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

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

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

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

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

Running Title: Selection for Translational Efficiency in GTAs

Keywords: GTA, Sphingomonadales, codon usage bias, tonB, addAB, head completion protein

Abstract

Gene transfer agents (GTAs) are virus-like elements that are encoded by some bacterial and archaeal genomes. The production of GTAs can be induced by the carbon depletion and results in host lysis and release of virus-like particles that contain mostly random fragments of the host DNA. The remaining members of a GTA-producing population act as GTA recipients by producing proteins needed for the GTA-mediated DNA acquisition. Here, we detect a strong codon usage bias in alphaproteobacterial RcGTA-like GTA genes, which likely improves the translational efficacy during their expression. While the strength of selection for translational efficiency fluctuates substantially among individual GTA genes and various taxonomic groups, it is especially pronounced in *Sphingomonadales*, whose members are known to inhabit nutrient-depleted environments. Additionally, the intensity of the selection acting on GTA genes negatively correlates with the carbon content of the encoded proteins, indicating the importance of controlling energetic cost of potentially frequent GTA production. By screening genomes for gene families with similar trends in codon usage biases to those in GTA genes, we found a gene that likely encodes “head completion protein” in some GTAs were it appeared missing, and 13 genes previously not implicated in GTA lifecycle. The latter genes are involved in a range of molecular processes, including the homologous recombination and response to the carbon starvation. Our findings highlight host-driven evolution of GTA genes expressed under nutrient-depleted conditions, and outline genes that are potentially involved in the previously hypothesized integration of GTA-delivered DNA into the host genome.

Introduction

Gene transfer agents are phage-like particles produced by multiple groups of bacteria and archaea (Lang et al. 2017). Unlike viruses, GTA particles tend to package random pieces of the host cell DNA instead of genes that encode GTAs themselves (Marrs 1974; Lang et al. 2012). Released GTA particles can deliver the packaged genetic material to other cells (Brimacombe et al. 2014), impacting exchange of genetic material in prokaryotic populations (McDaniel et al. 2010; Brimacombe et al. 2015; Quebatte et al. 2017). The benefits of GTA production and GTA-mediated DNA acquisition are not well understood. It has been hypothesized that GTAs may facilitate DNA repair (Marrs et al. 1977) or enable population-level exchange of traits needed under the conditions of a nutritional stress (Westbye et al. 2017; Kogay et al. 2020).

To date, at least three independently exapted GTAs are functionally characterized (Kogay et al. 2022). The most studied GTA system (RcGTA) belongs to the alphaproteobacterium *Rhodobacter capsulatus* (Marrs 1974). RcGTA is encoded by at least 24 genes that are distributed across 5 distinct genomic loci (Hynes et al. 2016; Shakya et al. 2017). Seventeen of the 24 genes are situated in one locus, which is dubbed the 'head-tail' cluster because it encodes most of the structural proteins of the RcGTA particles (Lang et al. 2017). RcGTA-like 'head-tail' clusters are present in many alphaproteobacterial genomes; they evolve slowly and are inferred to be inherited mostly vertically from a common ancestor of an alphaproteobacterial clade that spans multiple taxonomic orders (Lang and Beatty 2007; Shakya et al. 2017; Kogay et al. 2019). Additionally, multiple cellular genes regulate RcGTA production, release and reception (Hynes et al. 2016; Fogg 2019). It is likely that other, yet undiscovered, genes in *R. capsulatus* genome are involved in GTA lifecycle.

Expression of RcGTA is known to be triggered by specific environmental conditions, such as nutrient depletion (Westbye et al. 2017), under which a small fraction of the bacterial population becomes dedicated to GTA production (Fogg et al. 2012; Hynes et al. 2012). As a result, RcGTA-producing cells are likely to express GTA genes at high levels. Highly expressed genes that are involved in core biological processes, such as translational machinery, are known to exhibit strong codon usage bias (Roller et al. 2013). For example, codon usage in ribosomal proteins, which are highly expressed in almost all organisms, deviates most dramatically from the distribution of codons expected under their equal usage corrected for organismal GC content (Wright 1990). Such bias is primarily due to selection to match the pool of most abundant tRNA molecules in order to have the most efficient translation for proteins needed in high number of copies (Rocha 2004; Quax et al. 2015; Zhou et al. 2016). As a result, highly expressed genes tend to have codons that correspond to the most abundant tRNA molecules in the cell. Selection for such “translational efficiency” is ubiquitous among bacteria (Supek et al. 2010), and codon usage bias also extends to genes that are of high importance under specific environmental conditions (Supek et al. 2010; LaBella et al. 2021). Additionally, genes that encode interacting proteins, genes involved in the same pathway, or genes responsible for an adaptation of an organism to a specific ecological niche often exhibit similar codon usage biases (Fraser et al. 2004; LaBella et al. 2021).

Alphaproteobacterial RcGTA-like GTA genes (hereafter referred to as “GTA genes” for brevity) have a pronounced codon bias towards GC-rich codons (Shakya et al. 2017; Kogay et al. 2020). This bias is due to the encoded proteins consisting of energetically cheaper amino acids, hypothesized to be important due to GTA production during the carbon starvation (Kogay et al. 2020). In this study, we evaluated if the observed codon usage bias of GTA genes is due to the genes being under selection for the translation efficiency. For this purpose, we used two well-established metrics for assessment of codon bias and its match to the tRNA abundance: effective number of codons (ENC) (Wright 1990) and tRNA adaptation index (tAI) (dos Reis et al. 2003). We also searched for genes whose involvement in GTA production and regulation is currently unsuspected by screening GTA-encoding genomes for genes with codon usage patterns similar to those of GTA genes.

Results

Codon bias of GTA genes is under selection to match available tRNAs

To confirm the earlier observed presence of codon bias in GTA genes across alphaproteobacteria, we have calculated ENC for each reference GTA gene (see **Methods** for the definition) and compared them against the expected ENC of a gene in a genome under the null model of no codon bias, corrected for the genomic GC content (dos Reis et al. 2004). Indeed, we found that 1,543 out of 2,308 (66.8%) GTA reference genes detected across 208 GTA head-tail clusters deviate from the genome-specific null expectations by more than 10% (**Supplemental Figure S1**). However, there is a substantial variation in this deviation for different GTA genes (**Supplemental Figure S2**), and only in genes *g5* and *g8* the deviation is significantly higher than the genomic average (Kruskal-Wallis rank sum test, p-value < 2.2e-16; Dunn's test, p-value < 0.05, Benjamini-Hochberg correction).

To assess the match of the observed codon bias to available tRNA pool, we calculated tAI values of the reference GTA genes across 208 genomes and converted them to percentile tAI values (ptAI; see **Methods** for the definition) to allow for the intergenomic comparisons. Similar to the ENC values, the ptAI values also vary substantially across the genes and genomes (**Figure**

1), suggesting that there is a selection for translational efficiency, but the selection is not the same in different GTA genes and in different genomes. Because the GTA-containing alphaproteobacteria occupy diverse ecological niches and individual GTA proteins are expressed at variable levels (Bardy et al. 2020), the presence of differential selection pressure is expected. In the next two sections, we evaluate the strength of selection for translational efficiency in individual GTA genes and in specific taxonomic groups.

Selection for translational efficiency is uneven among GTA genes

The differences of ptAI values among the reference GTA genes across 208 genomes are statistically significant (Kruskal-Wallis rank sum test, p-value < 2.2e-16) (**Figure 1**). Particularly notable is a significant decline in ptAI values of the region encoding genes *g12* through *g15* (Dunn's test, p-value < 0.05, Benjamini-Hochberg correction)*,* which are located at the 3' end of the head-tail cluster and encode the tail components of a GTAs particle. In contrast, ptAI values of the genes *g5* (encoding major capsid protein) and *g11* (encoding tail tape measure protein) are significantly larger than ptAI values of other GTA genes (Dunn's test, p-values < 0.05, Benjamini-Hochberg correction). Notably, protein g5 is detected in the largest number of copies (145) per RcGTA particle than any other protein (Bardy et al. 2020), while proteins g12-g15 are present in a small number (1-6) copies per RcGTA particle (Bardy et al. 2020). Given that genes encoding proteins needed in a larger number of copies have a higher degree of adaptation to the tRNA pool (Plotkin and Kudla 2011), we hypothesize that the observed variation in ptAI values of GTA genes reflects the different number of GTA proteins in a GTA particle. Protein g11, however, is found in only 3 copies per RcGTA particle (Bardy et al. 2020) and therefore a demand for a larger copy number cannot explain its high ptAI values.

Variation of ptAI values could also be due to physical location of the genes in the GTA head-tail cluster. Similar to the operons (Lim et al. 2011), genes in the RcGTA head-tail cluster are co-transcribed from a single promoter upstream of the cluster (Lang and Beatty 2000; Fogg 2019). Because genes at the 3’ end of operons tend to have lower expression levels (Nishizaki et al. 2007), the low ptAI values of GTA genes *g12-g15* may be due to their distant location from the promoter.

Selection for translational efficiency is the strongest in *Sphingomonadales*’ genomes

In addition to differences in ptAI values across different GTA genes, there is also a large variance of ptAI values for each individual reference GTA gene across the 208 genomes (**Figure 1**). To evaluate if these differences represent variation in selection pressure in distinct taxonomic groups, we initially examined the ptAI values of gene *g5* that were grouped by alphaproteobacterial order. The *g5* gene was chosen due to its high abundance of the encoded protein in RcGTA particles (more copies than all other structural proteins combined) and for being the only gene with the highest detected deviations from the average genomic values for both ENC and ptAI. We found that ptAI values of the *g5* gene vary significantly among members of the four alphaproteobacterial orders (Kruskal-Wallis rank sum test, p-value < 0.05) (**Figure 2A**). In particular, *g5* genes from the *Sphingomonadales*’ genomes have significantly higher ptAI values than those from genomes of bacteria from other three orders (Mann-Whitney U test, p-value < 0.05, Benjamini-Hochberg correction). Twelve of the 14 *g5* genes with the highest overall ptAI values (> 90) (**Figure 2A**) also belong to the *Sphingomonadales* genomes. Beyond just *g5* gene, all reference GTA genes, as a group, have higher ptAI values in *Sphingomonadales* than in members of the three other alphaproteobacterial orders (Mann-

Whitney U test, p-value < 0.05, Benjamini-Hochberg correction) (**Supplemental Figure S3)**. These observations suggests that in *Sphingomonadales* in particular, there is a strong selection for efficient production of GTA particles. Because *Sphingomonadales* are known to live in nutrient-depleted environments (Balkwill et al. 2006), we suggest that GTA production is beneficial enough in those habitats to exert strong selection for translational efficiency.

The increase in translational efficiency of GTA genes is associated with a reduced energetic cost for production of the encoded proteins

Among the GTA proteins in four alphaproteobacterial orders, *Sphingomonadales'* GTA proteins also have the strongest skew in amino acid composition towards energetically cheaper amino acids (**Figure 2B**). To evaluate if selection for energy efficiency is linked to selection for translational efficiency, we examined the relationship between the ptAI values of GTA genes and the number of carbons in amino acid chains encoded by the *Sphingomonadales* GTA genes. We found that there is a significant negative correlation between the two (Pearson R = -0.19, N = 636, p-value < 0.05). We propose that in *Sphingomonadales* benefits associated with production of GTA particles in nutrient-limited conditions led not only to the selection for translational efficiency, but also to the selection for use of energetically cheaper amino acids in the GTA genes.

Similarities in translational efficiency trends point at 13 additional genes that are likely involved in GTA lifecycle

Within each genome, genes in the GTA 'head-tail' cluster are expected to exhibit similar patterns of selection for translational efficiency, due to their transcription from a single promoter

(Fogg 2019). Indeed, using phylogenetic generalized least squares (PGLS) method to correct for biases due to shared evolutionary histories, we found that all reference GTA genes (which are all from the 'head-tail' cluster) have a similar trend in ptAI values with a positive slope of the PGLS model fit. The trend is significant in all pairwise comparisons (**Supplemental Figure S4)**.

We conjecture that other genes in the other loci of a GTA "genome", as well as the host genes involved in GTA lifecycle, would exhibit similar patterns of selection for translational efficiency, allowing for discovery of yet unsuspected genes involved in GTA lifecycle. To identify such unknown genes that may be co-expressed with GTA genes, we examined correlations of ptAI values between reference GTA genes and 3,477 other gene families present in 208 alphaproteobacterial genomes. The PGLS analysis revealed 14 gene families, whose ptAI values correlate significantly with ptAI values of the reference GTA genes (**Table 1)**. One of 14 identified gene families is a homolog of *gafA*, which encodes a crucial transcription activator of GTA particles production in *Rhodobacter capsulatus* (Hynes et al. 2016; Fogg 2019). This gene is located outside of the RcGTA's head-tail cluster, and therefore was not included in the GTA reference set, but its discovery demonstrates the suitability of our approach to identify genes linked to the GTA lifecycle. Interestingly, *gafA* homologs were previously described only in the genomes of *Rhodobacterales* and some *Rhizobiales* (Hynes et al., 2016; Shakya et al., 2017; Fogg, 2019). However, with less stringent requirements for the similarity searches, we were able to identify this regulator in 196 of the 208 genomes (94.7%), spanning all major GTA-containing alphaproteobacterial orders.

Six of the remaining 13 gene families belong to the "replication, recombination, and repair" (L) functional category of the Clusters of Orthologous Groups (COG) classification (**Table 1**). Protein products of two of the 6 genes (*addA* and *addB*) form the heterodimeric helicase-nuclease complex that repairs double-stranded DNA breaks by homologous recombination and is functionally equivalent to the RecBCD complex (Kooistra et al. 1993). The knockout of AddAB complex is associated with a deficiency in RecA-dependent homologous recombination (Marsin et al. 2010). We hypothesize that the addAB pathway is involved in recombination of GTAs' genetic material with the host's genome. The main function of another protein from the L COG category, exodeoxyribonuclease VII large subunit (xseA), is to form a complex with xseB and degrade single-stranded DNA to oligonucleotides. However, expression of *xseA* gene without *xseB* gene leads to cell death (Jung et al. 2015). Because we did not find any correlation of codon bias for *xseB* gene, we speculate that instead of involvement in processing of GTA DNA, *xseA* gene product facilitates lysis of GTA producing cells and GTA particles release.

Three gene families are assigned 'molecular chaperones and related functions' function (COG category O) (**Table 1**). While GTAs encode their own chaperones that assist GTA protein folding (Bardy et al. 2020), it is well known that chaperones tend to be highly expressed in bacteria at times of stress and facilitate the survival of cells in rapidly changing environmental conditions (Genest et al. 2019). Because chaperones are essential in responding to the starvation-induced cellular stresses (Rockabrand et al. 1998), we conjecture that observed similarity in ptAI values of chaperone genes and the reference GTA genes is due to their expression being triggered by the similar environmental conditions.

One gene family, present only in members of *Sphingomonadales* order, encodes tonB energy transducer (**Table 1**). TonB-dependent transporters are involved in transport of diverse compounds, including carbohydrates, amino acids, lipids, vitamins and iron (Blanvillain et al. 2007; Eisenbeis et al. 2008; Tang et al. 2012). Because *Sphingomonadales* live in nutrient-limited environments, we hypothesize that as GTA-producing *Sphingomonadales* die, the nutrients released from the lysed cells are imported by TonB-dependent transporters by other members of the population. Similar to the quorum-sensing regulated expression of the gene encoding GTA receptor in *Rhodobacter capsulatus*' bacterial communities that produce GTAs (Brimacombe et al. 2014), *tonB* gene could be regulated to be expressed in the non-GTA-producing members of the population to aid their survival.

To evaluate if detected gene families interact with each other and with GTA genes, we have constructed the protein-protein interaction network of the 14 gene families, 12 GTA reference genes and 50 additional interactor proteins from the STRING database (**Figure 3**). Thirteen of the 14 families and all 12 reference GTA genes belong to two protein-protein interaction sub-networks (**Figure 3**), one of which contains all GTA reference genes, while the other is involved in a wide range of functions (**Table 1**). By carrying out the KEGG enrichment analysis, we found significant overrepresentation of four molecular pathways in the second protein-protein interaction network (**Supplemental Table S1**). Consistent with the 6 gene families assigned to the "replication, recombination, and repair" COG category, two of the pathways are 'homologous recombination' and 'mismatch repair', suggesting involvement of identified genes in integration of the genetic material delivered by GTAs into recipients' genomes. Two additional pathways, 'carotenoid biosynthesis' and 'terpenoid backbone biosynthesis', are less likely to be directly involved in the lifecycle of GTAs. Production of secondary metabolites is known to be protective against stress factors (Gershenzon and Dudareva 2007; Tyc et al. 2017), and carbon starvation leads to the upregulation of carotenoid biosynthesis pathway (Yang et al. 2015; Ram et al. 2020). Similar to the above-described genes encoding chaperones, we hypothesize that expression of 'carotenoid biosynthesis' and 'terpenoid backbone biosynthesis' genes is not related to GTA lifecycle, but is initiated by conditions that also activate production of GTAs.

A viral replacement of the head completion protein in Sphingomonadales' GTAs

Gene content of GTA head-tail clusters varies across alphaproteobacteria (Shakya et al. 2017). While some clusters do not contain homologs of all RcGTA genes, others include additional genes that are conserved across multiple clusters but have no known function (Shakya et al. 2017; Kogay et al. 2019). To predict whether any of these additional genes play a role in GTA production, we compared their ptAI values of genes found in at least 10 genomes to ptAI values of the reference GTA genes. One gene family, which is found only within GTA head-tail clusters of 11 genomes in one subclade of *Sphingomonadales*, has a significant positive correlation with 5 out of the 12 GTA reference genes (**Supplemental Table S2**). Interestingly, within *Sphingomonadales* GTA head-tail clusters this gene is located where the *g7* gene*,* which encodes a head completion protein, is found in the RcGTA head-tail cluster. Only seven of the 55 *Sphingomonadales* genomes in our dataset have detectable homologs of the *g7* gene. Among the remaining 48 genomes, 22 contain a gene encoding a protein of unknown function in the "gene *g7* locus", while 26 genomes don't have any gene in that locus.

The members of the identified gene family are substantially shorter than the RcGTA gene *g7* and have a different secondary structure (**Figure 4**), precluding the possibility that the identified protein is simply too divergent for a detectable amino acid similarity. However, we found viral head completion proteins that have similar protein length and similar secondary structures to both GTA head completion protein and the identified gene family (**Figure 4**). We conjecture that the gene encoding the head completion protein was replaced in some *Sphingomonadales* by a gene encoding an analogous viral protein.

Discussion

Our analyses of codon biases suggest presence of selection for translational efficiency in at least some alphaproteobacterial GTA systems. The strength of selection is the most pronounced (and therefore most easily detectable) in the major capsid protein gene, which is needed to be expressed to produce thousands of copies per GTA-producing bacterium. The strength of selection varies across taxonomic groups, which can be explained by diversity of environmental niches that GTA-containing alphaproteobacteria occupy. The selection is particularly prominent in *Sphingomonadales* order, whose members typically inhabit nutrient-limited conditions. Combined with an observation that production of GTAs is triggered by the nutritional stress (Westbye et al. 2017), our findings further underscore the earlier hypothesized importance of GTA systems in situations of nutrient scarcity (Kogay et al. 2020).

The strong bias towards codons with the most readily available tRNAs suggests that GTA genes are highly expressed in members of *Sphingomonadales* order. Additionally, the stronger selection for translation efficiency in GTA genes is associated with the larger decline in the carbon content of the proteins the genes encode. These findings suggest that benefits associated with GTA production are substantial enough to drive selection for both translational efficiency and low energetic costs of the translated proteins. We speculate that these modifications of GTA proteins allow the bacterial population under adverse conditions to increase both the speed of GTA particle production and the number of released GTA particles.

Exact role of GTA production by a small fraction of a bacterial population remains unresolved (reviewed in Lang et al. [2017]). The hypotheses proposed to-date include the adaptive advantages of GTA production in facilitating DNA repair (Marrs et al. 1977), disseminating beneficial genes (McDaniel et al. 2010), and decreasing population density during the carbon starvation periods (Westbye et al. 2017; Kogay et al. 2020). It is known that in *R. capsulatus*, the GTA genes are regulated in both producing and recipient cells by cellular master regulators that control expression of many other genes (Leung et al. 2012; Mercer et al. 2012). Therefore, we hypothesized that genes that are located outside of the GTA head-tail cluster, but are involved in GTA production, GTA reception and processing of the GTA-delivered DNA, would have similar signatures of selection for translational efficiency as GTA genes. Gratifyingly, our genome-wide screen for such patterns detected the direct GTA activator gene, *gafA* (Fogg 2019). We also identified multiple genes not yet implicated in GTA lifecycle. Some of these genes are involved in recombination and mismatch repair, providing computational support for the hypothesis that DNA brought to the cell via GTA particles becomes incorporated into the recipient's genome (Brimacombe et al. 2014). Another gene with similar selection pressures is *tonB*, whose protein product is involved in transport of organic molecules. This raises an intriguing possibility that GTA production in a population under the carbon-depleted environment may also provide surviving cells with ability to import as nutrients the debris of lysed GTA-producer cells.

Alphaproteobacterial GTAs likely originated millions of years ago from a lysogenic phage, and since then they were mostly vertically inherited by many alphaproteobacterial lineages (Lang and Beatty 2007; Shakya et al. 2017). However, as many other regions of a typical bacterial genome (Soucy et al. 2015), it is very likely that over time GTAs experienced gene replacements via horizontal gene transfer (Shakya et al. 2017). Instances of horizontal gene transfer between GTAs and phages have been already documented (Hynes et al. 2016; Zhan et al. 2016). By examining the patterns of selection for translational efficiency, we identified another case of likely ancient gene exchange with viruses that resulted in the replacement of the gene encoding head completion protein in some *Sphingomonadales*. Curiously, the gene currently has no significant primary sequence similarity to any gene in GenBank. Many other unannotated ORFs in alphaproteobacterial head-tail clusters outside of *Rhodobacterales* (Shakya et al. 2017) may also have functional roles in their respective GTA regions. Notably, when alphaproteobacterial RcGTA-like genomic regions appear incomplete due to lack of many homologs to genes required for GTA production in *R. capsulatus*, it could be due our inability to recognize some genes due to their replacements with analogous genes. Because such incomplete RcGTA-like clusters are abundant in alphaprotebacteria (Shakya et al. 2017), GTAs could be morphologically diverse and even more widespread across alphaproteobacteria than we currently estimate (Kogay et al. 2019).

Methods

Dataset of representative alphaproteobacterial genomes with GTA head-tail clusters

As an initial data set, we selected 212 representative alphaproteobacterial genomes previously predicted to contain GTAs (Kogay et al. 2020). The gene annotations of the genomes were downloaded from the RefSeq database (O'Leary et al. 2016) in October 2020. GTA head-tail clusters (Lang et al. 2017) were predicted using the GTA-Hunter program (Kogay et al. 2019). Because GTA-Hunter identifies only 11 out of the 17 genes in the RcGTA's head-tail cluster and also requires genes to align with their RcGTA homologs by at least 60% of their length, some GTA genes were likely missed by GTA-Hunter. To look for these potential false negatives, additional BLASTP searches with the e-value cutoff of 0.1 (Altschul et al. 1997) were performed using 17 RcGTA head-tail cluster genes as queries and protein-coding genes in 212 genomes as a database. Only matches located within the genomic regions designated as GTA gene clusters by GTA-Hunter were kept. In four genomes, calculations of genes' adaptation to tRNA pool did not converge (see below for details). As a result, only 208 genomes were retained in the reported analyses (**Supplemental Table S3**).

Identification of gene families in 208 alphaproteobacterial genomes

Within each genome, protein-coding genes less than 300 nucleotides in length were excluded in order to reduce the stochasticity of codon usage bias values due to the insufficient number of codons. The remaining protein-coding genes were clustered into gene families using Orthofinder v2.4 (Emms and Kelly 2019) with DIAMOND (Buchfink et al. 2015) for the sequence similarity search. Only gene families detected in at least 40 genomes were retained to ensure statistical power.

Some alphaproteobacterial GTA head-tail cluster regions contain protein-coding ORFs that do not have significant similarity to the RcGTA homologs of the genes shown to be required for GTA production in RcGTA. Gene families of these ORFs were retrieved from the collection of gene families predicted for all protein-coding genes (regardless of their length) using Orthofinder v2.4 (Emms and Kelly 2019) with DIAMOND (Buchfink et al. 2015) for the amino acid sequence similarity search. Only gene families that are both located within the genomic region encoding GTA head-tail cluster and found in at least 10 genomes were retained.

Reference set of GTA genes

Although RcGTA head-tail cluster contains 17 genes, genes *g3.5* and *g10.1* are less than 300 nucleotides in length, and genes *g1* and *g7* are not detected widely across analyzed genomes. Additionally, codon usage patterns of gene *g9* were found to be very different from other GTA genes (see **"Examination of similarity in adaptation to the tRNA pool among GTA genes"** section below for details). Therefore, in our inferences about selection, we considered only 12 of the 17 GTA genes (**Supplemental Table S4**), which we designate throughout the manuscript as "GTA reference genes".

Reconstruction of the reference phylogenomic tree

Twenty-nine marker proteins that are present in a single copy in more than 95% of the 208 retained genomes were retrieved using AMPHORA2 (Wu and Scott 2012). Amino acid sequences within each of the 29 marker families were aligned using MAFFT-linsi v7.455 (Katoh and Standley 2013). The best substitution matrix for each family was determined by *ProteinModelSelection.pl* script downloaded from https://github.com/stamatak/standard-RAxML/tree/master/usefulScripts in October 2020. Individual alignments of the marker families were concatenated, but each alignment was treated as a separate partition with its own best substitution model in the subsequent phylogenetic reconstruction. The maximum likelihood tree was reconstructed using IQ-TREE v 1.6.7 (Nguyen et al. 2015).

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

For the retained genes in each genome, effective number of codons (ENC) (Wright 1990) and G+C content variation at the 3rd codon position in the synonymous sites (GC3s) were calculated using CodonW (http://codonw.sourceforge.net). The null model of no codon usage bias was calculated as described in dos Reis et al. (dos Reis et al. 2004) using an in-house script. For every gene, the deviation of its ENC from the null model was calculated using an in-house script. Genes that have observed ENC higher than expected ENC were excluded from analyses.

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

The tRNA genes in each genome were predicted using tRNAscan-SE v 2.06, using a model trained on bacterial genomes (Lowe and Eddy 1997; Chan et al. 2021) and the Infernal mode without HMM filter to improve the sensitivity of the search (Nawrocki and Eddy 2013). We used tRNA gene copy number as the proxy for tRNA abundance, relying on the previously reported observation that the two correlate strongly (Duret 2000; dos Reis et al. 2004). The adaptiveness of each codon (ω_i) to the tRNA pool was calculated using the stAIcalc program with the maximum hill climbing stringency (Sabi et al. 2017). The tRNA adaptation index (tAI) of each retained gene was calculated as the geometric mean of its ω_i values (dos Reis et al. 2003). Because the distribution of tAI values varies among genomes (LaBella et al. 2019) (**Supplemental Figure S5**), tAI values were converted to their relative percentile tRNA adaptation index (ptAI) within a genome. The ptAI values range between 0 and 100, and represent the percentage of analyzed genes in a genome that have a smaller tAI than a particular gene.

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

The ptAI values were retrieved for a subset of 13 GTA genes that are at least 300 nucleotide in length and are widely detected across all taxonomic groups. The linear regression analysis of ptAI values between all GTA gene pairs was conducted using the phylogenetic generalized least squares method (PGLS) (Martins and Hansen 1997). The reference phylogenomic tree was used to correct for the shared evolutionary history. The analysis was done using the 'caper' package (Orme 2018) and λ, δ and κ parameters were estimated using the maximum likelihood function. Because ptAI values of gene *g9* were not significantly correlated with the ptAI values of 8 out of the 12 other examined GTA genes at p-value cutoff of 0.001 (**Supplemental Table S5**), the gene *g9* was not included into the reference set of GTA genes.

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

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

To identify gene families that exhibit tRNA pool adaptation patterns similar to those of GTA genes, a linear regression model of ptAI values between these gene families and reference GTA genes was fit using the PGLS (Martins and Hansen 1997). The PGLS analysis was carried out using the 'caper' package (Orme 2018) and λ, δ and κ parameters were estimated using the maximum likelihood function. The reference phylogenomic tree was used to correct for the shared phylogenetic history. A gene family was designated to be associated with a GTA, if the obtained fit of the model was statistically significant across all reference GTA genes. If a significantly associated gene family contained paralogs, the PGLS analysis was repeated by using individual ptAI values across all possible combinations of paralogs (if the total number of combinations was < 1000) or across random 1000 combinations of paralogs (if the total number of combinations was > 1000). This was carried out to ensure that the detected signal was not due to sampling associated with selecting the median ptAI value.

Genes with a significant similarity in trend of ptAI values were annotated via eggNOG-mapper v2.1 (Cantalapiedra et al. 2021).

Protein-protein interaction of GTA genes and gene families with similar ptAI values

To identify protein-protein interaction networks, reference GTA genes and genes from families with similar tRNA pool adaptation patterns were retrieved from the *Sphingomonas sp.* MM1 genome, chosen for it being the only genome that contains all genes from the GTA reference gene set and all 14 gene families that are significantly associated with GTAs. The locus tags of the retrieved *Sphingomonas sp.* MM1 genes were used as queries against STRING database v 11.0b (last accessed July 2021) (Szklarczyk et al. 2021) with the medium confidence score cutoff and all active interaction sources. The retrieved protein-protein interaction network was visualized in STRING using the queries and up to 50 additional interactor proteins and displaying edges based on the STRING confidence scores. The KEGG pathways (Kanehisa et al. 2021) enrichment analysis was conducted via hypergeometric testing on the whole retrieved network, as implemented in STRING.

Analysis of other protein-coding genes situated within GTA head-tail clusters

For gene families within GTA head-tail clusters, ptAI values were retrieved and compared to ptAI values of the reference GTA gene set using PGLS analysis as described above.

For the only gene family with a significant association with GTA genes (**Supplemental Table S6**), the secondary structure of its proteins were predicted using Porter 5.0 (Torrisi et al. 2019). To retrieve available viral head completion proteins, the phrase 'head-completion protein' was used as a query against the UniProt database (accessed in August 2021) (UniProt Consortium 2021). Among the 24 manually annotated ("reviewed") matches from the Swiss-Prot sub-database of the UniProt database, only 2 viral matches (accession numbers P68656 and P68660) had length similar to the genes in the above described gene family. Both proteins belong to the λ phage gpW family, and for *Escherichia* phage λ protein 3D structure is available in PDB (Berman et al. 2000). The secondary structure of RcGTA's g7 protein, structural viral homolog of RcGTA's g7 from *Bacillus* phage SPP1 (gp16) (Bardy et al. 2020) and head completion protein of phage λ were retrieved from the PDB database (Berman et al. 2000) in August 2021.

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

Refinement of the *tonB* gene family using phylogenetic tree

To identify orthologs within the large *tonB* gene family, evolutionary history of the *tonB* gene family was reconstructed and evaluated. To do so, amino acid sequences of the *tonB* gene family were aligned using MAFFT-linsi v7.455 (Katoh and Standley 2013). The phylogeny was reconstructed in IQ-TREE v1.6.7 (Nguyen et al. 2015) using the best substitution model (LG+F+R6) detected by ModelFinder (Kalyaanamoorthy et al. 2017). The tree was visualized using the iTOL v6 (Letunic and Bork 2021). The phylogeny was used to subdivide the family into two families, whereas the five genes on very long branches served as an outgroup (**Supplemental Figure S6**).

Calculation of energetic cost associated with production of the encoded proteins

To quantify the energetic cost of proteins, the carbon content of their amino acids was used as a proxy and was calculated by counting the number of carbons in the amino acid side chains, as described in Kogay et al. (2020). The total number of carbons in each protein was normalized by the protein length.

Data access

A list of accession numbers of the GTA regions identified in the analyzed genomes, accession numbers of genes in gene families across analyzed genomes, raw data related to tAI and ENC calculations, slopes and p-values of associations detected in PGLS analyses, multiple sequence alignments and phylogenetic trees of phylogenomic markers dataset and *tonB* gene family dataset, and an in-house script for ENC calculations, are available in the FigShare repository at DOI 10.6084/m9.figshare.20082749.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

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

References

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.

Balkwill DL, Fredrickson JK, Romine MF. 2006. Sphingomonas and Related Genera. In *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass*, doi:10.1007/0-387-30747-8_23 (ed. M Dworkin, et al.), pp. 605-629. Springer New York, New York, NY.

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.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**: 235-242.

Blanvillain S, Meyer D, Boulanger A, Lautier M, Guynet C, Denance N, Vasse J, Lauber E, Arlat M. 2007. Plant carbohydrate scavenging through tonB-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS One* **2**: e224.

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.

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.

Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**: 9077-9096.

dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**: 5036-5044.

dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res* **31**: 6976-6985.

Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**: 287-289.

Eisenbeis S, Lohmiller S, Valdebenito M, Leicht S, Braun V. 2008. NagA-dependent uptake of N-acetyl-glucosamine and N-acetyl-chitin oligosaccharides across the outer membrane of Caulobacter crescentus. *J Bacteriol* **190**: 5230-5238.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**: 238.

Fogg PC, Westbye AB, Beatty JT. 2012. One for all or all for one: heterogeneous expression and host cell lysis are key to gene transfer agent activity in Rhodobacter capsulatus. *PLoS One* **7**: e43772.

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

Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A* **101**: 9033-9038.

Genest O, Wickner S, Doyle SM. 2019. Hsp90 and Hsp70 chaperones: Collaborators in protein remodeling. *J Biol Chem* **294**: 2109-2120.

Gershenzon J, Dudareva N. 2007. The function of terpene natural products in the natural world. *Nat Chem Biol* **3**: 408-414.

Hynes AP, Mercer RG, Watton DE, Buckley CB, Lang AS. 2012. DNA packaging bias and differential expression of gene transfer agent genes within a population during production and release of the Rhodobacter capsulatus gene transfer agent, RcGTA. *Mol Microbiol* **85**: 314-325.

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.

Jung H, Liang J, Jung Y, Lim D. 2015. Characterization of cell death in Escherichia coli mediated by XseA, a large subunit of exonuclease VII. *J Microbiol* **53**: 820-828.

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, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* **49**: D545-D551.

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

Kogay R, Koppenhofer S, Beatty JT, Lang AS, Kuhn JH, Zhaxybayeva O. 2022. Formal Recognition and Classification of Gene Transfer Agents as Viriforms. *bioRxiv* doi:10.1101/2022.06.10.494566: 2022.2006.2010.494566.

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.

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

LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2019. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet* **15**: e1008304.

LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2021. Signatures of optimal codon usage in metabolic genes inform budding yeast ecology. *PLoS Biol* **19**: e3001185.

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, 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.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**: W293-W296.

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.

Lim HN, Lee Y, Hussein R. 2011. Fundamental relationship between operon organization and gene expression. *Proc Natl Acad Sci U S A* **108**: 10626-10631.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955-964.

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.

Marsin S, Lopes A, Mathieu A, Dizet E, Orillard E, Guerois R, Radicella JP. 2010. Genetic dissection of Helicobacter pylori AddAB role in homologous recombination. *FEMS Microbiol Lett* **311**: 44-50.

Martins EP, Hansen TF. 1997. Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. *The American Naturalist* **149**: 646 - 667.

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.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933-2935.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268-274.

Nishizaki T, Tsuge K, Itaya M, Doi N, Yanagawa H. 2007. Metabolic engineering of carotenoid biosynthesis in Escherichia coli by ordered gene assembly in Bacillus subtilis. *Appl Environ Microbiol* **73**: 1355-1361.

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.

Orme D. 2018. The caper package: comparative analysis of phylogenetics and evolution in R. .

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32-42.

Quax TE, Claassens NJ, Soll D, van der Oost J. 2015. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell* **59**: 149-161.

Quebatte M, Christen M, Harms A, Korner 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.

Ram S, Mitra M, Shah F, Tirkey SR, Mishra S. 2020. Bacteria as an alternate biofactory for carotenoid production: A review of its applications, opportunities and challenges. *Journal of Functional Foods* **67**: 103867.

Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* **14**: 2279-2286.

Rockabrand D, Livers K, Austin T, Kaiser R, Jensen D, Burgess R, Blum P. 1998. Roles of DnaK and RpoS in starvation-induced thermotolerance of Escherichia coli. *J Bacteriol* **180**: 846-854.

Roller M, Lucic V, Nagy I, Perica T, Vlahovicek K. 2013. Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Res* **41**: 8842-8852.

Sabi R, Volvovitch Daniel R, Tuller T. 2017. stAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* **33**: 589-591.

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

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**: 472-482.

Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. 2010. Translational selection is ubiquitous in prokaryotes. *PLoS Genet* **6**: e1001004.

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.

Tang K, Jiao N, Liu K, Zhang Y, Li S. 2012. Distribution and functions of TonB-dependent transporters in marine bacteria and environments: implications for dissolved organic matter utilization. *PLoS One* **7**: e41204.

Torrisi M, Kaleel M, Pollastri G. 2019. Deeper Profiles and Cascaded Recurrent and Convolutional Neural Networks for state-of-the-art Protein Secondary Structure Prediction. *Sci Rep* **9**: 12374.

Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. 2017. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends Microbiol* **25**: 280-292.

UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**: D480-D489.

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.

Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87**: 23-29.

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

Yang Y, Liu B, Du X, Li P, Liang B, Cheng X, Du L, Huang D, Wang L, Wang S. 2015. Complete genome sequence and transcriptomics analyses reveal pigment biosynthesis and regulatory mechanisms in an industrial strain, Monascus purpureus YY-1. *Sci Rep* **5**: 8331.

Zhan Y, Huang S, Voget S, Simon M, Chen F. 2016. A novel roseobacter phage possesses features of podoviruses, siphoviruses, prophages and gene transfer agents. *Sci Rep* **6**: 30372.

Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, Liu Y. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A* **113**: E6117-E6125.

Tables

Table 1. Functional annotations of 14 gene families, whose ptAI values have a significantly similar trend to ptAI values of the reference GTA genes.

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

Figures

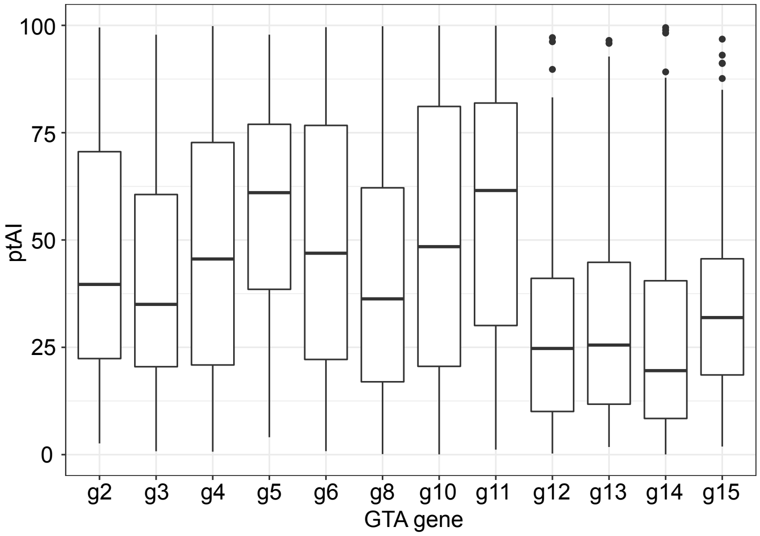


Figure 1. Distribution of ptAI values among reference GTA genes from GTA head-tail clusters in 208 alphaproteobacterial genomes. Line within a box displays the median ptAI value for a GTA gene across all genomes, in which the gene was detected. The boxes are bounded by first and third quartiles. Whiskers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

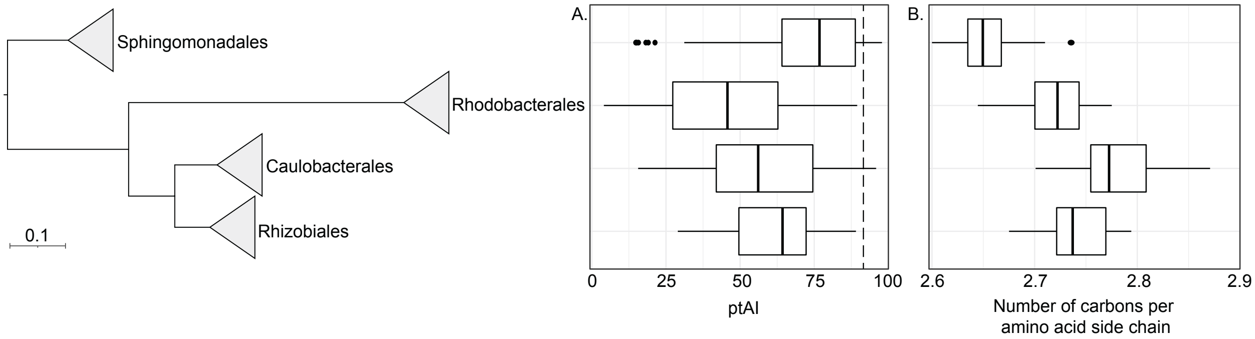


Figure 2. Distributions of (A) ptAI values in major capsid protein-encoding gene (*g5*) and (B) carbon content of amino acids in the g5 protein across four orders of the class *Alphaproteobacteria*. On both panels A and B, line within a box displays the median ptAI value for g5 representatives within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

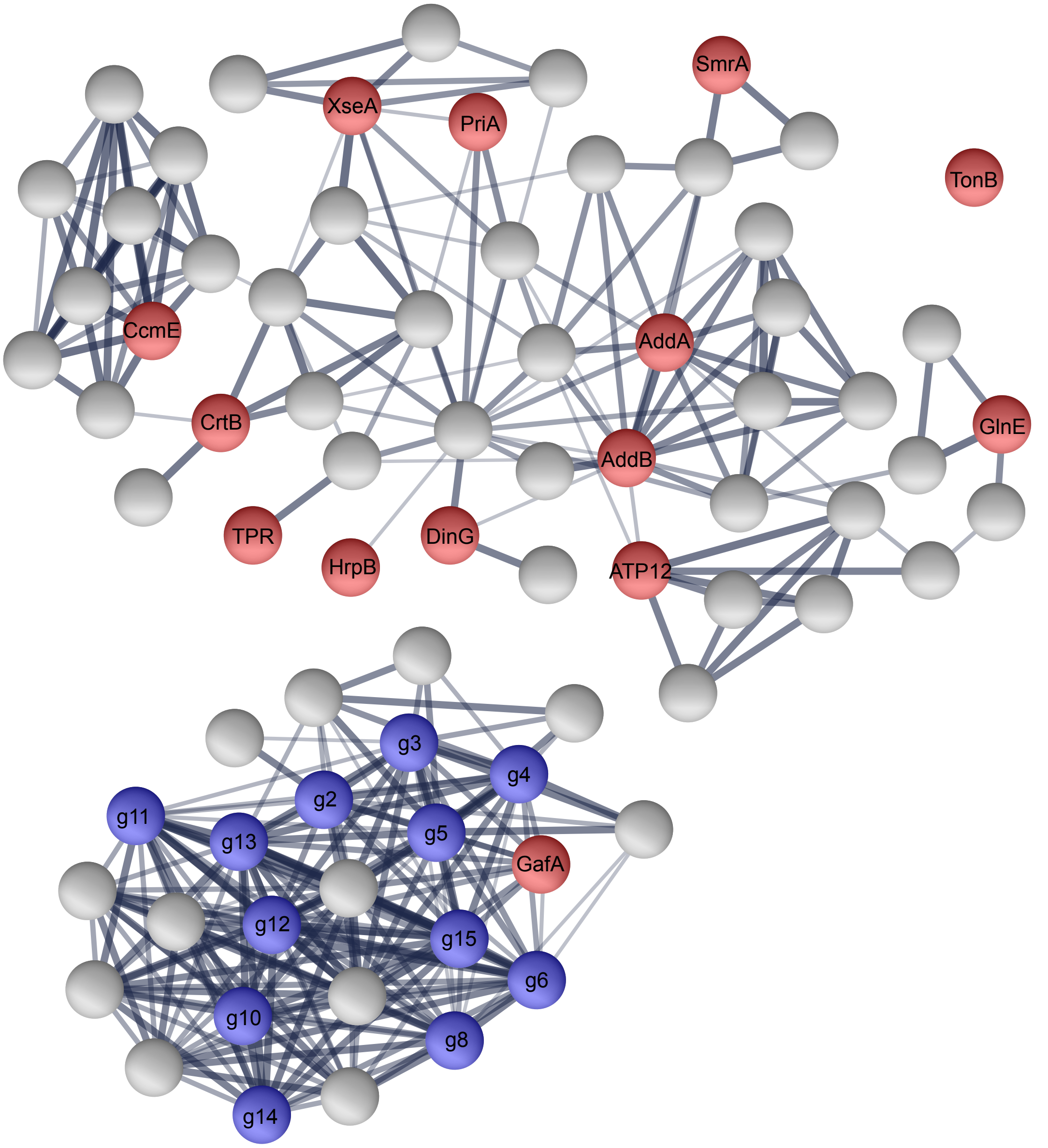


Figure 3. The protein-protein interactions among 12 GTA reference proteins and 14 proteins putatively co-expressed with GTAs. Nodes represent individual proteins. Blue-colored nodes correspond to GTA reference proteins and red-colored nodes correspond to 14 putatively co-expressed proteins. Gray-colored nodes represent additional proteins found through the STRING functional enrichment analysis. The thickness of the edges is proportional to the confidence score (varying between 0.4 and 1.0) of protein interactions reported by STRING.

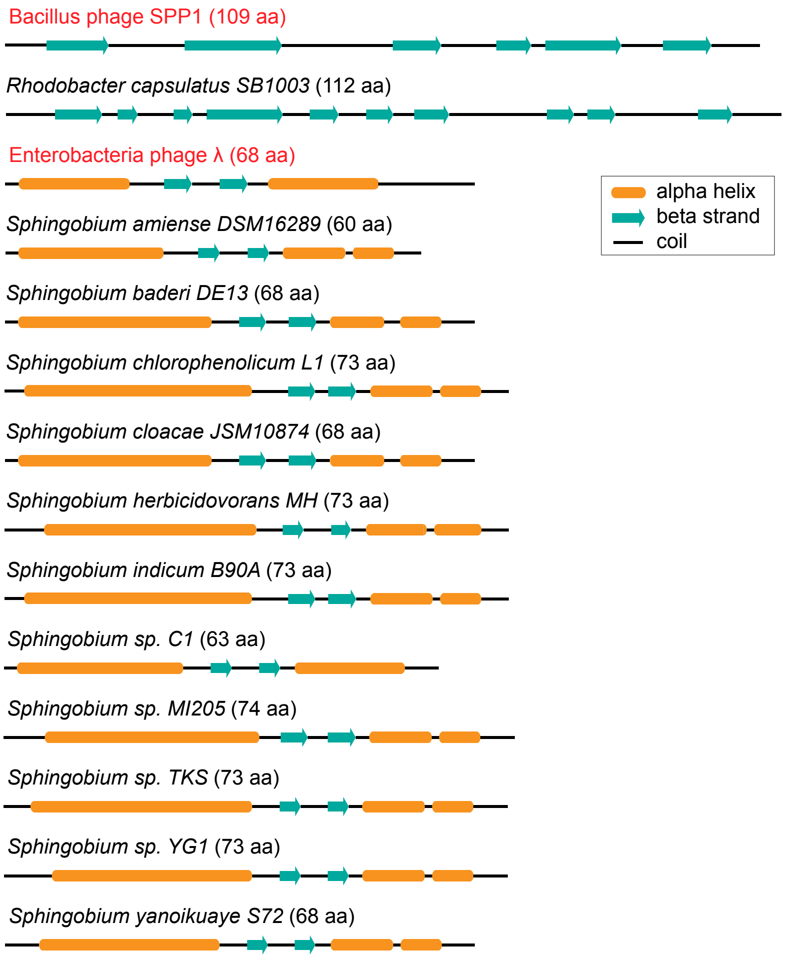

Bacillus phage SPP1 (109 aa)
Rhodobacter capsulatus SB1003 (112 aa)
Enterobacteria phage λ (68 aa)
Sphingobium amiense DSM16289 (60 aa)
Sphingobium baderi DE13 (68 aa)
Sphingobium chlorophenolicum L1 (73 aa)
Sphingobium cloacae JSM10874 (68 aa)
Sphingobium herbicidovorans MH (73 aa)
Sphingobium indicum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

Figure 4. Secondary structures of head completion proteins from phages and GTAs. The Enterobacteria phage lambda gpW and Bacillus phage SPP1 (highlighted in red) are two representatives of viral head completion proteins with major differences in secondary structures and proteins lengths. The secondary protein structures of RcGTA g7 (PDB ID 6TUI_8), Enterobacteria phage lambda gpW (1HYW), and Bacillus phage SPP1 (2KCA) were retrieved from the PDB database. The secondary protein structures of the putative head completion proteins from *Sphingomonadales* were predicted computationally. The secondary structures are scaled with respect to the protein lengths, which are listed in parentheses next to the taxonomic name.

Supplemental Material

for

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

by

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

Table of Contents:

Supplemental Tables S1-S6

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

KEGG Pathway ID	Pathway name	p-value (after Benjamini-Hochberg correction)
sphm00900	Terpenoid backbone biosynthesis	0.0102
sphm03440	Homologous recombination	0.0119
sphm00906	Carotenoid biosynthesis	0.0202
sphm03430	Mismatch repair	0.0349

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

Reference GTA gene	p-value	Slope
g2	0.82086	-0.03579
g3	0.59683	0.08254
g4	0.40305	0.13635
g5	0.21417	0.15219
g6	0.01450	0.41291
g8	0.83146	0.03253
g10	0.50256	0.17411
g11	0.00004	0.10553
g12	0.03593	0.25124
g13	0.85847	0.02868
g14	0.00170	0.59603
g15	0.00002	0.13779

Supplemental Table S3. List of selected 208 alphaproteobacterial genomes with GTA 'head-tail' clusters.

Genome name	Assembly ID
Agrobacterium fabrum strain 12D13	GCF_003667945.1
Agrobacterium rhizogenes strain K599	GCF_002005205.3
Agrobacterium sp RAC06	GCF_001713475.1
Agrobacterium tumefaciens strain 12D1	GCF_003667905.1
Agrobacterium tumefaciens strain 1D1108	GCF_003666425.1
Agrobacterium tumefaciens strain 1D1609	GCF_002943835.1
Agrobacterium vitis S4	GCF_000016285.1
Altererythrobacter dongtanensis strain KCTC 22672	GCF_001698205.1
Altererythrobacter mangrovi strain C9 11	GCF_002269345.1
Altererythrobacter marensis strain KCTC 22370	GCF_001028625.1
Altererythrobacter namhicola strain JCM 16345	GCF_001687545.1
Altererythrobacter sp B11	GCF_003569745.1
Antarctobacter heliothermus strain SMS3	GCF_002237555.1
Asticcacaulis excentricus M6	GCF_003966695.1
Aureimonas sp AU20	GCF_001442755.1
Azorhizobium caulinodans ORS 571	GCF_000010525.1
Blastochloris sp GI	GCF_003966715.1
Blastochloris viridis strain ATCC 19567	GCF_001402875.1
Blastomonas sp RAC04	GCF_001713435.1
Bosea sp AS 1	GCF_002220095.1
Bosea sp PAMC 26642	GCF_001562255.1
Bosea sp RAC05	GCF_001713455.1
Bosea sp Tri 49	GCF_003952665.1
Breoghania sp LA4	GCF_003432385.1
Brevundimonas naejangsanensis strain B1	GCF_000635915.2
Brevundimonas naejangsanensis strain BRV3	GCF_003627995.1
Brevundimonas sp DS20	GCF_001310255.1
Brevundimonas sp GW460	GCF_001636925.1
Brevundimonas sp LM2	GCF_002002865.1
Brevundimonas vesicularis strain FDAARGOS 289	GCF_002208825.2
Brucella canis strain 2009004498	GCF_001715365.1
Candidatus Filomicrobium marinum strain W	GCF_000981565.1
Caulobacter flavus strain RHGG3	GCF_003722335.1
Caulobacter henricii strain CB4	GCF_001414055.1
Caulobacter mirabilis strain FWC 38	GCF_002749615.1
Caulobacter segnis strain TK0059	GCF_003015125.1
Caulobacter sp K31	GCF_000019145.1
Caulobacter vibrioides strain CB2	GCF_002310295.2

Celeribacter ethanolicus strain TSPH2	GCF_002407265.1
Celeribacter indicus strain P73	GCF_000819565.1
Celeribacter manganoxidans strain DY25	GCF_002504165.1
Celeribacter marinus strain IMCC 12053	GCF_001308265.1
Chelatococcus daeguensis strain TAD1	GCF_001887265.1
Chelatococcus sp CO 6	GCF_001271345.1
Citromicrobium sp JL477	GCF_001304795.1
Croceicoccus naphthovorans strain PQ2	GCF_001028705.1
Devosia sp H5989	GCF_001185205.1
Dinoroseobacter shibae DFL 12	GCF_000018145.1
Erythrobacter atlanticus strain s21	GCF_001077815.2
Erythrobacter gangjinensis strain CGMCC115024	GCF_001886695.1
Erythrobacter litoralis HTCC2594	GCF_000013005.1
Erythrobacter seohaensis strain SW135	GCF_002795865.1
Erythrobacter sp YH 07	GCF_003355455.1
Gemmobacter sp HYN0069	GCF_003060865.1
Hartmannibacter diazotrophicus strain E19T	GCF_900231165.1
Hirschia baltica ATCC 49814	GCF_000023785.1
Hoeflea sp IMCC20628	GCF_001011155.1
Hyphomicrobium denitrificans 1NES1	GCF_000230975.2
Hyphomicrobium denitrificans ATCC 51888	GCF_000143145.1
Hyphomicrobium nitrativorans NL23	GCF_000503895.1
Hyphomicrobium sp MC1	GCF_000253295.1
Jannaschia sp CCS1	GCF_000013565.1
Ketogulonicigenium robustum strain B003	GCF_002117445.1
Ketogulonicigenium vulgare strain SKV	GCF_001693655.1
Leisingera methylohalidivorans DSM 14336	GCF_000511355.1
Loktanella vestfoldensis strain SMR4r	GCF_002158905.1
Maricaulis maris MCS10	GCF_000014745.1
Maritalea myrionectae strain HL27085	GCF_003433515.1
Marivivens sp JLT3646	GCF_001908835.1
Martelella endophytica strain YC6887	GCF_000960975.1
Martelella mediterranea DSM 17316	GCF_002043005.1
Martelella sp AD 3	GCF_001578105.1
Methylobacterium extorquens AM1	GCF_000022685.1
Methylobacterium phyllosphaerae strain CBMB27	GCF_001936175.1
Methylobacterium populi	GCF_002355515.1
Methylobacterium radiotolerans JCM 2831	GCF_000019725.1
Methylobacterium sp 17SD2 17	GCF_003173715.1
Methylobacterium sp AMS5	GCF_001542815.1
Methylobacterium sp DM1	GCF_003111705.1
Methyloceanibacter caenitepidi Gela4	GCF_000828475.1

Methylocystis rosea strain GW6	GCF_003855495.1
Methylocystis sp SC2	GCF_000304315.1
Methylosinus trichosporium OB3b	GCF_002752655.1
Microvirga ossetica strain V5	GCF_002741015.1
Microvirga sp 17	GCF_003151255.1
Neorhizobium galegae	GCF_000731295.1
Neorhizobium sp NCHU2750	GCF_003597675.1
Nitratireductor basaltis strain RR3 28	GCF_001953055.1
Nitrobacter hamburgensis X14	GCF_000013885.1
Nitrobacter winogradskyi Nb 255	GCF_000012725.1
Novosphingobium aromaticivorans DSM 12444	GCF_000013325.1
Novosphingobium resinovorum strain SA1	GCF_001742225.1
Novosphingobium sp P6W	GCF_000876675.2
Novosphingobium sp PP1Y	GCF_000253255.1
Ochrobactrum pituitosum strain AA2	GCF_002025625.1
Ochrobactrum sp A44	GCF_002278035.1
Octadecabacter antarcticus 307	GCF_000155675.2
Octadecabacter arcticus 238	GCF_000155735.2
Octadecabacter temperatus strain SB1	GCF_001187845.1
Oligotropha carboxidovorans OM4	GCF_000218585.1
Pannonibacter phragmitetus BB	GCF_003574985.1
Paracoccus aminophilus JCM 7686	GCF_000444995.1
Paracoccus aminovorans isolate JCM7685	GCF_900005615.1
Paracoccus contaminans	GCF_002105555.1
Paracoccus denitrificans PD1222	GCF_000203895.1
Paracoccus sp BM15	GCF_002847305.1
Paracoccus sp CBA4604	GCF_002865605.1
Paracoccus sp SC2 6	GCF_003324675.1
Paracoccus yeei strain TT13	GCF_002749495.1
Paracoccus zhejiangensis strain J6	GCF_002847445.1
Parvibaculum lavamentivorans DS 1	GCF_000017565.1
Parvularcula bermudensis HTCC2503	GCF_000152825.2
Pelagibaca abyssi strain JLT2014	GCF_001975705.1
Pelagibacterium halotolerans B2	GCF_000230555.1
Phaeobacter gallaeciensis strain JL2886	GCF_001678945.1
Phaeobacter gallaeciensis strain P63	GCF_002393525.1
Phaeobacter inhibens strain P70	GCF_002892125.1
Phaeobacter piscinae strain P13	GCF_002412045.1
Phaeobacter porticola strain P97	GCF_001888185.1
Phreatobacter sp S12	GCF_003008515.1
Pleomorphomonas sp SM30	GCF_003966995.1
Polymorphum gilvum SL003B 26A1	GCF_000192745.1

Porphyrobacter HT 58	GCF_002952215.1
Pseudorhodoplanes sinuspersici strain RIPI110	GCF_002119765.1
Rhizobium sp ACO	GCF_002600635.1
Rhizobium sp NT 26	GCF_000967425.1
Rhizobium sp Y9	GCF_002814035.1
Rhodobaca barguzinensis strain alga05	GCF_001870665.2
Rhodobacter blasticus strain 285	GCF_003071405.1
Rhodobacter capsulatus SB 1003	GCF_000021865.1
Rhodobacter sp CZR27	GCF_002407205.1
Rhodobacter sp LPB0142	GCF_001856665.1
Rhodobacter sphaeroides 241	GCF_003324715.1
Rhodobacter sphaeroides strain EBL0706	GCF_003429265.1
Rhodobacteraceae bacterium strain G7	GCF_002850435.1
Rhodobiaceae bacterium strain SMS8	GCF_003330885.1
Rhodopseudomonas palustris BisA53	GCF_000014825.1
Rhodopseudomonas palustris BisB18	GCF_000013745.1
Rhodopseudomonas palustris BisB5	GCF_000013685.1
Rhodopseudomonas palustris CGA009	GCF_000195775.1
Rhodopseudomonas palustris DX 1	GCF_000177255.2
Rhodopseudomonas palustris HaA2	GCF_000013365.1
Rhodopseudomonas palustris strain YSC3	GCF_003031245.1
Rhodovulum sp MB263	GCF_002073975.1
Rhodovulum sp P5	GCF_002079305.1
Rhodovulum sulfidophilum strain SNK001	GCF_001633145.1
Roseibacterium elongatum DSM 19469	GCF_000590925.1
Roseobacter denitrificans strain FDAARGOS 309	GCF_002983865.1
Roseobacter litoralis Och 149	GCF_000154785.2
Roseovarius sp AK1035	GCF_003288315.1
Ruegeria mobilis F1926	GCF_000376545.2
Ruegeria pomeroyi DSS 3	GCF_000011965.2
Ruegeria sp AD91A	GCF_003443535.1
Ruegeria sp TM1040	GCF_000014065.1
Shinella sp HZN7	GCF_001652565.1
Sinorhizobium sp RAC02	GCF_001713395.1
Sphingobium amiense DSM 16289	GCF_003967075.1
Sphingobium baderi strain DE 13	GCF_001456115.1
Sphingobium chlorophenolicum L 1	GCF_000147835.2
Sphingobium cloacae JCM 10874	GCF_002355855.1
Sphingobium herbicidovorans strain MH	GCF_002080435.1
Sphingobium indicum B90A	GCF_000264945.2
Sphingobium sp C1	GCF_002288285.1
Sphingobium sp EP60837	GCF_001658005.1

Sphingobium sp MI1205	GCF_001563285.1
Sphingobium sp RAC03	GCF_001713415.1
Sphingobium sp SYK 6	GCF_000283515.1
Sphingobium sp TKS	GCF_001563265.1
Sphingobium sp YBL2	GCF_000943805.1
Sphingobium sp YG1	GCF_003609795.1
Sphingobium yanoikuyae strain S72	GCF_002504085.1
Sphingomonas hengshuiensis strain WHSC 8	GCF_000935025.1
Sphingomonas melonis strain ZJ26	GCF_002504265.1
Sphingomonas panacis strain DCY99	GCF_001717955.1
Sphingomonas sp Cra20	GCF_002796605.1
Sphingomonas sp FARSPH	GCF_003355005.1
Sphingomonas sp KC8	GCF_002151445.1
Sphingomonas sp LK11	GCF_001971605.1
Sphingomonas sp LM7	GCF_002002925.1
Sphingomonas sp MM 1	GCF_000347675.2
Sphingomonas taxi strain ATCC 55669	GCF_000764535.1
Sphingopyxis fribergensis strain Kp52	GCF_000803645.1
Sphingopyxis granuli strain TFA	GCF_001559015.1
Sphingopyxis macrogoltabida strain 203	GCF_001314325.1
Sphingopyxis macrogoltabida strain EY 1	GCF_001307295.1
Sphingopyxis sp FD7	GCF_003609835.1
Sphingopyxis sp LPB0140	GCF_001889025.1
Sphingopyxis sp QXT 31	GCF_001984035.1
Sphingorhabdus flavimaris strain SMR4y	GCF_002218195.1
Sphingorhabdus sp Alg231	GCF_900149705.1
Sphingorhabdus sp M41	GCF_001586275.1
Sphingorhabdus sp YGSMI21	GCF_002776575.1
Sphingosinicella microcystinivorans B9	GCF_003967095.1
Starkeya novella DSM 506	GCF_000092925.1
Sulfitobacter pseudonitzschiae strain SMR1	GCF_002222635.1
Sulfitobacter sp AM1 D1	GCF_001886735.1
Sulfitobacter sp D7	GCF_003611275.1
Sulfitobacter sp SK012	GCF_003352085.1
Sulfitobacter sp SK025	GCF_003352105.1
Tabrizicola sp K13M18	GCF_003940805.1
Tateyamaria omphalii strain DOK1 4	GCF_001969365.1
Thalassococcus sp SH 1	GCF_003008555.1
Thiobacimonas profunda strain JLT2016	GCF_001969385.1
Thioclava nitratireducens strain 25B10	GCF_001940525.2
Variibacter gotjawalensis	GCF_002355335.1
Xanthobacter autotrophicus Py2	GCF_000017645.1

Yangia pacifica strain YSBP01	GCF_003111685.1
Yangia sp CCB	GCF_001687105.1

Supplemental Table S4. Decisions behind choosing the RcGTA 'head-tail' cluster homologs for the reference GTA gene set. The GTA genes selected for the reference set are highlighted in orange. Functional annotations are based on the descriptions in the RefSeq database records, unless noted otherwise.

GTA gene	RcGTA RefSeq ID	RcGTA functional annotation	Included?	Reason for a gene exclusion
g1	WP_013067406.1	small terminase [1]		Homologs are detected only in *Rhodobacterales* order
g2	WP_031321187.1	terminase family protein	Yes	
g3	WP_013067408.1	phage portal protein	Yes	
g3.5	WP_031323538.1	hypothetical protein		The gene is < 300 nucleotides in length
g4	WP_013067410.1	HK97 family phage prohead protease	Yes	
g5	WP_037091462.1	phage major capsid protein	Yes	
g6	WP_013067412.1	adaptor protein [2]	Yes	
g7	WP_013067413.1	head-tail adaptor protein		Detected only in 10% of *Sphingomonadales* genomes
g8	WP_013067414.1	tail terminator protein [2]	Yes	
g9	WP_013067415.1	phage major tail protein, TP901-1 family		Patterns of the selection for translational efficiency are inconsistent with those of other GTA genes
g10	WP_013067416.1	gene transfer agent family protein	Yes	
g10.1	WP_013067417.1	phage tail assembly chaperone		The gene is < 300 nucleotides in length
g11	WP_013067418.1	phage tail tape measure protein	Yes	
g12	WP_013067419.1	distal tail protein [2]	Yes	
g13	WP_013067420.1	baseplate hub protein [2]	Yes	
g14	WP_013067421.1	peptidase	Yes	
g15	WP_013067422.1	glycoside hydrolase/phage tail family protein	Yes	

1. Sherlock, D., J.X. Leong, and P.C.M. Fogg, *Identification of the First Gene Transfer Agent (GTA) Small Terminase in Rhodobacter capsulatus and Its Role in GTA Production and Packaging of DNA.* J Virol, 2019. **93**: e01328-19.
2. Bardy, P., et al., *Structure and mechanism of DNA delivery of a gene transfer agent.* Nat Commun, 2020. **11**: 3034.

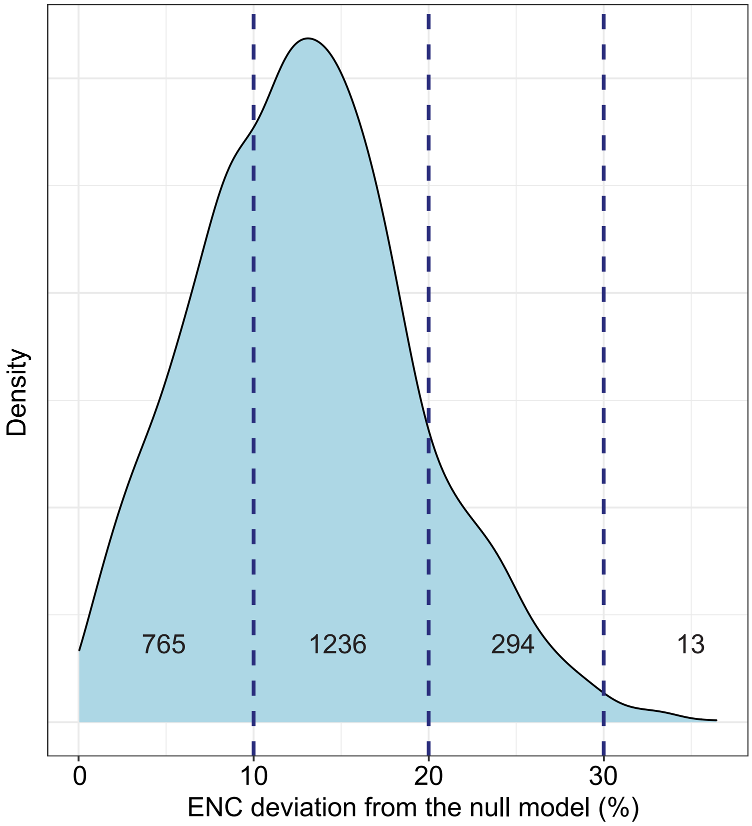
Supplemental Table S5. Significance and slope of the fit of the phylogenetic generalized least squares (PGLS) models between the reference GTA genes and the *g9* gene.

Reference GTA gene	p-value	Slope
g2	0.01287	0.18307
g3	0.00145	0.20857
g4	0.00099	0.25720
g5	2.23E-06	0.28841
g6	0.00511	0.23485
g8	0.00645	0.20851
g10	0.12627	0.08297
g11	0.00061	0.21435
g12	0.00031	0.31015
g13	0.00162	0.26218
g14	0.03469	0.14122
g15	0.00138	0.27938

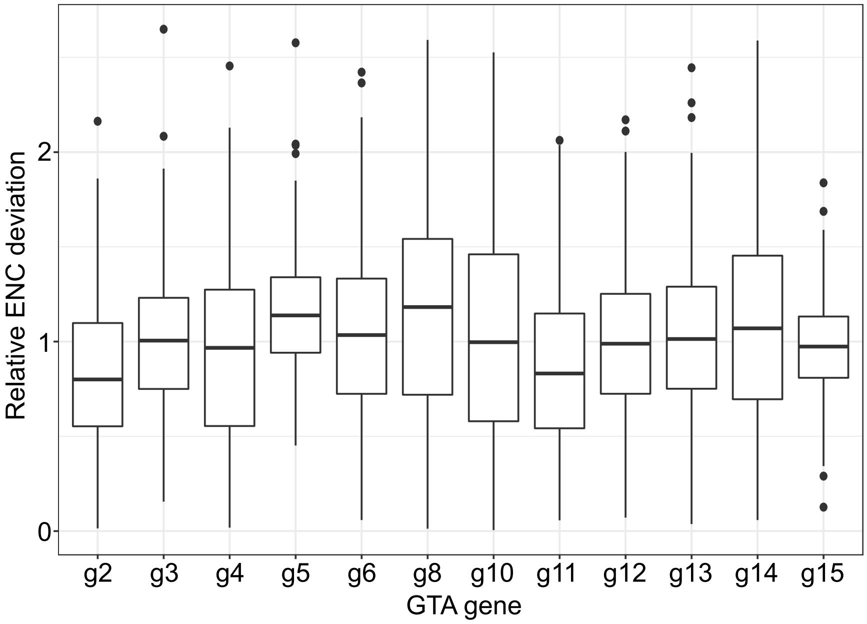
Supplemental Table S6. GenBank accession numbers of the putative g7 protein found in 11 *Sphingomonadales* genomes.

Organism name	RefSeq ID
Sphingobium chlorophenolicum L 1	WP_041390235.1
Sphingobium cloacae JCM 10874	WP_096362181.1
Sphingobium sp YG1	WP_120250369.1
Sphingobium amiense DSM 16289	WP_174522205.1
Sphingobium sp TKS	WP_082748471.1
Sphingobium sp MI1205	WP_083535884.1
Sphingobium indicum B90A	WP_007684091.1
Sphingobium baderi strain DE 13	WP_156415295.1
Sphingobium herbicidovorans strain MH	WP_037467270.1
Sphingobium sp C1	WP_017184510.1
Sphingobium yanoikuyae strain S72	WP_097384099.1

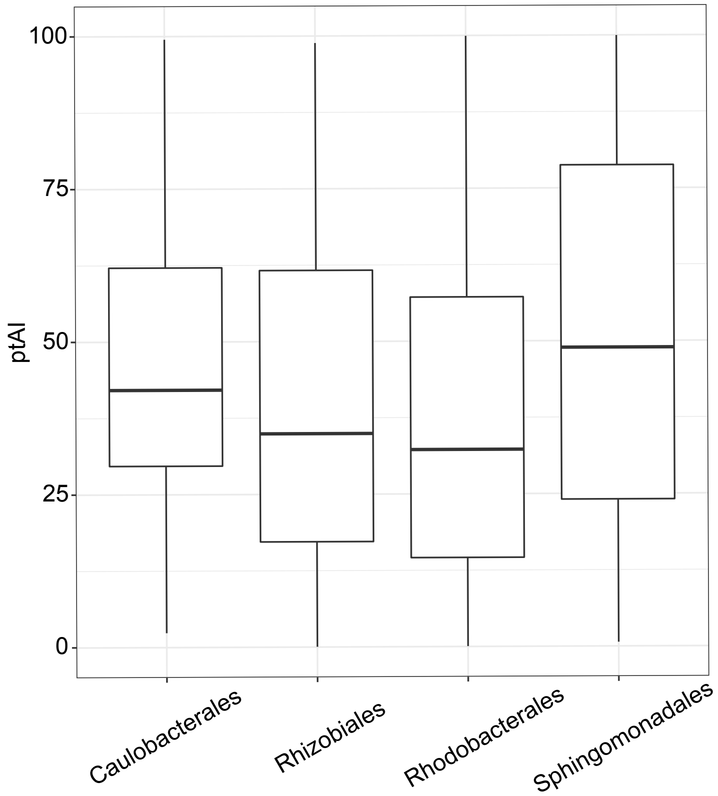
Supplemental Figures S1-S6



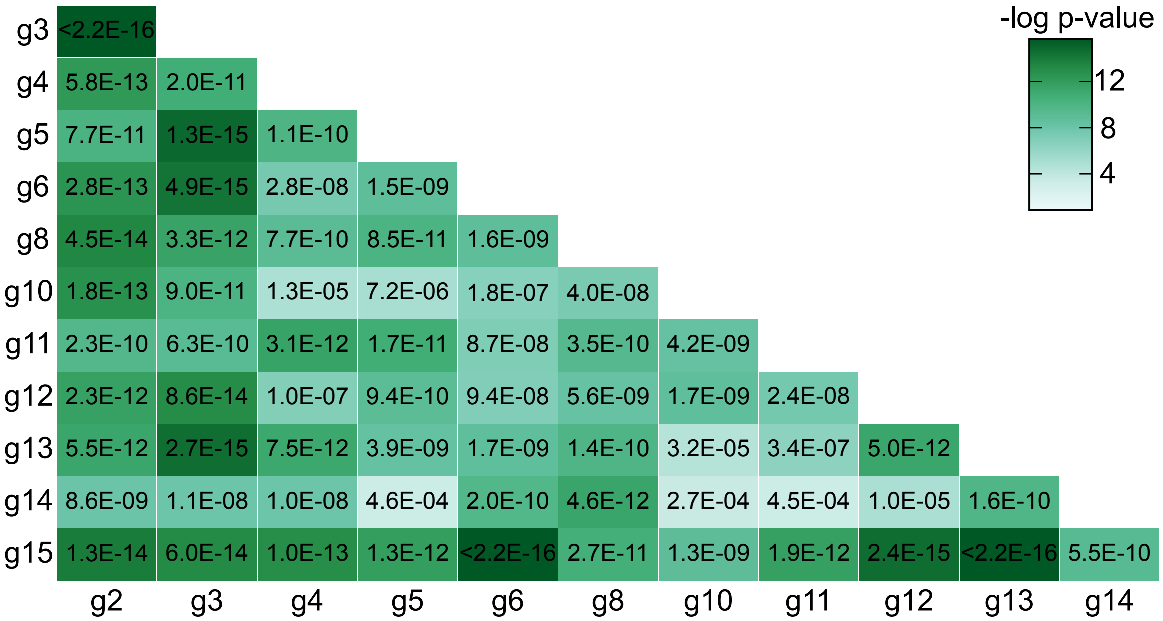
Supplemental Figure S1. Distribution of deviations of the effective number of codon (ENC) values from the expected ENC values under the null model of no codon bias. The distribution contains deviations for 2,308 reference GTA genes found in the 208 genomes. Numbers on the plot designate the number of reference GTA genes in an interval delineated by dashed lines.



Supplemental Figure S2. Deviation of the effective number of codon (ENC) values for individual reference GTA genes in comparison to the genomic average. The deviation of the ENC from the expectation under the null model for each GTA gene was normalized by the average ENC deviation of its genome. Line within a box displays the median normalized ENC value for a GTA gene across all genomes. The boxes are bounded by first and third quartiles. Whiskers represent ptAI values within 1.5*interquartile range. Dots outside of whiskers are outliers.

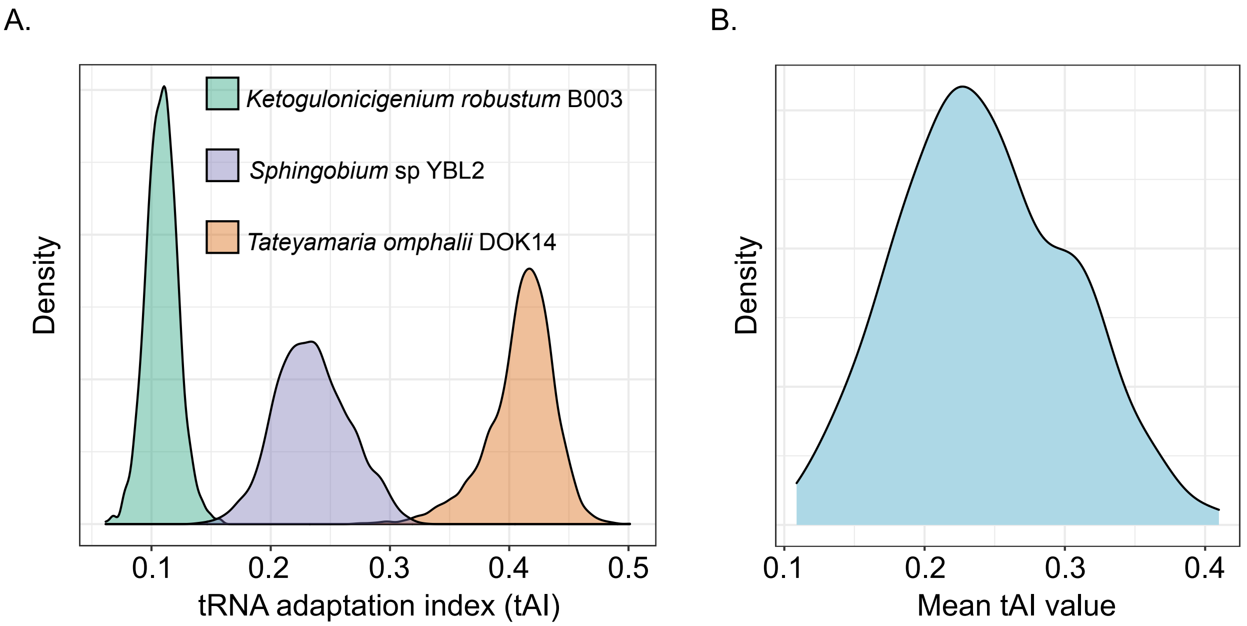


Supplemental Figure S3. Distributions of ptAI values in all reference GTA genes across four orders of the class *Alphaproteobacteria*. Line within a box displays the median ptAI value for a GTA gene across all genomes. The boxes are bounded by first and third quartiles. Whiskers represent ptAI values within 1.5*interquartile range. Dots outside of whiskers are outliers.

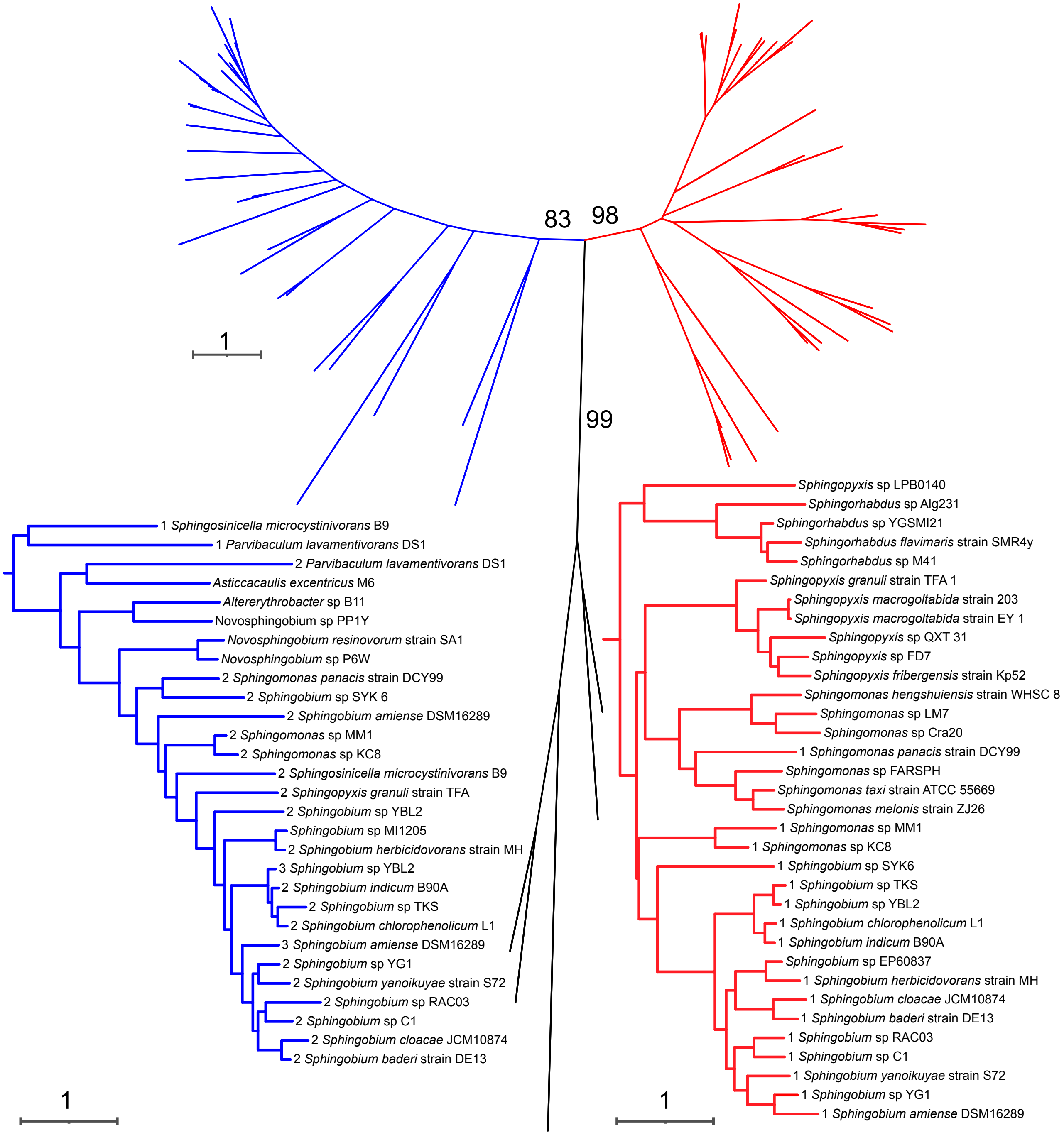


	g2	g3	g4	g5	g6	g8	g10	g11	g12	g13	g14
g3	<2.2E-16										
g4	5.8E-13	2.0E-11									
g5	7.7E-11	1.3E-15	1.1E-10								
g6	2.8E-13	4.9E-15	2.8E-08	1.5E-09							
g8	4.5E-14	3.3E-12	7.7E-10	8.5E-11	1.6E-09						
g10	1.8E-13	9.0E-11	1.3E-05	7.2E-06	1.8E-07	4.0E-08					
g11	2.3E-10	6.3E-10	3.1E-12	1.7E-11	8.7E-08	3.5E-10	4.2E-09				
g12	2.3E-12	8.6E-14	1.0E-07	9.4E-10	9.4E-08	5.6E-09	1.7E-09	2.4E-08			
g13	5.5E-12	2.7E-15	7.5E-12	3.9E-09	1.7E-09	1.4E-10	3.2E-05	3.4E-07	5.0E-12		
g14	8.6E-09	1.1E-08	1.0E-08	4.6E-04	2.0E-10	4.6E-12	2.7E-04	4.5E-04	1.0E-05	1.6E-10	
g15	1.3E-14	6.0E-14	1.0E-13	1.3E-12	<2.2E-16	2.7E-11	1.3E-09	1.9E-12	2.4E-15	<2.2E-16	5.5E-10

Supplemental Figure S4. PGLS model fit among ptAI values of the reference GTA gene pairs. Each pairwise comparison is represented by a rectangle that is color-coded according to the p-values from the PGLS analysis of the reference GTA gene pairs. The numerical p-values are listed within each rectangle.



Supplemental Figure S5. Distribution of tAI values in protein-coding genes of the analyzed genomes. Only genes at least 300 nucleotides in length were included. **A.** Distribution of tAI values of genes in three representative alphaproteobacterial genomes, selected to have the lowest, the median, and the highest mean tAI value among 208 genomes. **B.** Distribution of the average genomic tAI values across 208 genomes.


83
98
1
99
1 Sphingosinicella microcystinivorans B9
1 Parvibaculum lavamentivorans DS1
2 Parvibaculum lavamentivorans DS1
Asticcacaulis excentricus M6
Altererythrobacter sp B11
Novosphingobium sp PP1Y
Novosphingobium resinovorum strain SA1
Novosphingobium sp P6W
2 Sphingomonas panacis strain DCY99
2 Sphingobium sp SYK 6
2 Sphingobium amiense DSM16289
2 Sphingomonas sp MM1
2 Sphingomonas sp KC8
2 Sphingosinicella microcystinivorans B9
2 Sphingopyxis granuli strain TFA
2 Sphingobium sp YBL2
Sphingobium sp MI1205
2 Sphingobium herbicidovorans strain MH
3 Sphingobium sp YBL2
2 Sphingobium indicum B90A
2 Sphingobium sp TKS
2 Sphingobium chlorophenolicum L1
3 Sphingobium amiense DSM16289
2 Sphingobium sp YG1
2 Sphingobium yanoikuyae strain S72
2 Sphingobium sp RAC03
2 Sphingobium sp C1
2 Sphingobium cloacae JCM10874
2 Sphingobium baderi strain DE13
1
Sphingopyxis sp LPB0140
Sphingorhabdus sp Alg231
Sphingorhabdus sp YGSMI21
Sphingorhabdus flavimaris strain SMR4y
Sphingorhabdus sp M41
Sphingopyxis granuli strain TFA 1
Sphingopyxis macrogoltabida strain 203
Sphingopyxis macrogoltabida strain EY 1
Sphingopyxis sp QXT 31
Sphingopyxis sp FD7
Sphingopyxis fribergensis strain Kp52
Sphingomonas hengshuiensis strain WHSC 8
Sphingomonas sp LM7
Sphingomonas sp Cra20
1 Sphingomonas panacis strain DCY99
Sphingomonas sp FARSPH
Sphingomonas taxi strain ATCC 55669
Sphingomonas melonis strain ZJ26
1 Sphingomonas sp MM1
1 Sphingomonas sp KC8
1 Sphingobium sp SYK6
1 Sphingobium sp TKS
1 Sphingobium sp YBL2
1 Sphingobium chlorophenolicum L1
1 Sphingobium indicum B90A
Sphingobium sp EP60837
1 Sphingobium herbicidovorans strain MH
1 Sphingobium cloacae JCM10874
1 Sphingobium baderi strain DE13
1 Sphingobium sp RAC03
1 Sphingobium sp C1
1 Sphingobium yanoikuyae strain S72
1 Sphingobium sp YG1
1 Sphingobium amiense DSM16289
1

Supplemental Figure S6. The evolutionary history of the *tonB* gene family. The unrooted tree indicates the presence of two separate orthologous groups, whose branches are highlighted in red and blue colors. The branches shown in the black were used as an outgroup due to their extreme lengths (and therefore very distant relationships to other members of the family). The bootstrap values are shown only for selected branches. To provide taxa names, the two subtrees in red and blue are shown as rooted clades. Paralogs from the same genome are denoted by numbers (1-3) preceding a taxon name. The ptAI values of the tonB gene from the "red" clade taxa have a significant PGLS model fit with the ptAI values of the reference GTA genes.

Description of Supplemental Data available via FigShare
DOI 10.6084/m9.figshare.20082749

Gene Families in 208 genomes:

orthogroups.tsv.zip: Gene families in 208 alphaproteobacterial genomes; the families were constructed using only genes that are at least 300 nucleotides in length. Each line in the file represents one gene family (an orthogroup). In each line, the individual gene family members are identified by RefSeqID of a genome joined by an underscore with RefSeqID of protein sequence of the gene.

GTA gene predictions:

gta_regions.xlsx: Predicted GTA 'head-tail' clusters in the initial dataset of 212 genomes. The data in the columns for individual GTA genes show their RefSeq accession numbers; empty cells indicate that a gene was not detected in a genome. The 208 genomes that were retained for the selection analyses are highlighted in green.

Effective Number of Codons (ENC) calculations:

codonW_enc_gc3s.zip: Effective number of codons (ENC) and GC3s values for genes in 208 alphaproteobacterial genomes that are at least 300 nucleotides in length. Each genome is represented by one file. The individual genes are identified by RefSeqID of a protein.

enc_deviation_gta_genes.xlsx: Deviation (in %) of Effective Number of Codons (ENC) values of the reference GTA genes in 208 genomes from the null model of no codon bias. Empty cells reflect either absence of a GTA gene from a genome or if its observed ENC was higher than expected (and, therefore, unreliable due to sampling of codons in a finite length of a gene sequence).

rel_enc.xlsx: Deviation of the ENC values of the reference GTA genes in 208 genomes normalized by the average ENC deviation of all genes in a genome.

tRNA Adaptation Index (tAI) calculations:

stAIcalc_wi.zip: Codon adaptation indices (ω_i; i=1-64) estimated by stAIcalc for genes in 208 alphaproteobacterial genomes that are at least 300 nucleotides in length. Each genome is represented by one file. For each genome, codons from all annotated genes were combined to calculate ω_i values for each codon.

stAIcalc_tAI.zip: tRNA adaptation (tAI) values for genes in 208 alphaproteobacterial genomes that are at least 300 nucleotides in length. Each genome is represented by one file. For each gene in a genome, calculated tAI value is listed. The individual genes are identified by RefSeqID of a genome joined by an underscore with RefSeqID of protein sequence of the gene.

ptAI_gta_genes.xlsx: Percentile tAI (ptAI) values of GTA genes of at least 300 nucleotides and with a broad taxonomic representation in 208 genomes. Empty cells reflect absence of a GTA gene in a genome.

Phylogenetic generalized least squares (PGLS) analysis:

orthogroups_PGLS.xlsx: PGLS model fit (slope and p-value) between individual reference GTA genes and other gene families across 208 genomes. Fourteen gene families (listed in Table 1) that have a significant model fit across all reference GTA genes are highlighted in yellow.

Phylogenetic Analyses:

reference_aln_tree.zip: Concatenated alignment of 29 phylogenetic marker genes found in 208 alphaproteobacterial genomes in FASTA format (reference_alignment.fasta). Reference phylogenomic tree reconstructed from the alignment in Newick format (reference_tree.nwk).

tonB_aln_tree.zip: Alignment of the *tonB* gene homologs (OG0002642) detected in alphaproteobacterial genomes (in FASTA format; tonB_alignment.fasta) and their phylogenetic relationships (in Newick format; tonB_tree.nwk).

Code:

exp_enc_deviation.py: Python script that calculates the expected effective number of codons (ENC) based on the GC3s content and the deviation from the expectations under the null model of no codon bias.