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 random 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.

Introduction

Multiple bacteria and archaea produce Gene Transfer Agents (GTAs) – the viriforms whose function and mode of selection to maintain them remain unsolved (Lang, et al. 2017; Kogay, et al. 2022; Kuhn and Koonin 2023). These domesticated virus-derived elements are encoded by genes in their host's genome and, when produced, resemble tailed double-stranded DNA (dsDNA) viruses (phages). In contrast to viruses, GTAs do not package the genes that encode them, and instead contain random fragments of the producing host's genome (Lang and Beatty 2001; Kogay, et al. 2022). Experimentally, GTAs are most studied and characterized in the alphaproteobacteria *Rhodobacter capsulatus* (RcGTA) and *Caulobacter crescentus* (Marrs 1974; Lang and Beatty 2000; Gozzi, et al. 2022), but they are also produced by several additional bacterial and archaeal species (Lang, et al. 2012). Many more prokaryotes encode GTA-like genes (Lang, et al. 2002; Lang and Beatty 2007; Québatte, et al. 2017; Shakya, et al. 2017; Tamarit, et al. 2018; George, et al. 2022; Fallon and Carroll 2023), and the presence of GTA-like genes in almost 60% of publicly available alphaproteobacterial genomes (Kogay, et al. 2019) suggests that GTA production is more widespread than currently appreciated.

RcGTA production is a population-level phenomenon: it is triggered by nutrient depletion (Westbye, O'Neill, et al. 2017) and is regulated by quorum sensing (Leung, et al. 2012). Only a small subset of the population acts as RcGTA producers; the remaining cells become recipients, by displaying specific polysaccharide receptors for RcGTAs adsorption (Brimacombe, et al. 2013) and expressing competence genes (Brimacombe, et al. 2015). Genetic pieces delivered by RcGTAs to a recipient cell can be integrated into the cell's genome via homologous recombination (Brimacombe, et al. 2014).

The benefits of GTA production and of acquiring GTA-packaged DNA in a microbial population are not fully understood. Since their discovery, GTAs were hypothesized to mediate DNA repair (Marrs, et al. 1977), and recently this hypothesis was confirmed by experimental demonstration of GTA-mediated DNA repair via homologous recombination in *C. crescentus* (Gozzi, et al. 2022). Moreover, facilitation of DNA damage repair appears to improve the survival of *C. crescentus* populations in nutrient limited conditions (Gozzi, et al. 2022), possibly due to a reduction in mutational load. Beyond the repair of already existing genes, released GTA particles could enable exchange of beneficial traits in a microbial population (Lang and Beatty 2000; McDaniel, et al. 2010) and provide nutrients to surrounding cells as the programmed cell death phenomenon does (Allocati, et al. 2015; Kogay, et al. 2020), although these hypotheses remain to be experimentally verified. Despite these putative population-level benefits, GTA-producing cells lyse and therefore leave no progeny, making it impossible for selection that maintains GTA production to act on the level of individual cells. Better understanding of GTA production and reception cycle and of genes underlying it will likely help us elucidate ecological role of GTAs in microbial communities, and details of the population-level selection that preserves the trait.

The RcGTA is encoded and regulated by at least 24 genes that are distributed across five different loci (Hynes, et al. 2016). Seventeen genes are located in one locus that is commonly referred as the head-tail cluster (Lang, et al. 2017) (**Figure 1A**). The locus encodes the majority of structural proteins required for the RcGTA particle assembly (Bardy, et al. 2020). Products of many additional "host" genes are critical for the regulation of the RcGTA particle production, DNA uptake, and DNA integration. For example, the CckA-ChpT-CtrA phosphorelay system, which controls the cell cycle and DNA replication, modulates production of RcGTA particles and

their release (Mercer, et al. 2012; Farrera-Calderon, et al. 2021). Serine acetyltransferase (cysE1), which is required for biofilm formation, plays a critical role for the optimal receipt of RcGTAs (Sherlock and Fogg 2022). Capsular polysaccharides, which serve as RcGTA receptors, are synthesized under control of GtaR/I quorum-sensing genes (Brimacombe, et al. 2013). Competence machinery proteins ComEC, ComF and ComM facilitate entry of DNA into cells (Brimacombe, et al. 2015). Integration of the incoming genetic material into the host genome via homologous recombination is facilitated by DprA and RecA (Brimacombe, et al. 2014). It is likely that products of multiple other "host" gene families important for the proper functioning of GTAs remain to be discovered, and in this study we use a comparative genomics approach to search for such genes.

Genes that are involved in the similar molecular processes, or co-expressed together, tend to co-evolve with each other (Clark, et al. 2012; Steenwyk, et al. 2022), and, vice versa, the protein-protein interactions can be unveiled by finding co-evolving genes that encode the interacting proteins (Kim, et al. 2004; Brunette, et al. 2019). The co-evolution among genes can be effectively identified via Evolutionary Rate Covariation (ERC) approach (Goh, et al. 2000; Sato, et al. 2005; Clark, et al. 2012). Evolutionary rate covariation measures the degree of correlation of changes in evolutionary rates across the phylogenies of a pair of proteins, assuming that functionally related proteins have similar selection pressures, resulting in coordinated changes in substitution rates (Clark, et al. 2012; Steenwyk, et al. 2022). Because GTA head-tail cluster has resided within GTA-containing alphaproteobacterial genomes for at least 700 million years (Shakya, et al. 2017) and, as mentioned above, GTA production is tightly integrated into the molecular circuits of the GTA-carrying bacteria, evolutionary rates of gene families involved in GTA lifecycle are expected to correlate with the rates of the GTA genes.

Head-tail cluster genes are easily detectable across genomes in a large clade of alphaproteobacteria (Lang and Beatty 2007; Shakya, et al. 2017) and therefore provide a rich dataset for comparative analyses of evolutionary rates. In this study, we examined the evolutionary rate covariation patterns of protein-coding genes encoded in 293 representative alphaproteobacterial genomes that contain either complete or nearly complete GTA head-tail clusters. We found that GTA head-tail cluster genes co-evolve with 59 gene families, 55 of which have not been previously linked to GTAs. Thus, we dramatically expand the list of genes that could be important for GTAs' functionality and hence could provide insights into GTAs' role in bacterial populations. By combining our findings with the existing knowledge about GTAs, we propose a model that explains persistence of GTA production in bacterial populations.

Results

Alphaproteobacterial GTA genes co-evolve with each other

The ERC method was developed for and tested on eukaryotic genes (Clark, et al. 2012; Steenwyk, et al. 2022) and, to our knowledge, has not been applied to bacterial genomes. Therefore, before using the approach to identify genes co-evolving with GTA genes, we evaluated it on GTA genes found in the 293 alphaproteobacterial genomes. Because genes in the GTA head-tail cluster have a common promoter (Lang and Beatty 2000; Fogg 2019) and the gene products are functionally related (i.e., produce a GTA particle that has DNA packaged into its head), we expected strong co-evolution among GTA genes. Indeed, we found that 51 out of 55 possible pairs of the 11 reference GTA genes (see **Methods** for definition) have significantly similar co-variation of evolutionary rates, as measured by the Pearson's coefficient (p-value < 0.05 after Bonferroni correction). Each reference GTA gene co-evolves with at least 7 other

reference GTA genes, with 6 of them co-evolving with all 10 other reference GTA genes (**Figure 1B**). These findings suggest that the ERC method adequately identifies co-evolving genes in alphaproteobacterial genomes.

Co-evolving alphaproteobacterial genes encode functionally related proteins

Co-evolving genes of eukaryotes identified through ERC analyses were shown to be either functionally related or involved in similar biological processes (Clark, et al. 2012; Steenwyk, et al. 2022). To examine how robustly the ERC method can identify functionally related gene pairs in our dataset of alphaproteobacterial genomes, we (i) evaluated rates of covariation within functional modules of the GTA head-tail cluster and (ii) examined a relationship between co-evolution and literature- and experiment-based functional inferences for a subset of gene families nearly universally found across GTA-containing alphaproteobacteria.

GTA head-tail cluster encodes three modules that are responsible for distinct functional stages of GTA production: DNA packaging, head morphogenesis, and tail morphogenesis (Hynes, et al. 2016; Lang, et al. 2017; Bardy, et al. 2020) (**Figure 1A**). Phage genes within the same functional class are more likely to interact with one another (Rajagopala, et al. 2011). We found that the Pearson's correlation coefficient is significantly higher (and, therefore, co-evolutionary signal is significantly stronger) for reference GTA genes within each module than between the reference GTA genes from different modules (Mann-Whitney U-test, p-value = 0.002) (**Figure 1C**). These findings suggest that the strength of co-evolutionary signal measured by the ERC analysis correlates with the degree of physical and functional interactions among GTA genes.

Expanding our analysis beyond GTA genes, we examined protein-coding genes in a model marine bacterium *Phaeobacter inhibens*, which we chose as the representative GTA-containing alphaproteobacterium for three reasons: first, its genome encodes the largest number (1,370) of genes from 1,470 gene families nearly universally found across GTA-containing alphaproteobacteria; second, the information about the interactions of *P. inhibens*' proteins encoded by 1,320 out of the 1,370 genes is available in the STRING database (Szklarczyk, et al. 2021); and third, experimental relative fitnesses of *P. inhibens*'genes are catalogued in Fitness Browser (Price, et al. 2018). Using the ERC analysis on the 1,320 genes, we identified 10,514 co-evolving gene pairs (**Figure S1**). The co-evolution network, in which nodes correspond to genes and edges between them designate the presence of significant evolutionary rate covariation (**Figure S2**), is significantly similar to both the network of the pairwise interactions of the encoded proteins and the network of fitness effects (p-value < 0.001, permutation test) (**Figures 2A** and **2B**). Moreover, co-evolving genes are more likely to belong to the same Clusters of Orthologous Groups (COG) category (assortativity = 0.096; p-value < 0.001, permutation test) (**Figure 2C**). These comparisons show that co-evolving genes that encode well-characterized proteins (defined as present in the STRING and COG databases) and a subset of genes needed for specific environmental conditions (as determined by the Fitness Browser database) indeed tend to encode functionally related proteins. For example, gene pairs with the two largest Pearson's coefficients encode proteins that are involved in the same biological processes (**Figure 3**): *imuB* and *dnaE2* genes (Pearson's coefficient r = 0.80) are located in the same operon and are involved in SOS-induced mutagenesis and translesion synthesis (Abella, et al. 2004; Erill, et al. 2006), while *addA* and *addB* genes (Pearson's r = 0.74) are known to assemble into the

heterodimeric complex to facilitate homologous recombination and DNA repair (Kooistra, et al. 1993; Saikrishnan, et al. 2012).

It is worth noting that some of the genes identified as co-evolving in our analysis are not designated as encoding interacting proteins in the STRING database. However, given incompleteness of our knowledge about functionality of proteins encoded in a bacterial genome, these genes may represent functional connections yet unidentified in STRING. Indeed, ERC analysis has been used to uncover novel protein-protein interactions, especially between hypothetical proteins (Brunette, et al. 2019; Raza, et al. 2019; Forsythe, et al. 2021). Here are two examples of proteins identified as co-evolving in our analyses and likely interacting based on what's known about their functions, but not designated as such in the STRING database. The *yfgC* gene, which encodes a periplasmic metalloprotease involved in assembly of outer membrane proteins (Narita, et al. 2013), co-evolves with both the *lptD* (Pearson's r = 0.51) and *bamB* (Pearson's r = 0.51) genes (**Figure 4**). The YfgC protein plays a crucial role in the assembly of LptD, an outer membrane protein that participates in the lipopolysaccharide assembly (Narita, et al. 2013). The YfgC also interacts with the β-barrel-assembly machinery (BAM) complex, which consists of four lipoproteins, including BamB, and facilitates the assembly and integration of proteins into the outer membrane (Narita, et al. 2013; Han, et al. 2016).

GTA genes co-evolve with at least fifty-nine other gene families

By analyzing 1,470 gene families almost universally present among the 293 representative GTA-containing alphaproteobacteria, we identified 59 gene families that co-evolve with at least 5 reference GTA genes (**Table S1**) (see **Methods** for selection criteria).

Notably, four of the 59 gene families - encoding tail fiber protein (DUF2793), competence proteins (comEC and comF), and DNA-protecting protein that facilitates homologous recombination (dprA) (**Table 1**) - have already been shown to play important roles in the RcGTA lifecycle (Brimacombe, et al. 2014; Brimacombe, et al. 2015; Hynes, et al. 2016). Because tail fiber proteins are a part of the RcGTA particle and thus physically interact with other RcGTA proteins (Hynes, et al. 2016), DUF2793's co-evolution with other structural GTA genes is expected. The protein products of the remaining three genes (ComEC, ComF, and DprA) are required for the acquisition of DNA delivered by RcGTAs, and interact physically only with DNA molecules (Brimacombe, et al. 2014; Brimacombe, et al. 2015). This finding demonstrates that the ERC approach indeed could predict genes functionally linked to the GTA lifecycle.

The remaining 55 gene families are involved in various functions (**Tables 1, 2** and **S1**). While 33 of the 55 gene families (**Table 2**) offer exciting opportunities for future research into GTA lifecycle, 22 gene families (**Table 1**) can be either directly linked to the GTA lifecycle by being involved in DNA repair or are likely to be under similar selection pressure as GTA genes due to shared ecological importance (stress response, biofilm formation, oxidative respiration, and cofactor biosynthesis), as elaborated below.

Three of the 22 genes – *mutY* (encoding a glycosylase), *phrB* (encoding a DNA photolyase), and *tatD* (encoding an exonuclease) – play roles in repair of DNA damage induced by various oxidative agents and UV light (Krokan, et al. 1997; Kim and Sundin 2001; Chen, et al. 2014). Glycosylase actively modulates homologous recombination (Spek, et al. 2002), which could be important for facilitating integration of GTA-derived genetic material into the recipient's genome (Marrs 1974; Gozzi, et al. 2022). Photolyase and tatD exonuclease do not directly participate in homologous recombination, but are involved in DNA repair process (Kim

and Sundin 2001; Chen, et al. 2014; Gozzi, et al. 2022) and thus are likely to be under the similar selection pressure as GTA genes.

Two genes (*mazG* and *ydiU*) identified in our screen are involved in stress response. The product of the *mazG* gene modulates the programmed cell death in *Escherichia coli* and regulates intracellular level of ppGpp, the universal 'alarmone', which was previously implicated in the regulation of GTAs production (Gross, et al. 2006; Westbye, O'Neill, et al. 2017). The ppGpp molecule is involved in a cellular response to a variety of stress conditions, including nutritional stress. The product of the *ydiU* gene mediates UMPylation of bacterial chaperones, improving bacterial fitness under the stressful environmental settings (Yang, et al. 2020).

Twelve genes identified in our analyses have relevance to biofilms. The *hrpB* gene, which encodes ATP-dependent RNA helicase, has been shown to be important for biofilm formation and adhesion on surfaces (Granato, et al. 2016). The *ccmA* gene, *cycH* gene, and a gene from COG5387 family (*ATP12*) are involved in oxidative respiration, which promotes bacterial survival in the biofilms (Schinner, et al. 2020; Martin-Rodriguez 2022). Moreover, as a group, 59 co-evolving gene families are enriched in three metabolic pathways relevant to biofilms: cofactor biosynthesis, folate biosynthesis and sulfur relay system (hypergeometric test, p-value < 0.05, Bonferroni correction) (**Table 1**). Two genes from the 'cofactor biosynthesis' pathway (*moeA* and *moaC*) are involved in the molybdenum cofactor biosynthesis. Both metabolism of folate and molybdenum cofactors are important for biofilm formation (Andreae, et al. 2014; Wong, et al. 2018). The sulfur relay pathway is involved in the tRNA modifications (Leimkuhler, et al. 2017), which are implicated in fitness of bacteria within a biofilm (Schinner, et al. 2020). Notably, five additional genes that encode tRNA modification enzymes (*mnmE*, *tilS*, *SUA5*, *dusA*, *tadA*) are inferred to co-evolve with GTA genes (**Table 1**).

Discussion

By showing that GTA head-tail cluster genes, and especially genes within the same functional module of the cluster, tend to co-evolve with each other, and by examining function-coevolution relationship among proteins encoded in a model marine bacterium, we demonstrated that the ERC method is an effective approach to uncover functional relationships among protein-coding genes in bacteria, extending the method's applicability beyond eukaryotes. Applying the method to GTA-encoding alphaproteobacterial genomes that span >700 million years of evolution (Shakya, et al. 2017), we detected a significant evolutionary rate covariation of GTA head-tail cluster genes with 59 protein-coding genes. Multiple genes in this dataset are involved in stress response, DNA repair, homologous recombination, and biofilm formation. These functions are consistent with the accumulating experimental and computational evidence about GTA production, regulation and function in *R. capsulatus* and *C. crescentus*, and with previous hypotheses and models of GTA production triggered under environmental stress (Westbye, O'Neill, et al. 2017; Gozzi, et al. 2022), GTAs being involved in DNA repair in recipient cell (Gozzi, et al. 2022; Kogay and Zhaxybayeva 2022), and, most recently, of GTA production occurring in biofilms (Sherlock and Fogg 2022). Our discovery that 12 biofilm-implicated genes co-evolve with GTA genes further highlight the potential importance of biofilm settings for GTA production.

Alphaproteobacteria in general, and GTA-producing *R. capsulatus* and *C. crescentus* in particular, are known to form biofilms (Rossy, et al. 2019; Zhang, et al. 2019; Sherlock and Fogg 2022). Biofilms provide a microbial community with benefits that cannot be achieved by the individual cells, such as protection against antibiotics (Stewart and Costerton 2001) and viral infections (Vidakovic, et al. 2018; Simmons, et al. 2020). Despite either shown or hypothesized

benefits of GTA-disseminated DNA to the recipient cells (Lang, et al. 2017; Westbye, Beatty, et al. 2017; Gozzi, et al. 2022), zero relative fitness of the lysed GTA-producing cells implies that the trait of encoding GTAs can only be favored by selection when the benefits of receiving GTAs are confined to either clonal or very closely related cells that also possess the genes encoding the suicidal GTA production trait. Encoding genes for synthesis of specific polysaccharide receptors, which facilitate the adsorption of GTAs (Brimacombe, et al. 2013) and co-regulating the receptor production in one fraction of a population with GTA production in another fraction of a population, can limit GTAs' targets only to clonal cells or very closely related species. Additionally, success of homologous recombination declines exponentially with the increase in genomic sequence divergence (Vulic, et al. 1997; Vos 2009), further restricting the usefulness of GTA particles for DNA repair to closely related cells.

Within such groups of closely related cells, GTAs can be viewed as "public goods". However, any public goods system faces an inevitable rise of cheaters (Smith and Schuster 2019), and a population of GTA producers would be susceptible to cheaters that would not produce GTA particles but still have surface polysaccharides that serve as GTA receptors. Consistent with these conjectures, we observe both the pseudogenization and complete loss of GTA systems in multiple alphaproteobacterial lineages (Lang and Beatty 2007; Shakya, et al. 2017; Kogay, et al. 2019). Pseudogenized GTA gene clusters may represent recently emerged cheater lineages, while the absence of GTA gene clusters could be a result of cheater takeover in a species and consequent loss of GTA genes due to the deletion bias (Mira, et al. 2001). (It should be noted that, in both cases, the loss of GTA production in these lineages could also be attributed to acquisition of alternative molecular mechanisms to cope with nutritional stress and DNA repair, or inhabiting niches where GTA production costs outweigh its benefits.)

Mathematical modeling showed that maintenance of GTA production trait can be difficult in mixed populations of GTA producers and cheaters, at least under some conditions (Redfield and Soucy 2018). Yet, despite the likely appearance of cheaters and an observation of recurrent GTA loss in multiple lineages, the GTA production trait has been persisting in alphaproteobacteria for hundreds of millions of years (Shakya, et al. 2017), suggesting that some kind of population-level selection is successful. However, details of how such selection operates remain unknown.

One possible solution for an "altruistic" trait to persist in a population over time is to have the population segregated into small sub-populations, an evolutionary scenario first modeled by Wilson (Wilson 1975) and subsequently shown to be equivalent to multi-level selection models that emphasize close relatedness of sub-population members (Price 1970; Lehtonen 2016). Under this model, the cheaters arise stochastically and therefore are found in many but not all sub-populations. Notably, sub-populations without cheaters outcompete sub-populations with cheaters by having overall higher total productivity due to benefits of the public goods produced by the altruistic trait. We hypothesize that GTA systems persist over time because GTA production occurs in such spatially structured populations. We further hypothesize that *it is biofilms* that facilitate the division of GTA-producing populations into isolated sub-populations. A biofilm would ensure that clonal or closely related cells are in spatial proximity, plus they would protect the GTA-producing sub-population from being invaded by other cells, including cheaters. Dense packing of cells would also ensure that the released GTA particles are not dispersed in the environment and reach their recipients (Sherlock and Fogg 2022). Biofilms would also trap the organic debris of the lysed GTA-producer cells within the biofilm (Bayles 2007). DNA of these lysed cells could be used as a part of a biofilm matrix (Bayles 2007), enhancing the sub-population isolation and protection from environmental hazards (Devaraj, et

al. 2019). Other environmental and cellular debris could increase localized nutrient availability. Previous experimental work on biofilms supports some of our conjectures. In the GTA-producing *Caulobacter crescentus* the biofilm structure promotes clonal cells to reside in proximity (Rossy, et al. 2019). Biofilm formation is also increased during the ecological competition, providing microbes with the protection to resist invasion by different strains (Oliveira, et al. 2015).

Building on the current knowledge and previous models of GTA evolution and function, we propose the following model of selection acting on a structured, biofilm-forming bacterial population and maintaining GTA production (**Figure 5**). When a population experiences starvation, nutritional stress increases the generation of reactive oxygen species, which induces DNA damage (McBee, et al. 2017). In a structured population, the fate of each sub-population (i.e., of individual biofilms) depends on the genetic make-up of its cells. In a sub-population of GTA producers (GTA^+ cells), a small fraction of cells is "sacrificed", and their DNA is delivered to the remaining cells or hoarded as part of the biofilm matrix, while the cellular debris are utilized as additional nutrients. The resources are protected from invaders by the biofilm structure. As a result, the sub-population counteracts the negative effects of nutrient scarcity and DNA damage, and thus experiences, at worst, only a small population decline due to the lysed GTA^+ cells. However, if cheaters arise within the biofilm boundaries, the composition of such sub-population under multiple episodes of nutritional stress will change to a higher fraction of GTA non-producers (GTA^- cells), as these types of cells will not be lost due to lysis during GTA production and yet will experience all benefits of the public goods released by GTA^+ cells. Eventually, the sub-population will lose GTA production trait. In a sub-population of GTA^- cells, mutational load due to DNA damage and limited nutrients will result in a decline of the sub-population size. Therefore, GTA^+ sub-populations will have higher relative fitness in comparison

to GTA$^-$ sub-populations, resulting in the larger overall number of GTA$^+$ cells in the combined population. Thus, the GTA production trait will be maintained in the population as a whole.

Although bits of experimental evidence used in the above model come mostly from the research on GTAs in *Rhodobacter capsulatus* and *Caulobacter crescentus*, we hypothesize that our model is applicable to other GTA-containing alphaproteobacterial species, because the detected co-evolution patterns span multiple diverse alphaproteobacterial clades.

Materials and Methods

Identification of 293 representative alphaproteobacterial genomes with GTA regions

Initially, 1,642 alphaproteobacterial genomes and annotations of their protein-coding genes were retrieved from the NCBI's Assembly and RefSeq database (accessed June 2022) (O'Leary, et al. 2016) (**Table S2**). In these genomes, GTA regions were predicted using GTA-Hunter program with default parameters (Kogay, et al. 2019). Because GTA-Hunter looks only for 11 out of the 17 genes of the RcGTA's head-tail cluster, the remaining GTA genes were identified via BLASTP searches (E-value < 0.1) (Altschul, et al. 1997), using as queries the curated set of GTA regions from Kogay, et al. (2019). Only BLASTP matches that are located within the GTA-Hunter-predicted GTA regions were added. Using this procedure, GTA regions were identified in 701 genomes.

To avoid presence of multiple highly similar GTA regions in downstream analyses, the 701 genomes were clustered into 392 Operational Taxonomic Units (OTUs) using the Average Nucleotide Identity (ANI) cutoff of 95%, calculated via fastANI v1.1 (Jain, et al. 2018). Within each OTU, GTA regions were examined for "completeness", defined as having 14 out of the 17 head-tail cluster genes (genes *g1*, *g3.5* and *g7* were excluded because they are not easily detected

across GTA-contaning alphaproteobacterial clades (Shakya, et al. 2017)). Incomplete GTA regions were discarded. This criterion reduced the number of OTUs to 293. Within each of the 293 OTUs, only one, randomly selected, genome and its GTA region were retained for subsequent analyses (**Table S3**).

Reconstruction of the reference phylogenomic tree

From the set of 120 marker genes widely used for phylogenomic taxonomy (Parks, et al. 2018), 84 gene families were detected in a single copy in at least 95% of the 293 genomes using AMPHORA2 (Wu and Scott 2012). Amino acid sequences of each of these 84 gene families were aligned using MAFFT v7.505 with the 'linsi' option (Katoh and Standley 2013). The alignments were concatenated, and the best substitution model for each alignment and the optimal partition scheme were established via ModelFinder (Kalyaanamoorthy, et al. 2017). The maximum-likelihood phylogeny was reconstructed using IQ-TREE v2.2 (Minh, et al. 2020). The tree was rooted using the *Emcibacterales* and *Sphingomonadales* taxonomic orders, using the previously observed branching order of the *Alphaproteobacteria* as a guide (Shakya, et al. 2017; Kogay, et al. 2019).

Selection of Reference GTA genes

The amino acid sequences of the 14 GTA genes from the GTA regions of the 293 genomes were retrieved and aligned using MAFFT v7.505 with the 'linsi' option. Phylogenetic trees were reconstructed from each alignment using IQ-TREE v2.2 (Minh, et al. 2020) under the best substitution model identified by ModelFinder (Kalyaanamoorthy, et al. 2017). Each tree was compared to the reference tree using the normalized quartet scores calculated in ASTRAL v5.7.8

(Zhang, et al. 2018). Eleven GTA genes that exhibited a high congruency with the reference phylogeny (quartet score > 0.8) (**Figure S3**) were designated as "reference GTA genes" and are referred as such throughout the manuscript.

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

Gene families were defined as orthologous groups identified in Broccoli v1.2 (Derelle, et al. 2020), using DIAMOND (Buchfink, et al. 2015) for protein similarity searches and the maximum-likelihood method for phylogenetic reconstructions. Gene families that are present in a single copy in at least 50% of the 293 GTA-containing genomes (1,470 in total) were retained for further analyses.

To assign COG functional annotations to the gene families, one randomly selected representative from each family was used as a query against the eggNOG database v5.0.2 and processed through eggNOG-mapper v2.1.9 workflow (Cantalapiedra, et al. 2021). For gene families found to be co-evolving with the GTA region (see below), additional annotations were sought out using PaperBLAST (accessed in December 2022) (Price and Arkin 2017) and CD-searches against CDD database v3.20 (accessed in December 2022) (Lu, et al. 2020).

Inference of evolutionary rate covariation

Amino acid sequences of each gene family and each reference GTA gene were aligned using MAFFT v.7.505 with the 'linsi' option (Katoh and Standley 2013) . For gene families that are not found in all 293 genomes, the absent taxa were pruned from the reference tree using functions from the *ete3* package (Huerta-Cepas, et al. 2016). For each gene set, the topology of taxa relationships was constrained to the reference phylogeny and branch lengths were estimated

via IQ-TREE v2.2 (Minh, et al. 2020), using the best substitution model suggested by ModelFinder (Kalyaanamoorthy, et al. 2017). The trees were rooted using relationships in the reference phylogeny as a guide.

Covariations of evolutionary rates among 1,470 gene families and 11 GTA reference genes were examined rates using the CovER pipeline, as implemented in PhyKIT v1.11.12 (Steenwyk, et al. 2021; Steenwyk, et al. 2022). Within the pipeline, the following steps were carried out. For each pair, their trees were pruned to retain only shared taxa. All trees were corrected for the differences in mutation rates and divergence times among taxa; this was accomplished by dividing the length of each branch by the length of the corresponding branch of the reference tree. Branches with the normalized length > 5 were removed from further analyses, and the retained branch lengths were Z-transformed. For every pair, Pearson correlation coefficient was calculated. A pair of genes was designated as co-evolving, if the Pearson correlation coefficient was positive and had p-value < 0.05 after Bonferroni correction for multiple testing.

The above-described co-variation analysis was carried out on three datasets: among 11 GTA reference genes (11 x 10/2 = 55 comparisons), between GTA genes and 1,470 gene families (11 x 1,470 = 16,170 comparisons), and among 1,320 gene families present in one representative GTA-containing genome, *Phaeobacter inhibens* (1,320 x 1,319/2 = 870,540 comparisons). The *P. inhibens* analysis resulted in 10,514 co-evolving gene pairs (**Figure S1**), and the information was assembled into a co-evolution network, in which nodes represent 1,320 gene families and edges depict 10,514 co-evolution relationships. To annotate nodes in the *Phaeobacter inhibens*' co-evolution network with functional categories, the sub-network of 1,040 of 1,320 gene families

with an unambiguous COG assignment was extracted. The COG functional annotations of genes were assigned as labels.

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

Reconstruction of protein-protein interaction network in *Phaeobacter inhibens*

Amino acid sequences of the 1,320 above-described *Phaeobacter inhibens* genes were used as queries against the STRING database v11.5 (Szklarczyk, et al. 2021), with the high confidence score cutoff and all available sources. This search resulted in 8,612 interacting protein-protein pairs. The information was assembled into a *Phaeobacter inhibens* protein-protein interactions (PPI) network, in which nodes represent 1,320 genes and edges depict 8,612 PPIs.

Reconstruction of co-fitness network in *Phaeobacter inhibens*

Phaeobacter inhibens' gene pairs with similar fitnesses across a wide range of different experimental conditions were retrieved from the Fitness Browser (Price, et al. 2018) (accessed August 2022). Gene pairs were designated as co-fit if they either had a co-fitness value > 0.75, or they had a co-fitness value > 0.60 and were conserved in other bacterial species (Price, et al. 2018). This search resulted in 569 co-fit gene pairs. The information was assembled into a

Phaeobacter inhibens co-fitness network, in which nodes represent 1,320 genes and edges depict 569 co-fitness associations.

Comparison of networks and identification of subnetworks

The similarity between the *Phaeobacter inhibens* co-evolution network with either PPI or co-fitness networks was assessed by calculating Jaccard index (Jaccard 1912). The null distribution of Jaccard indices was created by randomly re-shuffling of the evolutionary rate covariation network 1,000 times.

Tendency of nodes to connect the nodes from the same COG category was measured by an assortativity coefficient, which was calculated for the *Phaeobacter inhibens* co-evolution network of 1,147 genes with an unambiguous COG assignment using *igraph* v1.3.5 (Csardi and Nepusz 2006). The permutation test for assortativity was performed by random shuffling of COG labels 1,000 times.

All networks were visualized using *igraph* v1.3.5 (Csardi and Nepusz 2006).

KEGG pathways enrichment analysis for 59 genes that co-evolve with GTAs

Each of the 59 gene families was assigned a KEGG Orhology (KO) label using BlastKOALA (Kanehisa, et al. 2016). Significantly enriched pathways were identified by the hypergeometric test (p-value < 0.05 with Bonferroni correction for multiple testing), as implemented in the *clusterProfiler* package v4.4.4 (Wu, et al. 2021).

Data Availability

The genomes used in this study are publicly available via NCBI Assembly (https://www.ncbi.nlm.nih.gov/assembly) database. The accession numbers of these genomes are listed in **Table S2**. The following datasets, which were derived from the genomes, are available in the FigShare repository (DOI 10.6084/m9.figshare.23929551): multiple sequence alignment of GTA genes; unconstrained and constrained phylogenetic trees of GTA genes; concatenated alignment of phylogenomic markers and reconstructed reference phylogenomic tree; amino acid sequences of 1,470 gene families; constrained phylogenetic trees of 1,470 gene families; covariation of evolutionary rates for all performed pairwise gene comparisons; *Phaeobacter inhibens*' co-evolutionary network and its subnetwork of nodes with unambiguous GOG assignment, protein-protein interaction network, fitness network, and a list of COG functional category assignments for nodes.

Acknowledgements

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

References

Abella M, Erill I, Jara M, Mazon G, Campoy S, Barbe J. 2004. Widespread distribution of a lexA-regulated DNA damage-inducible multiple gene cassette in the Proteobacteria phylum. Mol Microbiol 54:212-222.

Allocati N, Masulli M, Di Ilio C, De Laurenzi V. 2015. Die for the community: an overview of programmed cell death in bacteria. Cell Death Dis 6:e1609.

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 25:3389-3402.

Andreae CA, Titball RW, Butler CS. 2014. Influence of the molybdenum cofactor biosynthesis on anaerobic respiration, biofilm formation and motility in Burkholderia thailandensis. Res Microbiol 165:41-49.

Bardy P, Fuzik T, Hrebik D, Pantucek R, Thomas Beatty J, Plevka P. 2020. Structure and mechanism of DNA delivery of a gene transfer agent. Nat Commun 11:3034.

Bayles KW. 2007. The biological role of death and lysis in biofilm development. Nat Rev Microbiol 5:721-726.

Brimacombe CA, Ding H, Beatty JT. 2014. *Rhodobacter capsulatus* DprA is essential for RecA-mediated gene transfer agent (RcGTA) recipient capability regulated by quorum-sensing and the CtrA response regulator. Mol Microbiol 92:1260-1278.

Brimacombe CA, Ding H, Johnson JA, Beatty JT. 2015. Homologues of genetic transformation DNA import genes are required for *Rhodobacter capsulatus* gene transfer agent recipient capability regulated by the response regulator CtrA. J Bacteriol 197:2653-2663.

Brimacombe CA, Stevens A, Jun D, Mercer R, Lang AS, Beatty JT. 2013. Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). Mol Microbiol 87:802-817.

Brunette GJ, Jamalruddin MA, Baldock RA, Clark NL, Bernstein KA. 2019. Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes. Proc Natl Acad Sci U S A 116:19593-19599.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59-60.

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.

Chen YC, Li CL, Hsiao YY, Duh Y, Yuan HS. 2014. Structure and function of TatD exonuclease in DNA repair. Nucleic Acids Res 42:10776-10785.

Clark NL, Alani E, Aquadro CF. 2012. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res 22:714-720.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal, complex systems 1695:1-9.

Derelle R, Philippe H, Colbourne JK. 2020. Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment. Mol Biol Evol 37:3389-3396.

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.

Erill I, Campoy S, Mazon G, Barbe J. 2006. Dispersal and regulation of an adaptive mutagenesis cassette in the bacteria domain. Nucleic Acids Res 34:66-77.

Fallon AM, Carroll EM. 2023. Virus-like Particles from Wolbachia-Infected Cells May Include a Gene Transfer Agent. Insects 14:516.

Farrera-Calderon RG, Pallegar P, Westbye AB, Wiesmann C, Lang AS, Beatty JT. 2021. The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer agent production Is bidirectional and regulated by cyclic di-GMP. J Bacteriol 203:e00525-00520.

Fogg PCM. 2019. Identification and characterization of a direct activator of a gene transfer agent. Nat Commun 10:595.

Forsythe ES, Williams AM, Sloan DB. 2021. Genome-wide signatures of plastid-nuclear coevolution point to repeated perturbations of plastid proteostasis systems across angiosperms. Plant Cell 33:980-997.

George EE, Tashyreva D, Kwong WK, Okamoto N, Horak A, Husnik F, Lukes J, Keeling PJ. 2022. Gene Transfer Agents in Bacterial Endosymbionts of Microbial Eukaryotes. Genome Biol Evol 14: evac099.

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. J Mol Biol 299:283-293.

Gozzi K, Tran NT, Modell JW, Le TBK, Laub MT. 2022. Prophage-like gene transfer agents promote Caulobacter crescentus survival and DNA repair during stationary phase. PLoS Biol 20:e3001790.

Granato LM, Picchi SC, Andrade Mde O, Takita MA, de Souza AA, Wang N, Machado MA. 2016. The ATP-dependent RNA helicase HrpB plays an important role in motility and biofilm formation in Xanthomonas citri subsp. citri. BMC Microbiol 16:55.

Gross M, Marianovsky I, Glaser G. 2006. MazG -- a regulator of programmed cell death in Escherichia coli. Mol Microbiol 59:590-601.

Han L, Zheng J, Wang Y, Yang X, Liu Y, Sun C, Cao B, Zhou H, Ni D, Lou J, et al. 2016. Structure of the BAM complex and its implications for biogenesis of outer-membrane proteins. Nat Struct Mol Biol 23:192-196.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33:1635-1638.

Hynes AP, Shakya M, Mercer RG, Grull MP, Bown L, Davidson F, Steffen E, Matchem H, Peach ME, Berger T, et al. 2016. Functional and evolutionary characterization of a gene transfer agent's multilocus “genome". Mol Biol Evol 33:2530-2543.

Jaccard P. 1912. The distribution of the flora in the alpine zone. New phytologist 11:37-50.

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.

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.

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.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Kim JJ, Sundin GW. 2001. Construction and analysis of photolyase mutants of Pseudomonas aeruginosa and Pseudomonas syringae: contribution of photoreactivation, nucleotide excision repair, and mutagenic DNA repair to cell survival and mutability following exposure to UV-B radiation. Appl Environ Microbiol 67:1405-1411.

Kim WK, Bolser DM, Park JH. 2004. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics 20:1138-1150.

Kogay R, Koppenhofer S, Beatty JT, Kuhn JH, Lang AS, Zhaxybayeva O. 2022. Formal recognition and classification of gene transfer agents as viriforms. Virus Evol 8:veac100.

Kogay R, Neely TB, Birnbaum DP, Hankel CR, Shakya M, Zhaxybayeva O. 2019. Machine-learning classification suggests that many alphaproteobacterial prophages may instead be gene transfer agents. Genome Biol Evol 11:2941-2953.

Kogay R, Wolf YI, Koonin EV, Zhaxybayeva O. 2020. Selection for reducing energy cost of protein production drives the GC content and amino acid composition bias in gene transfer agents. mBio 11:e01206-01220.

Kogay R, Zhaxybayeva O. 2022. Selection for Translational Efficiency in Genes Associated with Alphaproteobacterial Gene Transfer Agents. mSystems 7:e0089222.

Kooistra J, Haijema BJ, Venema G. 1993. The *Bacillus subtilis* addAB genes are fully functional in *Escherichia coli*. Mol Microbiol 7:915-923.

Krokan HE, Standal R, Slupphaug G. 1997. DNA glycosylases in the base excision repair of DNA. Biochem J 325 (Pt 1):1-16.

Kuhn JH, Koonin EV. 2023. Viriforms-A New Category of Classifiable Virus-Derived Genetic Elements. Biomolecules 13:289.

Lang AS, Beatty JT. 2001. The gene transfer agent of Rhodobacter capsulatus and "constitutive transduction" in prokaryotes. Arch Microbiol 175:241-249.

Lang AS, Beatty JT. 2000. Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. Proc Natl Acad Sci U S A 97:859-864.

Lang AS, Beatty JT. 2007. Importance of widespread gene transfer agent genes in alpha-proteobacteria. Trends Microbiol 15:54-62.

Lang AS, Taylor TA, Beatty JT. 2002. Evolutionary implications of phylogenetic analyses of the gene transfer agent (GTA) of Rhodobacter capsulatus. J Mol Evol 55:534-543.

Lang AS, Westbye AB, Beatty JT. 2017. The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. Annu Rev Virol 4:87-104.

Lang AS, Zhaxybayeva O, Beatty JT. 2012. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol 10:472-482.

Lehtonen J. 2016. Multilevel Selection in Kin Selection Language. Trends Ecol Evol 31:752-762.

Leimkuhler S, Buhning M, Beilschmidt L. 2017. Shared Sulfur Mobilization Routes for tRNA Thiolation and Molybdenum Cofactor Biosynthesis in Prokaryotes and Eukaryotes. Biomolecules 7:5.

Leung MM, Brimacombe CA, Spiegelman GB, Beatty JT. 2012. The GtaR protein negatively regulates transcription of the gtaRI operon and modulates gene transfer agent (RcGTA) expression in *Rhodobacter capsulatus*. Mol Microbiol 83:759-774.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265-D268.

Marrs B. 1974. Genetic recombination in *Rhodopseudomonas capsulata*. Proc Natl Acad Sci U S A 71:971-973.

Marrs B, Wall JD, Gest H. 1977. Emergence of the biochemical genetics and molecular biology of photosynthetic bacteria. Trends Biochem Sci 2:105-108.

Martin-Rodriguez AJ. 2022. Respiration-induced biofilm formation as a driver for bacterial niche colonization. Trends Microbiol. 31:120-134.

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.

McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. 2010. High frequency of horizontal gene transfer in the oceans. Science 330:50.

Mercer RG, Quinlan M, Rose AR, Noll S, Beatty JT, Lang AS. 2012. Regulatory systems controlling motility and gene transfer agent production and release in *Rhodobacter capsulatus*. FEMS Microbiol Lett 331:53-62.

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.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589-596.

Narita S, Masui C, Suzuki T, Dohmae N, Akiyama Y. 2013. Protease homolog BepA (YfgC) promotes assembly and degradation of beta-barrel membrane proteins in Escherichia coli. Proc Natl Acad Sci U S A 110:E3612-3621.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733-745.

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.

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.

Price GR. 1970. Selection and covariance. Nature 227:520-521.

Price MN, Arkin AP. 2017. PaperBLAST: Text Mining Papers for Information about Homologs. mSystems 2: e00039-00017.

Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y, et al. 2018. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature 557:503-509.

Québatte M, Christen M, Harms A, Körner J, Christen B, Dehio C. 2017. Gene transfer agent promotes evolvability within the fittest subpopulation of a bacterial pathogen. Cell Syst 4:611-621 e616.

Rajagopala SV, Casjens S, Uetz P. 2011. The protein interaction map of bacteriophage lambda. BMC Microbiol 11:213.

Raza Q, Choi JY, Li Y, O'Dowd RM, Watkins SC, Chikina M, Hong Y, Clark NL, Kwiatkowski AV. 2019. Evolutionary rate covariation analysis of E-cadherin identifies Raskol as a regulator of cell adhesion and actin dynamics in Drosophila. PLoS Genet 15:e1007720.

Redfield RJ, Soucy SM. 2018. Evolution of Bacterial Gene Transfer Agents. Front Microbiol 9:2527.

Rossy T, Nadell CD, Persat A. 2019. Cellular advective-diffusion drives the emergence of bacterial surface colonization patterns and heterogeneity. Nat Commun 10:2471.

Saikrishnan K, Yeeles JT, Gilhooly NS, Krajewski WW, Dillingham MS, Wigley DB. 2012. Insights into Chi recognition from the structure of an AddAB-type helicase-nuclease complex. EMBO J 31:1568-1578.

Sato T, Yamanishi Y, Kanehisa M, Toh H. 2005. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21:3482-3489.

Schinner S, Engelhardt F, Preusse M, Thoming JG, Tomasch J, Haussler S. 2020. Genetic determinants of Pseudomonas aeruginosa fitness during biofilm growth. Biofilm 2:100023.

Shakya M, Soucy SM, Zhaxybayeva O. 2017. Insights into origin and evolution of alpha-proteobacterial gene transfer agents. Virus Evol 3:vex036.

Sherlock D, Fogg PCM. 2022. Loss of the Rhodobacter capsulatus Serine Acetyl Transferase Gene, cysE1, Impairs Gene Transfer by Gene Transfer Agents and Biofilm Phenotypes. Appl Environ Microbiol 88:e0094422.

Simmons EL, Bond MC, Koskella B, Drescher K, Bucci V, Nadell CD. 2020. Biofilm Structure Promotes Coexistence of Phage-Resistant and Phage-Susceptible Bacteria. mSystems 5.

Smith P, Schuster M. 2019. Public goods and cheating in microbes. Curr Biol 29:R442-R447.

Spek EJ, Vuong LN, Matsuguchi T, Marinus MG, Engelward BP. 2002. Nitric oxide-induced homologous recombination in Escherichia coli is promoted by DNA glycosylases. J Bacteriol 184:3501-3507.

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 37:2325–2331.

Steenwyk JL, Phillips MA, Yang F, Date SS, Graham TR, Berman J, Hittinger CT, Rokas A. 2022. An orthologous gene coevolution network provides insight into eukaryotic cellular and genomic structure and function. Sci Adv 8:eabn0105.

Stewart PS, Costerton JW. 2001. Antibiotic resistance of bacteria in biofilms. Lancet 358:135-138.

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49:D605-D612.

Tamarit D, Neuvonen MM, Engel P, Guy L, Andersson SGE. 2018. Origin and evolution of the *Bartonella* gene transfer agent. Mol Biol Evol 35:451-464.

Vidakovic L, Singh PK, Hartmann R, Nadell CD, Drescher K. 2018. Dynamic biofilm architecture confers individual and collective mechanisms of viral protection. Nat Microbiol 3:26-31.

Vos M. 2009. Why do bacteria engage in homologous recombination? Trends Microbiol 17:226-232.

Vulic M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci U S A 94:9763-9767.

Westbye AB, Beatty JT, Lang AS. 2017. Guaranteeing a captive audience: coordinated regulation of gene transfer agent (GTA) production and recipient capability by cellular regulators. Curr Opin Microbiol 38:122-129.

Westbye AB, O'Neill Z, Schellenberg-Beaver T, Beatty JT. 2017. The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. Microbiology (Reading) 163:1355-1363.

Wilson DS. 1975. A theory of group selection. Proc Natl Acad Sci U S A 72:143-146.

Wong EHJ, Ng CG, Goh KL, Vadivelu J, Ho B, Loke MF. 2018. Metabolomic analysis of low and high biofilm-forming Helicobacter pylori strains. Sci Rep 8:1409.

Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033-1034.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2:100141.

Yang Y, Yue Y, Song N, Li C, Yuan Z, Wang Y, Ma Y, Li H, Zhang F, Wang W, et al. 2020. The YdiU Domain Modulates Bacterial Stress Signaling through Mn(2+)-Dependent UMPylation. Cell Rep 32:108161.

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.

Zhang W, Ding W, Li YX, Tam C, Bougouffa S, Wang R, Pei B, Chiang H, Leung P, Lu Y, et al. 2019. Marine biofilms constitute a bank of hidden microbial diversity and functional potential. Nat Commun 10:517.

Tables

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

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

[@]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.

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

Gene name	Representative GenBank accession No.	GenBank Functional Annotation
prmC	QNR62505.1	peptide chain release factor N(5)-glutamine methyltransferase
bioC	ADE83961.1	conserved hypothetical protein
mhpC	ADE84522.1	hydrolase, alpha/beta fold family
miaA	ADE85369.1	tRNA delta(2)-isopentenylpyrophosphate transferase
-	ADE86839.1	protein-L-isoaspartate O-methyltransferase-2
phnP	ADE85005.1	metallo-beta-lactamase family protein
ispDF	ADE85540.1	bifunctional 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase/2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
glmU	ADE85246.1	bifunctional UDP-N-acetylglucosamine diphosphorylase/glucosamine-1-phosphate N-acetyltransferase
yfgZ	ADE86835.1	glycine cleavage T protein-2
yhiN	ADE84019.1	HI0933-like protein
-	ADE87182.1	conserved hypothetical protein
pcnB	ADE83954.1	CCA-adding enzyme
yfiH	ADE86534.1	protein of unknown function DUF152
-	ADE86535.1	protein of unknown function DUF185
alr	ADE85319.1	alanine racemase
rspA	ADE86886.1	mandelate racemase/muconate lactonizing enzyme family protein
ppiD	ADE86084.1	peptidyl-prolyl cis-trans isomerase D
aroE	ADE83833.1	shikimate 5-dehydrogenase
glnE	ADE86144.1	glutamate-ammonia-ligase adenylyltransferase
rns	ADE84242.1	ribonuclease T2 family protein
gluQ	ADE85705.1	glutamyl-Q tRNA(Asp) synthetase
MA20_39615	ADE85459.1	protein of unknown function DUF985
ptpA	ADE86851.1	protein-tyrosine-phosphatase
rne	ADE85895.1	ribonuclease E
ptr1	ADE86148.1	oxidoreductase, short-chain dehydrogenase/reductase family
queG	ADE86491.1	4Fe-4S ferredoxin, iron-sulfur cluster binding protein
tadB	ADE84271.1	type II secretion system protein
-	ADE84017.1	NAD-dependent epimerase/dehydratase family protein
nnrD	ADE85417.1	YjeF-related protein family
pepA	ADE86425.1	leucyl aminopeptidase-2
-	ADE84176.1	peptidase, S58 family
pepN	ADE84605.1	aminopeptidase N
MA20_18095	ADE84243.1	alcohol dehydrogenase, zinc-binding domain protein

Figures

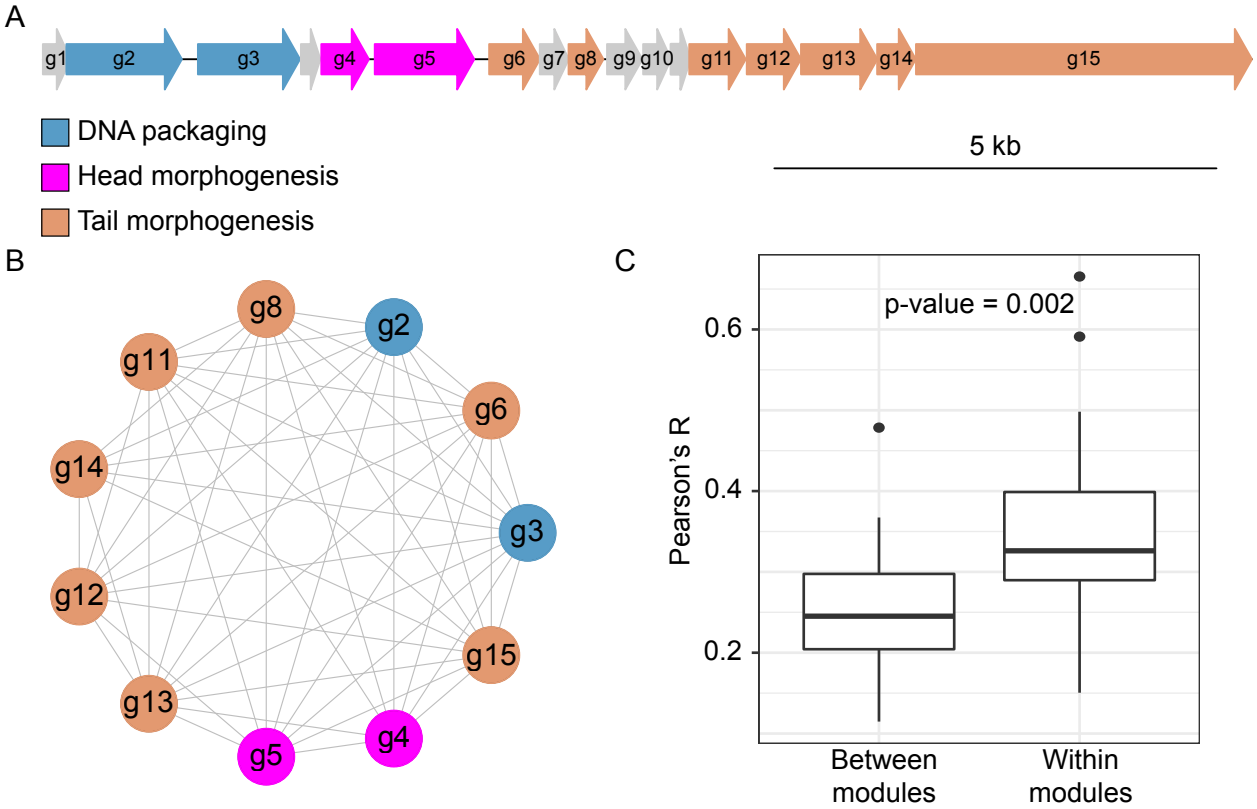

Figure 1. Co-evolution of 11 reference GTA genes from head-tail clusters. (A) The head-tail cluster of the RcGTA, as encoded in *R. capsulatus* strain SB 1003 genome (GenBank accession CP001312.1; spanning locus tags RCAP_rcc01682 - RCAP_rcc01698). Each gene is represented by an arrow and is drawn proportional 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.

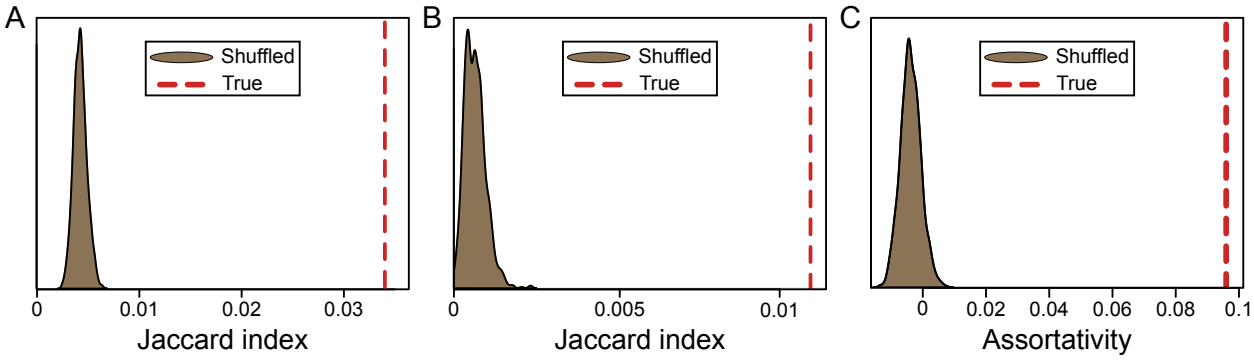

Figure 2. Strengths of correlations between the covariation evolutionary rate of genes and function of the proteins the genes encode, as measured by permutation tests. (A) Comparison of the co-evolution network and the PPI network of *Phaeobacter inhibens*. The distribution in brown corresponds to distances between the PPI network and 1,000 co-evolution networks in which edges were randomly shuffled. The dashed red line indicates the Jaccard index from the non-shuffled network comparison. **(B)** Comparison of the co-evolution network and the fitness effects network of *Phaeobacter inhibens*. The distribution in brown corresponds to distances between the network of fitness effects and 1,000 co-evolution networks in which edges were randomly shuffled. The dashed red line indicates the Jaccard index from the non-shuffled network comparison. **(C)** Positive assortativity between co-evolution and COG functional category assignment. The distribution in brown corresponds assortativity coefficient values for 1,000 networks in each COG labels were randomly shuffled. The dashed red line indicates the assortativity coefficient of the non-shuffled network.

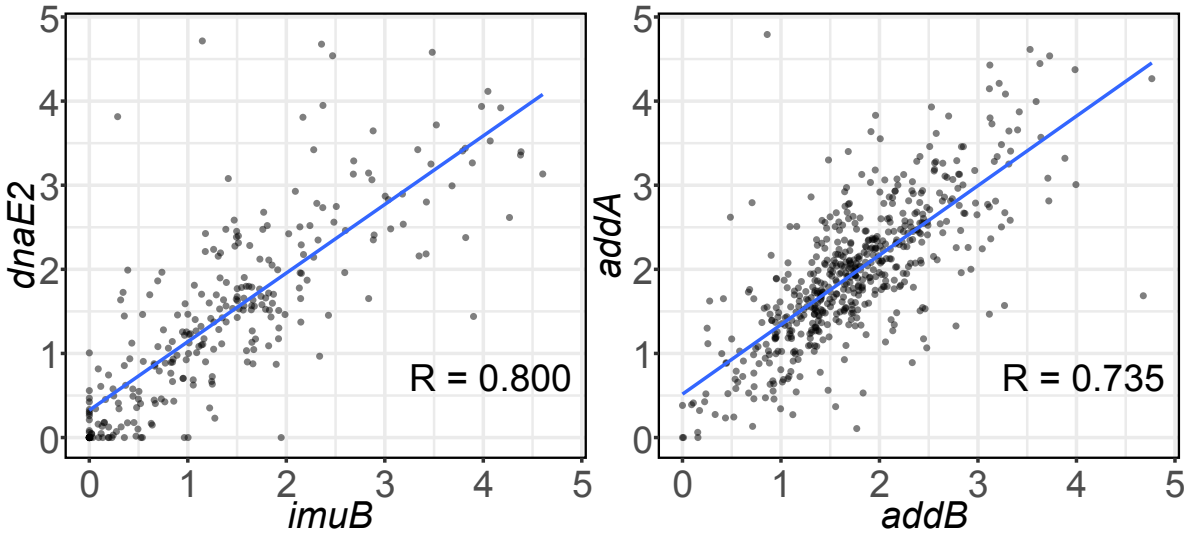

Figure 3. Covariation of evolutionary rates for two gene pairs with the largest Pearson's coefficients. Each dot represents the normalized branch length of the proteins shown on x- and y-axes. Blue line depicts the linear regression line. The value of Pearson's R is shown on each panel.

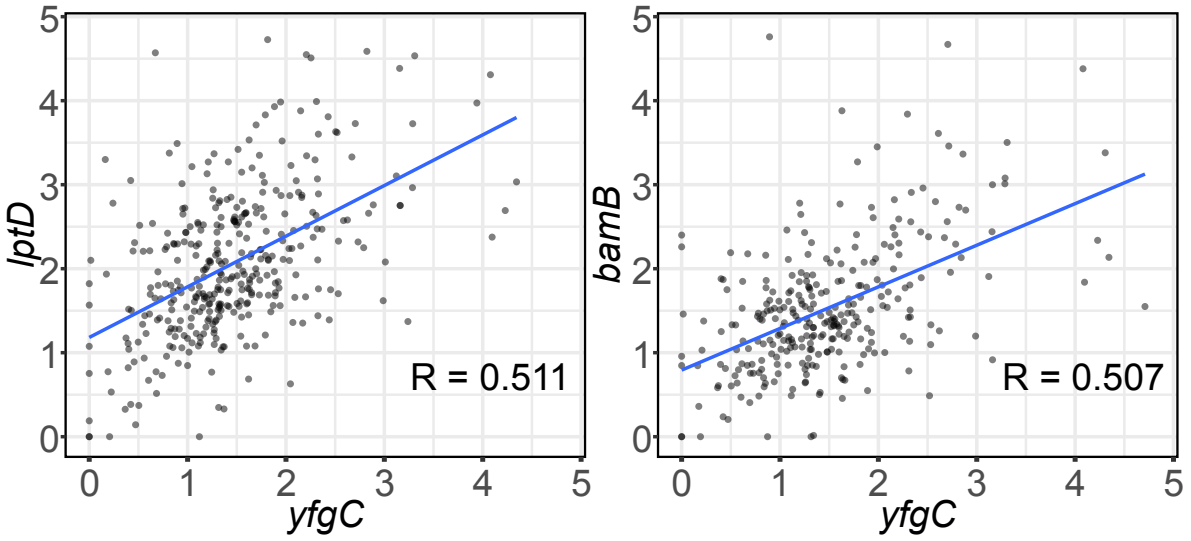

Figure 4. Covariation of evolutionary rates for two gene pairs with no evidence of interactions for their protein products in the STRING database. Each dot represents the normalized branch length of the proteins shown on x- and y-axes. Blue line depicts the linear regression line. The value of Pearson's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

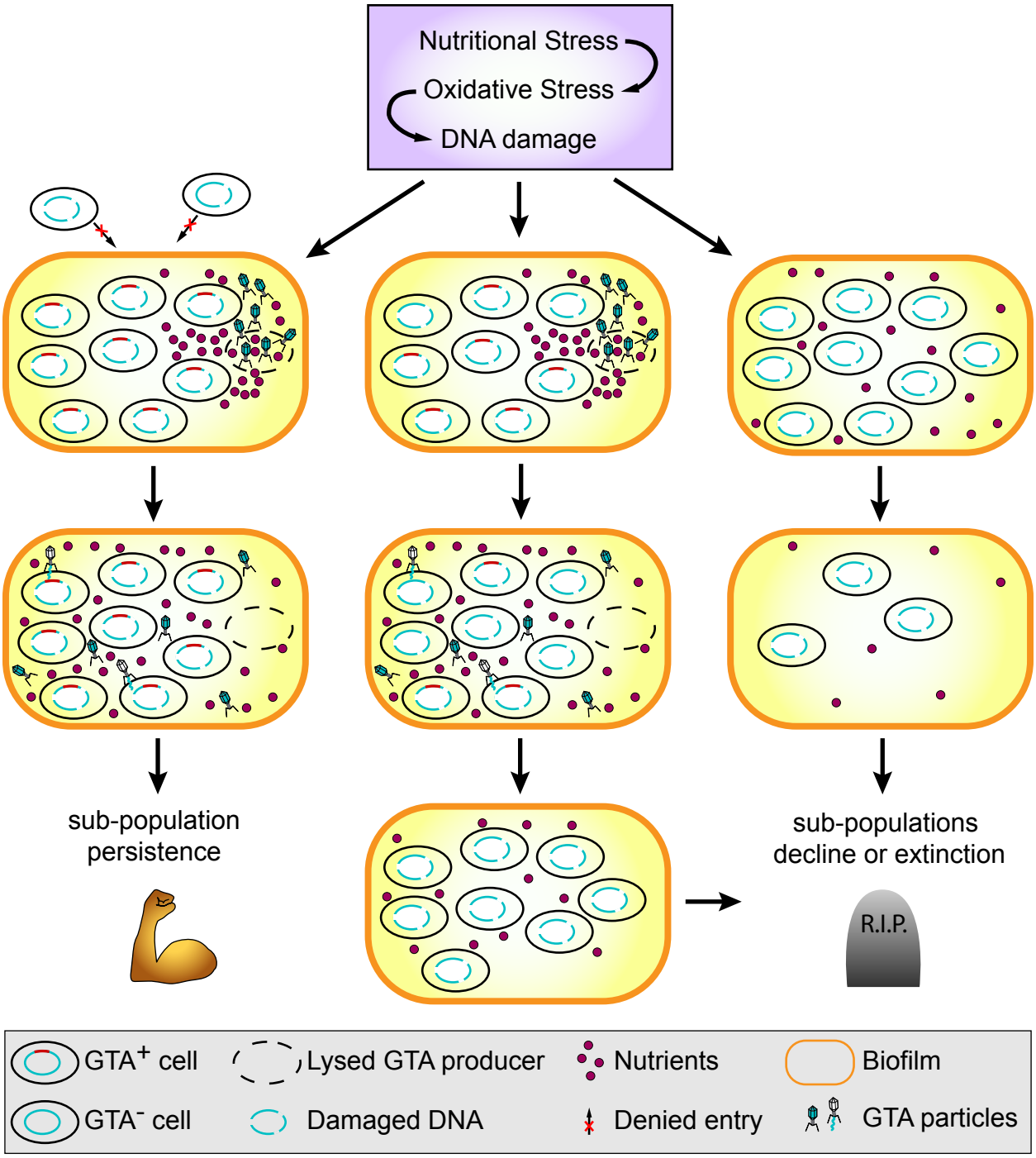

Figure 5. The proposed model of between-group selection that preserves the trait of GTA production in a bacterial population of closely related cells. See **Discussion** section for detailed description.