Selection for translational efficiency in genes associated with alphaproteobacterial gene transfer agents

Roman Kogay[1] and Olga Zhaxybayeva[1,2,#]

[1]*Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire, USA*

[2]*Department of Computer Science, Dartmouth College, Hanover, New Hampshire, USA*

[#]**Corresponding author:** olga.zhaxybayeva@dartmouth.edu

Running Title: Selection for Translational Efficiency in GTAs

Keywords: GTA, *Sphingomonadales*, codon usage bias, nutrient depletion, homologous recombination, tonB, addAB, head completion protein

Abstract

Gene transfer agents (GTAs) are virus-like elements that are encoded by some bacterial and archaeal genomes. The production of GTAs can be induced by the carbon depletion and results in host lysis and release of virus-like particles that contain mostly random fragments of the host DNA. The remaining members of a GTA-producing population act as GTA recipients by producing proteins needed for the GTA-mediated DNA acquisition. Here, we detect a codon usage bias towards codons with more readily available tRNAs in the RcGTA-like GTA genes of alphaproteobacterial genomes. Such bias likely improves the translational efficacy during GTA gene expression. While the strength of codon usage bias fluctuates substantially among individual GTA genes and across taxonomic groups, it is especially pronounced in *Sphingomonadales*, whose members are known to inhabit nutrient-depleted environments. By screening genomes for gene families with similar trends in codon usage biases to those in GTA genes, we found a gene that likely encodes head completion protein in some GTAs were it appeared missing, and 13 genes previously not implicated in GTA lifecycle. The latter genes are involved in various molecular processes, including the homologous recombination and transport of scarce organic matter. Our findings provide insights into the role of selection for translational efficiency in evolution of GTA genes, and outline genes that are potentially involved in the previously hypothesized integration of GTA-delivered DNA into the host genome.

42 ## Importance

43 Horizontal gene transfer (HGT) is a fundamental process that drives evolution of
44 microorganisms. HGT can result a rapid dissemination of beneficial genes within and among
45 microbial communities, and can be achieved via multiple mechanisms. One peculiar HGT
46 mechanism involves viruses "domesticated" by some bacteria and archaea (their hosts). These
47 so-called gene transfer agents (GTAs) are encoded in hosts' genomes, produced under starvation
48 conditions, and cannot propagate themselves as viruses. We show that GTA genes are under
49 selection to improve efficiency of their translation when the host activates GTA production. The
50 selection is especially pronounced in bacteria that occupy nutrient-depleted environments.
51 Intriguingly, several genes involved in DNA incorporation into a genome are under similar
52 selection pressure, suggesting that they may facilitate integration of GTA-delivered DNA into
53 the host genome. Our findings underscore the potential importance of GTAs as a mechanism of
54 HGT under nutrient-limited conditions, which are widespread in microbial habitats.
55

56 ## Introduction

57 Gene transfer agents are phage-like particles produced by multiple groups of bacteria and
58 archaea (1). Unlike viruses, GTA particles tend to package random pieces of the host cell DNA
59 instead of genes that encode GTAs themselves (2, 3). Released GTA particles can deliver the
60 packaged genetic material to other cells (4), impacting exchange of genetic material in
61 prokaryotic populations (5-7). The benefits of GTA production and GTA-mediated DNA
62 acquisition are not well understood. It has been hypothesized that GTAs may facilitate DNA
63 repair (8), enable population-level exchange of traits needed under the conditions of a nutritional
64 stress via horizontal gene transfer (HGT) (5) or decrease population density during the carbon
65 starvation periods (9).

66 To date, at least three independently exapted GTAs are functionally characterized (10).
67 The most studied GTA system (RcGTA) belongs to the alphaproteobacterium *Rhodobacter*
68 *capsulatus* (2). RcGTA is encoded by at least 24 genes that are distributed across 5 distinct
69 genomic loci (11, 12). Seventeen of the 24 genes are situated in one locus, which is dubbed the
70 'head-tail' cluster because it encodes most of the structural proteins of the RcGTA particles (1).
71 RcGTA-like 'head-tail' clusters are present in many alphaproteobacterial genomes; they evolve
72 slowly and are inferred to be inherited mostly vertically from a common ancestor of an
73 alphaproteobacterial clade that spans multiple taxonomic orders (12-14). Additionally, multiple
74 cellular genes regulate RcGTA production, release and reception (11, 15). It is likely that other,
75 yet undiscovered, genes in *R. capsulatus* genome are involved in GTA lifecycle.

76 Expression of RcGTA is known to be triggered by nutrient depletion (16), under which a
77 small fraction of the *R. capsulatus* population becomes dedicated to GTA production (17, 18).
78 As a result, RcGTA-producing cells likely express GTA genes at high levels. By extension,

79 RcGTA-like GTA genes in other alphaproteobacteria (hereafter referred to as "GTA genes" for
80 brevity) also likely to be highly expressed in GTA-producing cells of alphaproteobacterial
81 populations.

82 Highly expressed genes that are involved in core biological processes, such as
83 translational machinery, are known to exhibit a strong codon usage bias (19). For example,
84 codon usage in ribosomal proteins, which are highly expressed in almost all organisms, deviates
85 most dramatically from the distribution of codons expected under their equal usage corrected for
86 organismal GC content (20). Such bias is primarily due to selection to match the pool of most
87 abundant tRNA molecules in order to have the most efficient translation for proteins needed in
88 high number of copies (21-23). As a result, highly expressed genes tend to have codons that
89 correspond to the most abundant tRNA molecules in the cell. This type of selection is known as
90 the "selection for translational efficiency" and is ubiquitous among bacteria (24).

91 Besides constitutively highly expressed genes, selection for translational efficiency also
92 acts on genes that are highly expressed under specific environmental conditions that
93 microorganisms experience (19, 24, 25). For instance, genes that utilize galactose have higher
94 codon usage biases in budding yeasts that live in dairy-associated habitats than in yeasts that
95 occupy alcohol-associated habitats (25). Additionally, genes that encode interacting proteins and
96 genes involved in the same pathway often exhibit similar codon usage biases (25, 26).

97 In earlier work, we have discovered that alphaproteobacterial GTA genes have a striking
98 bias towards GC-rich codons in comparison to the rest of the genome (9, 12). However, this bias
99 is different from the codon usage bias: skewed composition of the encoded proteins towards
100 containing energetically cheaper amino acids caused the first two positions of the codons to be
101 enriched in Gs and Cs, due to the structure of the genetic code (9). In this study, we examined

102 GTA genes in 208 alphaproteobacterial genomes and assessed if there is additionally a codon
103 *usage* bias due to the genes being under selection for the translation efficiency. For this purpose,
104 we used two well-established metrics for assessment of codon usage bias and its match to the
105 tRNA abundance: effective number of codons (ENC) (20) and tRNA adaptation index (tAI) (27).
106 ENC quantifies how equally synonymous codons are used in a gene, and varies from 20 (when
107 only one codon is used per each amino acid; strong bias) to 61 (when all codons are used
108 equally; no bias). The tAI measures how optimally the codon usage of each gene fits the
109 available tRNA pool by correlating frequency of each codon in the gene with the abundance of
110 its cognate tRNA. The degree of adaptation of a gene is gauged by comparing its tAI value to the
111 tAI values of all other genes in a genome. We also searched for genes whose involvement in
112 GTA production and regulation is currently unsuspected by screening GTA-encoding genomes
113 for genes with codon usage patterns similar to those of GTA genes.
114

115 ## Results

116 **Codon usage bias of GTA genes and its match to available tRNAs varies across GTA genes**
117 **and GTA-containing genomes**

118 To assess the presence of codon usage bias in GTA genes across alphaproteobacteria, we
119 have calculated ENC for each "reference GTA gene" (see **Methods** for the definition) and
120 compared them against the expected ENC of a gene in a genome under the null model of no
121 codon usage bias, corrected for the genomic GC content (28). Indeed, we found that 1,543 out of
122 2,308 (66.8%) reference GTA genes detected across 208 GTA head-tail clusters deviate from the
123 genome-specific null expectations by more than 10% (**Supplemental Figure S1**). However,
124 there is a substantial variation in this deviation for different GTA genes (**Supplemental Figure**

125 **S2**), and only in genes *g5* and *g8* the deviation is significantly higher than the genomic average
126 (Kruskal-Wallis rank sum test, p-value < 2.2e-16; Dunn's test, p-value < 0.05, Benjamini-
127 Hochberg correction).
128 To assess the match of the observed codon usage bias to available tRNA pool, we
129 calculated tAI values of the reference GTA genes across 208 genomes and converted them to
130 percentile tAI values (ptAI; see **Methods** for the definition) to allow for the intergenomic
131 comparisons. Similar to the ENC values, the ptAI values also vary substantially across the genes
132 and genomes (**Figure 1**), suggesting that the strength of selection for translational efficiency
133 should be examined in individual GTA genes and in specific taxonomic groups, which we
134 investigate in the next two sections.
135
136 **Selection for translational efficiency is uneven among GTA genes**
137 The differences of ptAI values among the reference GTA genes are statistically
138 significant (Kruskal-Wallis rank sum test, p-value < 2.2e-16) (**Figure 1**). Particularly notable is a
139 significant decline in ptAI values of the region encoding genes *g12* through *g15* (Dunn's test, p-
140 value < 0.05, Benjamini-Hochberg correction), which are located at the 3' end of the head-tail
141 cluster and encode the tail components of a GTAs particle. In contrast, ptAI values of the genes
142 *g5* (encoding major capsid protein) and *g11* (encoding tail tape measure protein) are significantly
143 higher than ptAI values of other GTA genes (Dunn's test, p-values < 0.05, Benjamini-Hochberg
144 correction). Notably, protein g5 is detected in the largest number of copies (145) per RcGTA
145 particle than any other protein (29), while proteins g12-g15 are present in a small number of
146 copies (1-6) per RcGTA particle (29). Given that genes encoding proteins needed in a larger
147 number of copies have a higher degree of adaptation to the tRNA pool (30), we hypothesize that

148 the observed variation in ptAI values of GTA genes reflects the different number of GTA
149 proteins in a GTA particle. Protein g11, however, is found in only 3 copies per RcGTA particle
150 (29) and therefore a demand for a larger copy number cannot explain its high ptAI values.
151 Variation of ptAI values could also be due to physical location of the genes in the GTA
152 head-tail cluster. Similar to the operons (31), genes in the RcGTA head-tail cluster are co-
153 transcribed from a single promoter upstream of the cluster (15, 32). Because genes at the 3' end
154 of operons tend to have lower expression levels (33), the low ptAI values of GTA genes *g12-g15*
155 may be due to their distant location from the promoter.
156
157 **Selection for translational efficiency is the strongest in *Sphingomonadales*' genomes**
158 In addition to variability in ptAI values across different GTA genes, ptAI values of
159 individual GTA genes vary substantially across the 208 genomes (**Figure 1**). To evaluate if these
160 differences represent variation in selection pressure in distinct taxonomic groups, we initially
161 examined the ptAI values of gene *g5* that were grouped by alphaproteobacterial order. The *g5*
162 gene was chosen due to its high abundance of the encoded protein in RcGTA particles (more
163 copies than all other structural proteins combined) and for being the only gene with the highest
164 detected deviations from the average genomic values for both ENC and ptAI. We found that ptAI
165 values of the *g5* gene vary significantly among members of the four alphaproteobacterial orders
166 (Kruskal-Wallis rank sum test, p-value < 0.05) (**Figure 2A**). In particular, *g5* genes from the
167 *Sphingomonadales*' genomes have significantly higher ptAI values than those from genomes of
168 bacteria from other three orders (Mann-Whitney U test, p-value < 0.05, Benjamini-Hochberg
169 correction). Twelve of the fourteen *g5* genes with the highest overall ptAI values (> 90) (**Figure**
170 **2A**) also belong to the *Sphingomonadales* genomes. Beyond just *g5* gene, all reference GTA

171 genes, as a group, have higher ptAI values in *Sphingomonadales* than in members of the three
172 other alphaproteobacterial orders (Mann-Whitney U test, p-value < 0.05, Benjamini-Hochberg
173 correction) (**Supplemental Figure S3**). These observations suggests that in *Sphingomonadales*
174 in particular, there is a strong selection for efficient production of GTA particles. Because
175 *Sphingomonadales* are known to live in nutrient-depleted environments (34), we suggest that
176 GTA production is especially beneficial in those habitats to exert strong selection for
177 translational efficiency.
178
179 **The increase in translational efficiency of GTA genes is associated with a reduced energetic**
180 **cost for production of the encoded proteins**
181 Among the GTA proteins in four alphaproteobacterial orders, *Sphingomonadales'* GTA
182 proteins also have the strongest skew in amino acid composition towards energetically cheaper
183 amino acids (**Figure 2B**). To evaluate if selection for energy efficiency is linked to selection for
184 translational efficiency, we examined the relationship between the ptAI values of GTA genes and
185 the number of carbons in amino acid chains encoded by the *Sphingomonadales* GTA genes. We
186 found that there is a significant negative correlation between them (Pearson R = -0.19, N = 636,
187 p-value < 0.05). We propose that in *Sphingomonadales* benefits associated with production of
188 GTA particles in nutrient-limited conditions led not only to the selection for translational
189 efficiency, but also to the selection for use of energetically cheaper amino acids in the GTA
190 genes.
191

192 **Fourteen gene families have translational efficiency trends similar to those of GTA genes**

193 While the strength of selection for translational efficiency acting on GTA genes varies
194 across gene and genomes, we found that the combinations of ptAI values across all reference
195 GTA genes in a genome have similar trends across the 208 genomes. The similarity is significant
196 in all pairwise reference GTA gene comparisons (**Supplemental Figure S4)**, as determined
197 using phylogenetic generalized least squares (PGLS) method. We conjecture that ptAI values of
198 genes in the other loci of a GTA "genome", as well as the host genes involved in GTA lifecycle,
199 would exhibit similar trends to the ptAI values of reference GTA genes, allowing for discovery
200 of yet unsuspected genes involved in GTA lifecycle. To identify such unknown genes that may
201 be co-expressed with GTA genes, we examined correlations of ptAI values between reference
202 GTA genes and 3,477 other gene families present in 208 alphaproteobacterial genomes. The
203 PGLS analysis revealed 14 gene families, whose ptAI values correlate significantly with ptAI
204 values of the reference GTA genes (**Table 1)**.

205 One of 14 identified gene families is a homolog of *gafA*, which encodes a crucial
206 transcription activator of GTA particles production in *Rhodobacter capsulatus* (11, 15). This
207 gene is located outside of the RcGTA's head-tail cluster, and therefore was not included in the
208 set of reference GTA genes, but its discovery demonstrates the suitability of our approach to
209 identify genes linked to the GTA lifecycle. Interestingly, *gafA* homologs were previously
210 described only in the genomes of *Rhodobacterales* and some *Rhizobiales* (11, 12, 15). However,
211 with different criteria in the OrthoFinder-based similarity searches, we were able to identify this
212 regulator in 196 of the 208 genomes (94.7%), spanning all GTA-containing alphaproteobacterial
213 orders. The evolutionary history of the *gafA* homologs is largely congruent with the reference
214 phylogenomic tree (normalized quartet score of 0.87) and even more so with the phylogeny of

215 the concatenated GTA reference genes (normalized quartet score of 0.92) (tree topologies are
216 available at https://doi.org/10.6084/m9.figshare.20082749), suggesting that the *gafA* gene had
217 co-evolved with the GTA 'head-tail' cluster since the last common ancestor of RcGTA-like
218 GTAs.

219 The remaining 13 gene families belong to several functional categories of the Clusters of
220 Orthologous Groups (COG) classification (**Table 1**). While proteins encoded by some of these
221 genes can be postulated to be involved in GTA lifecycle (exemplified below by the *addAB*, *xseA*,
222 and *tonB* genes), similarity of codon usage biases between other genes and reference GTA genes
223 can be explained by their expression at similar environmental conditions (exemplified by three
224 genes from 'molecular chaperones and related functions' COG category).

225 Protein products of the *addA* and *addB* genes form the heterodimeric helicase-nuclease
226 complex that repairs double-stranded DNA breaks by homologous recombination and is
227 functionally equivalent to the RecBCD complex (35). The knockout of the AddAB complex is
228 associated with a deficiency in RecA-dependent homologous recombination (36). We
229 hypothesize that the *addAB* pathway is involved in recombination of GTAs' genetic material
230 with the host's genome.

231 The main function of exodeoxyribonuclease VII large subunit (xseA), which is encoded
232 by the *xseA* gene, is to form a complex with xseB and degrade single-stranded DNA to
233 oligonucleotides. However, expression of *xseA* gene without *xseB* gene leads to cell death (37).
234 Because we did not find any correlation of codon usage bias between the *xseB* gene and GTA
235 genes, we speculate that instead of involvement in processing of GTA DNA, *xseA* gene product
236 facilitates lysis of GTA producing cells and release of GTA particles.

237 The *tonB* gene encodes tonB energy transducer. TonB-dependent transporters are
238 involved in transport of diverse compounds, including carbohydrates, amino acids, lipids,
239 vitamins and iron (38-40). Similar to the quorum-sensing regulated expression of the gene
240 encoding GTA receptor in the non-GTA-producing cells of a *Rhodobacter capsulatus* population
241 (4), the *tonB* gene could also be regulated to be expressed in the non-GTA-producing cells to aid
242 the uptake of the nutrients released from the lysed cells via TonB-dependent transporters. The
243 *tonB* gene is currently detected only in members of *Sphingomonadales* order, suggesting that
244 such nutrient uptake is most relevant in the nutrient-limited environments.

245 Three genes from the 'molecular chaperones and related functions' COG category are
246 less likely to be directly involved in GTA lifecycle, because GTAs already encode their own
247 chaperones that assist GTA protein folding (29). However, it is well known that chaperones tend
248 to be highly expressed in bacteria at times of stress and facilitate the survival of cells in rapidly
249 changing environmental conditions (41). Because chaperones are essential in responding to the
250 starvation-induced cellular stresses (42), we conjecture that observed similarity in ptAI values of
251 the reference GTA genes is due to their expression being triggered by the similar environmental
252 conditions.

253 To evaluate if the detected gene families interact with each other and with GTA genes,
254 we have constructed the protein-protein interaction network of the 14 gene families, 12 GTA
255 reference genes and 50 additional interactor proteins from the STRING database (**Figure 3**).
256 Thirteen of the 14 families and all 12 reference GTA genes belong to two protein-protein
257 interaction sub-networks (**Figure 3**), one of which contains all GTA reference genes, while the
258 other is involved in a wide range of functions (**Table 1**). By carrying out the KEGG enrichment
259 analysis, we found significant overrepresentation of four molecular pathways in the second

260 protein-protein interaction network (**Supplemental Table S1**). Consistent with the 6 of the 13
261 gene families being assigned to the "replication, recombination, and repair" COG category, two
262 of the KEGG pathways are 'homologous recombination' and 'mismatch repair', further
263 corroborating involvement of identified genes in integration of the genetic material delivered by
264 GTAs into recipients' genomes. Two additional pathways, 'carotenoid biosynthesis' and
265 'terpenoid backbone biosynthesis', are less likely to be directly involved in the lifecycle of
266 GTAs. Production of secondary metabolites is known to be protective against stress factors (43,
267 44), and carbon starvation leads to the upregulation of carotenoid biosynthesis pathway (45, 46).
268 Similar to the above-described genes encoding chaperones, we hypothesize that expression of
269 'carotenoid biosynthesis' and 'terpenoid backbone biosynthesis' genes is not related to GTA
270 lifecycle, but is initiated by conditions that also activate production of GTAs.
271

272 **A replacement of the head completion protein in *Sphingomonadales*' GTAs**

273 Gene content of GTA head-tail clusters varies across alphaproteobacteria (12). While
274 some clusters do not contain homologs of all RcGTA genes, others include additional genes that
275 are conserved across multiple clusters but have no known function (12, 13). To predict whether
276 any of these additional genes play a role in GTA production, we compared their ptAI values of
277 genes found in at least 10 genomes to ptAI values of the reference GTA genes. One gene family,
278 which is found only within GTA head-tail clusters of 11 genomes in one subclade of
279 *Sphingomonadales* (GenBank accessions are available at
280 https://doi.org/10.6084/m9.figshare.20082749), has a significant positive correlation with 5 out
281 of the 12 GTA reference genes (**Supplemental Table S2**). Interestingly, within
282 *Sphingomonadales* GTA head-tail clusters this gene is located where the *g7* gene, which encodes

283 a head completion protein, is found in the RcGTA head-tail cluster (**Supplemental Figure S5**).
284 Only seven of the 55 *Sphingomonadales* genomes in our dataset have detectable homologs of the
285 *g7* gene. Among the remaining 48 genomes, 22 contain a gene encoding a protein of unknown
286 function in the "gene *g7* locus", while 26 genomes don't have any gene in that locus.
287 The members of the identified gene family are substantially shorter than the RcGTA gene
288 *g7* and have a different secondary structure (**Figure 4**), precluding the possibility that the
289 identified protein is simply too divergent for a detectable amino acid similarity. However, we
290 found viral head completion proteins that have similar protein length and similar secondary
291 structures to both GTA head completion protein and the identified gene family (**Figure 4**). We
292 conjecture that the gene encoding the head completion protein was replaced in some
293 *Sphingomonadales* by a gene encoding an analogous viral protein.
294

295 ## Discussion

296 Our analyses of codon usage biases suggest that alphaproteobacterial GTA systems are
297 under selection for an optimal translation of GTA proteins from GTA genes achieved by using
298 codons with more readily available tRNAs. The strength of such selection for translational
299 efficiency is the most pronounced (and therefore most easily detectable) in the major capsid
300 protein gene, which is needed to be expressed to produce thousands of copies per GTA-
301 producing bacterium. Additionally, the strength of the selection for translational efficiency varies
302 across taxonomic groups, but is particularly prominent in *Sphingomonadales* order, whose
303 members typically inhabit nutrient-limited conditions. We hypothesize that the observed
304 variation in the selection strength depends on severity and duration of the nutrient scarcity
305 experienced by a population capable of producing GTAs. On the one hand, a long-term exposure

306 to nutrient-depleted conditions would trigger a more efficient and/or more frequent production of
307 GTA particles, which would lead to a greater survival of communities with a better translational
308 efficiency of GTA systems and thus a higher codon usage bias in the GTA genes. On the other
309 hand, if GTAs are needed only in rare occasions due to the stable and abundant nutrient supplies,
310 the selection for translational efficiency would be weak and would result in a lower codon usage
311 bias. Combined with an observation that production of GTAs is triggered by the nutritional stress
312 (16), our findings that the selection is the strongest in alphaproteobacteria that inhabit nutrient-
313 limited environments further underscore the earlier hypothesized importance of GTA systems in
314 situations of nutrient scarcity (9).

315 Additionally, the stronger selection for translation efficiency in GTA genes is associated
316 with a larger decline in the carbon content of the proteins the genes encode. These findings
317 suggest that benefits associated with GTA production are substantial enough to drive selection
318 for both translational efficiency and low energetic costs of the translated proteins. We speculate
319 that these modifications of GTA proteins allow the bacterial population under adverse conditions
320 to increase both the speed of GTA particle production and the number of released GTA particles.

321 We hypothesized that genes that are located outside of the GTA head-tail cluster, but are
322 involved in GTA lifecycle, including processing and integration of the GTA-delivered DNA,
323 would have signatures of selection for translational efficiency similar to those of GTA genes.
324 Gratifyingly, our genome-wide screen for such patterns detected the direct GTA activator gene,
325 *gafA* (15). We also identified multiple genes not yet implicated in GTA lifecycle. Several of
326 these genes are involved in recombination and mismatch repair, providing bioinformatic
327 evidence for the hypothesis that GTAs facilitate HGT by distributing genetic fragments that
328 become incorporated into recipients' genomes via homologous recombination (4). Involvement

329 of other genes with similar selection pressures in GTA lifecycle is speculative and needs to be
330 investigated experimentally. But the putative co-expression of *xseA* and *tonB* genes with GTA
331 genes raises an intriguing possibility that, in addition to HGT, GTA production may provide an
332 extra benefit in a nutrient-depleted environment: scavenging of scarce organic matter from GTA-
333 producing cells. The lysis of the GTA-producing cells could be mediated by XseA and their
334 debris could be imported as nutrients by the surviving cells via TonB-dependent transporters.
335 Alphaproteobacterial GTAs likely originated millions of years ago from a lysogenic
336 phage, and since then they were mostly vertically inherited by many alphaproteobacterial
337 lineages (12, 14). However, similar to the HGT influence onto many other regions of a typical
338 bacterial genome (47), it is very likely that over time GTAs experienced gene replacements via
339 HGT (12). Instances of HGT between GTAs and phages have been already documented (11, 48).
340 By examining the patterns of selection for translational efficiency, we identified another case of
341 likely ancient gene exchange with viruses that resulted in the replacement of the gene encoding
342 head completion protein in some *Sphingomonadales*. Curiously, the gene currently has no
343 significant primary sequence similarity to any gene in GenBank. Many other unannotated ORFs
344 in alphaproteobacterial head-tail clusters outside of *Rhodobacterales* (12) may also have
345 functional roles in their respective GTA regions. Notably, when alphaproteobacterial RcGTA-
346 like genomic regions appear incomplete due to lack of many homologs to genes required for
347 GTA production in *R. capsulatus*, it could be due our inability to recognize some genes due to
348 their replacements with analogous genes. Because such incomplete RcGTA-like clusters are
349 abundant in alphaprotebacteria (12), GTAs could be morphologically diverse and even more
350 widespread across alphaproteobacteria than we currently estimate (13).
351

352 ## Methods

353 **Dataset of representative alphaproteobacterial genomes with GTA head-tail clusters**

354 As an initial data set, we selected 212 representative alphaproteobacterial genomes
355 previously predicted to contain GTAs (9). The gene annotations of the genomes were
356 downloaded from the RefSeq database (49) in October 2020. GTA head-tail clusters (1) were
357 predicted using the GTA-Hunter program (13). Because GTA-Hunter identifies only 11 out of
358 the 17 genes in the RcGTA's head-tail cluster and also requires genes to align with their RcGTA
359 homologs by at least 60% of their length, some GTA genes were likely missed by GTA-Hunter.
360 To look for these potential false negatives, additional BLASTP (50) searches with the e-value
361 cutoff of 0.1 were performed using 17 RcGTA head-tail cluster genes as queries and protein-
362 coding genes in 212 genomes as a database. Only matches located within the genomic regions
363 designated as GTA gene clusters by GTA-Hunter were kept. In four genomes, calculations of
364 genes' adaptation to tRNA pool (see **"Evaluation of the adaptiveness of protein-coding genes**
365 **to the tRNA pool"** section below for details) did not converge. As a result, only 208 genomes
366 were retained in the reported analyses (GenBank accessions are available at
367 https://doi.org/10.6084/m9.figshare.20082749).

368

369 **Identification of gene families in 208 alphaproteobacterial genomes**

370 Within each genome, protein-coding genes less than 300 nucleotides in length were
371 excluded in order to reduce the stochasticity of codon usage bias values due to the insufficient
372 number of codons. The remaining protein-coding genes were clustered into gene families using
373 Orthofinder v2.4 (51) with default parameters and DIAMOND (52) for the amino acid sequence

374 similarity search. Only gene families detected in at least 40 genomes were retained to ensure
375 statistical power.
376 Some alphaproteobacterial GTA head-tail cluster regions contain protein-coding ORFs
377 that do not have significant similarity to the RcGTA homologs of the genes shown to be required
378 for GTA production in RcGTA. Gene families of these ORFs were retrieved from the collection
379 of gene families predicted for all protein-coding genes (regardless of their length) using
380 Orthofinder v2.4 (51) with default parameters and DIAMOND (52) for the amino acid sequence
381 similarity search. Only gene families that are both located within the genomic region encoding
382 GTA head-tail cluster and found in at least 10 genomes were retained.
383
384 **Reference set of GTA genes**
385 Although RcGTA head-tail cluster contains 17 genes, genes *g3.5* and *g10.1* are less than
386 300 nucleotides in length, and genes *g1* and *g7* are not detected widely across analyzed genomes.
387 Additionally, codon usage patterns of gene *g9* were found to be very different from other GTA
388 genes (see **"Examination of similarity in adaptation to the tRNA pool among GTA genes"**
389 section below for details). Therefore, in our inferences about selection, we considered only 12 of
390 the 17 GTA genes (**Supplemental Table S3**), which we designate throughout the manuscript as
391 "GTA reference genes".
392 Amino acid sequences of GTA reference genes were aligned individually using MAFFT-
393 linsi v7.455 (53) and then concatenated into a single alignment. Each gene was treated as a
394 separate partition in the alignment and the best substitution model for each gene was determined
395 by ModelFinder (54). The maximum likelihood tree was reconstructed using IQ-TREE v1.6.7
396 (55) and the support values were calculated via 1,000 ultrafast bootstrap replicates (56).

397

398 **Reconstruction of the reference phylogenomic tree**

399 Twenty-nine marker proteins that are present in a single copy in more than 95% of the
400 208 retained genomes were retrieved using AMPHORA2 (57). Amino acid sequences within
401 each of the 29 marker families were aligned using MAFFT-linsi v7.455 (53). The best
402 substitution matrix for each family was determined by *ProteinModelSelection.pl* script
403 downloaded from https://github.com/stamatak/standard-RAxML/tree/master/usefulScripts in
404 October 2020. Individual alignments of the marker families were concatenated, but each
405 alignment was treated as a separate partition with its own best substitution model in the
406 subsequent phylogenetic reconstruction. The maximum likelihood tree was reconstructed using
407 IQ-TREE v1.6.7 (55) and the support values were calculated using 1,000 ultrafast bootstrap
408 replicates (56).

409

410 **Evaluation of codon usage bias in protein-coding genes using "effective number of codons"**
411 **metric**

412 For the retained genes in each genome, effective number of codons (ENC) (20) and G+C
413 content variation at the 3rd codon position in the synonymous sites (GC3s) were calculated using
414 CodonW (http://codonw.sourceforge.net). The null model of no codon usage bias was calculated
415 as described in dos Reis et al. (28) using an in-house script (available in the FigShare repository;
416 see below). For every gene, the deviation of its ENC from the null model was calculated using
417 the in-house script. Genes that have observed ENC higher than the expected were excluded from
418 analyses.

419

420 **Evaluation of the adaptiveness of protein-coding genes to the tRNA pool**

421 The tRNA genes in each genome were predicted using tRNAscan-SE v 2.06, using a
422 model trained on bacterial genomes (58, 59) and the Infernal mode without HMM filter to
423 improve the sensitivity of the search (60). tRNA gene copy number was used as the proxy for
424 tRNA abundance, following the previously reported observation that the two correlate strongly
425 (28, 61). The adaptiveness of each codon (ω_i) to the tRNA pool was calculated using the stAIcalc
426 program with the maximum hill climbing stringency (62). The tRNA adaptation index (tAI) of
427 each retained gene was calculated as the geometric mean of its ω_i values (27). Because the
428 distribution of tAI values varies among genomes (63) (**Supplemental Figure S6**), tAI values
429 were converted to their relative percentile tRNA adaptation index (ptAI) within a genome. The
430 ptAI values range between 0 and 100, and represent the percentage of analyzed genes in a
431 genome that have a smaller tAI than a particular gene.

432

433 **Examination of similarity in adaptation to the tRNA pool among GTA genes**

434 The ptAI values were retrieved for a subset of 13 GTA genes that are at least 300
435 nucleotide in length and are widely detected across all taxonomic groups. The linear regression
436 analysis of ptAI values between all GTA gene pairs was conducted using the phylogenetic
437 generalized least squares method (PGLS) (64). The reference phylogenomic tree was used to
438 correct for the shared evolutionary history. The analysis was done using the 'caper' package (65)
439 and λ, δ and κ parameters were estimated using the maximum likelihood function. Because ptAI
440 values of gene *g9* were not significantly correlated with the ptAI values of 8 out of the 12 other
441 examined GTA genes at p-value cutoff of 0.001 (**Supplemental Table S4**), the gene *g9* was not
442 included into the reference set of GTA genes.

443

444 **Identification of genes with ptAI values similar to that of the GTA genes**

445 For each gene family, the "within-genome" ptAI values were retrieved. For gene families
446 with at least two paralogs, the ptAI values for all paralogs from a particular genome were
447 replaced with their median ptAI value.

448 To identify gene families that exhibit tRNA pool adaptation patterns similar to those of
449 GTA genes, a linear regression model of ptAI values between these gene families and reference
450 GTA genes was fit using the PGLS (64). The PGLS analysis was carried out using the 'caper'
451 package (65) and λ, δ and κ parameters were estimated using the maximum likelihood function.
452 The reference phylogenomic tree was used to correct for the shared phylogenetic history. For
453 gene families found in at least 40 genomes, a gene family was designated to be associated with a
454 GTA, if obtained fit of the model was statistically significant across all reference GTA genes.
455 Because gene families found in less than 40 genomes were kept only if the genes are located
456 within the genomic regions encoding GTA head-tail clusters (see the "Identification of gene
457 families in 208 alphaproteobacterial genomes" section), a more relaxed criterion was adopted for
458 such gene families: a gene family was designated to be associated with a GTA if the fit of the
459 model was statistically significant across at least 40% of reference GTA genes. If a significantly
460 associated gene family contained paralogs, the PGLS analysis was repeated by using individual
461 ptAI values across all possible combinations of paralogs (if the total number of combinations
462 was < 1,000) or across random 1,000 combinations of paralogs (if the total number of
463 combinations was > 1,000). This was carried out to ensure that the detected signal was not due to
464 sampling associated with selecting the median ptAI value.

465 Genes with a significant similarity in trend of ptAI values were annotated via eggNOG-
466 mapper v2.1 (66).
467
468 **Protein-protein interaction of GTA genes and gene families with similar ptAI values**
469 To identify protein-protein interaction networks, reference GTA genes and genes from
470 families with similar tRNA pool adaptation patterns were retrieved from the *Sphingomonas* sp.
471 MM1 genome, chosen for it being the only genome that contains all genes from the GTA
472 reference gene set and all 14 gene families listed in **Table 1**. The locus tags of the retrieved
473 *Sphingomonas* sp. MM1 genes were used as queries against STRING database v 11.0b (last
474 accessed July 2021) (67) with the medium confidence score cutoff and all active interaction
475 sources. The retrieved protein-protein interaction network was visualized in STRING using the
476 queries and up to 50 additional interactor proteins, and displaying edges based on the STRING
477 confidence scores. The KEGG pathways (68) enrichment analysis was conducted via
478 hypergeometric testing on the whole retrieved network, as implemented in STRING.
479
480 **Analysis of other protein-coding genes situated within GTA head-tail clusters**
481 For gene families within GTA head-tail clusters, ptAI values were retrieved and
482 compared to ptAI values of the reference GTA gene set using PGLS analysis as described above.
483 For the only gene family with a significant association with GTA genes, the secondary structure
484 of its proteins were predicted using Porter v5.0 (69). To retrieve available viral head completion
485 proteins, the phrase 'head-completion protein' was used as a query against the UniProt database
486 (accessed in August 2021) (70). Among the 24 manually annotated ("reviewed") matches from
487 the Swiss-Prot sub-database of the UniProt database, only 2 viral matches (accession numbers

488 P68656 and P68660) had length similar to the genes in the above described gene family. Both
489 proteins belong to the λ phage gpW family, and for *Escherichia* phage λ protein 3D structure is
490 available in PDB (71). The secondary structure of RcGTA's g7 protein, structural viral homolog
491 of RcGTA's g7 from *Bacillus* phage SPP1 (gp16) (29) and head completion protein of phage λ
492 were retrieved from the PDB database (71) in August 2021.

493 In 48 *Sphingomonadales* genomes without a detectable homolog of RcGTA gene *g7*, the
494 genomic space either between the homologs of the RcGTA genes *g6* and *g8*, or, in genomes
495 without *g6* homolog, between homologs of the RcGTA genes *g5* and *g8*, was searched for
496 presence of open reading frames.

497

498 **Refinement of the *tonB* gene family using phylogenetic tree**

499 To identify orthologs within the large *tonB* gene family, evolutionary history of the *tonB*
500 gene family was reconstructed and evaluated. To do so, amino acid sequences of the *tonB* gene
501 family were aligned using MAFFT-linsi v7.455 (53). The phylogeny was reconstructed in IQ-
502 TREE v1.6.7 (55) using the best substitution model (LG+F+R6) detected by ModelFinder (54).
503 The tree was visualized using the iTOL v6 (72). The phylogeny was used to subdivide the family
504 into two families, whereas the five genes on very long branches served as an outgroup (tree
505 topology is available at https://doi.org/10.6084/m9.figshare.20082749).

506

507 **Calculation of energetic cost associated with production of the encoded proteins**

508 To quantify the energetic cost of proteins, the carbon content of their amino acids was
509 used as a proxy and was calculated by counting the number of carbons in the amino acid side

510 chains, as described in Kogay et al. (9). The total number of carbons in each protein was
511 normalized by the protein length.
512
513 **Retrieval and phylogenetic analyses of *gafA* homologs**
514 Amino acid sequences of *gafA* homologs in 196 alphaproteobacterial genomes were
515 detected via Orthofinder (gene family OG0001218). Only homologs found in single copy in a
516 genome (194 in total) were retained for phylogenetic analysis. These homologs were aligned
517 using MAFFT-linsi v7.455 (53). The best substitution model (LG+F+R6) was determined by
518 ModelFinder (54) and the maximum-likelihood tree was reconstructed by IQ-TREE v1.6.7 (55)
519 with the number of iterations to stop set to 500. The support values were calculated using 1,000
520 ultrafast bootstrap replicates (56). Both GTA reference tree and reference phylogenomic tree
521 were pruned to match the taxa in *gafA* phylogeny. The normalized quartet scores were calculated
522 using ASTRAL v5.7.8 (73).
523
524 **Data availability**
525 The following data are available in the FigShare repository under DOI
526 10.6084/m9.figshare.20082749 (https://doi.org/10.6084/m9.figshare.20082749): accession
527 numbers of 208 analyzed alphaproteobacterial genomes; accession numbers of the GTA regions
528 identified in the analyzed genomes; accession numbers of genes in gene families across analyzed
529 genomes; raw data related to tAI and ENC calculations; an in-house script for ENC calculations;
530 slopes and p-values of associations detected in PGLS analyses; accession numbers of the
531 putative g7 proteins in *Sphingomonadales* genomes; multiple sequence alignments and

532 phylogenetic trees of *tonB* and *gafA* gene families, concatenated phylogenomic markers, and
533 concatenated GTA reference genes.
534

535 Competing interest statement

536 The authors declare no competing interests.
537

538 Acknowledgements

539 This work was supported in part by the Simons Foundation Investigators in Mathematical
540 Modeling of Living Systems program (Award #327936 to O.Z.).
541

542 References

543 1. Lang AS, Westbye AB, Beatty JT. 2017. The distribution, evolution, and roles of gene
544 transfer agents in prokaryotic genetic exchange. *Annu Rev Virol* 4:87-104.
545 2. Marrs B. 1974. Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad*
546 *Sci U S A* 71:971-3.
547 3. Lang AS, Zhaxybayeva O, Beatty JT. 2012. Gene transfer agents: phage-like elements of
548 genetic exchange. *Nat Rev Microbiol* 10:472-82.
549 4. Brimacombe CA, Ding H, Beatty JT. 2014. *Rhodobacter capsulatus* DprA is essential for
550 RecA-mediated gene transfer agent (RcGTA) recipient capability regulated by quorum-
551 sensing and the CtrA response regulator. *Mol Microbiol* 92:1260-78.
552 5. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. 2010. High
553 frequency of horizontal gene transfer in the oceans. *Science* 330:50.

554 6. Brimacombe CA, Ding H, Johnson JA, Beatty JT. 2015. Homologues of genetic
555 transformation DNA import genes are required for *Rhodobacter capsulatus* gene transfer
556 agent recipient capability regulated by the response regulator CtrA. *J Bacteriol* 197:2653-
557 63.
558 7. Quebatte M, Christen M, Harms A, Korner J, Christen B, Dehio C. 2017. Gene transfer
559 agent promotes evolvability within the fittest subpopulation of a bacterial pathogen. *Cell*
560 *Syst* 4:611-621 e6.
561 8. Marrs B, Wall JD, Gest H. 1977. Emergence of the biochemical genetics and molecular
562 biology of photosynthetic bacteria. *Trends Biochem Sci* 2:105-108.
563 9. Kogay R, Wolf YI, Koonin EV, Zhaxybayeva O. 2020. Selection for reducing energy
564 cost of protein production drives the GC content and amino acid composition bias in gene
565 transfer agents. *mBio* 11:e01206-20.
566 10. Kogay R, Koppenhofer S, Beatty JT, Lang AS, Kuhn JH, Zhaxybayeva O. 2022. Formal
567 recognition and classification of gene transfer agents as viriforms. *bioRxiv*
568 doi:10.1101/2022.06.10.494566:https://doi.org/10.1101/2022.06.10.494566.
569 11. Hynes AP, Shakya M, Mercer RG, Grull MP, Bown L, Davidson F, Steffen E, Matchem
570 H, Peach ME, Berger T, Grebe K, Zhaxybayeva O, Lang AS. 2016. Functional and
571 evolutionary characterization of a gene transfer agent's multilocus “genome". *Mol Biol*
572 *Evol* 33:2530-43.
573 12. Shakya M, Soucy SM, Zhaxybayeva O. 2017. Insights into origin and evolution of alpha-
574 proteobacterial gene transfer agents. *Virus Evol* 3:vex036.

575 13. Kogay R, Neely TB, Birnbaum DP, Hankel CR, Shakya M, Zhaxybayeva O. 2019.
576 Machine-learning classification suggests that many alphaproteobacterial prophages may
577 instead be gene transfer agents. *Genome Biol Evol* 11:2941-2953.
578 14. Lang AS, Beatty JT. 2007. Importance of widespread gene transfer agent genes in alpha-
579 proteobacteria. *Trends Microbiol* 15:54-62.
580 15. Fogg PCM. 2019. Identification and characterization of a direct activator of a gene
581 transfer agent. *Nat Commun* 10:595.
582 16. Westbye AB, O'Neill Z, Schellenberg-Beaver T, Beatty JT. 2017. The *Rhodobacter*
583 *capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega
584 subunit. *Microbiology (Reading)* 163:1355-1363.
585 17. Fogg PC, Westbye AB, Beatty JT. 2012. One for all or all for one: heterogeneous
586 expression and host cell lysis are key to gene transfer agent activity in *Rhodobacter*
587 *capsulatus*. *PLoS One* 7:e43772.
588 18. Hynes AP, Mercer RG, Watton DE, Buckley CB, Lang AS. 2012. DNA packaging bias
589 and differential expression of gene transfer agent genes within a population during
590 production and release of the *Rhodobacter capsulatus* gene transfer agent, RcGTA. *Mol*
591 *Microbiol* 85:314-25.
592 19. Roller M, Lucic V, Nagy I, Perica T, Vlahovicek K. 2013. Environmental shaping of
593 codon usage and functional adaptation across microbial communities. *Nucleic Acids Res*
594 41:8842-52.
595 20. Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* 87:23-9.

596 21. Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy,
597 specialization, and efficient decoding for translation optimization. *Genome Res* 14:2279-
598 86.
599 22. Quax TE, Claassens NJ, Soll D, van der Oost J. 2015. Codon bias as a means to fine-tune
600 gene expression. *Mol Cell* 59:149-61.
601 23. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, Liu Y. 2016. Codon usage is an
602 important determinant of gene expression levels largely through its effects on
603 transcription. *Proc Natl Acad Sci U S A* 113:E6117-E6125.
604 24. Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. 2010. Translational selection is
605 ubiquitous in prokaryotes. *PLoS Genet* 6:e1001004.
606 25. LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2021. Signatures of
607 optimal codon usage in metabolic genes inform budding yeast ecology. *PLoS Biol*
608 19:e3001185.
609 26. Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among
610 interacting proteins. *Proc Natl Acad Sci U S A* 101:9033-8.
611 27. dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene
612 expression and codon usage bias from microarray data for the whole *Escherichia coli* K-
613 12 genome. *Nucleic Acids Res* 31:6976-85.
614 28. dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a
615 test for translational selection. *Nucleic Acids Res* 32:5036-44.
616 29. Bardy P, Fuzik T, Hrebik D, Pantucek R, Thomas Beatty J, Plevka P. 2020. Structure and
617 mechanism of DNA delivery of a gene transfer agent. *Nat Commun* 11:3034.

618 30. Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences
619 of codon bias. *Nat Rev Genet* 12:32-42.
620 31. Lim HN, Lee Y, Hussein R. 2011. Fundamental relationship between operon
621 organization and gene expression. *Proc Natl Acad Sci U S A* 108:10626-31.
622 32. Lang AS, Beatty JT. 2000. Genetic analysis of a bacterial genetic exchange element: the
623 gene transfer agent of *Rhodobacter capsulatus*. *Proc Natl Acad Sci U S A* 97:859-64.
624 33. Nishizaki T, Tsuge K, Itaya M, Doi N, Yanagawa H. 2007. Metabolic engineering of
625 carotenoid biosynthesis in *Escherichia coli* by ordered gene assembly in *Bacillus subtilis*.
626 *Appl Environ Microbiol* 73:1355-61.
627 34. Balkwill DL, Fredrickson JK, Romine MF. 2006. Sphingomonas and related genera, p
628 605-629. *In* Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (ed),
629 The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass doi:10.1007/0-387-
630 30747-8_23. Springer New York, New York, NY.
631 35. Kooistra J, Haijema BJ, Venema G. 1993. The *Bacillus subtilis* addAB genes are fully
632 functional in *Escherichia coli*. *Mol Microbiol* 7:915-23.
633 36. Marsin S, Lopes A, Mathieu A, Dizet E, Orillard E, Guerois R, Radicella JP. 2010.
634 Genetic dissection of *Helicobacter pylori* AddAB role in homologous recombination.
635 *FEMS Microbiol Lett* 311:44-50.
636 37. Jung H, Liang J, Jung Y, Lim D. 2015. Characterization of cell death in *Escherichia coli*
637 mediated by XseA, a large subunit of exonuclease VII. *J Microbiol* 53:820-8.
638 38. Tang K, Jiao N, Liu K, Zhang Y, Li S. 2012. Distribution and functions of TonB-
639 dependent transporters in marine bacteria and environments: implications for dissolved
640 organic matter utilization. *PLoS One* 7:e41204.

641 39. Eisenbeis S, Lohmiller S, Valdebenito M, Leicht S, Braun V. 2008. NagA-dependent
642 uptake of N-acetyl-glucosamine and N-acetyl-chitin oligosaccharides across the outer
643 membrane of *Caulobacter crescentus*. *J Bacteriol* 190:5230-8.
644 40. Blanvillain S, Meyer D, Boulanger A, Lautier M, Guynet C, Denance N, Vasse J, Lauber
645 E, Arlat M. 2007. Plant carbohydrate scavenging through tonB-dependent receptors: a
646 feature shared by phytopathogenic and aquatic bacteria. *PLoS One* 2:e224.
647 41. Genest O, Wickner S, Doyle SM. 2019. Hsp90 and Hsp70 chaperones: Collaborators in
648 protein remodeling. *J Biol Chem* 294:2109-2120.
649 42. Rockabrand D, Livers K, Austin T, Kaiser R, Jensen D, Burgess R, Blum P. 1998. Roles
650 of DnaK and RpoS in starvation-induced thermotolerance of *Escherichia coli*. *J Bacteriol*
651 180:846-54.
652 43. Gershenzon J, Dudareva N. 2007. The function of terpene natural products in the natural
653 world. *Nat Chem Biol* 3:408-14.
654 44. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. 2017. The ecological role of volatile
655 and soluble secondary metabolites produced by soil bacteria. *Trends Microbiol* 25:280-
656 292.
657 45. Yang Y, Liu B, Du X, Li P, Liang B, Cheng X, Du L, Huang D, Wang L, Wang S. 2015.
658 Complete genome sequence and transcriptomics analyses reveal pigment biosynthesis
659 and regulatory mechanisms in an industrial strain, *Monascus purpureus* YY-1. *Sci Rep*
660 5:8331.
661 46. Ram S, Mitra M, Shah F, Tirkey SR, Mishra S. 2020. Bacteria as an alternate biofactory
662 for carotenoid production: A review of its applications, opportunities and challenges.
663 *Journal of Functional Foods* 67:103867.

664 47. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of
665 life. *Nat Rev Genet* 16:472-82.
666 48. Zhan Y, Huang S, Voget S, Simon M, Chen F. 2016. A novel roseobacter phage
667 possesses features of podoviruses, siphoviruses, prophages and gene transfer agents. *Sci*
668 *Rep* 6:30372.
669 49. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B,
670 Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova
671 O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta
672 T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson
673 P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick
674 LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE,
675 Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference
676 sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional
677 annotation. *Nucleic Acids Res* 44:D733-45.
678 50. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997.
679 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
680 *Nucleic Acids Res* 25:3389-402.
681 51. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for
682 comparative genomics. *Genome Biol* 20:238.
683 52. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using
684 DIAMOND. *Nat Methods* 12:59-60.
685 53. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
686 improvements in performance and usability. *Mol Biol Evol* 30:772-80.

687 54. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017.
688 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*
689 14:587-589.
690 55. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and
691 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol*
692 *Evol* 32:268-74.
693 56. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2:
694 improving the ultrafast bootstrap approximation. *Mol Biol Evol* 35:518-522.
695 57. Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with
696 AMPHORA2. *Bioinformatics* 28:1033-4.
697 58. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer
698 RNA genes in genomic sequence. *Nucleic Acids Res* 25:955-64.
699 59. Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and
700 functional classification of transfer RNA genes. *Nucleic Acids Res* 49:9077-9096.
701 60. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches.
702 *Bioinformatics* 29:2933-5.
703 61. Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-
704 adapted for optimal translation of highly expressed genes. *Trends Genet* 16:287-9.
705 62. Sabi R, Volvovitch Daniel R, Tuller T. 2017. stAIcalc: tRNA adaptation index calculator
706 based on species-specific weights. *Bioinformatics* 33:589-591.
707 63. LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2019. Variation and
708 selection on codon usage bias across an entire subphylum. *PLoS Genet* 15:e1008304.

709 64. Martins EP, Hansen TF. 1997. Phylogenies and the comparative method: a general
710 approach to incorporating phylogenetic information into the analysis of interspecific data.
711 *The American Naturalist* 149:646 - 667.
712 65. Orme D. 2018. The caper package: comparative analysis of phylogenetics and evolution
713 in R. .
714 66. Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021.
715 eggNOG-mapper v2: functional annotation, orthology assignments, and domain
716 prediction at the metagenomic scale. *Mol Biol Evol* 38:5825-5829.
717 67. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT,
718 Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. 2021. The STRING database in
719 2021: customizable protein-protein networks, and functional characterization of user-
720 uploaded gene/measurement sets. *Nucleic Acids Res* 49:D605-D612.
721 68. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG:
722 integrating viruses and cellular organisms. *Nucleic Acids Res* 49:D545-D551.
723 69. Torrisi M, Kaleel M, Pollastri G. 2019. Deeper profiles and cascaded recurrent and
724 convolutional neural networks for state-of-the-art protein secondary structure prediction.
725 *Sci Rep* 9:12374.
726 70. UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021.
727 *Nucleic Acids Res* 49:D480-D489.
728 71. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN,
729 Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235-42.
730 72. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for
731 phylogenetic tree display and annotation. *Nucleic Acids Res* 49:W293-W296.

732 73. Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species
733 tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.
734

735 # Tables

736 **Table 1.** Functional annotations of 14 gene families, whose ptAI values have a significantly

737 similar trend to ptAI values of the reference GTA genes.

Gene	RefSeq ID of a Representative Protein	RefSeq Record Annotation	COG Category	COG functional category description
gafA	WP_121690074.1	DUF6456-domain containing protein	K	Transcription
addA	WP_121690807.1	Double-strand break repair helicase AddA	L	Replication, recombination, and repair
addB	WP_121690808.1	Double-strand break repair protein AddB	L	Replication, recombination, and repair
xseA	WP_092770100.1	Exodeoxyribonuclease VII large subunit	L	Replication, recombination, and repair
dinG	WP_067681212.1	ATP-dependent DNA helicase	KL	Transcription; Replication, recombination, and repair
hrpB	WP_121690814.1	ATP-dependent helicase HrpB	L	Replication, recombination, and repair
priA	WP_121691035.1	Primosomal protein N'	L	Replication, recombination, and repair
glnE	WP_121690099.1	Bifunctional [glutamine synthetase] adenylyltransferase/[glutamine synthetase]-adenylyl-L-tyrosine phosphorylase	OT	Molecular chaperones and related functions; Signal transduction mechanism
ccmE	WP_010971299.1	Cytochrome c maturation protein CcmE	O	Molecular chaperones and related functions
ATP12	WP_092769070.1	ATP12 family chaperone protein	O	Molecular chaperones and related functions
tonB	WP_119082607.1	Energy transducer TonB	M	Cell wall/membrane/envelope biogenesis
TPR	WP_162687979.1	Tetratricopeptide repeat protein	M	Cell wall/membrane/envelope biogenesis
smrA	WP_010970599.1	Smr/MutS family protein	S	Function unknown
crtB	WP_121690324.1	Phytoene/Squalene synthase family protein	I	Lipid transport and metabolism

738

739 ## Figure Legends

740 **Figure 1. Distribution of ptAI values among reference GTA genes from GTA head-tail**
741 **clusters in 208 alphaproteobacterial genomes.** Line within a box displays the median ptAI
742 value for a GTA gene across all genomes, in which the gene was detected. The boxes are
743 bounded by first and third quartiles. Whiskers represent ptAI values within 1.5*interquartile
744 range. Dots outside of whiskers are outliers.

745

746 **Figure 2. Distributions of (A) ptAI values in the major capsid protein-encoding gene (*g5*)**
747 **and (B) carbon content of amino acids in the g5 protein across four orders of the class**
748 ***Alphaproteobacteria*.** On both panels A and B, line within a box displays the median ptAI value
749 for g5 representatives within an order. The boxes are bounded by first and third quartiles.
750 Whiskers represent ptAI values within 1.5*interquartile range. Dots outside of whiskers are
751 outliers. The phylogenetic tree on the Y-axis is the reference phylogenomic tree (see **Methods**
752 for details), in which branches are collapsed at the taxonomic order level. Dashed line in Panel A
753 marks ptAI value of 90.

754

755 **Figure 3. The protein-protein interactions among 12 GTA reference proteins and 14**
756 **proteins putatively co-expressed with GTAs.** Nodes represent individual proteins. Blue-
757 colored nodes correspond to GTA reference proteins and red-colored nodes correspond to 14
758 putatively co-expressed proteins. Gray-colored nodes represent additional proteins found through
759 the STRING functional enrichment analysis. The thickness of the edges is proportional to the
760 STRING's confidence score of protein interactions (varying between 0.4 [thin line] and 1.0
761 [thick line]).

762 **Figure 4. Secondary structures of head completion proteins from phages and GTAs.** The
763 Enterobacteria phage lambda gpW and Bacillus phage SPP1 (highlighted in red) are two
764 representatives of viral head completion proteins with major differences in lengths and secondary
765 structures. The secondary structures of *R. capsulatus* (PDB ID 6TUI_8), Enterobacteria phage
766 lambda gpW (1HYW), and Bacillus phage SPP1 (2KCA) proteins were retrieved from the PDB
767 database. The secondary structures of the putative head completion proteins from
768 *Sphingomonadales* were predicted computationally. The secondary structures are scaled with
769 respect to the protein lengths, which are listed in parentheses next to the taxonomic names.
770

Supplemental Material Captions

Supplemental Tables S1-S4 and Figures S1-S6 are provided as PDF files.

Supplemental Table S1. Four molecular pathways significantly over-represented in the protein-protein interaction network shown in Figure 3. The pathway information was obtained from KEGG database.

Supplemental Table S2. Significance and slope of the fit of the phylogenetic generalized least squares (PGLS) models between the reference GTA genes and putative head-completion protein in *Sphingomonadales*. Statistically significant associations (p-value <0.05) are highlighted in orange.

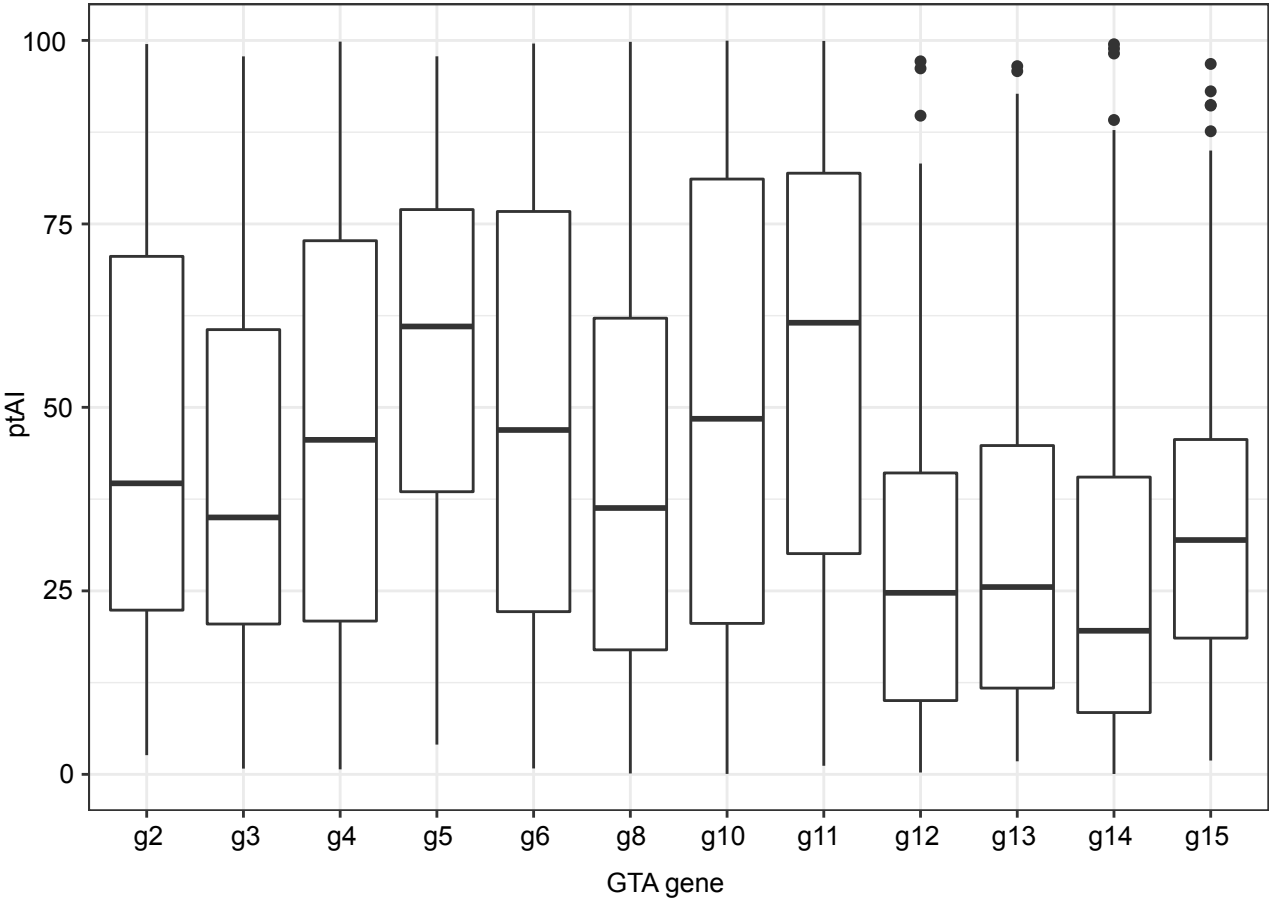
Supplemental Table S3. Decisions behind choosing the RcGTA 'head-tail' cluster homologs for the reference GTA gene set. The GTA genes selected for the reference set are highlighted in orange. Functional annotations are based on the descriptions in the RefSeq database records, unless noted otherwise.

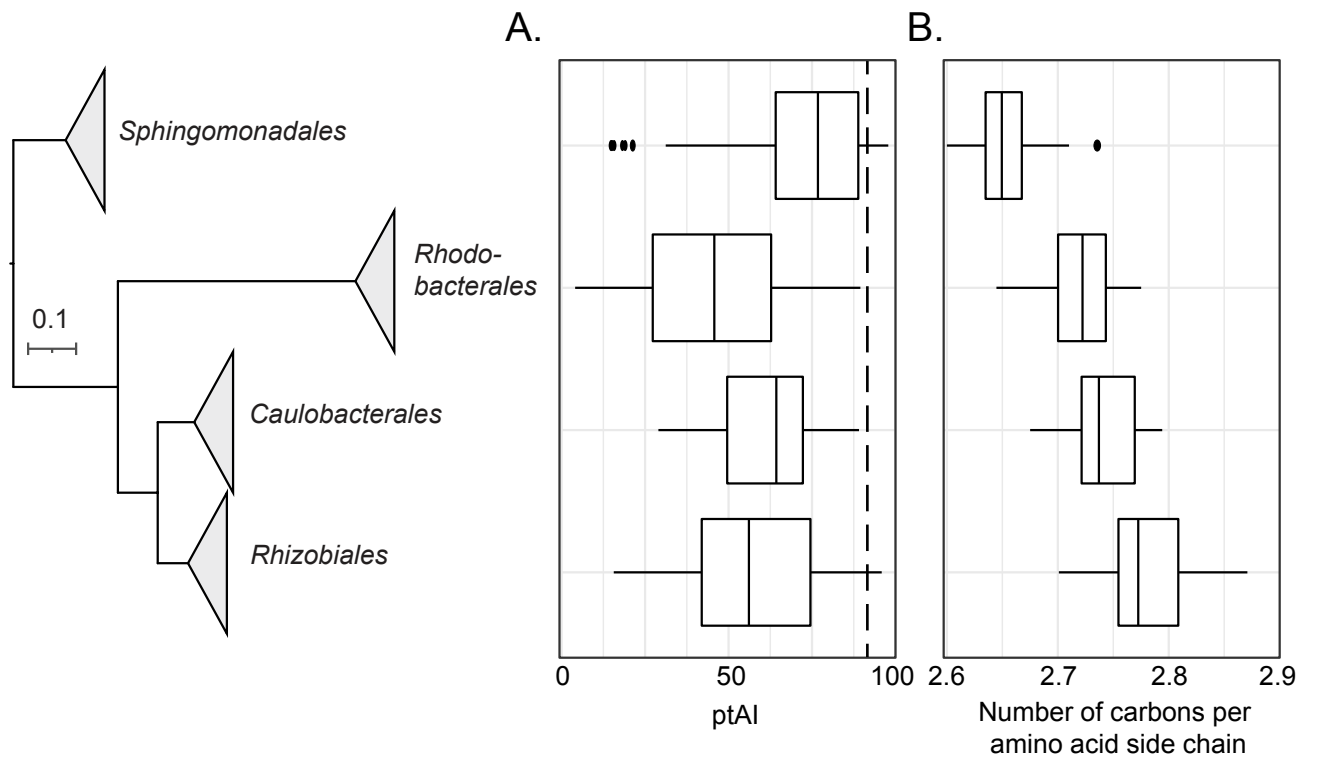
Supplemental Table S4. Significance and slope of the fit of the phylogenetic generalized least squares (PGLS) models between the reference GTA genes and the *g9* gene.

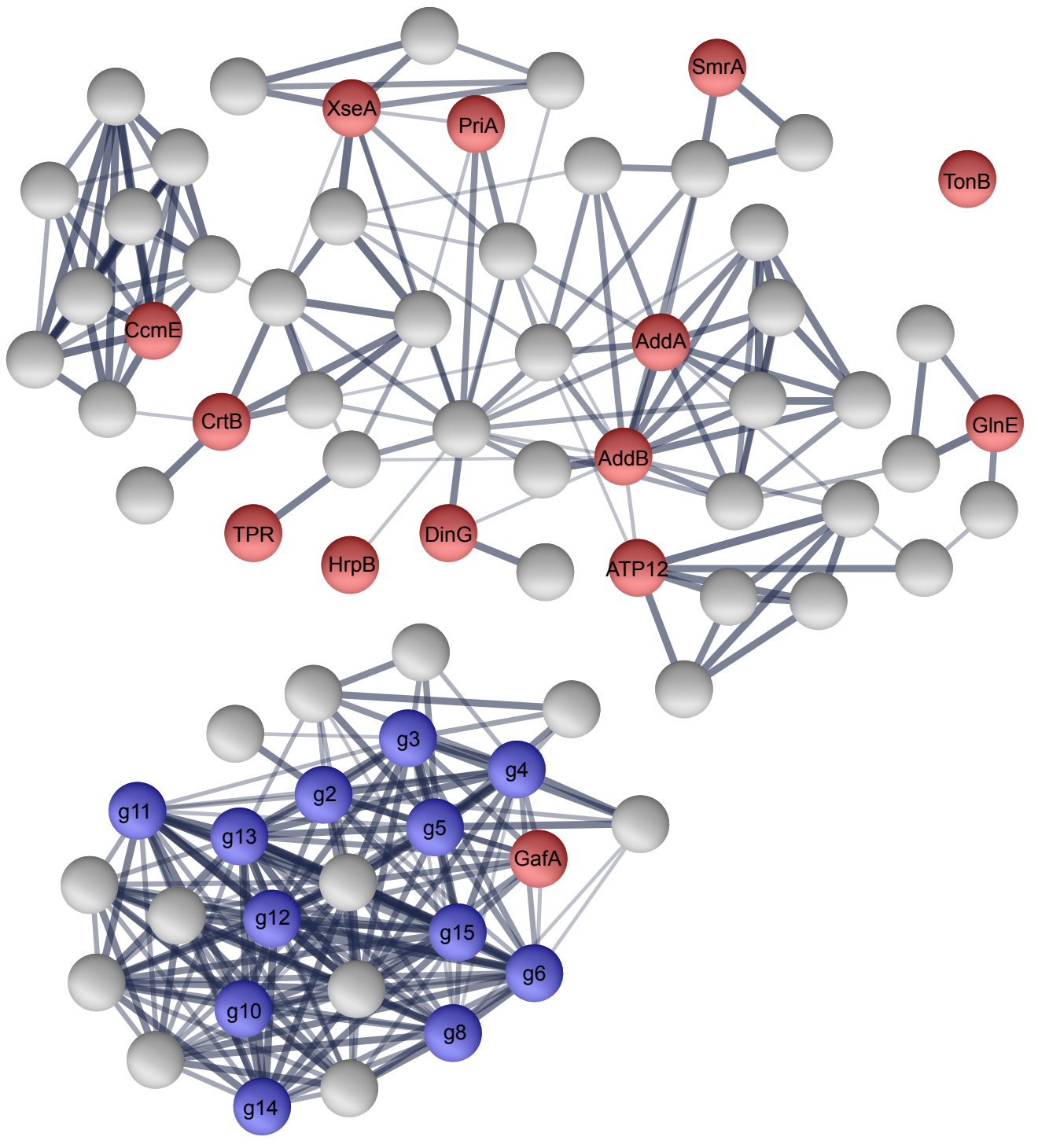
Supplemental Figure S1. Distribution of deviations of the effective number of codon (ENC) values from the expected ENC values under the null model of no codon bias. The distribution contains deviations for 2,308 reference GTA genes found in the 208 genomes.

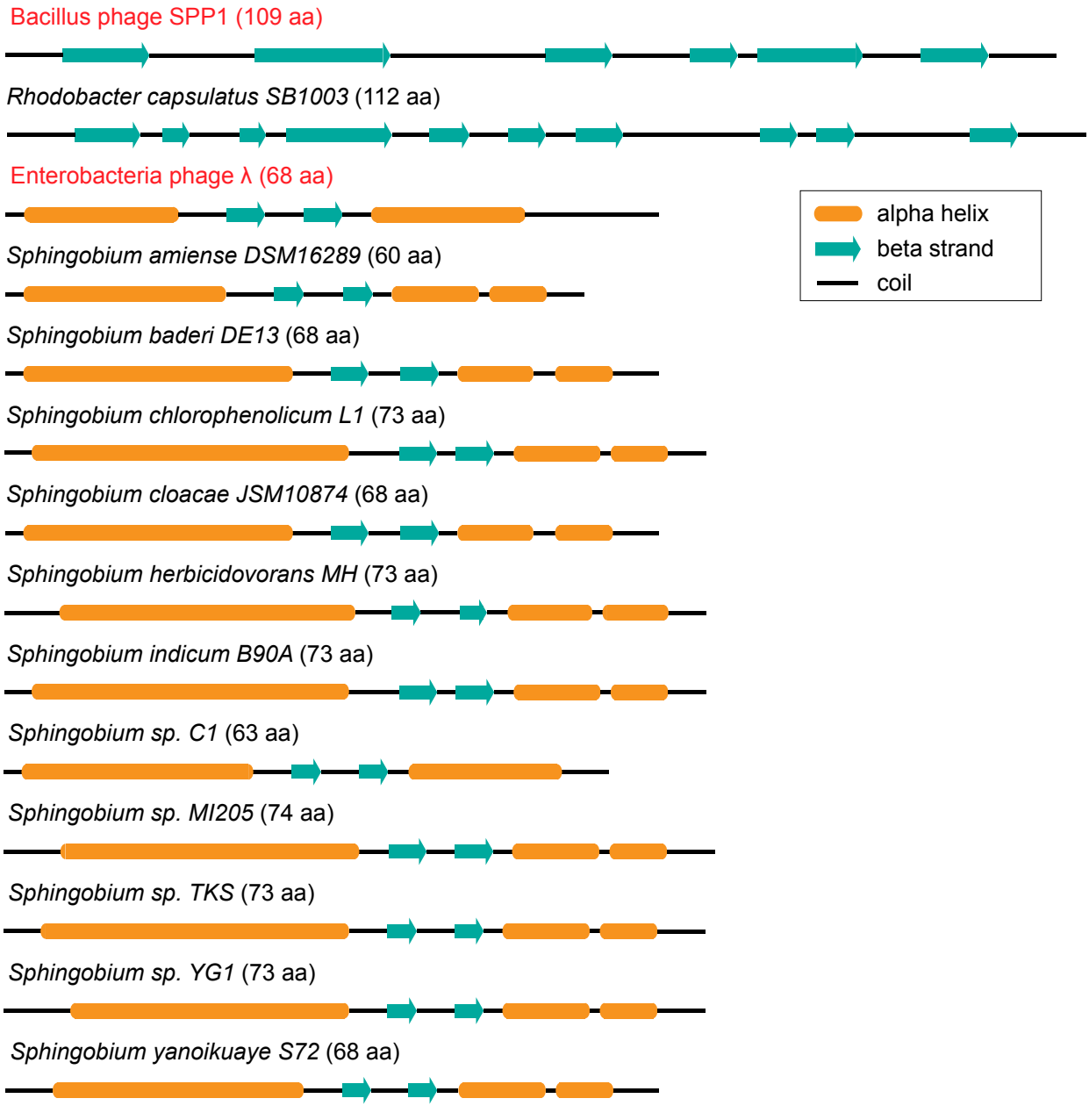
794 Numbers on the plot designate the number of reference GTA genes in an interval delineated by
795 dashed lines.
796
797 **Supplemental Figure S2. Deviation of the effective number of codon (ENC) values for**
798 **individual reference GTA genes in comparison to the genomic average.** The deviation of the
799 ENC from the expectation under the null model for each GTA gene was normalized by the
800 average ENC deviation of its genome. Line within a box displays the median normalized ENC
801 value for a GTA gene across all genomes. The boxes are bounded by first and third quartiles.
802 Whiskers represent ptAI values within 1.5*interquartile range. Dots outside of whiskers are
803 outliers.
804
805 **Supplemental Figure S3. Distributions of ptAI values in all reference GTA genes across**
806 **four orders of the class *Alphaproteobacteria*.** Line within a box displays the median ptAI value
807 for a GTA gene across all genomes. The boxes are bounded by first and third quartiles. Whiskers
808 represent ptAI values within 1.5*interquartile range. Dots outside of whiskers are outliers.
809
810 **Supplemental Figure S4. PGLS model fit among ptAI values of the reference GTA gene**
811 **pairs.** Each pairwise comparison is represented by a rectangle that is color-coded according to
812 the p-values from the PGLS analysis of the reference GTA gene pairs. The numerical p-values
813 are listed within each rectangle.
814
815 **Supplemental Figure S5. Gene neighborhood of the RcGTA gene *g7* and of the putative *g7***
816 **replacement gene in 11 *Sphingomonadales* spp.** Only region corresponding to the RcGTA

817 ‘head-tail’ cluster is depicted, with each gene represented by an arrow scaled relative to its length
818 within each cluster. The RcGTA genes (*g1-g15*) are color-coded and their homologs in
819 *Sphingomonadales* are shown in the same color. Putative *g7* replacements in *Sphingomonadales*
820 are shown in magenta and marked with an arrow. Pseudogenes are colored in black, while genes
821 without an established relationship to GTA production are shown in gray. Phylogenetic tree is a
822 subtree extracted from the reference phylogeny. The scale bar corresponds to the number of
823 substitutions per site.
824
825 **Supplemental Figure S6. Distribution of tAI values in protein-coding genes of the analyzed**
826 **genomes.** Only genes at least 300 nucleotides in length were included. **A.** Distribution of tAI
827 values of genes in three representative alphaproteobacterial genomes, selected to have the lowest,
828 the median, and the highest mean tAI value among 208 genomes. **B.** Distribution of the average
829 genomic tAI values across 208 genomes.
830
831
832


100
75
50
25
0
ptAI
g2
g3
g4
g5
g6
g8
g10
g11
g12
g13
g14
g15
GTA gene


A.
B.
Sphingomonadales
Rhodo-
bacterales
0.1
Caulobacterales
Rhizobiales
0
50
100
ptAI
2.6
2.7
2.8
2.9
Number of carbons per
amino acid side chain


SmrA
XseA
PriA
TonB
CcmE
AddA
CrtB
GlnE
AddB
TPR
DinG
HrpB
ATP12
g3
g4
g2
g11
g13
g5
GafA
g12
g15
g6
g10
g8
g14


Bacillus phage SPP1 (109 aa)
Rhodobacter capsulatus SB1003 (112 aa)
Enterobacteria phage λ (68 aa)
Sphingobium amiense DSM16289 (60 aa)
Sphingobium baderi DE13 (68 aa)
Sphingobium chlorophenolicum L1 (73 aa)
Sphingobium cloacae JSM10874 (68 aa)
Sphingobium herbicidovorans MH (73 aa)
Sphingobium indicum B90A (73 aa)
Sphingobium sp. C1 (63 aa)
Sphingobium sp. MI205 (74 aa)
Sphingobium sp. TKS (73 aa)
Sphingobium sp. YG1 (73 aa)
Sphingobium yanoikuaye S72 (68 aa)
alpha helix
beta strand
coil