Co-evolution of gene transfer agents and their alphaproteobacterial hosts

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: Co-evolution of gene transfer agents with their hosts

Keywords: biofilm, horizontal gene transfer, DNA repair, nutritional stress, multi-level selection, social evolution

Abstract

Gene transfer agents (GTAs) are enigmatic elements that resemble small viruses and are known to be produced during nutritional stress by some bacteria and archaea. The production of GTAs is regulated by quorum sensing, under which a small fraction of the population acts as GTA producers, while the rest become GTA recipients. In contrast to canonical viruses, GTAs cannot propagate themselves because they package pieces of the producing cell's genome. In alphaproteobacteria, GTAs are mostly vertically inherited and reside in their hosts' genomes for hundreds of millions of years. While GTAs' ability to transfer genetic material within a population and their long-term preservation suggests an increased fitness of GTA-producing microbes, the associated benefits and type of selection that maintains GTAs are poorly understood. By comparing rates of evolutionary change in GTA genes to the rates in gene families abundantly present across 293 alphaproteobacterial genomes, we detected 59 gene families that likely co-evolve with GTA genes. These gene families are predominantly involved in stress response, DNA repair, and biofilm formation. We hypothesize that biofilm formation enables the physical proximity of GTA-producing cells, limiting GTA-derived benefits only to a group of closely related cells. We further conjecture that population structure of biofilm-forming sub-populations ensures that the trait of GTA production is maintained despite the inevitable rise of "cheating" genotypes. Because release of GTA particles kills the producing cell, maintenance of GTAs is an exciting example of social evolution in a microbial population.

40 ## Importance

41 Gene transfer agents (GTAs) are viruses domesticated by some archaea and bacteria as vehicles
42 for carrying pieces of the host genome. Produced under certain environmental conditions, GTA
43 particles can deliver DNA to neighboring, closely related cells. Function of GTAs remains
44 uncertain. While making GTAs is suicidal for a cell, GTA-encoding genes are widespread in
45 genomes of alphaproteobacteria. Such GTA persistence implies functional benefits but raises
46 question about how selection maintains this lethal trait. By showing that GTA genes co-evolve
47 with genes involved in stress response, DNA repair, and biofilm formation, we provide support
48 for the hypothesis that GTAs facilitate DNA exchange during the stress conditions and present a
49 model for how GTAs persist in biofilm-forming bacterial populations despite being lethal.

50

51 # Introduction

52 Multiple bacteria and archaea produce Gene Transfer Agents (GTAs) – the viriforms
53 whose function and mode of selection to maintain them remain unsolved (1-3). These
54 domesticated virus-derived elements are encoded by genes in their host's genome and, when
55 produced, resemble tailed double-stranded DNA (dsDNA) viruses (phages). In contrast to
56 viruses, GTAs do not package the genes that encode them, and instead contain fragments of the
57 producing host's genome (1, 4-7). Experimentally, GTAs are most studied and characterized in
58 the alphaproteobacteria *Rhodobacter capsulatus* (RcGTA) and *Caulobacter crescentus* (6, 8, 9),
59 but they are also produced by several additional bacterial and archaeal species (10). Many more
60 prokaryotes encode GTA-like genes (5, 11-18), and the presence of GTA-like genes in almost
61 60% of publicly available alphaproteobacterial genomes (19) suggests that GTA production is
62 more widespread than currently appreciated.

63 RcGTA production is a population-level phenomenon: it is triggered by nutrient depletion
64 (20) and is regulated by quorum sensing (21). Only a small subset of the population acts as
65 RcGTA producers; the remaining cells become recipients, by displaying specific polysaccharide
66 receptors for RcGTAs adsorption (22) and expressing competence genes (23). Genetic pieces
67 delivered by RcGTAs to a recipient cell can be integrated into the cell's genome via homologous
68 recombination (24).

69 The benefits of GTA production and of acquiring GTA-packaged DNA in a microbial
70 population are not fully understood. Since their discovery, GTAs were hypothesized to mediate
71 DNA repair (25), and recently this hypothesis was confirmed by experimental demonstration of
72 GTA-mediated DNA repair via homologous recombination in *C. crescentus* (6). Moreover,
73 facilitation of DNA damage repair appears to improve the survival of *C. crescentus* populations

74 in nutrient limited conditions (6), possibly due to a reduction in mutational load. Beyond the
75 repair of already existing genes, released GTA particles could enable exchange of beneficial
76 traits in a microbial population (8, 26) and provide nutrients to surrounding cells as the
77 programmed cell death phenomenon does (27, 28), although these hypotheses remain to be
78 experimentally verified. Despite these putative population-level benefits, GTA-producing cells
79 lyse and therefore leave no progeny, making it impossible for selection that maintains GTA
80 production to act on the level of individual cells. Better understanding of GTA production and
81 reception cycle and of genes underlying it will likely help us elucidate ecological role of GTAs in
82 microbial communities, and details of the population-level selection that preserves the trait.
83 The RcGTA is encoded and regulated by at least 24 genes that are distributed across five
84 different loci (29). Seventeen genes are located in one locus that is commonly referred as the
85 head-tail cluster (2) (**Figure 1A**). The locus encodes the majority of structural proteins required
86 for the RcGTA particle assembly (30). Products of many additional "host" genes are critical for
87 the regulation of the RcGTA particle production, DNA uptake, and DNA integration. For
88 example, the CckA-ChpT-CtrA phosphorelay system, which controls the cell cycle and DNA
89 replication in other model organisms (31-33), modulates production of RcGTA particles and their
90 release (34, 35). Serine acetyltransferase (cysE1), which is required for biofilm formation, plays
91 a critical role for the optimal receipt of RcGTAs (36). Capsular polysaccharides, which serve as
92 RcGTA receptors, are synthesized under control of GtaR/I quorum-sensing genes (22).
93 Competence machinery proteins ComEC, ComF and ComM facilitate entry of DNA into cells
94 (23). Integration of the incoming genetic material into the host genome via homologous
95 recombination is facilitated by DprA and RecA (24). It is likely that products of multiple other

96 “host” gene families important for the proper functioning of GTAs remain to be discovered, and
97 in this study we use a comparative genomics approach to search for such genes.

98 Genes that are involved in the similar molecular processes, or co-expressed together, tend
99 to co-evolve with each other (37, 38), and, vice versa, the protein-protein interactions can be
100 unveiled by finding co-evolving genes that encode the interacting proteins (39, 40). The co-
101 evolution among genes can be effectively identified via Evolutionary Rate Covariation (ERC)
102 approach (37, 41, 42). Evolutionary rate covariation measures the degree of correlation of
103 changes in evolutionary rates across the phylogenies of a pair of proteins, assuming that
104 functionally related proteins have similar selection pressures, resulting in coordinated changes in
105 substitution rates (37, 38). Because GTA head-tail cluster has resided within GTA-containing
106 alphaproteobacterial genomes for at least 700 million years (13) and, as mentioned above, GTA
107 production is tightly integrated into the molecular circuits of the GTA-carrying bacteria,
108 evolutionary rates of gene families involved in GTA lifecycle are expected to correlate with the
109 rates of the GTA genes.

110 Head-tail cluster genes are easily detectable across genomes in a large clade of
111 alphaproteobacteria (11, 13) and therefore provide a rich dataset for comparative analyses of
112 evolutionary rates. In this study, we examined the evolutionary rate covariation patterns of
113 protein-coding genes encoded in 293 representative alphaproteobacterial genomes that contain
114 either complete or nearly complete GTA head-tail clusters. We found that GTA head-tail cluster
115 genes co-evolve with 59 gene families, 55 of which have not been previously linked to GTAs.
116 Thus, we dramatically expand the list of genes that could be important for GTAs’ functionality
117 and hence could provide insights into GTAs’ role in bacterial populations. By combining our

118 findings with the existing knowledge about GTAs, we propose a model that explains persistence
119 of GTA production in bacterial populations.
120

121 # Results

122 **Alphaproteobacterial GTA genes co-evolve with each other**

123 The ERC method was developed for and tested on eukaryotic genes (37, 38) and, to our
124 knowledge, has not been applied to bacterial genomes. Therefore, before using the approach to
125 identify genes co-evolving with GTA genes, we evaluated it on GTA genes found in the 293
126 alphaproteobacterial genomes. Because genes in the GTA head-tail cluster have a common
127 promoter (8, 43) and the gene products are functionally related (i.e., produce a GTA particle that
128 has DNA packaged into its head), we expected strong co-evolution among GTA genes. Indeed,
129 we found that 51 out of 55 possible pairs of the 11 reference GTA genes (see **Methods** for
130 definition) have significantly similar co-variation of evolutionary rates, as measured by the
131 Pearson's coefficient (p-value < 0.05 after Bonferroni correction). Each reference GTA gene co-
132 evolves with at least 7 other reference GTA genes, with 6 of them co-evolving with all 10 other
133 reference GTA genes (**Figure 1B**). These findings suggest that the ERC method adequately
134 identifies co-evolving genes in alphaproteobacterial genomes.
135

136 **Co-evolving alphaproteobacterial genes encode functionally related proteins**

137 Co-evolving genes of eukaryotes identified through ERC analyses were shown to be
138 either functionally related or involved in similar biological processes (37, 38). To examine how
139 robustly the ERC method can identify functionally related gene pairs in our dataset of
140 alphaproteobacterial genomes, we (i) evaluated rates of covariation within functional modules of

141 the GTA head-tail cluster and (ii) examined a relationship between co-evolution and literature-
142 and experiment-based functional inferences for a subset of gene families nearly universally
143 found across GTA-containing alphaproteobacteria.

144 GTA head-tail cluster encodes three modules that are responsible for distinct functional
145 stages of GTA production: DNA packaging, head morphogenesis, and tail morphogenesis (2, 29,
146 30) (**Figure 1A**). Phage genes within the same functional class are more likely to interact with
147 one another (44). We found that the Pearson's correlation coefficient is significantly higher (and,
148 therefore, co-evolutionary signal is significantly stronger) for reference GTA genes within each
149 module than between the reference GTA genes from different modules (Mann-Whitney U-test, p-
150 value = 0.002) (**Figure 1C**). These findings suggest that the strength of co-evolutionary signal
151 measured by the ERC analysis correlates with the degree of physical and functional interactions
152 among GTA genes.

153 Expanding our analysis beyond GTA genes, we examined protein-coding genes in a
154 model marine bacterium *Phaeobacter inhibens*, whose genome encodes the largest number
155 (1,370) of genes from 1,470 gene families nearly universally found across GTA-containing
156 alphaproteobacteria, and in *Caulobacter crescentus* and *Dinoroseobacter shibae*, which are
157 experimentally shown to produce GTAs (5, 6). The information about the interactions of most
158 proteins encoded by the protein-coding genes in the genomes of these bacteria is available in the
159 STRING database (45), and the experimental relative fitnesses of many genes are cataloged in
160 Fitness Browser (46). Using the ERC analysis on the 1,320, 1,199, and 1,255 genes, we
161 identified 10,514, 9,373 and 9,675 co-evolving gene pairs for *P. inhibens*, *C. crescentus*, and *D.*
162 *shibae*, respectively (**Figure S1**). The co-evolution networks, in which nodes correspond to
163 genes and edges designate the presence of significant evolutionary rate covariation (see FigShare

164 repository), are significantly similar to both the networks of the pairwise interactions of the
165 encoded proteins and the networks of fitness effects for all three genomes (p-value < 0.001,
166 permutation test) (**Figure 2**). Moreover, co-evolving genes are more likely to belong to the same
167 Clusters of Orthologous Groups (COG) category (assortativity of 0.096 for *P. inhibens*, 0.083 for
168 *C. crescentus*, and 0.094 for *D. shibae*; p-value < 0.001 in permutation tests for all three
169 genomes) (**Figure 2**). These comparisons show that co-evolving genes that encode well-
170 characterized proteins (defined as present in the STRING and COG databases) and a subset of
171 genes needed for specific environmental conditions (as determined by the Fitness Browser
172 database) indeed tend to encode functionally related proteins. For example, *P. inhibens'* gene
173 pairs with the two largest Pearson's coefficients encode proteins that are involved in the same
174 biological processes (**Figure 3**): *imuB* and *dnaE2* genes (Pearson's coefficient r = 0.800) are
175 located in the same operon and are involved in SOS-induced mutagenesis and translesion
176 synthesis (47, 48), while *addA* and *addB* genes (Pearson's r = 0.735) are known to assemble into
177 the heterodimeric complex to facilitate homologous recombination and DNA repair (49, 50).
178 It is worth noting that some of the genes identified as co-evolving in our analysis are not
179 designated as encoding interacting proteins in the STRING database. However, given
180 incompleteness of our knowledge about functionality of proteins encoded in a bacterial genome,
181 these genes may represent functional connections yet unidentified in STRING. Indeed, ERC
182 analysis has been used to uncover novel protein-protein interactions, especially between
183 hypothetical proteins (40, 51, 52). Here are two examples of proteins identified as co-evolving in
184 our analyses and likely interacting based on what's known about their functions, but not
185 designated as such in the STRING database. The *yfgC* gene, which encodes a periplasmic
186 metalloprotease involved in assembly of outer membrane proteins (53), co-evolves with both the

187 *lptD* and *bamB* genes (**Figure 4**). The YfgC protein plays a crucial role in the assembly of LptD,
188 an outer membrane protein that participates in the lipopolysaccharide assembly (53). The YfgC
189 also interacts with the β-barrel-assembly machinery (BAM) complex, which consists of four
190 lipoproteins, including BamB, and facilitates the assembly and integration of proteins into the
191 outer membrane (53, 54).
192
193 **GTA genes co-evolve with at least fifty-nine other gene families**
194 By analyzing 1,470 gene families almost universally present among the 293
195 representative GTA-containing alphaproteobacteria, we identified 59 gene families that co-
196 evolve with at least 5 reference GTA genes (**Table S1**) (see **Methods** for selection criteria).
197 Notably, four of the 59 gene families - encoding tail fiber protein (DUF2793),
198 competence proteins (comEC and comF), and DNA-protecting protein that facilitates
199 homologous recombination (dprA) (**Table 1**) - have already been shown to play important roles
200 in the RcGTA lifecycle (23, 24, 29). Because tail fiber proteins are a part of the RcGTA particle
201 and thus physically interact with other RcGTA proteins (29), DUF2793's co-evolution with other
202 structural GTA genes is expected. The protein products of the remaining three genes (ComEC,
203 ComF, and DprA) are required for the acquisition of DNA delivered by RcGTAs, and interact
204 physically only with DNA molecules (23, 24). This finding demonstrates that the ERC approach
205 indeed could predict genes functionally linked to the GTA lifecycle.
206 The remaining 55 gene families are involved in various functions (**Tables 1, 2** and **S1**).
207 While 32 of the 55 gene families (**Table 2**) offer exciting opportunities for future research into
208 GTA lifecycle, 23 gene families (**Table 1**) can be either directly linked to the GTA lifecycle by
209 being involved in DNA repair or are likely to be under similar selection pressure as GTA genes

210 due to shared ecological importance (stress response, biofilm formation, oxidative respiration,
211 and cofactor biosynthesis), as elaborated below.

212 Three of the 22 genes – *mutY* (encoding a glycosylase), *phrB* (encoding a DNA
213 photolyase), and *tatD* (encoding an exonuclease) – play roles in repair of DNA damage induced
214 by various oxidative agents and UV light (55-57). Glycosylase actively modulates homologous
215 recombination (58), which could be important for facilitating integration of GTA-derived genetic
216 material into the recipient's genome (6, 9). Photolyase and tatD exonuclease do not directly
217 participate in homologous recombination, but are involved in DNA repair process (6, 55, 56) and
218 thus are likely to be under the similar selection pressure as GTA genes.

219 Two genes (*mazG* and *ydiU*) identified in our screen are involved in stress response. The
220 product of the *mazG* gene modulates the programmed cell death in *Escherichia coli* and regulates
221 intracellular level of ppGpp, the universal 'alarmone', which was previously implicated in the
222 regulation of GTAs production (20, 59). The ppGpp molecule is involved in a cellular response
223 to a variety of stress conditions, including nutritional stress. The product of the *ydiU* gene
224 mediates UMPylation of bacterial chaperones, improving bacterial fitness under the stressful
225 environmental settings (60).

226 Thirteen genes identified in our analyses have relevance to biofilms. While these genes
227 are linked to the formation of biofilms, they are also involved in a range of other molecular
228 functions. The *hrpB* gene, which encodes ATP-dependent RNA helicase, has been shown to be
229 important for both biofilm formation and adhesion on surfaces (61). The *ccmA* gene, *cycH* gene,
230 and a gene from COG5387 family (*ATP12*) are involved in oxidative respiration, which
231 promotes bacterial survival in the biofilms (62, 63). Moreover, as a group, 59 co-evolving gene
232 families are enriched in three metabolic pathways relevant to biofilms: cofactor biosynthesis,

233 folate biosynthesis and sulfur relay system (hypergeometric test, p-value < 0.05, Bonferroni
234 correction) (**Table 1**). Two genes from the 'cofactor biosynthesis' pathway (*moeA* and *moaC*) are
235 involved in the molybdenum cofactor biosynthesis. Both metabolism of folate and molybdenum
236 cofactors are important for biofilm formation (64, 65). The sulfur relay pathway is involved in
237 the tRNA modifications (66), which are implicated in fitness of bacteria within a biofilm (62).
238 Notably, five additional genes that encode tRNA modification enzymes (*mnmE*, *tilS*, *SUA5*,
239 *dusA*, *tadA*) are inferred to co-evolve with GTA genes (**Table 1**). Finally, the *tadB* gene is
240 involved in the assembly of the tad pili, which have been shown to be involved in promoting the
241 surface colonization and biofilm cohesion (67, 68). Interestingly, in *D. shibae tad* genes are co-
242 regulated with GTA genes, further corroborating linkage between tad pili and GTAs (33).
243

244 ## Discussion

245 By showing that GTA head-tail cluster genes, and especially genes within the same
246 functional module of the cluster, tend to co-evolve with each other, and by examining function-
247 coevolution relationship among proteins encoded in a model marine bacterium, we demonstrated
248 that the ERC method is an effective approach to uncover functional relationships among protein-
249 coding genes in bacteria, extending the method's applicability beyond eukaryotes. Applying the
250 method to GTA-encoding alphaproteobacterial genomes that span >700 million years of
251 evolution (13), we detected a significant evolutionary rate covariation of GTA head-tail cluster
252 genes with 59 protein-coding genes. Multiple genes in this dataset are involved in stress
253 response, DNA repair, homologous recombination, and biofilm formation. These functions are
254 consistent with the accumulating experimental and computational evidence about GTA
255 production, regulation and function in *R. capsulatus* and *C. crescentus*, and with previous

256 hypotheses and models of GTA production triggered under environmental stress (6, 20), GTAs
257 being involved in DNA repair in recipient cell (6, 69), and, most recently, of GTA production
258 occurring in biofilms (36). Our discovery that 13 biofilm-implicated genes co-evolve with GTA
259 genes further highlight the potential importance of biofilm settings for GTA production.

260 Alphaproteobacteria in general, and GTA-producing *R. capsulatus* and *C. crescentus* in
261 particular, are known to form biofilms (36, 70-72). Biofilms provide a microbial community with
262 benefits that cannot be achieved by the individual cells, such as protection against antibiotics
263 (73) and viral infections (74, 75). Despite either shown or hypothesized benefits of GTA-
264 disseminated DNA to the recipient cells (2, 6, 76), zero relative fitness of the lysed GTA-
265 producing cells implies that the trait of encoding GTAs can only be favored by selection when
266 the benefits of receiving GTAs are confined to either clonal or very closely related cells that also
267 possess the genes encoding the suicidal GTA production trait. Encoding genes for synthesis of
268 specific polysaccharide receptors, which facilitate the adsorption of GTAs (22) and co-regulating
269 the receptor production in one fraction of a population with GTA production in another fraction
270 of a population, can limit GTAs' targets only to clonal cells or very closely related species.
271 Additionally, success of homologous recombination declines exponentially with the increase in
272 genomic sequence divergence (77, 78), further restricting the usefulness of GTA particles for
273 DNA repair to closely related cells.

274 Within such groups of closely related cells, GTAs can be viewed as "population-level
275 goods". However, any population-level goods system faces an inevitable rise of cheaters (79),
276 and a population of GTA producers would be susceptible to cheaters that would not produce GTA
277 particles but still have surface polysaccharides that serve as GTA receptors. Consistent with these
278 conjectures, we observe both the pseudogenization and complete loss of GTA systems in

279 multiple alphaproteobacterial lineages (11, 13, 19). Pseudogenized GTA gene clusters may
280 represent recently emerged cheater lineages, while the absence of GTA gene clusters could be a
281 result of cheater takeover in a species and consequent loss of GTA genes due to the deletion bias
282 (80). (It should be noted that, in both cases, the loss of GTA production in these lineages could
283 also be attributed to acquisition of alternative molecular mechanisms to cope with nutritional
284 stress and DNA repair, or inhabiting niches where GTA production costs outweigh its benefits.)
285 Mathematical modeling showed that maintenance of GTA production trait can be difficult in
286 mixed populations of GTA producers and cheaters, at least under some conditions (81). Yet,
287 despite the likely appearance of cheaters and an observation of recurrent GTA loss in multiple
288 lineages, the GTA production trait has been persisting in alphaproteobacteria for hundreds of
289 millions of years (13), suggesting that some kind of population-level selection is successful.
290 However, details of how such selection operates remain unknown.

291 One possible solution for an "altruistic" trait to persist in a population over time is to
292 have the population segregated into small sub-populations, an evolutionary scenario first
293 modeled by Wilson (82) and subsequently shown to be equivalent to multi-level selection models
294 that emphasize close relatedness of sub-population members (83, 84). Under this model, the
295 cheaters arise stochastically and therefore are found in many but not all sub-populations.
296 Notably, sub-populations without cheaters outcompete sub-populations with cheaters by having
297 overall higher total productivity due to benefits of the population-level goods produced by the
298 altruistic trait. We hypothesize that GTA systems persist over time because GTA production
299 occurs in such spatially structured populations. We further hypothesize that *it is biofilms* that
300 facilitate the division of GTA-producing populations into isolated sub-populations. A biofilm
301 would ensure that clonal or closely related cells are in spatial proximity, plus they would protect

302 the GTA-producing sub-population from being invaded by other cells, including cheaters. Dense
303 packing of cells would also ensure that the released GTA particles are not dispersed in the
304 environment and reach their recipients (36). Biofilms would also trap the organic debris of the
305 lysed GTA-producer cells within the biofilm (85). DNA of these lysed cells could be used as a
306 part of a biofilm matrix (85), enhancing the sub-population isolation and protection from
307 environmental hazards (86). Other environmental and cellular debris could increase localized
308 nutrient availability. Previous experimental work on biofilms supports some of our conjectures.
309 In the GTA-producing *Caulobacter crescentus* the biofilm structure promotes clonal cells to
310 reside in proximity (71). Biofilm formation is also increased during the ecological competition,
311 providing microbes with the protection to resist invasion by different strains (87).
312 Building on the current knowledge and previous models of GTA evolution and function,
313 we propose the following model of selection acting on a structured, biofilm-forming bacterial
314 population and maintaining GTA production (**Figure 5**). When a population experiences
315 starvation, nutritional stress increases the generation of reactive oxygen species, which induces
316 DNA damage (88). In a structured population, the fate of each sub-population (i.e., of individual
317 biofilms) depends on the genetic make-up of its cells. In a sub-population of GTA producers
318 (GTA^+ cells), a small fraction of cells is "sacrificed", and their DNA is delivered to the
319 remaining cells or hoarded as part of the biofilm matrix, while the cellular debris are utilized as
320 additional nutrients. The resources are protected from invaders by the biofilm structure. As a
321 result, the sub-population counteracts the negative effects of nutrient scarcity and DNA damage,
322 and thus experiences, at worst, only a small population decline due to the lysed GTA^+ cells.
323 However, if cheaters arise within the biofilm boundaries, the composition of such sub-population
324 under multiple episodes of nutritional stress will change to a higher fraction of GTA non-

325 producers (GTA^- cells), as these types of cells will not be lost due to lysis during GTA production
326 and yet will experience all benefits of the population-level goods released by GTA^+ cells.
327 Eventually, the sub-population will lose GTA production trait. In a sub-population of GTA^- cells,
328 mutational load due to DNA damage and limited nutrients will result in a decline of the sub-
329 population size. Therefore, GTA^+ sub-populations will have higher relative fitness in comparison
330 to GTA^- sub-populations, resulting in the larger overall number of GTA^+ cells in the combined
331 population. Thus, the GTA production trait will be maintained in the population as a whole.
332 Although bits of experimental evidence used in the above model come mostly from the
333 research on GTAs in *Rhodobacter capsulatus* and *Caulobacter crescentus*, we hypothesize that
334 our model is applicable to other GTA-containing alphaproteobacterial species, because the
335 detected co-evolution patterns span multiple diverse alphaproteobacterial clades.
336

337 # Materials and Methods

338 **Identification of 293 representative alphaproteobacterial genomes with GTA regions**

339 Initially, 1,642 alphaproteobacterial genomes and annotations of their protein-coding
340 genes were retrieved from the NCBI's Assembly and RefSeq database (accessed June 2022) (89)
341 (**Table S2**). In these genomes, GTA regions were predicted using GTA-Hunter program with
342 default parameters (19). Because GTA-Hunter looks only for 11 out of the 17 genes of the
343 RcGTA's head-tail cluster, the remaining GTA genes were identified via BLASTP searches (E-
344 value < 0.1) (90), using as queries the curated set of GTA regions from (19). Only BLASTP
345 matches that are located within the GTA-Hunter-predicted GTA regions were added. Using this
346 procedure, GTA regions were identified in 701 genomes.

347 To avoid presence of multiple highly similar GTA regions in downstream analyses, the
348 701 genomes were clustered into 392 Operational Taxonomic Units (OTUs) using the Average
349 Nucleotide Identity (ANI) cutoff of 95%, calculated via fastANI v1.1 (91). Within each OTU,
350 GTA regions were examined for "completeness", defined as having 14 out of the 17 head-tail
351 cluster genes (genes *g1*, *g3.5* and *g7* were excluded because they are not easily detected across
352 GTA-contaning alphaproteobacterial clades (13)). Incomplete GTA regions were discarded. This
353 criterion reduced the number of OTUs to 293. Within each of the 293 OTUs, only one, randomly
354 selected, genome and its GTA region were retained for subsequent analyses (**Table S3**).
355
356 **Reconstruction of the reference phylogenomic tree**
357 From the set of 120 marker genes widely used for phylogenomic taxonomy (92), 84 gene
358 families were detected in a single copy in at least 95% of the 293 genomes using AMPHORA2
359 (93). Amino acid sequences of each of these 84 gene families were aligned using MAFFT v7.505
360 with the 'linsi' option (94). The alignments were concatenated, and the best substitution model
361 for each alignment and the optimal partition scheme were established via ModelFinder (95). The
362 maximum-likelihood phylogeny was reconstructed using IQ-TREE v2.2 (96). The tree was
363 rooted using the *Emcibacterales* and *Sphingomonadales* taxonomic orders, using the previously
364 observed branching order of the *Alphaproteobacteria* as a guide (13, 19).
365
366 **Selection of Reference GTA genes**
367 The amino acid sequences of the 14 GTA genes from the GTA regions of the 293
368 genomes were retrieved and aligned using MAFFT v7.505 with the 'linsi' option. Phylogenetic
369 trees were reconstructed from each alignment using IQ-TREE v2.2 (96) under the best

370 substitution model identified by ModelFinder (95). Each tree was compared to the reference tree
371 using the normalized quartet scores calculated in ASTRAL v5.7.8 (97). Eleven GTA genes that
372 exhibited a high congruency with the reference phylogeny (quartet score > 0.8) (**Figure S2**) were
373 designated as "reference GTA genes" and are referred as such throughout the manuscript.
374

375 **Identification and functional annotation of gene families in 293 GTA-containing genomes**

376 Gene families were defined as orthologous groups identified in Broccoli v1.2 (98), using
377 DIAMOND (99) for protein similarity searches and the maximum-likelihood method for
378 phylogenetic reconstructions. Gene families that are present in a single copy in at least 50% of
379 the 293 GTA-containing genomes (1,470 in total) were retained for further analyses.
380 To assign COG functional annotations to the gene families, one randomly selected
381 representative from each family was used as a query against the eggNOG database v5.0.2 and
382 processed through eggNOG-mapper v2.1.9 workflow (100). For gene families found to be co-
383 evolving with the GTA region (see below), additional annotations were sought out using
384 PaperBLAST (accessed in December 2022) (101) and CD-searches against CDD database v3.20
385 (accessed in December 2022) (102).
386

387 **Inference of evolutionary rate covariation**

388 Amino acid sequences of each gene family and each reference GTA gene were aligned
389 using MAFFT v.7.505 with the 'linsi' option (94) . For gene families that are not found in all 293
390 genomes, the absent taxa were pruned from the reference tree using functions from the *ete3*
391 package (103). For each gene set, the topology of taxa relationships was constrained to the
392 reference phylogeny and branch lengths were estimated via IQ-TREE v2.2 (96), using the best

393 substitution model suggested by ModelFinder (95). The trees were rooted using relationships in
394 the reference phylogeny as a guide.

395 Covariations of evolutionary rates among 1,470 gene families and 11 GTA reference
396 genes were examined rates using the CovER pipeline, as implemented in PhyKIT v1.11.12 (38,
397 104). Within the pipeline, the following steps were carried out. For each pair, their trees were
398 pruned to retain only shared taxa. All trees were corrected for the differences in mutation rates
399 and divergence times among taxa; this was accomplished by dividing the length of each branch
400 by the length of the corresponding branch of the reference tree. Branches with the normalized
401 length > 5 were removed from further analyses, and the retained branch lengths were Z-
402 transformed. For every pair, Pearson correlation coefficient was calculated. A pair of genes was
403 designated as co-evolving, if the Pearson correlation coefficient was positive and had p-value <
404 0.05 after Bonferroni correction for multiple testing.

405 The above-described co-variation analysis was carried out on five datasets: among 11
406 GTA reference genes (11 x 10/2 = 55 comparisons), between GTA genes and 1,470 gene families
407 (11 x 1,470 = 16,170 comparisons), among 1,320 gene families present in GTA-containing
408 genome *Phaeobacter inhibens* (1,320 x 1,319/2 = 870,540 comparisons), among 1,199 gene
409 families present in GTA-producing genome *Caulobacter crescentus* (1,199 x 1,198/2 = 718,201
410 comparisons), and among 1,255 detected in GTA-producing genome *Dinoroseobacter shibae*
411 (1,255 x 1,254 = 786,885). The *P. inhibens, C. crescentus and D.shibae* analyses resulted in
412 10,514, 9,373 and 9,675 co-evolving gene pairs, respectively (**Figure S1**). The gene pairs
413 information was assembled into co-evolution networks, in which nodes represent gene families
414 and edges depict co-evolution relationships. To annotate nodes in the co-evolution networks with
415 functional categories, the sub-networks of 1,040 out of 1,320 gene families (for *P. inhibens*), 953

416 out of 1,199 gene families (for *C. crescentus*), and 997 out of 1,255 gene families (for *D. shibae*),
417 that had unambiguous COG assignments were extracted. The COG functional annotations of
418 genes were assigned as labels.

419 To minimize the number of false positives and to shorten the list of candidate gene
420 families that co-evolve with GTA genes, the following criteria were applied in addition to
421 Pearson's R > 0 and Bonferroni-corrected p-value < 0.05: A gene family was required (i) to co-
422 evolve with at least 5 GTA reference genes and (ii) to be above the 95th percentile in the of 1,470
423 gene families ranked by both p-value and Pearson's R for at least 5 GTA reference genes. Under
424 these criteria, 59 gene families were retained for further analyses.

425

426 **Reconstruction of protein-protein interaction networks**

427 Amino acid sequences of the 1,320 (*P. inhibens*), 1,199 (*C. crescentus*), and 1,255 (*D.*
428 *shibae*) above-described gene families were used as queries against the STRING database v.11.5
429 (45), with the high confidence score cutoff and all available sources. This search resulted in
430 8,612, 8,385, and 8,366 interacting protein-protein pairs for *P. inhibens*, *C. crescentus*, and *D.*
431 *shibae*, respectively. The information was assembled into a *Phaeobacter inhibens* protein-protein
432 interactions (PPI) networks, in which nodes represent genes and edges depict PPIs.

433

434 **Reconstruction of co-fitness networks**

435 *P. inhibens*', *C. crescentus'*, and *D. shibae*'s gene pairs with similar fitnesses across a
436 wide range of different experimental conditions were retrieved from the Fitness Browser (46)
437 (accessed October 2023). Gene pairs were designated as co-fit if they either had a co-fitness
438 value > 0.75, or they had a co-fitness value > 0.60 and were conserved in other bacterial species

439 (46). This search resulted in 569, 2,657, and 540 co-fit gene pairs for *P. inhibens*, *C. crescentus*,
440 and *D. shibae*, respectively. The information was assembled into co-fitness networks, in which
441 nodes represent genes and edges depict co-fitness associations.
442
443 **Comparison of networks and identification of subnetworks**
444 The similarity between *P. inhibens' C. crescentus'*, and *D. shibae*' co-evolution networks
445 and either PPI or co-fitness networks was assessed by calculating Jaccard index (105), which
446 measures the fraction of edges shared between the networks. The null distribution of Jaccard
447 indices was created by randomly re-shuffling of the evolutionary rate covariation network 1,000
448 times.
449 Tendency of nodes from the same COG category to connect with each other was
450 measured by an assortativity coefficient calculated for the co-evolution subnetworks with
451 unambiguous COG assignments. The coefficient quantifies the connectivity of nodes belonging
452 to the same class and varies between 1 (only nodes of the same class are connected) and -1 (only
453 nodes that belong to different classes are connected). To assess the significance of the observed
454 assortativity values, the permutation test was performed by random shuffling of COG labels
455 1,000 times. The values of assortativity coefficients were calculated using *igraph* v1.3.5 (106).
456 All networks were visualized using *igraph* v1.3.5 (106).
457
458 **KEGG pathways enrichment analysis for 59 genes that co-evolve with GTAs**
459 Each of the 59 gene families was assigned a KEGG Orhology (KO) label using
460 BlastKOALA (107). Significantly enriched pathways were identified by the hypergeometric test

461 (p-value < 0.05 with Bonferroni correction for multiple testing), as implemented in the
462 *clusterProfiler* package v4.4.4 (108).
463

464 Data Availability

465 The genomes used in this study are publicly available via NCBI Assembly
466 (https://www.ncbi.nlm.nih.gov/assembly) database. The accession numbers of these genomes are
467 listed in **Table S2**. The following datasets, which were derived from the genomes, are available
468 in the FigShare repository (https://doi.org/10.6084/m9.figshare.23929551.v1): multiple sequence
469 alignment of GTA genes; unconstrained and constrained phylogenetic trees of GTA genes;
470 concatenated alignment of phylogenomic markers and reconstructed reference phylogenomic
471 tree; amino acid sequences of 1,470 gene families; constrained phylogenetic trees of 1,470 gene
472 families; covariation of evolutionary rates for all performed pairwise gene comparisons;
473 *Phaeobacter inhibens*', *Caulobacter crescentus', Dinoroseobacter shibae'*s co-evolutionary
474 networks and their subnetworks of nodes with unambiguous GOG assignment, protein-protein
475 interaction networks, fitness networks, and lists of COG functional category assignments for
476 nodes.
477

478 Acknowledgements

479 We thank Dr. Carey D. Nadell (Dartmouth College) for stimulating discussions about
480 biofilms and social evolution theory, and for critical reading of the manuscript. The work was
481 supported in part by Dartmouth Fellowship and Cramer funds to RK, and Dartmouth Dean of
482 Faculty funds to OZ.

483 # References

484 1. Kogay R, Koppenhofer S, Beatty JT, Kuhn JH, Lang AS, Zhaxybayeva O. 2022. Formal
485 recognition and classification of gene transfer agents as viriforms. Virus Evol 8:veac100.
486 2. Lang AS, Westbye AB, Beatty JT. 2017. The distribution, evolution, and roles of gene
487 transfer agents in prokaryotic genetic exchange. Annu Rev Virol 4:87-104.
488 3. Kuhn JH, Koonin EV. 2023. Viriforms-a new category of classifiable virus-derived
489 genetic elements. Biomolecules 13:289.
490 4. Lang AS, Beatty JT. 2001. The gene transfer agent of *Rhodobacter capsulatus* and
491 "constitutive transduction" in prokaryotes. Arch Microbiol 175:241-9.
492 5. Tomasch J, Wang H, Hall ATK, Patzelt D, Preusse M, Petersen J, Brinkmann H, Bunk B,
493 Bhuju S, Jarek M, Geffers R, Lang AS, Wagner-Dobler I. 2018. Packaging of
494 *Dinoroseobacter shibae* DNA into gene transfer agent particles is not random. Genome
495 Biol Evol 10:359-369.
496 6. Gozzi K, Tran NT, Modell JW, Le TBK, Laub MT. 2022. Prophage-like gene transfer
497 agents promote *Caulobacter crescentus* survival and DNA repair during stationary phase.
498 PLoS Biol 20:e3001790.
499 7. Hynes AP, Mercer RG, Watton DE, Buckley CB, Lang AS. 2012. DNA packaging bias
500 and differential expression of gene transfer agent genes within a population during
501 production and release of the *Rhodobacter capsulatus* gene transfer agent, RcGTA. Mol
502 Microbiol 85:314-25.
503 8. Lang AS, Beatty JT. 2000. Genetic analysis of a bacterial genetic exchange element: the
504 gene transfer agent of *Rhodobacter capsulatus*. Proc Natl Acad Sci U S A 97:859-64.
505 9. Marrs B. 1974. Genetic recombination in *Rhodopseudomonas capsulata*. Proc Natl Acad
506 Sci U S A 71:971-3.
507 10. Lang AS, Zhaxybayeva O, Beatty JT. 2012. Gene transfer agents: phage-like elements of
508 genetic exchange. Nat Rev Microbiol 10:472-82.
509 11. Lang AS, Beatty JT. 2007. Importance of widespread gene transfer agent genes in alpha-
510 proteobacteria. Trends Microbiol 15:54-62.
511 12. Lang AS, Taylor TA, Beatty JT. 2002. Evolutionary implications of phylogenetic analyses
512 of the gene transfer agent (GTA) of *Rhodobacter capsulatus*. J Mol Evol 55:534-43.
513 13. Shakya M, Soucy SM, Zhaxybayeva O. 2017. Insights into origin and evolution of alpha-
514 proteobacterial gene transfer agents. Virus Evol 3:vex036.
515 14. George EE, Tashyreva D, Kwong WK, Okamoto N, Horak A, Husnik F, Lukes J, Keeling
516 PJ. 2022. Gene transfer agents in bacterial endosymbionts of microbial eukaryotes.
517 Genome Biol Evol 14:evac099.
518 15. Fallon AM, Carroll EM. 2023. Virus-like particles from *Wolbachia*-infected cells may
519 include a gene transfer agent. Insects 14:516.
520 16. Québatte M, Christen M, Harms A, Körner J, Christen B, Dehio C. 2017. Gene transfer
521 agent promotes evolvability within the fittest subpopulation of a bacterial pathogen. Cell
522 Syst 4:611-621 e6.
523 17. Tamarit D, Neuvonen MM, Engel P, Guy L, Andersson SGE. 2018. Origin and evolution
524 of the *Bartonella* gene transfer agent. Mol Biol Evol 35:451-464.
525 18. Lindqvist LL, Jarmusch SA, Sonnenschein EC, Strube ML, Kim J, Nielsen MW, Kempen
526 PJ, Schoof EM, Zhang SD, Gram L. 2023. Tropodithietic Acid, a Multifunctional

527 Antimicrobial, Facilitates Adaption and Colonization of the Producer, Phaeobacter
528 piscinae. mSphere 8:e0051722.
529 19. Kogay R, Neely TB, Birnbaum DP, Hankel CR, Shakya M, Zhaxybayeva O. 2019.
530 Machine-learning classification suggests that many alphaproteobacterial prophages may
531 instead be gene transfer agents. Genome Biol Evol 11:2941-2953.
532 20. Westbye AB, O'Neill Z, Schellenberg-Beaver T, Beatty JT. 2017. The *Rhodobacter*
533 *capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega
534 subunit. Microbiology (Reading) 163:1355-1363.
535 21. Leung MM, Brimacombe CA, Spiegelman GB, Beatty JT. 2012. The GtaR protein
536 negatively regulates transcription of the gtaRI operon and modulates gene transfer agent
537 (RcGTA) expression in *Rhodobacter capsulatus*. Mol Microbiol 83:759-74.
538 22. Brimacombe CA, Stevens A, Jun D, Mercer R, Lang AS, Beatty JT. 2013. Quorum-
539 sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus*
540 gene transfer agent (RcGTA). Mol Microbiol 87:802-17.
541 23. Brimacombe CA, Ding H, Johnson JA, Beatty JT. 2015. Homologues of genetic
542 transformation DNA import genes are required for *Rhodobacter capsulatus* gene transfer
543 agent recipient capability regulated by the response regulator CtrA. J Bacteriol 197:2653-
544 63.
545 24. Brimacombe CA, Ding H, Beatty JT. 2014. *Rhodobacter capsulatus* DprA is essential for
546 RecA-mediated gene transfer agent (RcGTA) recipient capability regulated by quorum-
547 sensing and the CtrA response regulator. Mol Microbiol 92:1260-78.
548 25. Marrs B, Wall JD, Gest H. 1977. Emergence of the biochemical genetics and molecular
549 biology of photosynthetic bacteria. Trends Biochem Sci 2:105-108.
550 26. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. 2010. High
551 frequency of horizontal gene transfer in the oceans. Science 330:50.
552 27. Allocati N, Masulli M, Di Ilio C, De Laurenzi V. 2015. Die for the community: an
553 overview of programmed cell death in bacteria. Cell Death Dis 6:e1609.
554 28. Kogay R, Wolf YI, Koonin EV, Zhaxybayeva O. 2020. Selection for reducing energy cost
555 of protein production drives the GC content and amino acid composition bias in gene
556 transfer agents. mBio 11:e01206-20.
557 29. Hynes AP, Shakya M, Mercer RG, Grull MP, Bown L, Davidson F, Steffen E, Matchem
558 H, Peach ME, Berger T, Grebe K, Zhaxybayeva O, Lang AS. 2016. Functional and
559 evolutionary characterization of a gene transfer agent's multilocus “genome". Mol Biol
560 Evol 33:2530-43.
561 30. Bardy P, Fuzik T, Hrebik D, Pantucek R, Thomas Beatty J, Plevka P. 2020. Structure and
562 mechanism of DNA delivery of a gene transfer agent. Nat Commun 11:3034.
563 31. Quon KC, Yang B, Domian IJ, Shapiro L, Marczynski GT. 1998. Negative control of
564 bacterial DNA replication by a cell cycle regulatory protein that binds at the chromosome
565 origin. Proc Natl Acad Sci U S A 95:120-5.
566 32. Laub MT, Chen SL, Shapiro L, McAdams HH. 2002. Genes directly controlled by CtrA,
567 a master regulator of the Caulobacter cell cycle. Proc Natl Acad Sci U S A 99:4632-7.
568 33. Wang H, Ziesche L, Frank O, Michael V, Martin M, Petersen J, Schulz S, Wagner-Dobler
569 I, Tomasch J. 2014. The CtrA phosphorelay integrates differentiation and communication
570 in the marine alphaproteobacterium Dinoroseobacter shibae. BMC Genomics 15:130.

571 34. Mercer RG, Quinlan M, Rose AR, Noll S, Beatty JT, Lang AS. 2012. Regulatory systems
572 controlling motility and gene transfer agent production and release in *Rhodobacter*
573 *capsulatus*. FEMS Microbiol Lett 331:53-62.
574 35. Farrera-Calderon RG, Pallegar P, Westbye AB, Wiesmann C, Lang AS, Beatty JT. 2021.
575 The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer
576 agent production Is bidirectional and regulated by cyclic di-GMP. J Bacteriol
577 203:e00525-20.
578 36. Sherlock D, Fogg PCM. 2022. Loss of the *Rhodobacter capsulatus* serine acetyl
579 transferase gene, cysE1, impairs gene transfer by gene transfer agents and biofilm
580 phenotypes. Appl Environ Microbiol 88:e0094422.
581 37. Clark NL, Alani E, Aquadro CF. 2012. Evolutionary rate covariation reveals shared
582 functionality and coexpression of genes. Genome Res 22:714-20.
583 38. Steenwyk JL, Phillips MA, Yang F, Date SS, Graham TR, Berman J, Hittinger CT, Rokas
584 A. 2022. An orthologous gene coevolution network provides insight into eukaryotic
585 cellular and genomic structure and function. Sci Adv 8:eabn0105.
586 39. Kim WK, Bolser DM, Park JH. 2004. Large-scale co-evolution analysis of protein
587 structural interlogues using the global protein structural interactome map (PSIMAP).
588 Bioinformatics 20:1138-50.
589 40. Brunette GJ, Jamalruddin MA, Baldock RA, Clark NL, Bernstein KA. 2019. Evolution-
590 based screening enables genome-wide prioritization and discovery of DNA repair genes.
591 Proc Natl Acad Sci U S A 116:19593-19599.
592 41. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of
593 proteins with their interaction partners. J Mol Biol 299:283-93.
594 42. Sato T, Yamanishi Y, Kanehisa M, Toh H. 2005. The inference of protein-protein
595 interactions by co-evolutionary analysis is improved by excluding the information about
596 the phylogenetic relationships. Bioinformatics 21:3482-9.
597 43. Fogg PCM. 2019. Identification and characterization of a direct activator of a gene
598 transfer agent. Nat Commun 10:595.
599 44. Rajagopala SV, Casjens S, Uetz P. 2011. The protein interaction map of bacteriophage
600 lambda. BMC Microbiol 11:213.
601 45. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT,
602 Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. 2021. The STRING database in
603 2021: customizable protein-protein networks, and functional characterization of user-
604 uploaded gene/measurement sets. Nucleic Acids Res 49:D605-D612.
605 46. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA,
606 Lamson JS, Suh Y, Carlson HK, Esquivel Z, Sadeeshkumar H, Chakraborty R, Zane GM,
607 Rubin BE, Wall JD, Visel A, Bristow J, Blow MJ, Arkin AP, Deutschbauer AM. 2018.
608 Mutant phenotypes for thousands of bacterial genes of unknown function. Nature
609 557:503-509.
610 47. Abella M, Erill I, Jara M, Mazon G, Campoy S, Barbe J. 2004. Widespread distribution
611 of a lexA-regulated DNA damage-inducible multiple gene cassette in the Proteobacteria
612 phylum. Mol Microbiol 54:212-22.
613 48. Erill I, Campoy S, Mazon G, Barbe J. 2006. Dispersal and regulation of an adaptive
614 mutagenesis cassette in the bacteria domain. Nucleic Acids Res 34:66-77.
615 49. Kooistra J, Haijema BJ, Venema G. 1993. The *Bacillus subtilis* addAB genes are fully
616 functional in *Escherichia coli*. Mol Microbiol 7:915-23.

617 50. Saikrishnan K, Yeeles JT, Gilhooly NS, Krajewski WW, Dillingham MS, Wigley DB.
618 2012. Insights into Chi recognition from the structure of an AddAB-type helicase-
619 nuclease complex. EMBO J 31:1568-78.
620 51. Raza Q, Choi JY, Li Y, O'Dowd RM, Watkins SC, Chikina M, Hong Y, Clark NL,
621 Kwiatkowski AV. 2019. Evolutionary rate covariation analysis of E-cadherin identifies
622 Raskol as a regulator of cell adhesion and actin dynamics in Drosophila. PLoS Genet
623 15:e1007720.
624 52. Forsythe ES, Williams AM, Sloan DB. 2021. Genome-wide signatures of plastid-nuclear
625 coevolution point to repeated perturbations of plastid proteostasis systems across
626 angiosperms. Plant Cell 33:980-997.
627 53. Narita S, Masui C, Suzuki T, Dohmae N, Akiyama Y. 2013. Protease homolog BepA
628 (YfgC) promotes assembly and degradation of beta-barrel membrane proteins in
629 Escherichia coli. Proc Natl Acad Sci U S A 110:E3612-21.
630 54. Han L, Zheng J, Wang Y, Yang X, Liu Y, Sun C, Cao B, Zhou H, Ni D, Lou J, Zhao Y,
631 Huang Y. 2016. Structure of the BAM complex and its implications for biogenesis of
632 outer-membrane proteins. Nat Struct Mol Biol 23:192-6.
633 55. Kim JJ, Sundin GW. 2001. Construction and analysis of photolyase mutants of
634 *Pseudomonas aeruginosa* and *Pseudomonas syringae*: contribution of photoreactivation,
635 nucleotide excision repair, and mutagenic DNA repair to cell survival and mutability
636 following exposure to UV-B radiation. Appl Environ Microbiol 67:1405-11.
637 56. Chen YC, Li CL, Hsiao YY, Duh Y, Yuan HS. 2014. Structure and function of TatD
638 exonuclease in DNA repair. Nucleic Acids Res 42:10776-85.
639 57. Krokan HE, Standal R, Slupphaug G. 1997. DNA glycosylases in the base excision repair
640 of DNA. Biochem J 325 (Pt 1):1-16.
641 58. Spek EJ, Vuong LN, Matsuguchi T, Marinus MG, Engelward BP. 2002. Nitric oxide-
642 induced homologous recombination in Escherichia coli is promoted by DNA
643 glycosylases. J Bacteriol 184:3501-7.
644 59. Gross M, Marianovsky I, Glaser G. 2006. MazG -- a regulator of programmed cell death
645 in *Escherichia coli*. Mol Microbiol 59:590-601.
646 60. Yang Y, Yue Y, Song N, Li C, Yuan Z, Wang Y, Ma Y, Li H, Zhang F, Wang W, Jia H, Li
647 P, Li X, Wang Q, Ding Z, Dong H, Gu L, Li B. 2020. The YdiU domain modulates
648 bacterial stress signaling through Mn(2+)-dependent UMPylation. Cell Rep 32:108161.
649 61. Granato LM, Picchi SC, Andrade Mde O, Takita MA, de Souza AA, Wang N, Machado
650 MA. 2016. The ATP-dependent RNA helicase HrpB plays an important role in motility
651 and biofilm formation in *Xanthomonas citri* subsp. citri. BMC Microbiol 16:55.
652 62. Schinner S, Engelhardt F, Preusse M, Thoming JG, Tomasch J, Haussler S. 2020. Genetic
653 determinants of *Pseudomonas aeruginosa* fitness during biofilm growth. Biofilm
654 2:100023.
655 63. Martin-Rodriguez AJ. 2022. Respiration-induced biofilm formation as a driver for
656 bacterial niche colonization. Trends Microbiol doi:10.1016/j.tim.2022.08.007:120-134.
657 64. Wong EHJ, Ng CG, Goh KL, Vadivelu J, Ho B, Loke MF. 2018. Metabolomic analysis of
658 low and high biofilm-forming *Helicobacter pylori* strains. Sci Rep 8:1409.
659 65. Andreae CA, Titball RW, Butler CS. 2014. Influence of the molybdenum cofactor
660 biosynthesis on anaerobic respiration, biofilm formation and motility in *Burkholderia*
661 *thailandensis*. Res Microbiol 165:41-9.

662 66. Leimkuhler S, Buhning M, Beilschmidt L. 2017. Shared sulfur mobilization routes for
663 tRNA thiolation and molybdenum cofactor biosynthesis in prokaryotes and eukaryotes.
664 Biomolecules 7:5.
665 67. Mignolet J, Panis G, Viollier PH. 2018. More than a Tad: spatiotemporal control of
666 Caulobacter pili. Curr Opin Microbiol 42:79-86.
667 68. Sangermani M, Hug I, Sauter N, Pfohl T, Jenal U. 2019. Tad Pili Play a Dynamic Role in
668 Caulobacter crescentus Surface Colonization. mBio 10.
669 69. Kogay R, Zhaxybayeva O. 2022. Selection for translational efficiency in genes associated
670 with alphaproteobacterial gene transfer agents. mSystems 7:e0089222.
671 70. Zhang W, Ding W, Li YX, Tam C, Bougouffa S, Wang R, Pei B, Chiang H, Leung P, Lu
672 Y, Sun J, Fu H, Bajic VB, Liu H, Webster NS, Qian PY. 2019. Marine biofilms constitute
673 a bank of hidden microbial diversity and functional potential. Nat Commun 10:517.
674 71. Rossy T, Nadell CD, Persat A. 2019. Cellular advective-diffusion drives the emergence of
675 bacterial surface colonization patterns and heterogeneity. Nat Commun 10:2471.
676 72. Michael V, Frank O, Bartling P, Scheuner C, Goker M, Brinkmann H, Petersen J. 2016.
677 Biofilm plasmids with a rhamnose operon are widely distributed determinants of the
678 'swim-or-stick' lifestyle in roseobacters. ISME J 10:2498-513.
679 73. Stewart PS, Costerton JW. 2001. Antibiotic resistance of bacteria in biofilms. Lancet
680 358:135-8.
681 74. Simmons EL, Bond MC, Koskella B, Drescher K, Bucci V, Nadell CD. 2020. Biofilm
682 structure promotes coexistence of phage-resistant and phage-susceptible bacteria.
683 mSystems 5.
684 75. Vidakovic L, Singh PK, Hartmann R, Nadell CD, Drescher K. 2018. Dynamic biofilm
685 architecture confers individual and collective mechanisms of viral protection. Nat
686 Microbiol 3:26-31.
687 76. Westbye AB, Beatty JT, Lang AS. 2017. Guaranteeing a captive audience: coordinated
688 regulation of gene transfer agent (GTA) production and recipient capability by cellular
689 regulators. Curr Opin Microbiol 38:122-129.
690 77. Vulic M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: DNA
691 polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci
692 U S A 94:9763-7.
693 78. Vos M. 2009. Why do bacteria engage in homologous recombination? Trends Microbiol
694 17:226-32.
695 79. Smith P, Schuster M. 2019. Public goods and cheating in microbes. Curr Biol 29:R442-
696 R447.
697 80. Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial
698 genomes. Trends Genet 17:589-96.
699 81. Redfield RJ, Soucy SM. 2018. Evolution of bacterial gene transfer agents. Front
700 Microbiol 9:2527.
701 82. Wilson DS. 1975. A theory of group selection. Proc Natl Acad Sci U S A 72:143-6.
702 83. Lehtonen J. 2016. Multilevel selection in kin selection language. Trends Ecol Evol
703 31:752-762.
704 84. Price GR. 1970. Selection and covariance. Nature 227:520-1.
705 85. Bayles KW. 2007. The biological role of death and lysis in biofilm development. Nat Rev
706 Microbiol 5:721-6.

86. Devaraj A, Buzzo JR, Mashburn-Warren L, Gloag ES, Novotny LA, Stoodley P, Bakaletz LO, Goodman SD. 2019. The extracellular DNA lattice of bacterial biofilms is structurally related to Holliday junction recombination intermediates. Proc Natl Acad Sci U S A 116:25068-25077.
87. Oliveira NM, Martinez-Garcia E, Xavier J, Durham WM, Kolter R, Kim W, Foster KR. 2015. Biofilm formation as a response to ecological competition. PLoS Biol 13:e1002191.
88. McBee ME, Chionh YH, Sharaf ML, Ho P, Cai MW, Dedon PC. 2017. Production of superoxide in bacteria is stress- and cell state-dependent: a gating-optimized flow cytometry method that Minimizes ROS measurement artifacts with fluorescent dyes. Front Microbiol 8:459.
89. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733-45.
90. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-402.
91. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114.
92. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996-1004.
93. Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033-4.
94. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-80.
95. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589.
96. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530-1534.
97. Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:153.
98. Derelle R, Philippe H, Colbourne JK. 2020. Broccoli: combining phylogenetic and network analyses for orthology assignment. Mol Biol Evol 37:3389-3396.
99. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59-60.

100. Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825-5829.
101. Price MN, Arkin AP. 2017. PaperBLAST: text mining papers for information about homologs. mSystems 2: e00039-17.
102. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265-D268.
103. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol 33:1635-8.
104. Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen XX, Rokas A. 2021. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics 37:2325–31.
105. Jaccard P. 1912. The distribution of the flora in the alpine zone. New phytologist 11:37-50.
106. Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal, complex systems 1695:1-9.
107. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428:726-731.
108. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2:100141.

Tables

Table 1. Twenty-seven gene families that co-evolve with reference GTA genes and discussed throughout the manuscript.

Gene name	Representative GenBank accession number	Functional Annotation[@]	Number of genomes that encode the gene
DUF2793	ADE83936.1	tail fiber protein[1]	225
comEC	ADE86092.1	competence protein	280
comF	ADE83962.1	competence protein F	281
dprA	ADE86822.1	DNA-protecting protein DprA	288
phrB	ADE86685.1	deoxyribodipyrimidine photo-lyase	228
mutY	ADE83991.1	A/G-specific adenine glycosylase	290
tatD	ADE85006.1	TatD-related deoxyribonuclease family protein	291
mazG	ADE85524.1	MazG family protein	291
ydiU	ADE85462.1	protein of unknown function UPF0061	243
hrpB	ADE86856.1	ATP-dependent RNA helicase HrpB	276
ccmA	ADE85530.1	heme exporter protein A	293
cycH	ADE86047.1	cytochrome c-type biogenesis protein CycH	226
ATP12	ADE84099.1	ATP12 chaperone protein family	293
pdxA[*]	ADE86420.1	4-hydroxythreonine-4-phosphate dehydrogenase	290
coaE[*]	ADE83834.1	dephospho-CoA kinase	292
hemD[*]	ADE87233.1	uroporphyrinogen-III synthase	286
ribD[*]	ADE86804.1	riboflavin biosynthesis protein RibD	221
folk[*#]	ADE87043.1	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	289
moeA[*#]	AAV95421.1	molybdenum cofactor biosynthesis protein A	270
moaC[*#$]	ADE86569.1	molybdenum cofactor biosynthesis protein C-2	280
moeB[$]	ADE84219.1	molybdenum cofactor biosynthesis protein B-1	282
mnmE	ADE83829.1	tRNA modification GTPase TrmE	289
tilS	QNR64972.1	tRNA lysidine(34) synthetase TilS	282
SUA5	ADE84169.1	Sua5/YciO/YrdC/YwlC family protein	292
dusA	ADE85522.1	tRNA-dihydrouridine synthase A	216
tadA	ADE86235.1	tRNA-specific adenosine deaminase	293
tadB	ADE84271.1	type II secretion system protein	207

[@]As provided in the GenBank records, except when a reference is provided
[*]Biosynthesis of cofactors (KEGG pathway ko01240)
[#]Folate biosynthesis (ko00790)
[$]Sulfur relay system (ko04122)
[1]Hynes AP et al. 2016. Functional and evolutionary characterization of a gene transfer agent's multilocus "genome". Mol Biol Evol 33:2530-2543.

1 **Table 2. Thirty-two gene families that co-evolve with reference GTA genes, but without**
2 **known connection to the GTA production cycle.**

Gene name	Representative GenBank accession No.	GenBank Functional Annotation	Number of genomes that encode the gene
prmC	QNR62505.1	peptide chain release factor N(5)-glutamine methyltransferase	292
bioC	ADE83961.1	conserved hypothetical protein	292
mhpC	ADE84522.1	hydrolase, alpha/beta fold family	289
miaA	ADE85369.1	tRNA delta(2)-isopentenylpyrophosphate transferase	291
-	ADE86839.1	protein-L-isoaspartate O-methyltransferase-2	292
phnP	ADE85005.1	metallo-beta-lactamase family protein	288
ispDF	ADE85540.1	bifunctional 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase/2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	282
glmU	ADE85246.1	bifunctional UDP-N-acetylglucosamine diphosphorylase/glucosamine-1-phosphate N-acetyltransferase	288
yfgZ	ADE86835.1	glycine cleavage T protein-2	292
yhiN	ADE84019.1	HI0933-like protein	193
-	ADE87182.1	conserved hypothetical protein	243
pcnB	ADE83954.1	CCA-adding enzyme	291
yfiH	ADE86534.1	protein of unknown function DUF152	293
-	ADE86535.1	protein of unknown function DUF185	293
alr	ADE85319.1	alanine racemase	251
rspA	ADE86886.1	mandelate racemase/muconate lactonizing enzyme family protein	208
ppiD	ADE86084.1	peptidyl-prolyl cis-trans isomerase D	291
aroE	ADE83833.1	shikimate 5-dehydrogenase	291
glnE	ADE86144.1	glutamate-ammonia-ligase adenylyltransferase	285
rns	ADE84242.1	ribonuclease T2 family protein	260
gluQ	ADE85705.1	glutamyl-Q tRNA(Asp) synthetase	289
MA20_39615	ADE85459.1	protein of unknown function DUF985	201
ptpA	ADE86851.1	protein-tyrosine-phosphatase	172
rne	ADE85895.1	ribonuclease E	289
ptr1	ADE86148.1	oxidoreductase, short-chain dehydrogenase/reductase family	259
queG	ADE86491.1	4Fe-4S ferredoxin, iron-sulfur cluster binding protein	286

-	ADE84017.1	NAD-dependent epimerase/dehydratase family protein	279
nnrD	ADE85417.1	YjeF-related protein family	262
pepA	ADE86425.1	leucyl aminopeptidase-2	291
-	ADE84176.1	peptidase, S58 family	191
pepN	ADE84605.1	aminopeptidase N	269
MA20_18095	ADE84243.1	alcohol dehydrogenase, zinc-binding domain protein	262

1

Figures

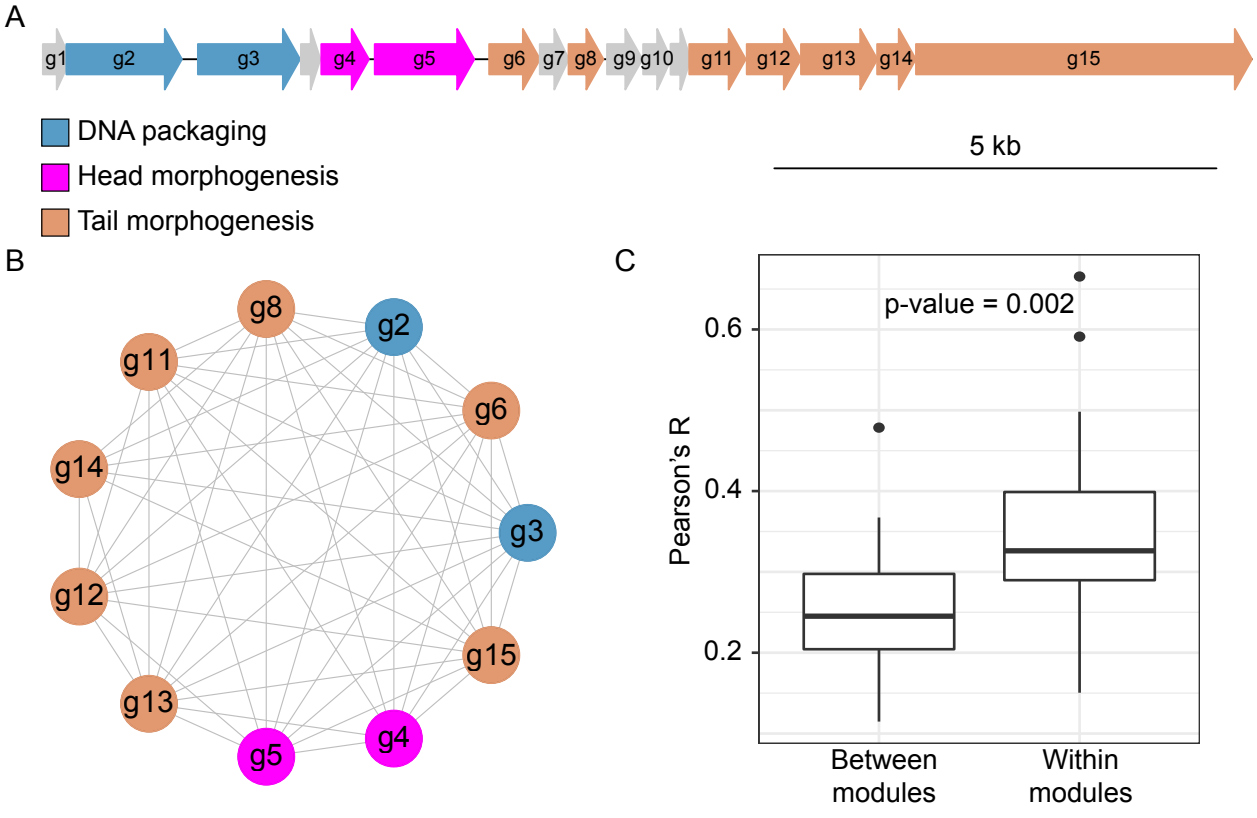

Figure 1. Co-evolution of 11 reference GTA genes from head-tail clusters. (A) The head-tail cluster of the RcGTA, as encoded in *R. capsulatus* strain SB 1003 genome (GenBank accession CP001312.1; spanning locus tags RCAP_rcc01682 - RCAP_rcc01698). Each gene is represented by an arrow and is drawn proportional to its length (see scale bar). Genes shown in non-gray color correspond to the 11 GTA reference genes used in the ERC analysis. Reference genes from each functional module are depicted in a distinct color. **(B)** Co-evolution network of the reference GTA genes. Nodes represent genes and are connected by edges if there is a significant evolutionary rate covariation (Pearson's R > 0 and p-value <0.05 after Bonferroni correction). **(C)** Comparison of the rate covariation strength in reference GTA genes that belong either to different functional modules ("Between Modules") or to the same functional module ("Within Modules). Boxplots represent median values that are bounded by first and third quartiles.

Whiskers illustrate data points that lie within 1.5 $*$ interquartile range. Significance was measured using Mann-Whitney U-test.

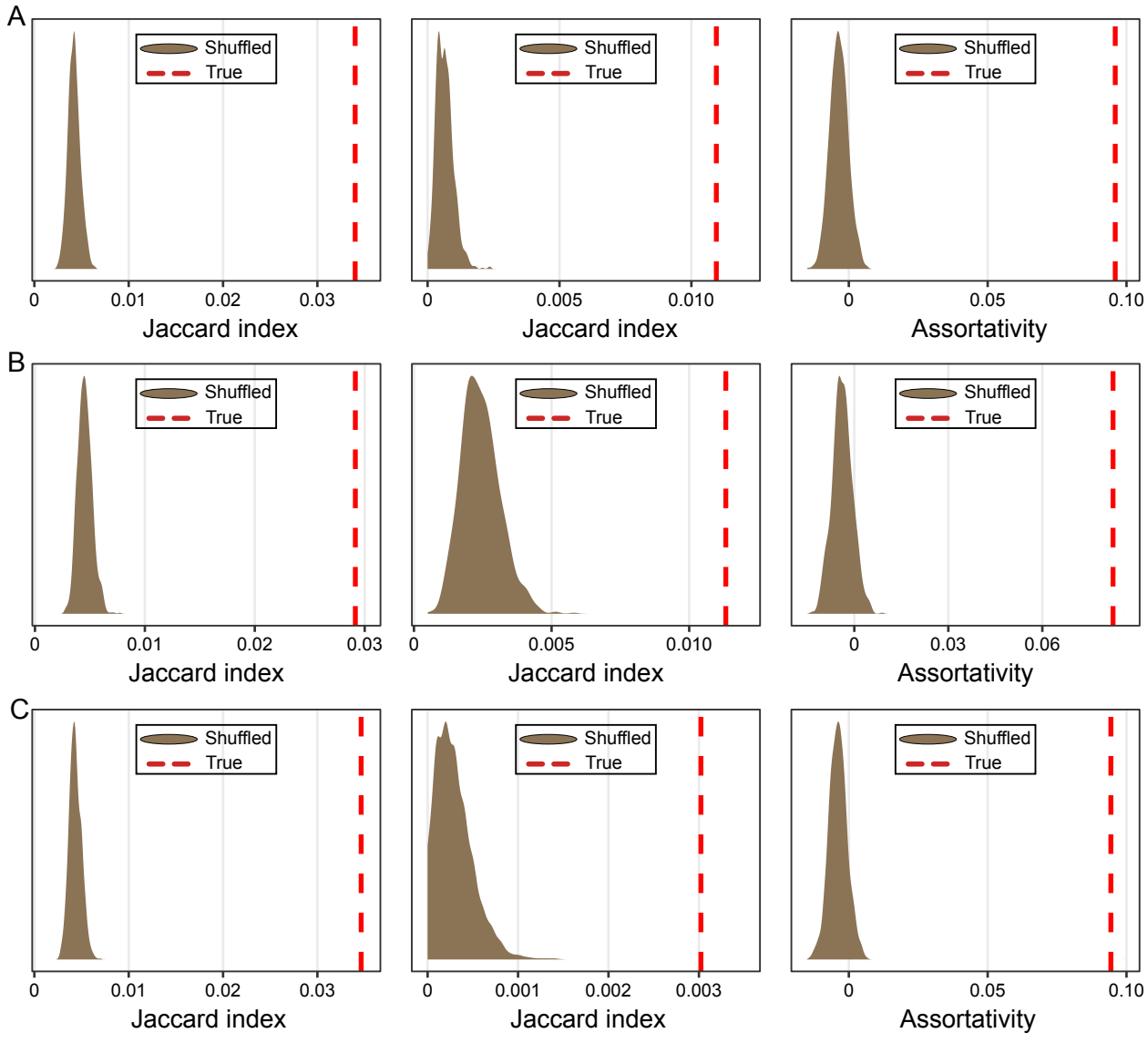

Figure 2. Strengths of correlations between the covariation evolutionary rate of genes and function of the proteins the genes encode, as measured by permutation tests. The analyses were carried out for **(A)** *Phaeobacter inhibens*, **(B)** *Caulobacter crescentus*, and **(C)** *Dinoroseobacter shibae*. Graphs on the left show the comparisons between the co-evolution and PPI networks, while graphs in the middle show the comparisons between co-evolution and co-fitness network. For these graphs, the distributions in brown represent distances between the PPI/co-fitness networks and 1,000 co-evolution networks, in which edges were randomly shuffled. The dashed red lines indicate the Jaccard index from the non-shuffled network

comparisons. Graphs on the right show positive assortativity between co-evolution and COG functional category assignment. The distributions in brown represent assortativity coefficient values for 1,000 networks, in which COG labels were randomly shuffled. The dashed red lines indicate the assortativity coefficient of the non-shuffled network.

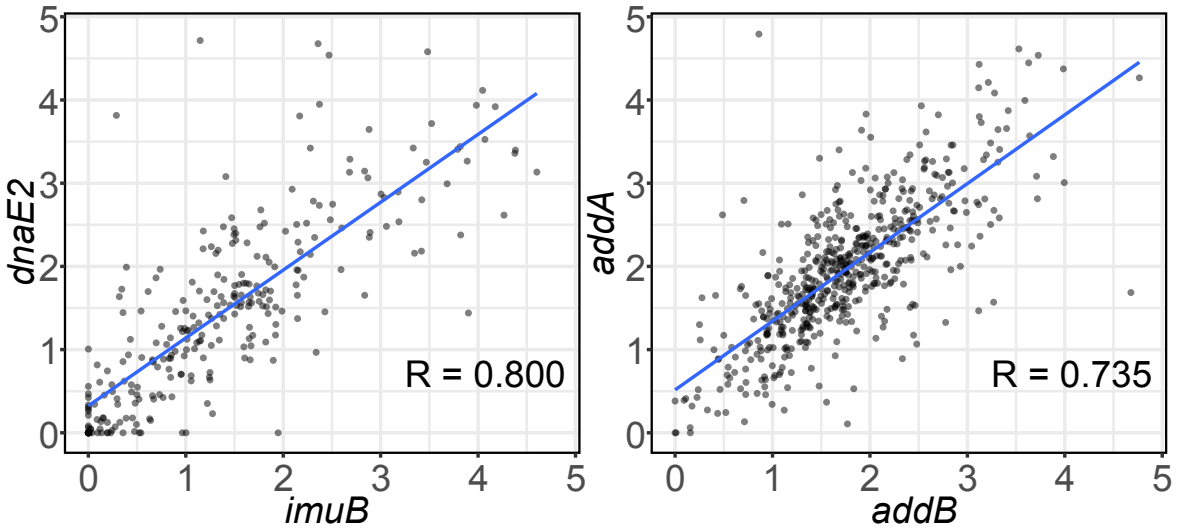

Figure 3. Covariation of evolutionary rates for two gene pairs in *P. inhibens* with the largest Pearson's coefficients. Each dot represents the normalized branch length of the proteins shown on x- and y-axes. Blue line depicts the linear regression line. The value of Pearson's R is shown on each panel.

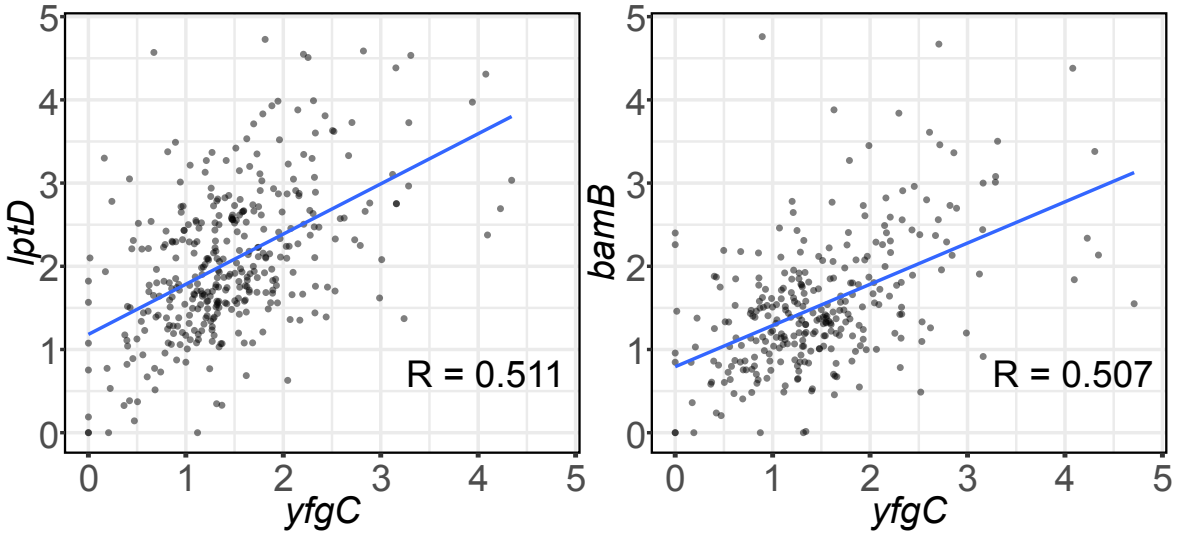

Figure 4. Covariation of evolutionary rates for two gene pairs in *P. inhibens* with no evidence of interactions for their protein products in the STRING database. Each dot represents the normalized branch length of the proteins shown on x- and y-axes. Blue line depicts the linear regression line. The value of Pearson's R is shown on each panel.

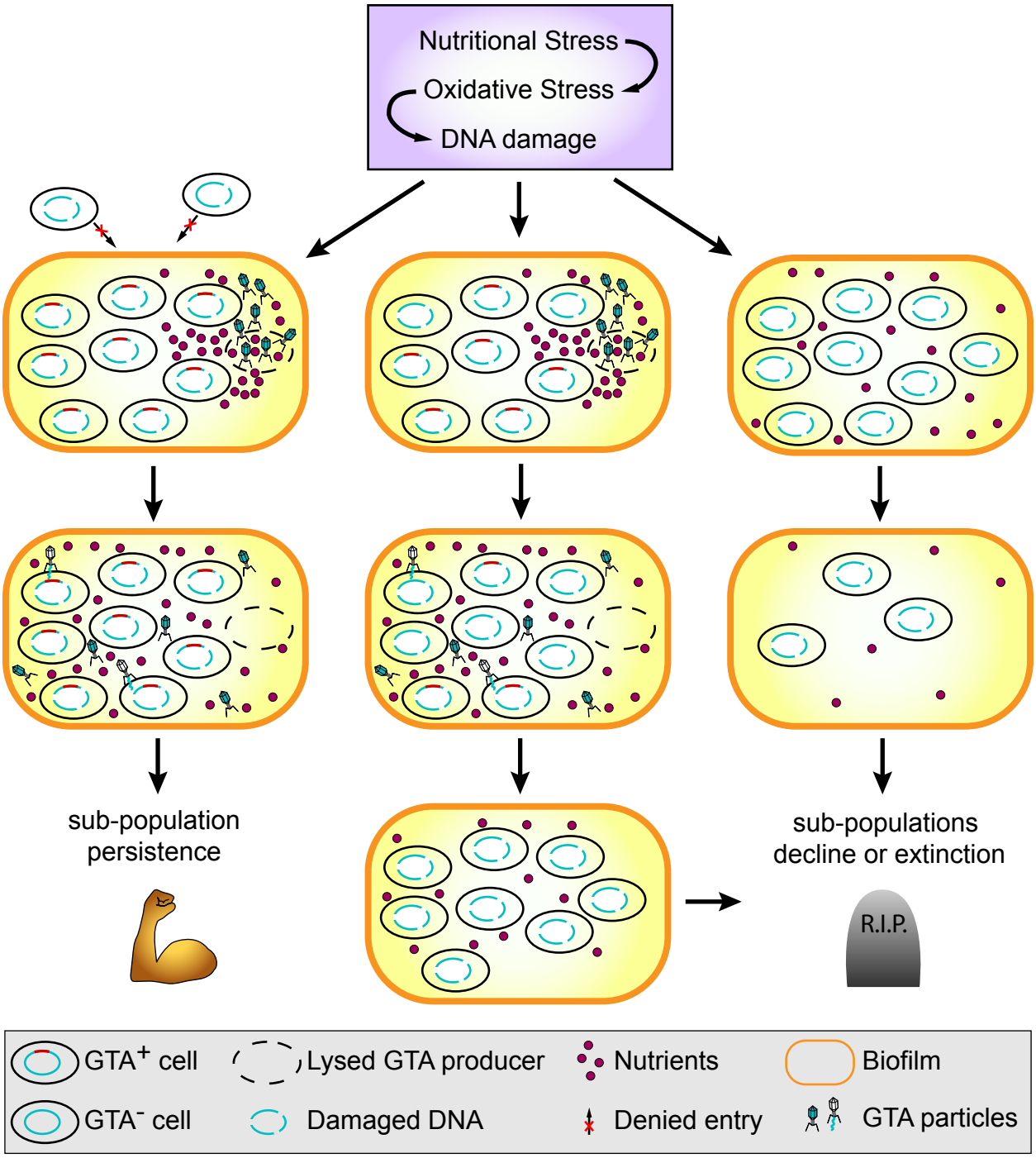

Figure 5. The proposed model of between-group selection that preserves the trait of GTA production in a bacterial population of closely related cells. See **Discussion** section for detailed description.