not required in these
382 lineages.

383 Our work supports a model in which the programmed ribosomal frameshift motif in *prfB*
384 was present in the last common ancestor of bacteria and autoregulates RF2 expression in most
385 bacterial species. In many species that have lost the motif, it is likely that high TGA stop codon
386 usage increased demand for RF2 and so RF2 autoregulation is no longer required. While our
387 work offers a comprehensive survey of the evolution and purpose of RF2 autoregulation, future
388 studies are needed to determine whether there are release factor or structural differences in the
389 ribosome that also affect frameshifting efficiency in diverse bacteria and may have selected for
390 loss of the *prfB* programmed frameshifting motif.

391

392

393 **Data Availability Statement**

394 Data for all genomes surveyed can be found in Table S1 and S2. Scripts for data acquisition and
395 analyses are available on GitHub at https://github.com/cassprince/prfB_evolution.

396

397 **Materials and Methods**

398

399 **Strains and media.** All strains were derived from *B. subtilis* 168 *trpC2* and *M. smegmatis* MC^2
400 155. *B. subtilis* was grown shaking in Lysogeny Broth media at 37˚C, and *M. smegmatis* was
401 grown shaking in 7H9 Middlebrook media at 37°C as indicated. Antibiotics were used at final
402 concentrations of 5 µg/mL chloramphenicol and 20 µg/mL kanamycin. Plasmids used in this study
403 are described in Table 1 and novel plasmid sequences are available on the project GitHub at
404 https://github.com/cassprince/prfB_evolution/blob/main/data/plasmid_sequences.fasta. All
405 experiments were performed in biological triplicate.

406

407 **Frameshift reporter lysates and western blots.** *M. smegmatis* overnight cultures were
408 normalized to an OD600 of 0.2. Cells were then harvested and pelleted after 12 hours. *B. subtilis*
409 overnight cultures were normalized to an OD600 of 0.05. Reporter expression was induced with
410 1mM IPTG when cultures reached an OD600 of 1. Cells were then harvested and pelleted after
411 30 minutes of induction. *B. subtilis* and *M. smegmatis* cell pellets were treated with lysis buffer
412 (10 mM Tris pH 8, 50 mM EDTA, 1 mg/mL lysozyme) for 10 minutes at 37°C. *M. smegmatis* cells
413 were further lysed using bead beating for five cycles of 20 seconds at 4350 rpm with 3 minutes
414 on ice between cycles. All lysates were mixed with SDS loading dye, boiled at 90°C for 5 minutes,
415 and cooled on ice. For *M. smegmatis*, the protein levels of the full-length reporter were
416 approximately 15x greater than those of the experimental *prfB* frameshifting reporters based on
417 band intensity. Therefore, the lysates for the full-length reporter strain were diluted 15x to
418 normalize the protein levels between reporters.

419 Proteins were run on a 12% SDS-PAGE gel for 70 minutes at 150 V. To measure *prfB*
420 overexpression levels and reporter frameshifting levels, proteins were transferred from SDS-
421 PAGE gels to a PVDF membrane (BioRad) for 100 minutes at 300 mAmps. The membrane was
422 blocked in 3% bovine serum albumin (BSA) overnight at 4°C. Anti-FLAG antibody conjugated to
423 horseradish peroxidase (Sigma SAB4200119) was added to the BSA for 1.5 hours at room
424 temperature. The membrane was washed with PBS-T three times for 5 minutes at room
425 temperature and developed with ECL substrate and enhancer (Biorad 170-5060). Band intensities
426 were quantified using ImageJ v1.53k (44). P-values for differences in band intensity were
427 calculated with the R stats v4.2.2 package using a Welch two-sampled *t*-test.
428
429 ***prfB* sequence acquisition and analyses.** *prfB* nucleotide sequences were downloaded from
430 all representative prokaryotic genomes in the NCBI RefSeq database as annotated by the NCBI
431 Prokaryotic Genome Annotation Pipeline (22). Only genomes with CheckM (45) contamination
432 scores below 10% were used, resulting in 12,751 genomes. The NCBI accession numbers,
433 species names, and taxids for all genomes used can be found in Table S1. *prfB* sequences were
434 aligned using MAFFT v7.453 (46). Python scripts were used to identify premature stop codons by
435 searching for any stop codon that was in-frame but not found in the final three nucleotides of the
436 sequence. The surrounding region was then extracted, and the identity of the stop codon was
437 recorded. If no premature stop codon was found, the expected region of the frameshifting motif
438 (based on multiple sequence alignment) was extracted. The scripts utilized the biopython v1.78
439 (47) package for sequence manipulation. The extracted regions were converted to sequence
440 logos using the logomaker v0.8 package (48).
441
442 **Phylogenetic analyses.** 16S rRNA sequences were identified and acquired using BLAST
443 v2.13.0. Sequences were aligned using MAFFT v7.453 (46). The alignments were applied to
444 FastTree v2.1.11 (49) to infer a maximum likelihood tree. Trees were visualized using the ggtree

445 v3.6.2 package (50). FastTree produces unrooted phylogenies, so trees were midpoint rooted
446 using the phangorn v2.11.1 package (51). The simplified tree in Figure 3A was produced by
447 randomly selecting a representative genome for each phylum and subsetting the large 16S tree
448 in Figure 3B. The identities of the randomly selected genomes can be found in Table S3.
449 Taxonomic classification was assigned to genomes using the NCBI Taxonomy database (52) and
450 taxonkit v0.14.1 (53). GC content for each genome was downloaded from NCBI. To determine
451 terminal stop codon usage, all coding sequences were downloaded as annotated by NCBI PGAP
452 for a random subset of 1000 genomes. The list of genomes in the subset can be found in Table
453 S2. A novel Python script recorded the last three nucleotides of each coding sequence per
454 genome, utilizing the biopython v1.78 package for sequence manipulation. P-values for
455 differences in GC content and terminal stop codon usage between “frameshift” and “no frameshift”
456 genomes were calculated with the R stats v4.2.2 package using a Welch two-sample t-test.
457
458
459 **Acknowledgements**
460 HAF and CRP were supported by NIH R35GM147049. CRP was supported by a Graduate
461 Research Fellowship from the National Science Foundation.
462
463
464
465
466

467 Table 1. Strains, Plasmids, and Primers

468

Strain	Description	Source
HAF1	*B. subtilis* wild type 168 trpC2	(54)
HAF477	*M. smegmatis* wild type MC^2 155	Kenneth Keiler
CP109	168 trpC2 pHF328 *sacA::$P_{hyperspank}$-3xFLAG-mcherry-gfp*	
CP127	168 trpC2 pHF328 *sacA::$P_{hyperspank}$-3xFLAG-mcherry-B.subtilis prfB frameshifting motif-gfp*	
CP271	168 trpC2 pHF328 *sacA::$P_{hyperspank}$-3xFLAG-mcherry-B.subtilis prfB recoded frameshifting motif-TGA-gfp*	
CP252	MC^2 155 pMV306hsp *P_{hsp60}-3xFLAG-mcherry-gfp* codon optimized for *M. smegmatis*	
CP253	MC^2 155 pMV306hsp *P_{hsp60}-3xFLAG-mcherry-B. subtilis prfB frameshifting motif-gfp* codon optimized for *M. smegmatis*	
CP267	MC^2 155 pMV306hsp *P_{hsp60}-3xFLAG-mcherry-B. subtilis prfB recoded frameshifting motif-TGA-gfp* codon optimized for *M. smegmatis*	

483

484

Plasmid	Description	Source

486

ECE743	empty vector, ori1030, XylR-P_{xylA} upstream of MCS, amp^r, mls^r (replicative)	(55)
pHF328	pDR111 $P_{hyperspank}$ and MCS cloned into ECE174 backbone at the BamHI and EcoRI sites. Integration at *sacA*.	This study
pCP105	pHF328 *sacA::$P_{hyperspank}$-3xFLAG-mcherry-gfp*	This study
pCP125	pHF328 *sacA::$P_{hyperspank}$-3xFLAG-mcherry-B.subtilis prfB FS-gfp*	This study
pCP269	pHF328 *sacA::$P_{hyperspank}$-3xFLAG-mcherry-B.subtilis prfB no FS motif-gfp*	This study
pCP234	pMV306hsp *P_{hsp60}-3xFLAG-mcherry-gfp* codon optimized for *M. smegmatis* (replicative)	This study
pCP238 (replicative)	pMV306hsp *P_{hsp60}-3xFLAG-mcherry-B. subtilis prfB FS-gfp* codon optimized for *M. smegmatis*	This study
pCP258 (replicative)	pMV306hsp *P_{hsp60}-3xFLAG-mcherry-B. subtilis prfB no FS-gfp* codon optimized for *M. smegmatis*	This study

Primer	Sequence
CP76-F	5'-ATAACAATTAAGCTTGGAGGAAAAAAAATGGATTATAAAGACGACGACG-3'
CP77-F	5'-ACCTTTAGACAGACCTGAATTCGAGCTCGGTACCC-3'
CP78-F	5'-CCGAGCTCGAATTCAGGTCTGTCTAAAGGTGAAGAACTG-3'

507 CP79-R 5’-TTGCATGCGGCTAGCTTATTTGTAGAGCTCATCCATGCCG-3’

508

509

510 # References

511

512 1. Tse H, Cai JJ, Tsoi H-W, Lam EP, Yuen K-Y. 2010. Natural selection retains overrepresented
513 out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics
514 11:491.

515 2. Atkins JF, Gesteland RF, Reid BR, Anderson CW. 1979. Normal tRNAs promote ribosomal
516 frameshifting. Cell 18:1119–1131.

517 3. Kastelein RA, Remaut E, Fiers W, van Duin J. 1982. Lysis gene expression of RNA phage MS2
518 depends on a frameshift during translation of the overlapping coat protein gene. Nature
519 295:35–41.

520 4. Larsen B, Gesteland RF, Atkins JF. 1997. Structural probing and mutagenic analysis of the
521 stem-loop required for Escherichia coli dnaX ribosomal frameshifting: programmed
522 efficiency of 50%11Edited By D. E. Draper. 1. Journal of Molecular Biology 271:47–60.

523 5. Meydan S, Klepacki D, Karthikeyan S, Margus T, Thomas P, Jones JE, Khan Y, Briggs J,
524 Dinman JD, Vázquez-Laslop N, Mankin AS. 2017. Programmed Ribosomal Frameshifting
525 Generates a Copper Transporter and a Copper Chaperone from the Same Gene. 2.
526 Molecular Cell 65:207–219.

527 6. Farabaugh PJ. 1996. Programmed translational frameshifting. Annu Rev Genet 30:507–528.

528 7. Dinman JD. 2012. Control of gene expression by translational recoding. Adv Protein Chem
529 Struct Biol 86:129–149.

530 8. Dinman JD. 2012. Mechanisms and implications of programmed translational frameshifting.
531 Wiley Interdiscip Rev RNA 3:661–673.

532 9. Advani VM, Dinman JD. 2016. Reprogramming the genetic code: The emerging role of
533 ribosomal frameshifting in regulating cellular gene expression. Bioessays 38:21–26.

534 10. Craigen WJ, Cook RG, Tate WP, Caskey CT. 1985. Bacterial peptide chain release factors:
535 conserved primary structure and possible frameshift regulation of release factor 2. Proc
536 Natl Acad Sci U S A 82:3616–3620.

537 11. Craigen WJ, Caskey CT. 1986. Expression of peptide chain release factor 2 requires high-
538 efficiency frameshift. 6076. Nature 322:273–275.

539 12. Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive Analysis of Stop Codon Usage
540 in Bacteria and Its Correlation with Release Factor Abundance *. 44. Journal of Biological
541 Chemistry 289:30334–30342.

542 13. Weiss RB, Dunn DM, Atkins JF, Gesteland RF. 1987. Slippery runs, shifty stops, backward
543 steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting. Cold Spring Harb
544 Symp Quant Biol 52:687–693.

545 14. Poole ES, Brown CM, Tate WP. 1995. The identity of the base following the stop codon
546 determines the efficiency of in vivo translational termination in Escherichia coli. 1. EMBO J
547 14:151–158.

548 15. Major LL, Poole ES, Dalphin ME, Mannering SA, Tate WP. 1996. Is the in-frame termination
549 signal of the Escherichia coli release factor-2 frameshift site weakened by a particularly
550 poor context? Nucleic Acids Res 24:2673–2678.

551 16. Poole ES, Major LL, Mannering SA, Tate WP. 1998. Translational termination in Escherichia
552 coli: three bases following the stop codon crosslink to release factor 2 and affect the
553 decoding efficiency of UGA-containing signals. Nucleic Acids Res 26:954–960.

554 17. Márquez V, Wilson DN, Tate WP, Triana-Alonso F, Nierhaus KH. 2004. Maintaining the
555 ribosomal reading frame: the influence of the E site during translational regulation of
556 release factor 2. Cell 118:45–55.

557 18. Weiss RB, Dunn DM, Dahlberg AE, Atkins JF, Gesteland RF. 1988. Reading frame switch
558 caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during
559 elongation of protein synthesis in Escherichia coli. The EMBO Journal 7:1503–1507.

560 19. Baranov PV, Gesteland RF, Atkins JF. 2002. Recoding: translational bifurcations in gene
561 expression. Gene 286:187–201.

562 20. Baranov PV, Gesteland RF, Atkins JF. 2002. Release factor 2 frameshifting sites in different
563 bacteria. 4. EMBO Rep 3:373–377.

564 21. Bekaert M, Atkins JF, Baranov PV. 2006. ARFA: a program for annotating bacterial release
565 factor genes, including prediction of programmed ribosomal frameshifting. 20.
566 Bioinformatics 22:2463–2465.

567 22. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F,
568 Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J,
569 Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2020. RefSeq:
570 expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model
571 curation. D1. Nucleic Acids Res 49:D1020–D1028.

572 23. Scolnick E, Tompkins R, Caskey T, Nirenberg M. 1968. Release factors differing in specificity
573 for terminator codons. Proc Natl Acad Sci U S A 61:768–774.

574 24. Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, Szöllősi GJ,
575 Williams TA. 2021. A rooted phylogeny resolves early bacterial evolution. 6542. Science
576 372:eabe0511.

577 25. Persson BC, Atkins JF. 1998. Does disparate occurrence of autoregulatory programmed
578 frameshifting in decoding the release factor 2 gene reflect an ancient origin with loss in
579 independent lineages? 13. J Bacteriol 180:3462–3466.

580 26. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN,
581 Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R,
582 Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:1–6.

583 27. Sirand-Pugnet P, Citti C, Barré A, Blanchard A. 2007. Evolution of mollicutes: down a bumpy
584 road with twists and turns. Research in Microbiology 158:754–766.

585 28. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ,
586 Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M,

587 Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF,
588 Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA, Venter JC.
589 1995. The minimal gene complement of Mycoplasma genitalium. Science 270:397–403.

590 29. Grosjean H, Breton M, Sirand-Pugnet P, Tardy F, Thiaucourt F, Citti C, Barré A, Yoshizawa S,
591 Fourmy D, Crécy-Lagard V de, Blanchard A. 2014. Predicting the Minimal Translation
592 Apparatus: Lessons from the Reductive Evolution of Mollicutes. PLOS Genetics
593 10:e1004363.

594 30. Inamine JM, Ho KC, Loechel S, Hu PC. 1990. Evidence that UGA is read as a tryptophan
595 codon rather than as a stop codon by Mycoplasma pneumoniae, Mycoplasma genitalium,
596 and Mycoplasma gallisepticum. J Bacteriol 172:504–506.

597 31. Citti C, Maréchal-Drouard L, Saillard C, Weil JH, Bové JM. 1992. Spiroplasma citri UGG and
598 UGA tryptophan codons: sequence of the two tryptophanyl-tRNAs and organization of the
599 corresponding genes. J Bacteriol 174:6471–6478.

600 32. Doi Y, Teranaka M, Yora K, Asuyama H. 1967. Mycoplasma- or PLT Group-like
601 Microorganisms Found in the Phloem Elements of Plants Infected with Mulberry Dwarf,
602 Potato Witches' Broom, Aster Yellows, or Paulownia Witches' Broom. Japanese Journal of
603 Phytopathology 33:259–266.

604 33. Hördt A, López MG, Meier-Kolthoff JP, Schleuning M, Weinhold L-M, Tindall BJ, Gronow S,
605 Kyrpides NC, Woyke T, Göker M. 2020. Analysis of 1,000+ Type-Strain Genomes
606 Substantially Improves Taxonomic Classification of Alphaproteobacteria. Front Microbiol
607 11:468.

608 34. Mora L, Heurgué-Hamard V, de Zamaroczy M, Kervestin S, Buckingham RH. 2007.
609 Methylation of bacterial release factors RF1 and RF2 is required for normal translation
610 termination in vivo. J Biol Chem 282:35638–35645.

611 35. Elliott T, Wang X. 1991. Salmonella typhimurium prfA mutants defective in release factor 1.
612 J Bacteriol 173:4144–4154.

613 36. Uno M, Ito K, Nakamura Y. 1996. Functional specificity of amino acid at position 246 in the
614 tRNA mimicry domain of bacterial release factor 2. Biochimie 78:935–943.

615 37. Baggett NE, Zhang Y, Gross CA. 2017. Global analysis of translation termination in E. coli.
616 PLoS Genet 13:e1006676.

617 38. Dinçbas-Renqvist V, Engström Å, Mora L, Heurgué-Hamard V, Buckingham R, Ehrenberg M.
618 2000. A post-translational modification in the GGQ motif of RF2 from Escherichia coli
619 stimulates termination of translation. EMBO J 19:6900–6907.

620 39. Wei Y, Wang J, Xia X. 2016. Coevolution between Stop Codon Usage and Release Factors in
621 Bacterial Species. Molecular Biology and Evolution 33:2357–2367.

622 40. Ho AT, Hurst LD. 2022. Variation in Release Factor Abundance Is Not Needed to Explain
623 Trends in Bacterial Stop Codon Usage. Molecular Biology and Evolution 39:msab326.

624 41. Naeem FM, Gemler BT, McNutt ZA, Bundschuh R, Fredrick K. 2024. Analysis of programmed
625 frameshifting during translation of prfB in Flavobacterium johnsoniae. 2. RNA 30:136–148.

626 42. McNutt ZA, Gandhi MD, Shatoff EA, Roy B, Devaraj A, Bundschuh R, Fredrick K. 2021.
627 Comparative Analysis of anti-Shine- Dalgarno Function in Flavobacterium johnsoniae and
628 Escherichia coli. Front Mol Biosci 8:787388.

629 43. Yamao F, Iwagami S, Azumi Y, Muto A, Osawa S, Fujita N, Ishihama A. 1988. Evolutionary
630 dynamics of tryptophan tRNAs in Mycoplasma capricolum. Mol Gen Genet 212:364–369.

631 44. Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image
632 analysis. Nat Methods 9:671–675.

633 45. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing
634 the quality of microbial genomes recovered from isolates, single cells, and metagenomes.
635 Genome Res 25:1043–1055.

636 46. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
637 Improvements in Performance and Usability. 4. Molecular Biology and Evolution 30:772–
638 780.

639 47. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff
640 F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for
641 computational molecular biology and bioinformatics. 11. Bioinformatics 25:1422–1423.

642 48. Tareen A, Kinney JB. 2020. Logomaker: beautiful sequence logos in Python. 7.
643 Bioinformatics 36:2272–2274.

644 49. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees
645 for Large Alignments. 3. PLOS ONE 5:e9490.

646 50. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and
647 annotation of phylogenetic trees with their covariates and other associated data. 1.
648 Methods in Ecology and Evolution 8:28–36.

649 51. Schliep KP. 2011. phangorn: phylogenetic analysis in R. 4. Bioinformatics 27:592–593.

650 52. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R,
651 O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I.
652 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database
653 (Oxford) 2020:baaa062.

654 53. Shen W, Ren H. 2021. TaxonKit: A practical and efficient NCBI taxonomy toolkit. Journal of
655 Genetics and Genomics 48:844–850.

656 54. Gaidenko TA, Kim T-J, Price CW. 2002. The PrpC Serine-Threonine Phosphatase and PrkC
657 Kinase Have Opposing Physiological Roles in Stationary-Phase Bacillus subtilis Cells. J
658 Bacteriol 184:6109–6114.

659 55. Popp PF, Dotzler M, Radeck J, Bartels J, Mascher T. 2017. The Bacillus BioBrick Box 2.0:
660 expanding the genetic toolbox for the standardized work with Bacillus subtilis. Sci Rep
661 7:15058.

662

Supplemental Figures

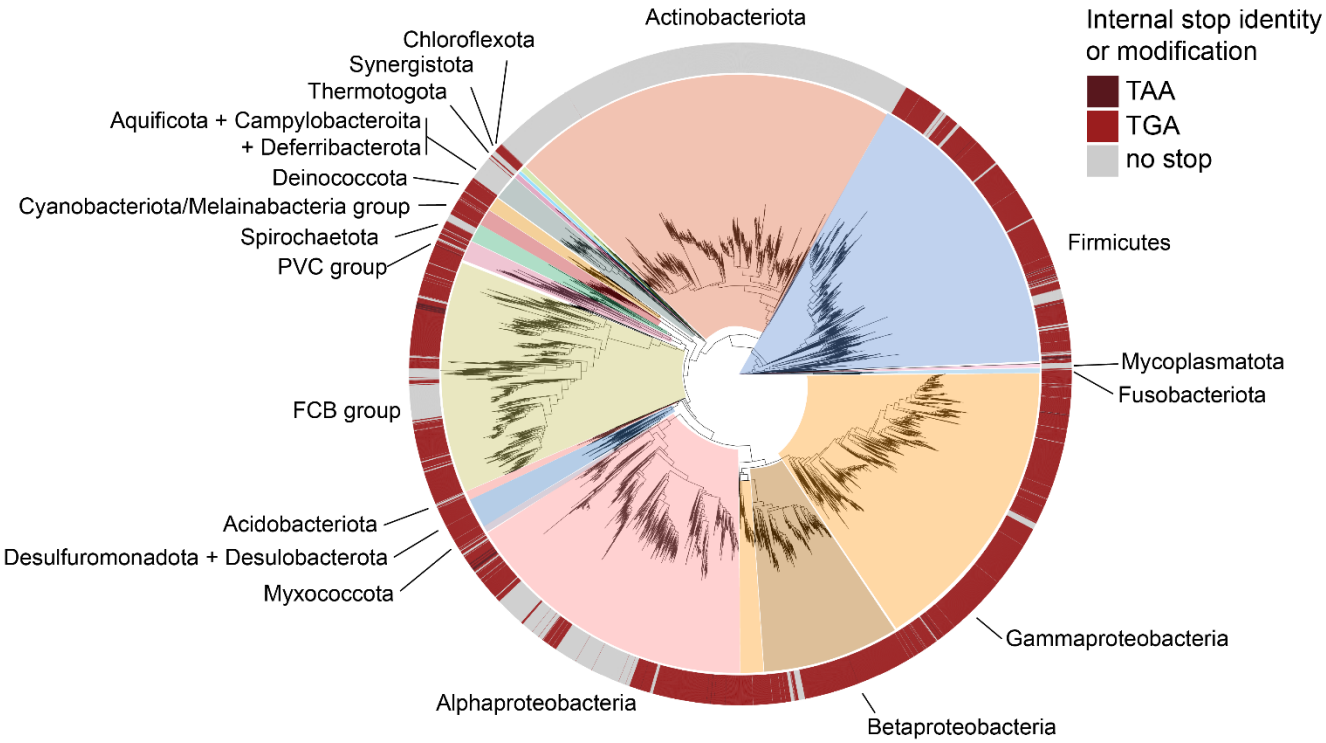


Figure S1. TAA is poorly represented and randomly distributed as the premature stop codon in the frameshifting motif. 16S maximum-likelihood phylogenetic tree from Figure 3B showing the distribution of premature stop codons in *prfB*. Phyla with more than 10 available and high-quality reference genomes are shown.

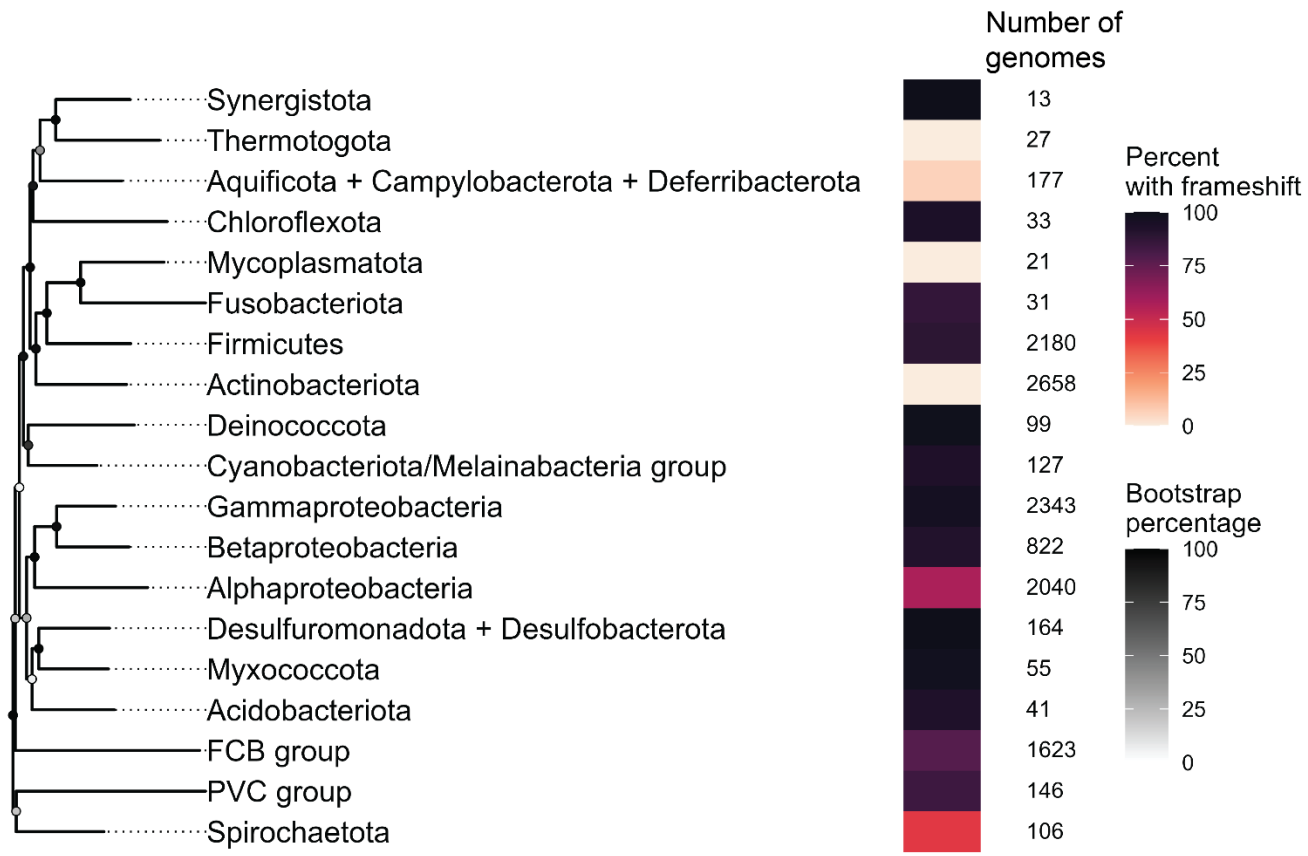


Figure S2. Percent of genomes containing the *prfB* frameshifting motif across phyla. A subtree of the large 16S tree was created using a random single representative genome for each phylum with more than 10 genomes. The number of analyzed genomes per phylum is located to the right of each bar. Darkened circles indicate bootstrap values for each node.

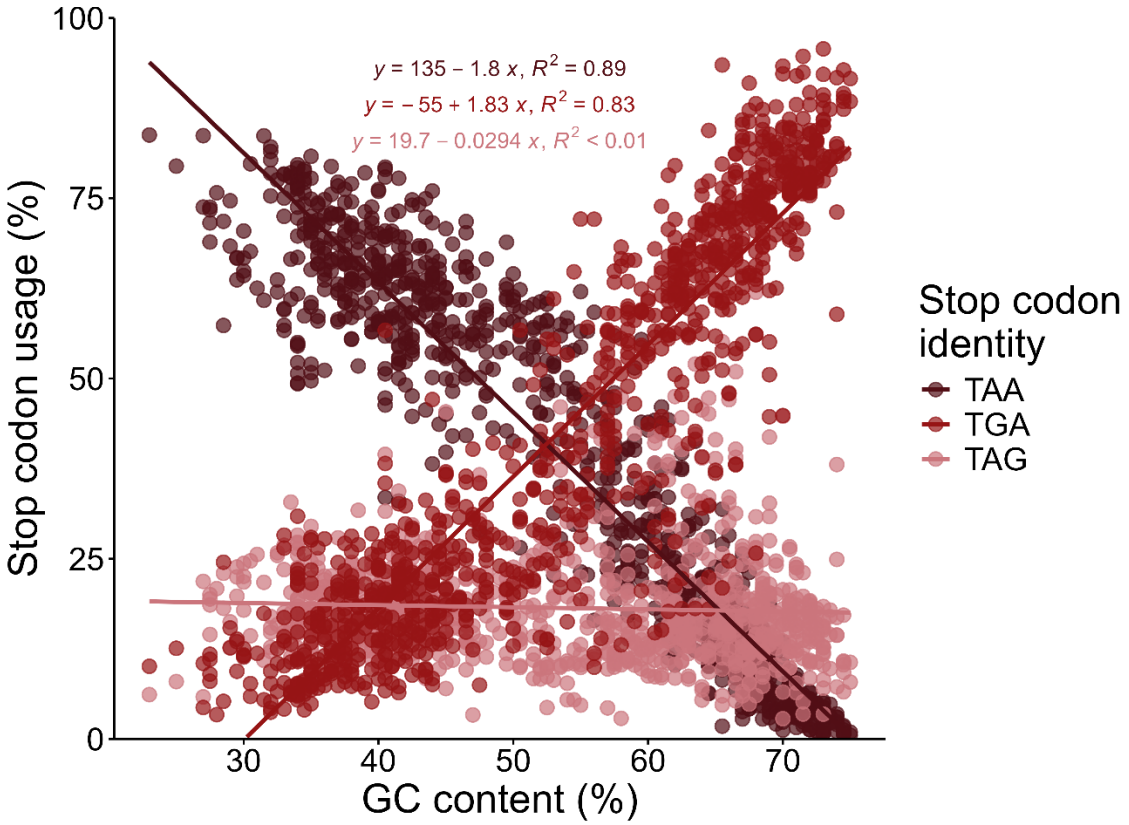


Figure S3. TGA stop codon usage increases with GC content. Terminal stop codon usage in all coding sequences for a subset of 1000 random reference genomes plotted verses GC content. Each genome has one point for each of the three stop codons.

Supplemental table legends

Table S1. Metadata of the 12,571 genomes used in this study.

Table S2. Total stop codon usage for the random subset of 1,000 genomes in Figure 5.

Table S3. Accession numbers for random representatives of phyla in Figure 3A and S2.