Fall 11-2023

Tracing Evolution of Gene Transfer Agents Using Comparative Genomics

Roman Kogay
Dartmouth College, roman.kogay.gr@dartmouth.edu

TRACING EVOLUTION OF GENE TRANSFER AGENTS USING COMPARATIVE GENOMICS

A Thesis
Submitted to the faculty
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

in

Ecology, Evolution, Environment and Society

by Roman Kogay

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

September 2023

Examining Committee:

_____
(chair) Olga Zhaxybayeva, Ph.D.

_____
Mark A. McPeek, Ph.D.

_____
Carey D. Nadell, Ph.D.

_____
George A. O'Toole, Ph.D.

_____
Andrew S. Lang, Ph.D.

_____
F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

`

Abstract

The accumulating evidence suggest that viruses and their components can be domesticated by their hosts, equipping them with convenient molecular toolkits for various functions. One of such domesticated system is Gene Transfer Agents (GTAs) that are produced by some bacteria and archaea. GTAs morphologically resemble small phage-like particles and contain random fragments of their host genome. They are produced only by a small fraction of the microbial population and are released through a lysis of the host cell. Bioinformatic analyses suggest that GTAs are especially abundant in the taxonomic class of *Alphaproteobacteria*, where they are vertically inherited and evolve as a part of their host genomes. In this work, we extensively analyze evolutionary patterns of alphaproteobacterial GTAs using comparative genomics, phylogenomics and machine learning methods. We initially develop an algorithm that validate the wide presence of GTA elements in alphaproteobacterial genomes, where they are generally mistaken for prophages due to their homology. Furthermore, we demonstrate that GTAs evolve under the selection that reduces the energetic cost of their production, indicating their importance for the conditions of the nutrient depletion. The genome-wide screenings of translational selection and coevolution signatures highlight the significance of GTAs as a stress-response adaptation for the horizontal gene transfer, revealing a set of previously unknown genes that could play a role in the GTA cycle. As production of GTAs leads to the host death, their maintenance is likely to be under a kin or group level selection. By combining our findings with accumulated body of knowledge, this work proposes a conceptual model illustrating the role of GTAs in bacterial populations and their persistence for hundreds of millions of years of evolution.

Acknowledgements

I feel incredibly lucky for having crossed paths with so many wonderful people throughout the years, all of whom contributed to both my personal and intellectual growth. In case I inadvertently did not mention your name here, please know that I am deeply grateful to every one of you.

First, I would like to thank my advisor Olga for her enormous support and guidance throughout my PhD study. Since our first Skype meeting on November 20, 2017, it was so clear to me that you are not only a great scientist, but also an exceptional mentor. Thank you for your unwavering patience, and reliability, which provided me with the reassuring sense that there was always someone supporting me throughout this journey. Thank you for creating such a perfect atmosphere, where I always felt that my research ideas, even when incorrect, were always welcome for discussion and exploration. Thank you for teaching me that science is not only about doing research and creating knowledge, but also about communication. I will always remember with a smile and a sense of amusement how our first writings were nearly entirely 'red-colored' after your revisions. Thank you for sharing your enthusiasm and positive attitude both inside and outside the lab. Words are not adequate to convey the depth of my gratitude towards you.

I would like to thank my committee members – Carey, George, Mark, and Andrew. Thank you for your advice and support over these years. I genuinely appreciate all your opinions and feedbacks on my projects and presentations. Thank you for sharing your expertise and perspectives on different topics of microbiology and evolution. I did not expect that by my fifth year here, the mere mention of GTAs would immediately make me also think about biofilms and group selection.

I would like to thank all my past and present lab mates for making the lab space (both in the virtual and physical worlds) such a wonderful and nurturing environment. Thank you for supporting me all these years and sharing your thoughts and ideas about my work. I'm especially grateful to Zhengshuang Hua and Anne Farrell, for sharing their knowledge and expertise on different topics of phylogenetics. As I joined the lab without any meaningful experience in that area, it is hard to overestimate the importance of your

support. I would like to also thank Jonathan Chiou and Jack Gourdeau. Working with you on your undergraduate projects has been very insightful and enjoyable experience to say the least, and I learned so much from you in the process.

I am very grateful to all collaborators with whom I have had the privilege and pleasure of working over the past years. Special thanks to my first collaborators Eugene Koonin and Yuri Wolf from NCBI. Collaborating with you have taught me a lot about looking at evolutionary problems and potential solutions from different, and occasionally unexpected angles. I would like to extend special gratitude to Weicheng Ma, Dan Rockmore and Soroush Vosoughi from the Computer Science department. Although our project is not included in this thesis, the collaboration with you at the crossroads of evolutionary biology and large language models has been an incredibly enlightening and rewarding experience.

Throughout my time in the graduate school, EEES, Dartmouth and Upper Valley communities have been so welcoming and inclusive. I am very thankful to all my friends that I met here for being with me throughout all my highs and lows. Special thanks to Anna, Amar, Kevin, Monika, Sarah and 'Matt Ayres fav crew'. Our board game nights, potlucks, and outdoor activities will forever remain in my memories.

I would like to also thank my friends from Kazakhstan. Although I thought it would be hard to maintain a meaningful friendship, while seeing each other so rarely, you have certainly shown me otherwise. Despite the thousands of kilometers that separate me from most of you, I always feel a strong and deep connection. Thank you especially to Aidana, Aika, Danat, Dauren, Karina, Lera, Vlad, Yerkin, and the whole 'KTL brotherhood'. I deeply value each of you and have full confidence that our friendship will continue for many years to come.

Finally, I want to thank my family for all unconditional love and support. Thank you for being my biggest cheerleaders in every moment of my life and encouraging me to pursue my goals and interests. Mom, Dad, and Katya, I love you so much.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 1

Introduction and Chapters Overview

Roman Kogay[1]

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

Horizontal gene transfer – a major mechanism for the microbial evolution

Thanks to the billions of years of evolution, a remarkable diversity of microbial organisms has emerged, colonizing almost every imaginable environmental habitat – from cold ecosystems to terrestrial hot springs (DeLong & Pace, 2001; Lozupone & Knight, 2007; Shu & Huang, 2022). Hence, microbes are arguably can be considered as the most dominant life forms in the biosphere, playing a critical role in a wide range of biogeochemical processes (Madsen, 2011). Consequently, it is essential to gain a better understanding of factors that govern their evolution and diversification in order to address arising challenges in many different fields including medicine (Frieri et al., 2017), climate change (Kirchman et al., 2009), agriculture (Sundin & Wang, 2018) and bioremediation (Pieper & Reineke, 2000).

Microbes can get novel traits and phenotypes via two primary mechanisms – either by mutation of the existing DNA material or through a horizontal gene transfer (HGT). In the latter, the microbial cells acquire the genetic material not from their parents, as opposed to the vertical inheritance, but from other, even very distantly related organisms (Gogarten & Townsend, 2005; Koonin et al., 2001). Despite the initial skepticism and underestimation, the modern genomic data unequivocally suggest that horizontal gene transfer is a crucial factor in the microbial evolution (Soucy et al., 2015). The main benefits of gene exchange in bacterial communities are associated with the rapid spread of advantageous alleles and genes, promoting adaptation to new and unstable ecological conditions (Arnold et al., 2022). For example, bacteria can rapidly become unsusceptible to antibiotics by horizontal acquisition of certain genes from antibiotic-resistant bacteria, posing a serious threat to the global healthcare (Andersson & Hughes, 2010). Systematic quantification of HGT events suggest that a lot of bacterial genomes exhibit a highly mosaic structure, with a considerable number of their genes being derived from diverse microbial lineages (Andreani et al., 2017; Jain et al., 2003; Zamani-Dahaj et al., 2016; Zhaxybayeva et al., 2006).

Three canonical molecular mechanisms that facilitate HGT are transformation, conjugation, and transduction (**Figure 1**). In transformation, bacteria uptake the genetic

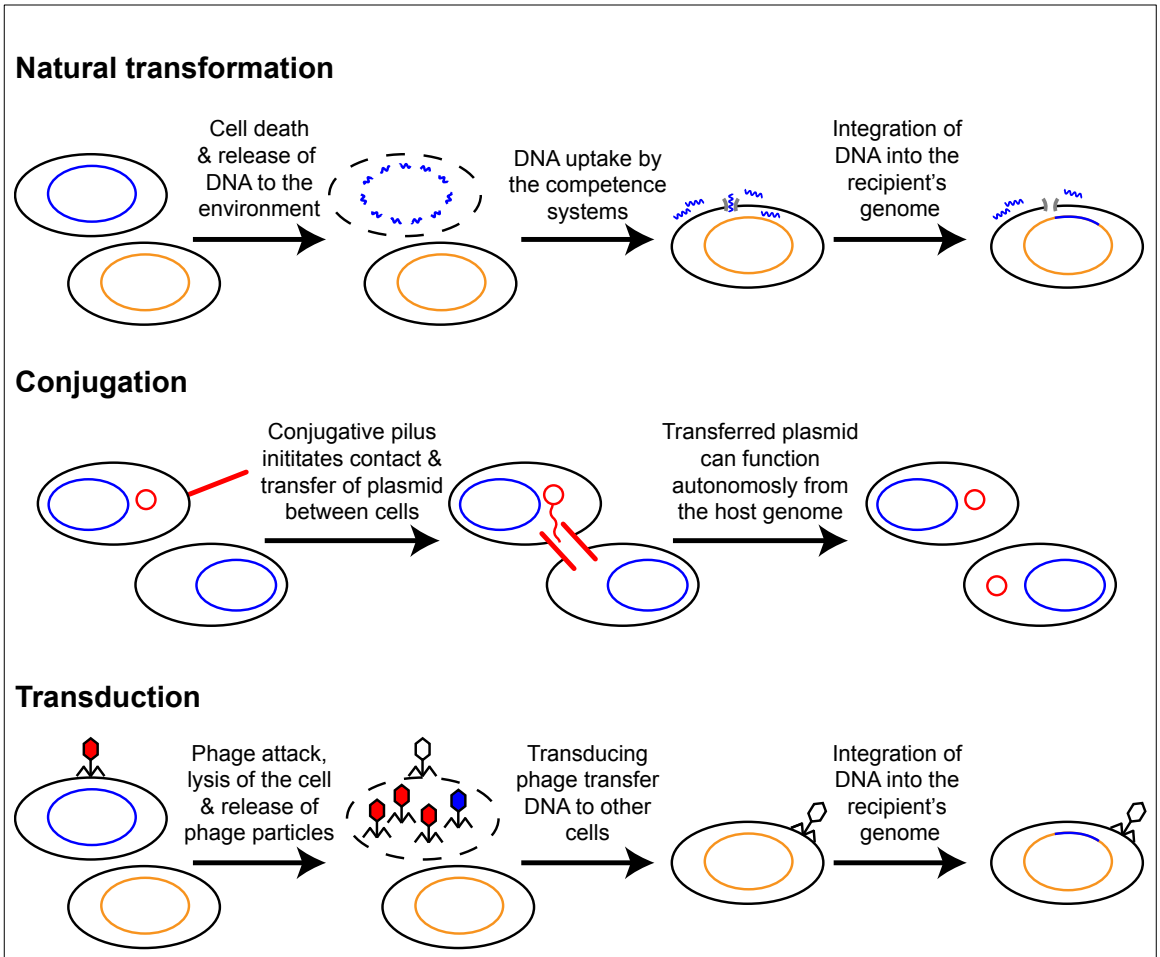

Figure 1. Three canonical mechanisms for HGT.

material from the environment by developing a physiological state known as a competence (Avery et al., 1944; Griffith, 1928). Although a natural transformation under the laboratory conditions was observed only for a limited number of bacterial species (Johnston et al., 2014), the competence proteins can be bioinformatically identified in the vast majority of sequenced bacterial genomes suggesting that it can be more common in natural environments (Pimentel & Zhang, 2018). During the transformation, bacteria import environmental DNA using the competence proteins, generating a ssDNA molecule (Chen & Dubnau, 2004). Consequently, the imported ssDNA can either be utilized as a source of deoxyribonucleotides or get integrated into the host genome via a homologous recombination pathway by recruiting DprA and RecA enzymes (Johnston et al., 2014; Mell & Redfield, 2014). As a success rate of the homologous recombination exponentially decreases with increasing divergence of DNA sequence (Majewski, 2001;

Majewski & Cohan, 1999), transformation primarily promotes the HGT between closely related organisms. Benefits linked to the transformation-mediated homologous recombination include the dissemination of advantageous alleles, reduction of the mutation load (Takeuchi et al., 2014), assistance in DNA damage repair (Michod et al., 1988), and introduction of novel genes via flanking homology (Kung et al., 2013).

In conjugation, donor and recipient cells establish a cell-to-cell contact via a pilus that facilitates transfer of the conjugative genetic elements (Lederberg & Tatum, 1946). Such conjugative elements generally possess all genes required for the assembly of a conjugative machinery and can additionally carry extra accessory genes. The content of the accessory genes is very variable in the nature and thus the conjugative elements considerably differ in their size. The smallest elements contain only genes necessary for their propagation, while the largest ones have a length of at least ~1 MB, encoding diverse metabolic pathways (Romanchuk et al., 2014). Upon the transfer into the recipient cells, conjugative elements can either stay as autonomous plasmids or integrate themselves into the host genome using the molecular mechanisms similar to those of viruses (Johnson & Grossman, 2015). Consequently, conjugation can occur between phylogenetically distant microbes and was even demonstrated to happen between *Bacteria* and *Archaea* that represent two distinct domains of life (Dodsworth et al., 2010). The conjugation is an extremely widespread process, as it has been detected in diverse microbial communities that were isolated from vastly different environments (Cury et al., 2017; Davison, 1999).

In transduction, the HGT is facilitated by bacterial and archaeal viruses. It can either happen due to the erroneous excision of integrated viruses, which will include flanking segments of the host genome or the accidental packaging of random fragments of the host DNA into viral capsids (Touchon et al., 2017). Consequently, the transducing viruses shuttle packaged genetic material to other cells in the population. Hence, transduction can equip bacteria with novel traits, including resistance to antibiotics, secretion of virulence factors and metabolism of novel molecules (Haaber et al., 2017; Lindell et al., 2004; Wagner & Waldor, 2002). As bacterial viruses (phages) are the most abundant entities in the biosphere and the rate of infection is extremely high (~10^{23}

infections per second), the transduction greatly influences the bacterial ecological and evolutionary dynamics (Clokie et al., 2011; Suttle, 2007).

The shared characteristic among described canonical processes for HGT is their ubiquitous occurrence in the nature, and their impact on microbes belonging to almost all established phylogenetic clades. However, other additional mechanisms for exchange of genes continue to be discovered, expanding the currently known repertoire. These relatively understudied systems include membrane vesicles, nanotubes, and gene transfer agents (Emamalipour et al., 2020; Soucy et al., 2015).

Membrane vesicles (MVs) are produced by all living cells and represent spherical buddings derived from the cell surfaces (Domingues & Nielsen, 2017). In addition to mediating functions similar to other extracellular vesicles, such as cell-to-cell communication, delivering different cargo and removing toxic compounds, MVs also participate in genetic exchange by carrying DNA molecules (Fulsundar et al., 2014; Grull et al., 2018). By coating the genetic material, MVs confer protection of the DNA against nucleases and facilitate HGT between cells (Domingues & Nielsen, 2017). Multiple studies indicate that the nature of DNA within the MVs can vary, including the genetic material originating from chromosomes, plasmids and phages (Orench-Rivera & Kuehn, 2016). Upon the entry into the recipient cell, the MVs-delivered DNA can get integrated into the genome by the homologous recombination (Domingues & Nielsen, 2017).

Similar to MVs, nanotubes are also membranous structures that mediate intercellular communication by physically connecting neighboring cells (Dubey & Ben-Yehuda, 2011). Unlike conjugation pili that generally facilitate transfer of mobile elements that directly encode them, nanotubes can transfer non-conjugative plasmids and participate in exchange of various cytoplasmic molecules, including nutrients and toxins (Abe et al., 2020). Interestingly, the recent study in *Bacillus subtilis* suggests that the nanotubes are formed when cells are dying, or even after the cell disintegration, challenging the previous belief that they have a physiological role in natural environments (Pospisil et al., 2020). Hence, further studies are needed to validate the role of nanotubes in HGT and their prevalence in microbial communities.

Gene transfer agents – phage-like particles for genetic exchange

Gene transfer agents are found in multiple microbial organisms

Another understudied molecular system that can facilitate exchange of genes and the main theme of this dissertation is gene transfer agents (GTAs) (Lang et al., 2012). GTAs were discovered in 1970s by studying the model alphaproteobacterium *Rhodobacter capsulatus*, when Barry Marrs found that a previously unknown small-sized vector is capable of transferring any genetic marker to other cells in the population (Marrs, 1974). While the molecular nature of that vector was initially unresolved, the subsequent studies have unveiled that GTAs resemble tiny phage-like particles that carry a small fragment of a linear duplex DNA molecule (Solioz & Marrs, 1977; Solioz et al., 1975). Since then, the production of GTAs was experimentally demonstrated in multiple bacterial species and at least one archaeon (Anderson et al., 1994; Bertani, 1999; Biers et al., 2008; Gozzi et al., 2022; Humphrey et al., 1997; Nagao et al., 2015; Rapp & Wall, 1987; Tomasch et al., 2018). Intriguingly, among the experimentally validated GTAs, there are five genetically unrelated groups that present in phylogenetically distant microbial clades. These data suggest that GTAs have originated multiple times from different genetic sources via the process of convergent evolution (Lang et al., 2017). While these distinct GTA groups have some molecular and genetic differences, they also share significant commonalities with each other. All GTA groups generally encapsulate mostly random pieces of the host genome, primarily propagate via the vertical inheritance and cannot self-replicate independently from their hosts due to the limited size of their capsid heads (Lang et al., 2017). These properties put GTAs in a stark contrast with mobile elements involved in conjugation and transduction that typically exhibit an autonomous mode of behavior (Frost et al., 2005).

As GTA-producing organisms die, individual selection becomes non-operational for that trait, and the benefits associated with GTAs production should function at the population level. While the nature of such benefits is not completely characterized, GTAs were hypothesized to be involved in exchange of advantageous genes and facilitating the repair of DNA damage (Lang et al., 2012; Marrs et al., 1977; McDaniel et al., 2010). Indeed, 48 years after their discovery, the beneficial role of GTAs in promoting DNA

repair processes was experimentally confirmed for the GTA system in the model organism *Caulobacter crescentus* (Gozzi et al., 2022). Hence, GTAs can be described as phage-derived elements that have been repurposed by their hosts to serve as the vehicles for genetic exchange within microbial populations.

The gene transfer agent of *Rhodobacter capsulatus* (RcGTA)

Currently, the most well-studied and well-characterized GTA system belongs to the *Rhodobacter capsulatus* (abbreviated as "RcGTA"). The production of RcGTA is governed by at least twenty four genes that are scattered across five different loci (**Figure 2**) (Hynes et al., 2016; Lang & Beatty, 2000; Westbye et al., 2016). One of these loci, known as the 'head-tail cluster' (~14 kb), harbors seventeen genes that are required for the assembly of DNA packaging machinery and construction of head and tail structures (Lang & Beatty, 2000). Comparative genomics analyses suggest that eighteen out of twenty-four genes have homologs in the viral databases, supporting the idea that RcGTA system has originated from phages many millions years ago (Shakya et al., 2017).

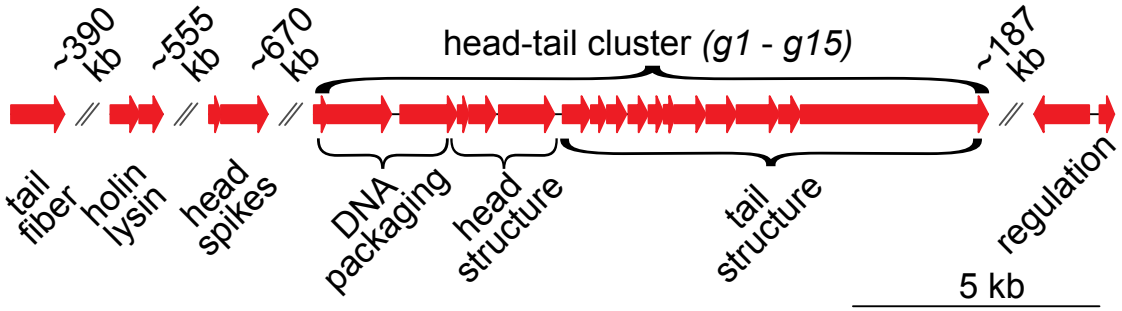

Figure 2. Gene map of five loci that encode production of RcGTAs. Each gene is illustrated as an arrow and their functional role is shown below them.

Despite that RcGTA shares significant morphological and genetic similarities with phages, there are notable differences in their molecular cycles (**Figure 3**). The production of RcGTA particles is increased during the stationary phase (Solioz et al., 1975) and is particularly pronounced under the conditions of the nutrient depletion (Westbye, O'Neill, et al., 2017). As a result, RcGTAs are produced only by a small fraction of the population, approximately 0.15 to 3% (P. C. Fogg et al., 2012; Hynes et al., 2012). The RcGTA

producers start actively expressing RcGTA genes, causing a subsequent lysis and release of particles that contain ~4.5 kb of encapsulated DNA (Yen et al., 1979). The remaining members of the population act as RcGTA recipients by producing specific polysaccharide receptors (Brimacombe et al., 2013). These receptors play a crucial role in the adsorption of RcGTA particles, making them essential for the successful gene transfer (Alim et al., 2023; Brimacombe et al., 2013). Consequently, the RcGTA-delivered DNA entry into cells is facilitated by the competence system (Brimacombe et al., 2015) and it is integrated into the recipient genome via the RecA-mediated homologous recombination (Brimacombe et al., 2014). These mechanisms indicate that RcGTA-mediated HGT incorporates properties of both transduction and transformation processes, where phage-like entities facilitate exchange of genes using the competence system and homologous recombination (Brimacombe et al., 2015).

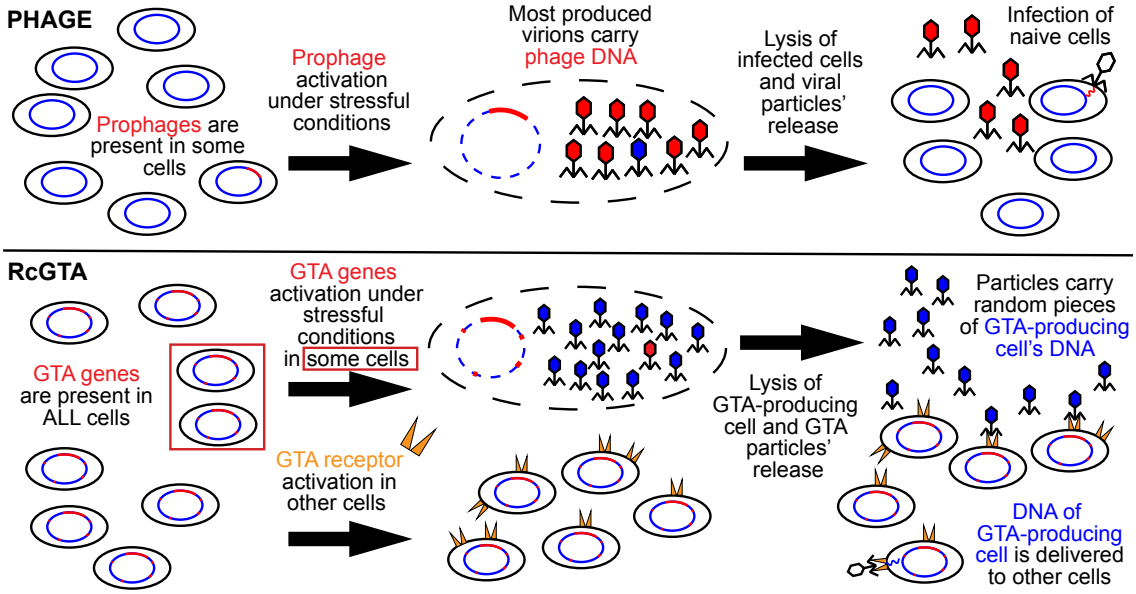

Figure 3. The differences between RcGTA and lysogenic phage cycles.

As production of RcGTA particles effectively kills the cell, it is critical to tightly control the expression of RcGTA genes. Indeed, numerous studies indicate that multiple genetic circuits are involved in this important task (Fogg, 2019; Lang & Beatty, 2000; Mercer & Lang, 2014; Mercer et al., 2012; Westbye, Beatty, et al., 2017). The expression of RcGTA genes is primarily regulated by the GtaR/GtaI-quorum sensing regulator and

the CckA-ChpT-CtrA phosphorelay system (Lang & Beatty, 2000; Leung et al., 2012; Mercer et al., 2012; Schaefer et al., 2002). Molecular studies suggest that both CtrA and GtaR indirectly control the expression of RcGTA genes using the intermediary proteins (Fogg, 2019; Leung et al., 2012). One of such proteins is GafA, which serves as a direct transcription regulator of two RcGTA loci – the head-tail cluster and the holin/endolysin genes that are crucial for the release of RcGTA particles (Fogg, 2019). The complete mechanism of the CtrA protein's action on RcGTA genes is not fully understood. However, it appears to play a role in regulating the production of RcGTAs by affecting levels of the cyclic dimeric GMP (c-di-GMP) second messenger (Farrera-Calderon et al., 2021; Pallegar et al., 2020). Such complex and deep RcGTA system's integration into the molecular circuits of its host signifies its importance for the producing organisms.

Evolution of RcGTA-like GTA systems

While homologs of the RcGTA head-tail cluster are found in viral databases (Lang & Beatty, 2000), they are also extensively detected in bacterial genomes and particularly pronounced in the taxonomic class of *Alphaproteobacteria* (Lang & Beatty, 2007; Lang et al., 2002; Shakya et al., 2017). Although some of such homologs are likely located in the prophage regions (Shakya et al., 2017), others may be part of the functional or decaying RcGTA-like GTA systems (here referred as "GTA systems" for brevity). Indeed, functionally validated GTA systems in alphaproteobacterial species of *Ruegeria pomeroyi* (Biers et al., 2008), *Rhodovulum sulfidophilum* (Nagao et al., 2015), *Dinorosobacter shibae* (Tomasch et al., 2018) and *Caulobacter crescentus* (Gozzi et al., 2022) are all encoded by homologs of RcGTA genes. The comparative genomics analyses indicate that the regions containing at least 9 homologs of the head-tail cluster (so-called 'large clusters') generally have evolutionary patterns similar to the core housekeeping genes, which significantly differ from those observed in common phages (Shakya et al., 2017). These large clusters of GTA genes are widely computationally detected in alphaproteobacterial genomes within four taxonomic orders – *Sphingomonadales*, *Rhizobiales*, *Caulobacterales* and *Rhodobacterales* (Lang & Beatty, 2007; Shakya et al., 2017). The molecular clock estimates that large clusters have originated at least 700 million years ago and have been evolving as part of their host genomes since that time

with very limited horizontal transfer events (Shakya et al., 2017). Their widespread presence raises an intriguing hypothesis that GTAs are much more widespread than it is currently acknowledged. Conversely, multiple species from the same alphaproteobacterial clade do not encode large clusters, suggesting that under some ecological conditions GTAs are not beneficial for their hosts and can get purged from the genomes (Shakya et al., 2017).

Overview of chapters

The main purpose of this thesis is to advance our understanding and knowledge of alphaproteobacterial GTAs by thoroughly investigating their evolutionary patterns using phylogenomics and comparative genomics methods.

In chapter 2, we conceptualize and implement a machine learning tool called "GTA-Hunter", which can quickly and accurately predict large clusters of GTA genes without relying on complex and time-consuming phylogenetic analyses. The efficiency of the GTA-Hunter arises from its ability to distinguish GTA proteins from their viral homologs by analyzing their amino acid composition dissimilarities. After running the GTA-Hunter on a collection of more than 1,400 alphaproteobacterial genomes, we detected putative GTA systems in approximately 57.5% of them. Our findings reveal that GTA systems are generally incorrectly annotated as prophages due to their homology.

In chapter 3, we seek to understand the nature of amino acid bias in GTAs that we have observed while analyzing GTA clusters predicted by GTA-Hunter. We discovered that GTA proteins are composed of energetically cheaper amino acids in comparison to their viral homologs. Interestingly, in instances when viruses horizontally acquire GTA genes, that amino acid bias disappears over the course of their evolution. This implies that reduction in the energetic cost of proteins plays an important role for GTAs, but not for phages. In *Alphaproteobacteria*, this bias is particularly pronounced in the taxonomic order of *Sphingomonadales*, whose members are known to predominantly live in nutrient depleted conditions. By analyzing patterns of substitutions in that taxon, we found that in many cases positive selection drives the reduction in the protein cost. These findings

support that GTAs are specialized bacterial adaptations that facilitate the survival of their host populations under the nutrient-limited conditions, which are very prevalent in nature.

While our findings indicate that selection favors nonsynonymous substitutions toward energetically cheaper amino acids, synonymous substitutions could also be under the selection pressure. Although synonymous codons encode the same amino acids, they are not used to the same extent because different organisms are enriched in different sets of codons. That phenomenon is known as the codon usage bias and is explained in part by the translational selection pressure to match the available pool of tRNA molecules (Plotkin & Kudla, 2011). In chapter 4, we examine signatures of codon usage bias in GTAs and their host genomes. We found that codon usage bias substantially fluctuates among individual GTA genes and different taxonomic groups but is especially notable in *Sphingomonadales* order. We further detected that codon usage bias has a significant negative correlation with the energetic cost of GTA proteins, indicating that increase in production of GTAs is also associated with the stronger selection on the carbon saving. By conducting genome-wide screening for gene families with the similar patterns in the codon usage bias, we found 13 genes that were not previously implicated to be involved in the GTA cycle. They are significantly enriched in 'homologous recombination', 'mismatch repair', 'carotenoid biosynthesis', and 'terpenoid backbone biosynthesis' molecular pathways. These results provide insights into the impact of translational selection on evolution of GTA genes across various taxonomic clades and outline a specific set of genes that are likely to be involved in the integration of GTA-delivered DNA into the recipient genome.

As GTAs evolve as part of their host genomes for hundreds of millions of years, GTA genes are expected to coevolve with other gene families involved in their cycle. In chapter 5, we study coevolutionary relationships between GTA genes and other gene families residing in their host genomes by comparing their evolutionary rates across phylogenetic trees. Our results suggest that GTA genes significantly coevolve with each other, and with 59 gene families, 4 of which have been previously experimentally validated to be involved in the GTA cycle. Other gene families are associated with various molecular processes, including DNA repair, stress response and biofilm

formation. By combining existing knowledge about GTAs, we outlined a model that explains their persistence in microbial populations.

Despite the shared ancestry with viruses, molecular and evolutionary data suggest that GTAs are strongly integrated into cellular functions of their hosts and provide benefits to them. In chapter 6, we propose to classify GTAs as 'viriforms' that per definition of International Committee on Virus Taxonomy (ICTV) represent virus-derived elements that have been exapted by their hosts to perform functions important for their lifecycle. Using phylogenetic analyses, we demonstrated that different GTAs groups have clearly distinct origins and follow different evolutionary trajectories comparatively to the viruses. We established specific criteria that characterize GTAs and outlined a classification scheme for three distinct GTAs lineages. The ICTV has approved and ratified the proposal in April 2023 (Zerbini et al., 2023).

References

Abe, K., Nomura, N., & Suzuki, S. (2020). Biofilms: hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism. *FEMS Microbiol Ecol*, *96*(5). https://doi.org/10.1093/femsec/fiaa031

Alim, N. T. B., Koppenhofer, S., Lang, A. S., & Beatty, J. T. (2023). Extracellular polysaccharide receptor and receptor-binding proteins of the *Rhodobacter capsulatus* bacteriophage-like gene transfer agent RcGTA. *Genes (Basel)*, *14*(5). https://doi.org/10.3390/genes14051124

Anderson, B., Goldsmith, C., Johnson, A., Padmalayam, I., & Baumstark, B. (1994). Bacteriophage-like particle of *Rochalimaea henselae*. *Mol Microbiol*, *13*(1), 67-73. https://doi.org/10.1111/j.1365-2958.1994.tb00402.x

Andersson, D. I., & Hughes, D. (2010). Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat Rev Microbiol*, *8*(4), 260-271. https://doi.org/10.1038/nrmicro2319

Andreani, N. A., Hesse, E., & Vos, M. (2017). Prokaryote genome fluidity is dependent on effective population size. *ISME J*, *11*(7), 1719-1721. https://doi.org/10.1038/ismej.2017.36

Arnold, B. J., Huang, I. T., & Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*, *20*(4), 206-218. https://doi.org/10.1038/s41579-021-00650-4

Avery, O. T., Macleod, C. M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of Pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *J Exp Med*, *79*(2), 137-158. https://doi.org/10.1084/jem.79.2.137

Bertani, G. (1999). Transduction-like gene transfer in the methanogen *Methanococcus voltae*. *J Bacteriol*, *181*(10), 2992-3002. https://doi.org/10.1128/JB.181.10.2992-3002.1999

Biers, E. J., Wang, K., Pennington, C., Belas, R., Chen, F., & Moran, M. A. (2008). Occurrence and expression of gene transfer agent genes in marine bacterioplankton. *Appl Environ Microbiol*, *74*(10), 2933-2939. https://doi.org/10.1128/AEM.02129-07

Brimacombe, C. A., Ding, H., & Beatty, J. T. (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*(6), 1260-1278. https://doi.org/10.1111/mmi.12628

Brimacombe, C. A., Ding, H., Johnson, J. A., & Beatty, J. T. (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*(16), 2653-2663. https://doi.org/10.1128/JB.00332-15

Brimacombe, C. A., Stevens, A., Jun, D., Mercer, R., Lang, A. S., & Beatty, J. T. (2013). Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). *Mol Microbiol*, *87*(4), 802-817. https://doi.org/10.1111/mmi.12132

Chen, I., & Dubnau, D. (2004). DNA uptake during bacterial transformation. *Nat Rev Microbiol*, *2*(3), 241-249. https://doi.org/10.1038/nrmicro844

Clokie, M. R., Millard, A. D., Letarov, A. V., & Heaphy, S. (2011). Phages in nature. *Bacteriophage*, *1*(1), 31-45. https://doi.org/10.4161/bact.1.1.14942

Cury, J., Touchon, M., & Rocha, E. P. C. (2017). Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res*, *45*(15), 8943-8956. https://doi.org/10.1093/nar/gkx607

Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid*, *42*(2), 73-91. https://doi.org/10.1006/plas.1999.1421

DeLong, E. F., & Pace, N. R. (2001). Environmental diversity of bacteria and archaea. *Syst Biol*, *50*(4), 470-478. https://www.ncbi.nlm.nih.gov/pubmed/12116647

Dodsworth, J. A., Li, L., Wei, S., Hedlund, B. P., Leigh, J. A., & de Figueiredo, P. (2010). Interdomain conjugal transfer of DNA from bacteria to archaea. *Appl Environ Microbiol*, *76*(16), 5644-5647. https://doi.org/10.1128/AEM.00967-10

Domingues, S., & Nielsen, K. M. (2017). Membrane vesicles and horizontal gene transfer in prokaryotes. *Curr Opin Microbiol*, *38*, 16-21. https://doi.org/10.1016/j.mib.2017.03.012

Dubey, G. P., & Ben-Yehuda, S. (2011). Intercellular nanotubes mediate bacterial communication. *Cell*, *144*(4), 590-600. https://doi.org/10.1016/j.cell.2011.01.015

Emamalipour, M., Seidi, K., Zununi Vahed, S., Jahanban-Esfahlan, A., Jaymand, M., Majdi, H., Amoozgar, Z., Chitkushev, L. T., Javaheri, T., Jahanban-Esfahlan, R., & Zare, P. (2020). Horizontal gene transfer: from evolutionary flexibility to

disease progression. *Front Cell Dev Biol*, *8*, 229. https://doi.org/10.3389/fcell.2020.00229

Farrera-Calderon, R. G., Pallegar, P., Westbye, A. B., Wiesmann, C., Lang, A. S., & Beatty, J. T. (2021). The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer agent production Is bidirectional and regulated by cyclic di-GMP. *J Bacteriol*, *203*(5), e00525-00520. https://doi.org/10.1128/JB.00525-20

Fogg, P. C., Westbye, A. B., & Beatty, J. T. (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*(8), e43772. https://doi.org/10.1371/journal.pone.0043772

Fogg, P. C. M. (2019). Identification and characterization of a direct activator of a gene transfer agent. *Nat Commun*, *10*(1), 595. https://doi.org/10.1038/s41467-019-08526-1

Frieri, M., Kumar, K., & Boutin, A. (2017). Antibiotic resistance. *J Infect Public Health*, *10*(4), 369-378. https://doi.org/10.1016/j.jiph.2016.08.007

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*, *3*(9), 722-732. https://doi.org/10.1038/nrmicro1235

Fulsundar, S., Harms, K., Flaten, G. E., Johnsen, P. J., Chopade, B. A., & Nielsen, K. M. (2014). Gene transfer potential of outer membrane vesicles of *Acinetobacter baylyi* and effects of stress on vesiculation. *Appl Environ Microbiol*, *80*(11), 3469-3483. https://doi.org/10.1128/AEM.04248-13

Gogarten, J. P., & Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*, *3*(9), 679-687. https://doi.org/10.1038/nrmicro1204

Gozzi, K., Tran, N. T., Modell, J. W., Le, T. B. K., & Laub, M. T. (2022). Prophage-like gene transfer agents promote *Caulobacter crescentus* survival and DNA repair

during stationary phase. *PLoS Biol*, *20*(11), e3001790. https://doi.org/10.1371/journal.pbio.3001790

Griffith, F. (1928). The Significance of Pneumococcal Types. *J Hyg (Lond)*, *27*(2), 113-159. https://doi.org/10.1017/s0022172400031879

Grull, M. P., Mulligan, M. E., & Lang, A. S. (2018). Small extracellular particles with big potential for horizontal gene transfer: membrane vesicles and gene transfer agents. *FEMS Microbiol Lett*, *365*(19). https://doi.org/10.1093/femsle/fny192

Haaber, J., Penades, J. R., & Ingmer, H. (2017). Transfer of antibiotic resistance in *Staphylococcus aureus*. *Trends Microbiol*, *25*(11), 893-905. https://doi.org/10.1016/j.tim.2017.05.011

Humphrey, S. B., Stanton, T. B., Jensen, N. S., & Zuerner, R. L. (1997). Purification and characterization of VSH-1, a generalized transducing bacteriophage of *Serpulina hyodysenteriae*. *J Bacteriol*, *179*(2), 323-329. https://doi.org/10.1128/jb.179.2.323-329.1997

Hynes, A. P., Mercer, R. G., Watton, D. E., Buckley, C. B., & Lang, A. S. (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*(2), 314-325. https://doi.org/10.1111/j.1365-2958.2012.08113.x

Hynes, A. P., Shakya, M., Mercer, R. G., Grull, M. P., Bown, L., Davidson, F., Steffen, E., Matchem, H., Peach, M. E., Berger, T., Grebe, K., Zhaxybayeva, O., & Lang, A. S. (2016). Functional and evolutionary characterization of a gene transfer agent's multilocus “genome". *Mol Biol Evol*, *33*(10), 2530-2543. https://doi.org/10.1093/molbev/msw125

Jain, R., Rivera, M. C., Moore, J. E., & Lake, J. A. (2003). Horizontal gene transfer accelerates genome innovation and evolution. *Mol Biol Evol*, *20*(10), 1598-1602. https://doi.org/10.1093/molbev/msg154

Johnson, C. M., & Grossman, A. D. (2015). Integrative and conjugative elements (ICEs): what they do and how they work. *Annu Rev Genet*, *49*, 577-601. https://doi.org/10.1146/annurev-genet-112414-055018

Johnston, C., Martin, B., Fichant, G., Polard, P., & Claverys, J. P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat Rev Microbiol*, *12*(3), 181-196. https://doi.org/10.1038/nrmicro3199

Kirchman, D. L., Moran, X. A., & Ducklow, H. (2009). Microbial growth in the polar oceans - role of temperature and potential impact of climate change. *Nat Rev Microbiol*, *7*(6), 451-459. https://doi.org/10.1038/nrmicro2115

Koonin, E. V., Makarova, K. S., & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, *55*, 709-742. https://doi.org/10.1146/annurev.micro.55.1.709

Kung, S. H., Retchless, A. C., Kwan, J. Y., & Almeida, R. P. (2013). Effects of DNA size on transformation and recombination efficiencies in *Xylella fastidiosa*. *Appl Environ Microbiol*, *79*(5), 1712-1717. https://doi.org/10.1128/AEM.03525-12

Lang, A. S., & Beatty, J. T. (2000). Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. *Proc Natl Acad Sci U S A*, *97*(2), 859-864. https://doi.org/10.1073/pnas.97.2.859

Lang, A. S., & Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*, *15*(2), 54-62. https://doi.org/10.1016/j.tim.2006.12.001

Lang, A. S., Taylor, T. A., & Beatty, J. T. (2002). Evolutionary implications of phylogenetic analyses of the gene transfer agent (GTA) of *Rhodobacter capsulatus*. *J Mol Evol*, *55*(5), 534-543. https://doi.org/10.1007/s00239-002-2348-7

Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol*, *4*(1), 87-104. https://doi.org/10.1146/annurev-virology-101416-041624

Lang, A. S., Zhaxybayeva, O., & Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol*, *10*(7), 472-482. https://doi.org/10.1038/nrmicro2802

Lederberg, J., & Tatum, E. L. (1946). Gene recombination in *Escherichia coli*. *Nature*, *158*(4016), 558. https://doi.org/10.1038/158558a0

Leung, M. M., Brimacombe, C. A., Spiegelman, G. B., & Beatty, J. T. (2012). The GtaR protein negatively regulates transcription of the gtaRI operon and modulates gene transfer agent (RcGTA) expression in *Rhodobacter capsulatus*. *Mol Microbiol*, *83*(4), 759-774. https://doi.org/10.1111/j.1365-2958.2011.07963.x

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., & Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc Natl Acad Sci U S A*, *101*(30), 11013-11018. https://doi.org/10.1073/pnas.0401526101

Lozupone, C. A., & Knight, R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A*, *104*(27), 11436-11440. https://doi.org/10.1073/pnas.0611525104

Madsen, E. L. (2011). Microorganisms and their roles in fundamental biogeochemical cycles. *Curr Opin Biotechnol*, *22*(3), 456-464. https://doi.org/10.1016/j.copbio.2011.01.008

Majewski, J. (2001). Sexual isolation in bacteria. *FEMS Microbiol Lett*, *199*(2), 161-169. https://doi.org/10.1111/j.1574-6968.2001.tb10668.x

Majewski, J., & Cohan, F. M. (1999). DNA sequence similarity requirements for interspecific recombination in Bacillus. *Genetics*, *153*(4), 1525-1533. https://doi.org/10.1093/genetics/153.4.1525

Marrs, B. (1974). Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad Sci U S A*, *71*(3), 971-973. https://doi.org/10.1073/pnas.71.3.971

Marrs, B., Wall, J. D., & Gest, H. (1977). Emergence of the biochemical genetics and molecular biology of photosynthetic bacteria. *Trends Biochem Sci*, *2*(5), 105-108. https://doi.org/10.1016/0968-0004(77)90173-6

McDaniel, L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., & Paul, J. H. (2010). High frequency of horizontal gene transfer in the oceans. *Science*, *330*(6000), 50. https://doi.org/10.1126/science.1192243

Mell, J. C., & Redfield, R. J. (2014). Natural competence and the evolution of DNA uptake specificity. *J Bacteriol*, *196*(8), 1471-1483. https://doi.org/10.1128/JB.01293-13

Mercer, R. G., & Lang, A. S. (2014). Identification of a predicted partner-switching system that affects production of the gene transfer agent RcGTA and stationary phase viability in *Rhodobacter capsulatus*. *BMC Microbiol*, *14*, 71. https://doi.org/10.1186/1471-2180-14-71

Mercer, R. G., Quinlan, M., Rose, A. R., Noll, S., Beatty, J. T., & Lang, A. S. (2012). Regulatory systems controlling motility and gene transfer agent production and release in *Rhodobacter capsulatus*. *FEMS Microbiol Lett*, *331*(1), 53-62. https://doi.org/10.1111/j.1574-6968.2012.02553.x

Michod, R. E., Wojciechowski, M. F., & Hoelzer, M. A. (1988). DNA repair and the evolution of transformation in the bacterium *Bacillus subtilis*. *Genetics*, *118*(1), 31-39. https://doi.org/10.1093/genetics/118.1.31

Nagao, N., Yamamoto, J., Komatsu, H., Suzuki, H., Hirose, Y., Umekage, S., Ohyama, T., & Kikuchi, Y. (2015). The gene transfer agent-like particle of the marine phototrophic bacterium *Rhodovulum sulfidophilum*. *Biochem Biophys Rep*, *4*, 369-374. https://doi.org/10.1016/j.bbrep.2015.11.002

Orench-Rivera, N., & Kuehn, M. J. (2016). Environmentally controlled bacterial vesicle-mediated export. *Cell Microbiol*, *18*(11), 1525-1536. https://doi.org/10.1111/cmi.12676

Pallegar, P., Pena-Castillo, L., Langille, E., Gomelsky, M., & Lang, A. S. (2020). Cyclic di-GMP-mediated regulation of gene transfer and motility in *Rhodobacter capsulatus*. *J Bacteriol*, *202*(2). https://doi.org/10.1128/JB.00554-19

Pieper, D. H., & Reineke, W. (2000). Engineering bacteria for bioremediation. *Curr Opin Biotechnol*, *11*(3), 262-270. https://doi.org/10.1016/s0958-1669(00)00094-x

Pimentel, Z. T., & Zhang, Y. (2018). Evolution of the natural transformation protein, ComEC, in bacteria. *Front Microbiol*, *9*, 2980. https://doi.org/10.3389/fmicb.2018.02980

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, *12*(1), 32-42. https://doi.org/10.1038/nrg2899

Pospisil, J., Vitovska, D., Kofronova, O., Muchova, K., Sanderova, H., Hubalek, M., Sikova, M., Modrak, M., Benada, O., Barak, I., & Krasny, L. (2020). Bacterial nanotubes as a manifestation of cell death. *Nat Commun*, *11*(1), 4963. https://doi.org/10.1038/s41467-020-18800-2

Rapp, B. J., & Wall, J. D. (1987). Genetic transfer in *Desulfovibrio desulfuricans*. *Proc Natl Acad Sci U S A*, *84*(24), 9128-9130. https://doi.org/10.1073/pnas.84.24.9128

Romanchuk, A., Jones, C. D., Karkare, K., Moore, A., Smith, B. A., Jones, C., Dougherty, K., & Baltrus, D. A. (2014). Bigger is not always better: transmission and fitness burden of approximately 1MB *Pseudomonas syringae* megaplasmid pMPPla107. *Plasmid*, *73*, 16-25. https://doi.org/10.1016/j.plasmid.2014.04.002

Schaefer, A. L., Taylor, T. A., Beatty, J. T., & Greenberg, E. P. (2002). Long-chain acyl-homoserine lactone quorum-sensing regulation of *Rhodobacter capsulatus* gene

transfer agent production. *J Bacteriol*, *184*(23), 6515-6521. https://doi.org/10.1128/JB.184.23.6515-6521.2002

Shakya, M., Soucy, S. M., & Zhaxybayeva, O. (2017). Insights into origin and evolution of alpha-proteobacterial gene transfer agents. *Virus Evol*, *3*(2), vex036. https://doi.org/10.1093/ve/vex036

Shu, W. S., & Huang, L. N. (2022). Microbial diversity in extreme environments. *Nat Rev Microbiol*, *20*(4), 219-235. https://doi.org/10.1038/s41579-021-00648-y

Solioz, M., & Marrs, B. (1977). The gene transfer agent of *Rhodopseudomonas capsulata*. Purification and characterization of its nucleic acid. *Arch Biochem Biophys*, *181*(1), 300-307. https://doi.org/10.1016/0003-9861(77)90508-2

Solioz, M., Yen, H. C., & Marris, B. (1975). Release and uptake of gene transfer agent by *Rhodopseudomonas capsulata*. *J Bacteriol*, *123*(2), 651-657. https://doi.org/10.1128/jb.123.2.651-657.1975

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nat Rev Genet*, *16*(8), 472-482. https://doi.org/10.1038/nrg3962

Sundin, G. W., & Wang, N. (2018). Antibiotic resistance in plant-pathogenic bacteria. *Annu Rev Phytopathol*, *56*, 161-180. https://doi.org/10.1146/annurev-phyto-080417-045946

Suttle, C. A. (2007). Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol*, *5*(10), 801-812. https://doi.org/10.1038/nrmicro1750

Takeuchi, N., Kaneko, K., & Koonin, E. V. (2014). Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. *G3 (Bethesda)*, *4*(2), 325-339. https://doi.org/10.1534/g3.113.009845

Tomasch, J., Wang, H., Hall, A. T. K., Patzelt, D., Preusse, M., Petersen, J., Brinkmann, H., Bunk, B., Bhuju, S., Jarek, M., Geffers, R., Lang, A. S., & Wagner-Dobler, I. (2018). Packaging of *Dinoroseobacter shibae* DNA into gene transfer agent

particles is not random. *Genome Biol Evol*, *10*(1), 359-369. https://doi.org/10.1093/gbe/evy005

Touchon, M., Moura de Sousa, J. A., & Rocha, E. P. (2017). Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr Opin Microbiol*, *38*, 66-73. https://doi.org/10.1016/j.mib.2017.04.010

Wagner, P. L., & Waldor, M. K. (2002). Bacteriophage control of bacterial virulence. *Infect Immun*, *70*(8), 3985-3993. https://doi.org/10.1128/IAI.70.8.3985-3993.2002

Westbye, A. B., Beatty, J. T., & Lang, A. S. (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. https://doi.org/10.1016/j.mib.2017.05.003

Westbye, A. B., Kuchinski, K., Yip, C. K., & Beatty, J. T. (2016). The gene transfer agent RcGTA contains head spikes needed for binding to the *Rhodobacter capsulatus* polysaccharide cell capsule. *J Mol Biol*, *428*(2 Pt B), 477-491. https://doi.org/10.1016/j.jmb.2015.12.010

Westbye, A. B., O'Neill, Z., Schellenberg-Beaver, T., & Beatty, J. T. (2017). The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology (Reading)*, *163*(9), 1355-1363. https://doi.org/10.1099/mic.0.000519

Yen, H. C., Hu, N. T., & Marrs, B. L. (1979). Characterization of the gene transfer agent made by an overproducer mutant of *Rhodopseudomonas capsulata*. *J Mol Biol*, *131*(2), 157-168. https://doi.org/10.1016/0022-2836(79)90071-8

Zamani-Dahaj, S. A., Okasha, M., Kosakowski, J., & Higgs, P. G. (2016). Estimating the frequency of horizontal gene transfer using phylogenetic models of gene gain and loss. *Mol Biol Evol*, *33*(7), 1843-1857. https://doi.org/10.1093/molbev/msw062

Zerbini, F. M., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., Garcia, M. L., Hendrickson, R. C., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Lobocka, M., Oksanen, H. M., Robertson, D. L., Rubino, L., Sabanadzovic, S., . . . Varsani, A. (2023). Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2023). *Arch Virol*, *168*(7), 175. https://doi.org/10.1007/s00705-023-05797-4

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., & Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res*, *16*(9), 1099-1108. https://doi.org/10.1101/gr.5322306

Chapter 2

Machine-learning classification suggests that many alphaproteobacterial prophages may instead be gene transfer agents

Roman Kogay[1], Taylor B. Neely[1,2], Daniel P. Birnbaum[1,3], Camille R. Hankel[1,4], Migun Shakya[1,5], and Olga Zhaxybayeva[1,6]

[1]Department of Biological Sciences, Dartmouth College, Hanover, NH, USA
[2]Present address: Amazon.com Inc., Seattle, WA, USA
[3]Present address: School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[4]Present address: Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA
[5]Present address: Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA
[6]Department of Computer Science, Dartmouth College, Hanover, NH, USA

Published in *Genome Biology and Evolution* on 27 September 2019
(DOI: 10.1093/gbe/evz206)

Supplementary Material is available online (DOI: 10.1093/gbe/evz206)

Author contributions

DPB, MS and OZ designed the study. TBN, CRH, MS, and RK developed the software. RK used the software to analyze genomes and performed follow-up analyses. RK and OZ interpreted results. RK and OZ wrote the initial draft with the input from TN and MS. All authors revised and finalized the manuscript.

Abstract

Many of the sequenced bacterial and archaeal genomes encode regions of viral provenance. Yet, not all of these regions encode bona fide viruses. Gene transfer agents (GTAs) are thought to be former viruses that are now maintained in genomes of some bacteria and archaea and are hypothesized to enable exchange of DNA within bacterial populations. In *Alphaproteobacteria*, genes homologous to the 'head-tail' gene cluster that encodes structural components of the *Rhodobacter capsulatus* GTA (RcGTA) are found in many taxa, even if they are only distantly related to *Rhodobacter capsulatus*. Yet, in most genomes available in GenBank RcGTA-like genes have annotations of typical viral proteins, and therefore are not easily distinguished from their viral homologs without additional analyses. Here, we report a 'support vector machine' classifier that quickly and accurately distinguishes RcGTA-like genes from their viral homologs by capturing the differences in the amino acid composition of the encoded proteins. Our open-source classifier is implemented in Python and can be used to scan homologs of the RcGTA genes in newly sequenced genomes. The classifier can also be trained to identify other types of GTAs, or even to detect other elements of viral ancestry. Using the classifier trained on a manually curated set of homologous viruses and GTAs, we detected RcGTA-like 'head-tail' gene clusters in 57.5% of the 1,423 examined alphaproteobacterial genomes. We also demonstrated that more than half of the in silico prophage predictions are instead likely to be GTAs, suggesting that in many alphaproteobacterial genomes the RcGTA-like elements remain unrecognized.

Introduction

Viruses that infect bacteria (phages) are extremely abundant in biosphere (Keen, 2015). Some of the phages integrate their genomes into bacterial chromosomes as part of their infection cycle and survival strategy. Such integrated regions, known as prophages, are very commonly observed in sequenced bacterial genomes. For example, Touchon et al. (2016) report that 46% of the examined bacterial genomes contain at least one prophage (Touchon et al., 2016). Yet, not all of the prophage-like regions represent bona fide viral genomes (Koonin & Krupovic, 2018). One such exception is a Gene Transfer

Agent, or GTA for short (reviewed most recently in (Lang et al., 2017) and (Grull et al., 2018)). Many of genes that encode GTAs have significant sequence similarity to phage genes, but the produced tailed phage-like particles generally package pieces of the host genome unrelated to the "GTA genome" (Hynes et al., 2012; Tomasch et al., 2018). Moreover, the particles are too small to package complete GTA genome (Lang et al., 2017). Hence, GTAs are different from lysogenic viruses, as they do not use the produced phage-like particles for the purpose of their propagation.

Currently, five genetically unrelated GTAs are known to exist in Bacteria and Archaea (Lang et al., 2017). The best studied GTA is produced by the alphaproteobacterium *Rhodobacter capsulatus* and is referred hereafter as the RcGTA. Since RcGTA's discovery 45 years ago (Marrs, 1974), the genes for the related, or RcGTA-like, elements have been found in many of the alphaproteobacterial genomes (Shakya et al., 2017). For a number of *Rhodobacterales* isolates that carry RcGTA-like genes, there is an experimental evidence of GTA particle production (Y. Fu et al., 2010; Nagao et al., 2015; Tomasch et al., 2018). Seventeen of the genes of the RcGTA "genome" are found clustered in one locus and encode proteins that are involved in DNA packaging and head-tail morphogenesis (**Figure 1** and **Supplementary Table S1**). This locus is referred to as a 'head-tail cluster'. The remaining seven genes of the RcGTA genome are distributed across four loci and are involved in maturation, release and regulation of RcGTA production (Hynes et al., 2016). Since the head-tail cluster resembles a typical phage genome with genes organized in modules similar to those of a λ phage genome (Lang et al., 2017), and since many of its genes have homologs in bona fide viruses and conserved phage gene families (Shakya et al., 2017), the cluster is usually designated as a prophage by algorithms designed to detect prophage regions in a genome (Shakya et al., 2017). The RcGTA's classification as a prophage raises a possibility that some of the 'in silico'-predicted prophages may instead represent genomic regions encoding RcGTA-like elements.

Presently, to distinguish RcGTA-like genes from the truly viral homologs one needs to examine evolutionary histories of the RcGTA-like and viral homologs and to compare gene content of a putative RcGTA-like element to the RcGTA "genome". These

analyses can be laborious and often require subjective decision making in interpretations of phylogenetic trees. An automated method that could quickly scan thousands of genomes is needed. Notably, the RcGTA-like genes and their viral homologs have different amino acid composition (**Figure 1** and **Supplementary Figure S1**). Due to the purifying selection acting on the RcGTA-like genes at least in the *Rhodobacterales* order (Lang et al., 2012) and of their overall significantly lower substitution rates when compared to viruses (Shakya et al., 2017), we hypothesize that the distinct amino acid composition of the RcGTA-like genes is preserved across large evolutionary distances, and therefore the RcGTA-like genes can be distinguished from their bona fide viral homologs by their amino acid composition.

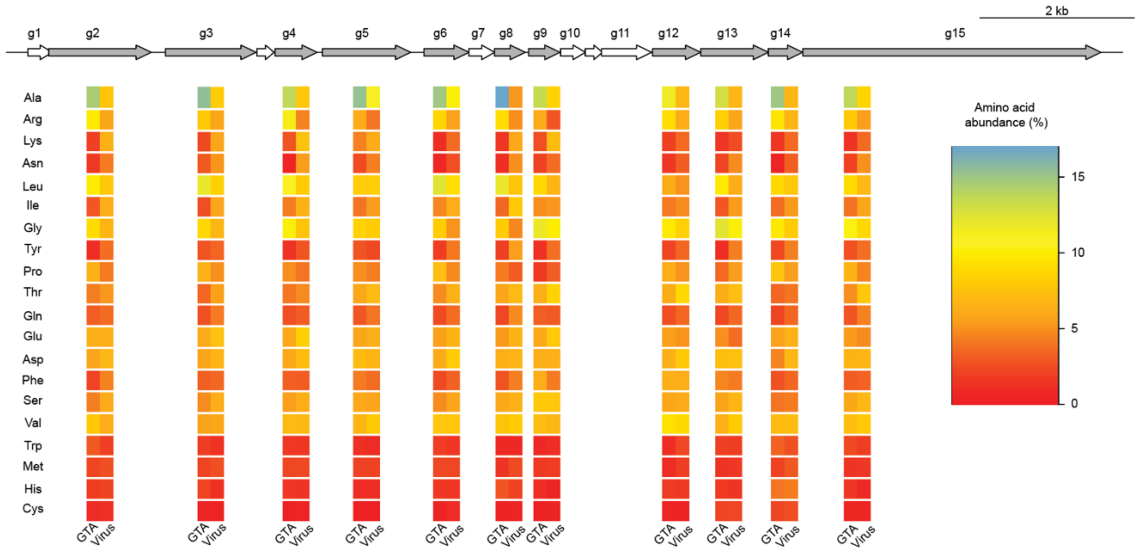

Figure 1. The 'head-tail' cluster of the *Rhodobacter capsulatus* GTA "genome" and the amino acid composition of viral and alphaproteobacterial homologs for some of its genes. Genes that are used in the machine learning classification are highlighted in grey. For those genes, the heatmap below a gene shows the relative abundance of each amino acid (rows) averaged across the RcGTA-like and viral homologs that were used in the classifier training (columns). The amino acids are sorted by the absolute difference in the average relative abundance between RcGTA-like and viral homologs, which was additionally averaged across 11 genes. The heatmaps of the amino acid composition in the individual homologs are shown in **Supplementary Figure S1**.

Support vector machine (SVM) is a machine learning algorithm that can quickly and accurately separate data into two classes from the differences in specific features within each class (Cortes & Vapnik, 1995). The SVM-based classifications have been successfully used to delineate protein families (e.g., DNA binding proteins (Bhardwaj et al., 2005), G-protein coupled receptors (Karchin et al., 2002), and herbicide resistance proteins (Meher et al., 2019)), to distinguish plastid and eukaryotic host genes (Kaundal et al., 2013), and to predict influenza host from DNA and amino acid oligomers found in the sequences of the flu virus (Xu et al., 2017). During the training step, the SVM constructs a hyperplane that best separates the two classes. During the classification step, data points that fall on one side of the hyperplane are assigned to one class, while those on the other side are assigned to the other class. In our case, the two classes of elements in need of separation are phages and GTAs, while their distinguishing features are several metrics that capture the amino acid composition of the encoding genes.

In this study, we developed, implemented, and cross-validated an SVM classifier that distinguishes RcGTA-like head-tail cluster genes from their phage homologs with high accuracy. We then applied the classifier to 1,423 alphaproteobacterial genomes to examine prevalence of putative RcGTA-like elements in this diverse taxonomic group and to assess how many of the RcGTA-like elements are mistaken for prophages in the in silico predictions.

Materials and Methods

The Support Vector Machine (SVM) classifier and its implementation

Let's denote as u a homolog of an RcGTA-like gene g that needs to be assigned to a class y, "GTA" ($y = -1$) or "virus" ($y = 1$). The assignment is carried out using a weighted soft-margin SVM classifier, which is trained on a dataset of m sequences $T^g = \{T_1^g, \ldots, T_m^g\}$ that are homologous to u (see **"SVM training data"** section below). The basis of the classification is the n-dimensional vector of features **x** associated with sequences u and T_i^g (see **"Generation of sequence features"** section below). Each sequence T_i^g is known to belong to a class y_i.

Using the training dataset T^g, we identify hyperplane that separates two classes as an optimal solution to the objective function:

$$\min\left(\frac{1}{2}||\mathbf{w}||^2 + \mathbf{C}\sum_{i=1}^{m}\xi_i\right) \text{ (eq. 1)}$$

subject to:

$$\forall_i\colon y_i(\mathbf{w}\mathbf{x_i} + b) \geq 1 - \xi_i, \text{where } \xi_i \geq 0, i = 1, \ldots, m \text{ (eq. 2)}$$

where $\mathbf{w}$ and b define the hyperplane $f(\mathbf{x}) = \mathbf{w}\mathbf{x_i} + b$ that divides the two classes, ξ_i is the slack variable that allows some training data points not to meet the separation requirement, and $\mathbf{C}$ is a regularization parameter, which is represented as an $m \times m$ diagonal matrix. The $\mathbf{C}$ matrix determines how lenient the soft-margin SVM is in allowing for genes to be misclassified: larger values "harden" the margin, while smaller values "soften" the margin by allowing more classification errors. The product $\mathbf{C}\xi$ represents the cost of misclassification. The most suitable values for the $\mathbf{C}$ matrix were determined empirically during cross-validation, as described in the **"Model training, cross validation, and assessment"** section below.

To solve equation 1, we represented this minimization problem in the Lagrangian dual form $L(\alpha)$:

$$\max_{\alpha_i} \quad L(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{i=j}^{m}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i\mathbf{x}_j) \text{ (eq. 3)}$$

subject to:

$$\forall_i: \sum_{i=1}^{m}\alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C, \qquad i = 1, \ldots, m$$

where K represents a kernel function. The minimization problem was solved using the convex optimization (CVXOPT) quadratic programming solver (Andersen et al., 2012).

The pseudocode of the algorithm for the weighted soft-margin SVM classifier training and prediction is shown in **Figure 2**.

SVM training data

To train the classifier, sets of "true viruses" (class $y = 1$) and "true GTAs" (class $y = -1$) were constructed separately for each RcGTA-like gene g. To identify the representatives of "true viruses", amino acid sequences of 17 genes from the RcGTA head-tail cluster were used as queries in BLASTP (E-value < 0.001; query and subject overlap by at least 60% of their length) and PSI-BLASTP searches (E-value < 0.001; query and subject overlap by at least 40% of their length; maximum of six iterations) of the viral RefSeq database release 90 (last accessed in November 2018; accession numbers of the viral entries are provided in **Supplementary Table S2**). BLASTP and PSI-BLAST executables were from the BLAST v. 2.6.0+ package (Altschul et al., 1997). The obtained homologs are listed in **Supplementary Table S3**. Due to few or no viral homologs for some of the queries, the final training sets T^g were limited to 11 out of 17 RcGTA-like head-tail cluster genes (g2, g3, g4, g5, g6, g8, g9, g12, g13, g14, g15; see **Supplementary Table S1** for functional annotations of these genes).

As the representatives of the "true GTAs", we used the RcGTA-like regions that were designated as such via phylogenetic and genome neighborhood analyses by Shakya et. al. (2017) (Shakya et al., 2017). To make sure that our "true GTAs" do not contain any other regions, we created a database of the 235 complete alphaprotebacterial genomes that were available in the RefSeq database prior to January 2014 (**Supplementary Table S4**). To identify the representatives of "true GTAs" in this database, amino acid sequences of 17 genes from the RcGTA head-tail cluster (Lang et al., 2017) were used as queries in BLASTP (E-value < 0.001; query and subject overlap by at least 60% of their length) and PSI-BLAST searches (E-value < 0.001; query and subject overlap by at least 40% of their length; maximum of six iterations) of the database. For each genome, the retrieved homologs were designated as an RcGTA-like head-tail cluster if at least 9 of the homologs had no more than 5,000 base pairs between any two adjacent genes. The maximum distance cutoff was based on the observed distances between the neighboring RcGTA head-tail cluster genes. This assignment was determined by clustering of the

1: Let $T = (T_1, \ldots, T_m)$ be an array of training sequences T_i, $1 \leq i \leq m$

2: Let $X = (x_i)$ be the feature sets for genes $T_i \in T$

3: Let $Y = (y_i)$ be the classes for genes $T_i \in T$

4: Let $W = (d_i)$ be the weights for genes $T_i \in T$

5: Let $y_i = -1$ if T_i is a GTA and $y_i = 1$ if it is a virus

6: Let $QUADPROG$ be a quadratic programming solver

7: **procedure** $SVMTrain(T, C)$

8: $Compute\ Lagrange - multipliers = QUADPROG(X, Y, C * W)$

9: Let $alphas = \{\alpha_i \in Lagrange - multipliers : \alpha_i > 10^{-5}\}$

10: Let $support\ vectors = \{T_i \in Lagrange - multipliers : \alpha_i > 10^{-5}\}$

11: **return** $alphas,\ support\ vectors$

12: **end procedure**

13:

14: Let u be an unclassified gene, where x_u is the feature set of u

15: **procedure** $SVMPredict(alphas,\ supportvectors,\ x_u)$

16: Let $score = 0$

17: for $\alpha_i \in alphas\ and\ T_i \in support\ vectors$ **do**

18: $score = score + (\alpha_i * y_i * K(x_i * x_u))$

19: **end for**

20: if $score < 0$ **then**

21: **return "**GTA"

22: else

23: **return** "virus"

24: **end if**

25: **end procedure**

Figure 2. The pseudocode of the SVM classifier algorithm that distinguishes RcGTA-like genes from the 'true' viruses. The algorithm is implemented in the GTA-Hunter software package (see "**Software Implementation**" section in **Materials and Methods**).

obtained homologs with the DBSCAN algorithm (Ester et al., 1996) using an in-house Python script (available in a **GitHub** repository; see **"Software Implementation"** section below). The resulting set of 88 "true GTAs" is provided in **Supplementary Table S5** and was verified to represent a subset of RcGTA-like elements that were identified by Shakya et. al. (2017) (Shakya et al., 2017)

Since GTA functionality has been extensively studied only in *Rhodobacter capsulatus* SB1003 (Lang et al., 2017) and horizontal gene transfer likely occurred multiple times between the putative GTAs and bacterial viruses (Hynes et al., 2016; Zhan et al., 2016), the bacterial homologs that were both too divergent from other bacterial RcGTA-like homologs and more closely related to the viral homologs were eliminated from the training sets to reduce possible noise in classification. To do so, for each of the 11 trainings sets T^g, all detected viral and bacterial homologs were aligned using MUSCLE v3.8.31 (Edgar, 2004) and then pairwise phylogenetic distances were estimated under PROTGAMMAJTT substitution model using RAxML version 8.2.11 (Stamatakis, 2014). For each bacterial homolog in a set T^g, the pairwise phylogenetic distances between it and all other bacterial homologs were averaged. This average distance was defined as an outlier distance (o) if it satisfied the inequality:

$$o > Q_3 + 1.5 * (Q_3 - Q_1) \text{ (eq. 4)}$$

where Q_1 and Q_3 are the first and third quartiles, respectively, of the distribution of the average distances for all bacterial homologs in the training set T^g. If an outlier distance was greater than the shortest distance from it to a viral homolog in the set T^g, the bacterial homolog was removed from the dataset. The alignments, list of removed sequences and the associated calculations are available in the **FigShare** repository.

Additionally, for each gene g, the sequences that had the same RefSeq ID (and therefore 100% amino acid identity) were removed from the training data sets. The final number of sequences in each training dataset are listed in **Table 1**.

Assignments of weights to the training set sequences

Highly similar training sequences can have an increased influence on the position of the hyperplane, as misclassification of two or more similar sequences can be

Table 1. Number of the RcGTA homologs in the "true GTA" and "true virus" training datasets.

Gene	"true GTAs"	"true viruses"
g2	69	1646
g3	65	769
g4	62	465
g5	67	627
g6	61	19
g8	62	96
g9	66	61
g12	63	12
g13	73	57
g14	67	124
g15	67	155

considered less optimal than misclassification of only one sequence. This could be a problem for our classifier, since there is generally a highly unequal representation of taxonomic groups in the RefSeq database. To correct for this taxonomic bias, a weighting scheme was introduced into the soft-margin of the SVM classifier during training. To do so, sequences in each training set $T^g = \{T_1, \ldots, T_m\}$ were aligned in MUSCLE v3.8.31 (Edgar, 2004) (The alignments are available in the **FigShare** repository). For each pair of sequences in a training set T^g, phylogenetic distances were calculated in RAxML version 8.2.11 (Stamatakis, 2014) under the best substitution model (PROTGAMMAAUTO; the selected substitution matrices are listed in the **Supplementary Table S6**). The farthest-neighbor hierarchical clustering method was used to group sequences with distances below a specified threshold t. Weight d_i of each sequence in a group was defined as a reciprocal of the number of genes in the group. These weights are used to adjust the cost of misclassification by multiplying C_{ii} for each sequence T_i by d_i. The most suitable value of t was determined empirically during cross-validation, as described in the **"Model training, cross validation, and assessment"** section below.

Generation of sequence features

To use amino acid sequences in the SVM classifier, each sequence was transformed to an n-dimensional vector of compositional features. Three metrics that

capture different aspects of sequence composition were implemented: frequencies of "words" of size k (k-mers), pseudo amino-acid composition (PseAAC), and physicochemical properties of amino acids.

In the first feature type, amino acid sequence of a gene is broken into a set of overlapping subsequences of size k, and frequencies of these n unique k-mers form a feature vector **x**. Values of k equal to 1, 2, 3, 4, 5 and 6 were evaluated for prediction accuracy (see the **"Model training, cross validation, and assessment"** section below).

The second feature type, pseAAC, has n = (20+λ) dimensions and take into account frequencies of 20 amino acids, as well as correlations of hydrophobicity, hydrophilicity and side-chain mass of amino acids that are □ positions apart in the sequence of the gene (after (Chou, 2001). More precisely, PseAAC feature set **x** of a sequence of length L consisting of amino acids $R_1R_2\ldots R_L$ is defined as follows:

$$x_i = \begin{cases} \dfrac{r_i}{\sum_{i=1}^{20} r_i + \omega \sum_{k=1}^{\lambda} s_k}, & \text{if } 1 \le i \le 20, \\ \dfrac{\omega s_{j-20}}{\sum_{i=1}^{20} r_i + \omega \sum_{k=1}^{\lambda} s_k}, & \text{if } 21 \le j \le 20 + \lambda \end{cases} \quad (\text{eq. } 5)$$

where r_i is the frequency of the i-th amino acid (out of 20 possible), ω is a weight constant for the order effect that was set to 0.05, and s_k (k = 1, …, □) are sequence order-correlation factors. These factors are defined as

$$s_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k} \quad (\text{eq. } 6)$$

where

$$J_{i,j} = \frac{1}{3}\left[\left(H_1(R_j) - H_1(R_i)\right)^2 + \left(H_2(R_j) - H_2(R_i)\right)^2 + \left(M(R_j) - M(R_i)\right)^2\right] \quad (\text{eq. } 7)$$

and $H_1(R_i)$, $H_2(R_i)$, and $M(R_i)$ denote the hydrophobicity, hydrophilicity, and side-chain mass of amino acid R_i, respectively. The $H_1(R_i)$, $H_2(R_i)$, and $M(R_i)$ scores were subjected to a conversion as described in the following equation:

$$\begin{cases} H_1(i) = \dfrac{H_1^0(i) - \sum_{i=1}^{20} \dfrac{H_1^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20}\left[H_1^0(i) - \sum_{i=1}^{20} \dfrac{H_1^0(i)}{20}\right]^2}{20}}} \\ H_2(i) = \dfrac{H_2^0(i) - \sum_{i=1}^{20} \dfrac{H_2^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20}\left[H_2^0(i) - \sum_{i=1}^{20} \dfrac{H_2^0(i)}{20}\right]^2}{20}}} \\ M(i) = \dfrac{M^0(i) - \sum_{i=1}^{20} \dfrac{M^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20}\left[M^0(i) - \sum_{i=1}^{20} \dfrac{M^0(i)}{20}\right]^2}{20}}} \end{cases} \quad (\text{eq. } 8),$$

where $H_1^0(i)$ is the original hydrophobicity value of the i-th amino acid, $H_2^0(i)$ is hydrophilicity value, and $M^0(i)$ is the mass of its side chain. Values of λ equal to 3 and 6 were evaluated for prediction accuracy (see the **"Model training, cross validation, and assessment"** section below).

The third feature type relies on classification of amino acids into 19 overlapping classes of physicochemical properties (**Supplementary Table S7**; after (Kaundal et al., 2013)). For a given sequence, each of its encoded amino acids was counted towards one of the 19 classes, and the overall scores for each class were normalized by the length of the sequence to form n = 19-dimensional feature vector **x**.

Model training, cross validation, and assessment

For each GTA gene, parameter, and feature type, the accuracy of the classifier was evaluated using a five-fold cross-validation scheme, in which a dataset was randomly divided into five different sub-samples. Four parts were combined to form the training set, while the fifth part was used as the validation set and its SVM-assigned

classifications compared to the known classes. This step was repeated five times, so that every set was tested as a known class at least once.

For each class y ("GTA" and "Virus"), the results were evaluated by their accuracy scores, defined as the number of correctly classified homologs divided by the total number of homologs that were tested. The cross-validation procedure was repeated ten times to reduce the partitioning bias, and the generated results were averaged, resulting in an Average Accuracy Score (AAS) for each gene and each class. To ensure that "GTA" and "Virus" classes had equal impact on the accuracy assessment, each class was assigned a weight of 0.5. The final, Weighted Accuracy Score (WAS) was calculated as:

$$\mathrm{WAS}^{\mathrm{g}} = 100 * (\mathrm{AAS}^{\mathrm{g}}_{\mathrm{GTA}} * 0.5 + \mathrm{AAS}^{\mathrm{g}}_{\mathrm{Virus}} * 0.5) \quad (\mathrm{eq.}\, 9)$$

The most suitable "softness" of the SVM margin was determined by trying all possible combinations of several raw diagonal values of the matrix $\mathbf{C}$ (0.01, 0.1, 1, 100, 10000) and the threshold t (0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1). The set of parameters and features that resulted in the highest WAS was defined as the optimal set for a gene g. If multiple parameter and feature sets resulted in the equally highest WAS, we applied the following parameter selection criteria, in the priority order listed, until only one parameter set was left: first, we selected parameter set(s) with k-mer size that on average performed better than other k-mer sizes; second, we avoided parameter set(s) that included PseAAC and physicochemical composition features; third, we selected parameter set(s) with the value of $\mathbf{C}$ that gives the highest average accuracy across the remaining parameter sets; and finally, we opted for the parameter set with the value of t that also gives the highest WAS across the remaining parameter sets. Additionally, we evaluated classifier accuracy using the Matthews correlation coefficient (MCC) (Matthews, 1975).

Selection of alphaproteobacterial genomes for testing the presence of RcGTA-like genes

From the alphaproteobacterial genomes deposited to the RefSeq database between January 2014 and January 2019, we selected 636 complete and 789 high-quality draft

genomes, with the latter defined as genome assemblies with N50 length >400 kbp. The taxonomy of each genome was assigned using the GTDB-Tk toolkit (Parks et al., 2018). The GTDB assignment is based on the combination of Average Nucleotide Identity (Jain et al., 2018) and phylogenetic placement on the reference tree (as implemented in the pplacer program (Matsen et al., 2010). Three of the 1,425 genomes could not be reliably placed into a known alphaproteobacterial order, and hence were left unclassified. Two of the 1,425 genomes were removed from further analyses due to their classification outside of the *Alphaproteobacteria* class, resulting in 635 complete and 788 high-quality genomes in our dataset (**Supplementary Table S8**).

Detection of RcGTA-like genes and head-tail clusters in *Alphaproteobacteria*

The compiled training datasets of the RcGTA-like genes (see the **"SVM training data"** section) were used as queries in BLASTP (E-value < 0.001; query and subject overlap by at least 60% of their length) searches of amino acid sequences of all annotated genes from the 1,423 alphaproteobacterial genomes. Acquired homologs of unknown affiliation (sequences u) were then assigned to either "GTA" or "virus" category by running the SVM classifier with the identified optimal parameters for each gene g (**Table 2**).

The proximity of the individually predicted RcGTA-like genes in each genome was evaluated by running the DBSCAN algorithm (Ester et al., 1996) implemented in an in-house Python script (available in a **GitHub** repository; see **"Software Implementation"** section below). The retrieved homologs were designated as an RcGTA-like head-tail cluster only if at least 6 of the RcGTA-like genes had no more than 8,000 base pairs between any two adjacent genes. The maximum distance cutoff was increased from the 5,000 base pairs used for the clustering of homologs in the training datasets (see **"SVM Training Data"** section) because the SVM classifier evaluates only 11 of the 17 RcGTA-like head-tail cluster homologs and therefore the distances between some of the identified RcGTA-like genes can be larger.

To reduce the bias arising from the overrepresentation of particular taxa in the estimation of the RcGTA-like cluster abundance in *Alphaproteobacteria*, the 1,423 genomes were grouped into Operational Taxonomic Units (OTUs) by computing pairwise

Average Nucleotide Identity (ANI) using the FastANI v1.1 program (Jain et al., 2018) and defining boundaries between OTUs at the 95% threshold. Since not all OTUs consist uniformly of genomes that were either all with or all without the RcGTA-like clusters, each RcGTA-like cluster in an OTU was assigned a weight of "1/[number of genomes in an OTU]". The abundance of the RcGTA-like clusters in different alphaproteobacterial orders was corrected by summing up the weighted numbers of RcGTA-like clusters.

Table 2. The combinations of features and parameters that showed the highest weighted accuracy score (WAS) in cross-validation. The listed parameter sets were used in predictions of the RcGTA-like genes in 1,423 alphaproteobacterial genomes. See **Materials and Methods** for the procedure on selecting one parameter set in the cases where multiple parameter sets had the identical highest WAS.

Gene	Weighted Accuracy Score, WAS (%)	Matthews Correlation Coefficient, MCC	k-mer (size)	PseAAC (value of λ)	Grouping based on physicochemical properties of amino acids	C	t
g2	100	1	2	-[1]	-	10000	0.02
g3	100	1	3	-	-	10000	0.02
g4	100	1	3	3	-	10000	0.02
g5	100	1	3	-	-	100	0.02
g6	95.9	0.88	4	-	+	0.1	0.02
g8	99.4	0.98	2	3	-	0.1	0.03
g9	100	1	2	-	-	100	0.1
g12	95.6	0.90	5	-	-	10000	0.05
g13	99.1	0.98	2	-	-	100	0
g14	99.6	0.99	6	6	-	0.01	0.03
g15	99.7	0.99	2	-	-	10000	0.02

[1] throughout the table, "-" denotes that the feature type was not used

Software Implementation

The above described SVM classifier, generation of sequence features, and preparation and weighting of training data are implemented in a Python program called "GTA-Hunter". The source code of the program is available via GitHub at https://github.com/ecg-lab/GTA-Hunter-v1. The repository also contains training data for the detection of the RcGTA-like heat-tail cluster genes, examples of how to run the

program, and the script for clustering of the detected RcGTA-like genes using the DBSCAN algorithm.

Assessment of prevalence of the RcGTA-like clusters among putative prophages

Putative prophages in the 1,423 alphaproteobacterial genomes were predicted using the PHASTER web server ((Arndt et al., 2016); accessed in January, 2019). The PHASTER program was chosen due to its solid performance in benchmarking studies (de Sousa et al., 2018) and its useful scoring system that ranks predictions based on a prophage region completeness (Song et al., 2019). To restrict our evaluation to likely functional prophages, only predicted prophages with the PHASTER score >90 (i.e., classified as "intact" prophages) were retained for further analyses. The proportion of these predicted intact prophages classified by the GTA-Hunter as "GTA"s was calculated by comparing the overlap between the genomic locations of the predicted intact prophages and the putative RcGTA-like regions.

Construction of the alphaproteobacterial reference phylogeny

From the set of 120 phylogenetically informative proteins (Parks et al., 2017), 83 protein families that are present in a single copy in >95% of 1,423 alphaproteobacterial genomes were extracted using hmmsearch (E-value < 10-7) via modified AMPHORA2 scripts (Wu & Scott, 2012) (**Supplementary Table S9**). For each protein family, homologs from Escherichia coli str. K12 substr. DH10B and Pseudomonas aeruginosa PAO1 genomes (also retrieved using hmmsearch, as described above) were added to be used as an outgroup in the reconstructed phylogeny. The amino acid sequences of each protein family were aligned using MUSCLE v3.8.31 (Edgar, 2004). Individual alignments were concatenated, keeping each alignment as a separate partition in further phylogenetic analyses (Chernomor et al., 2016). The most suitable substitution model for each partition was selected using ProteinModelSelection.pl script downloaded from https://github.com/stamatak/standard-RAxML/tree/master/usefulScripts. Gamma distribution with 4 categories was used to account for rate heterogeneity among sites (Yang, 1994). The maximum likelihood phylogenetic tree was reconstructed with IQ-TREE v 1.6.7 (Nguyen et al., 2015). One thousand ultrafast bootstrap replicates were

used to get support values for each branch (Hoang et al., 2018; Minh et al., 2013). The concatenated sequence alignment in PHYLIP format and the reconstructed phylogenetic tree in Newick format are available in the **FigShare** repository.

Examination of conditions associated with the decreased fitness of the knock-out mutants of the RcGTA-like head-tail cluster genes

From the three genomes that are known to contain RcGTA-like clusters (*Caulobacter crescentus* NA100, *Dinoroseobacter shibae* DFL-12, and *Phaeobacter inhibens* BS107), fitness experiments data associated with the knock-out mutants of the RcGTA-like head-tail cluster genes were retrieved from the Fitness Browser ((Price et al., 2018); accessed in May, 2019 via http://fit.genomics.lbl.gov/cgi-bin/myFrontPage.cgi). Price et al. (2018) defined gene fitness as the log2 change in abundance of knock-out mutants in that gene during the experiment. For our analyses, the significantly decreased fitness of each mutant was defined as a deviation from the fitness of 0 with a $|\mathrm{t} - \mathrm{score}| \geq 4$. The conditions associated with the significantly decreased fitness were compared across the RcGTA-like head-tail cluster genes in all three genomes.

Data deposition

Sequence alignments and phylogenetic trees are available in a **FigShare** repository at DOI 10.6084/m9.figshare.8796419. The Python source code of the described classifier and additional scripts used in the analyses are available via a **GitHub** repository at https://github.com/ecg-lab/GTA-Hunter-v1.

Results

GTA-Hunter is an effective way to distinguish RcGTA-like genes from their viral homologs

The performance of the developed SVM classifier depends on values of parameters that determine type and composition of sequence features, specify acceptable levels of misclassification, and account for biases in taxonomic representation of the sequences in the training sets. To find the most effective set of parameters, for each of the

11 RcGTA-like head-tail genes with the sufficient number of homologs available (**Figure 1**; also, see **Materials and Methods** for details) we evaluated the performance of 1,435 different combinations of the parameters using a cross-validation technique (**Supplementary Table S10**).

Generally, the classifiers that only use k-mers as the feature have higher median WAS values than the classifiers that solely rely either on physicochemical properties of amino acids or on pseudo amino acid composition (PseAAC) (**Supplementary Figure S2** and **Supplementary Table S10**), indicating that the conservation of specific amino acids blocks is important in delineation of RcGTA-like genes from their viral counterparts. However, the WAS values are lower for the large k-mer sizes (**Supplementary Figure S2**), likely due to the feature vectors becoming too sparse. Consequently, parameter combinations with values of k above 6 were not tested. The WAS values are also lower for k=1, likely due to the low informativeness of the feature. The lowest observed WAS values involve usage of physicochemical properties of proteins as a feature (**Supplementary Figure S2** and **Supplementary Table S10**), suggesting the conservation of physicochemical properties of amino acids among proteins of similar function in viruses and RcGTA-like regions despite their differences in the amino acid composition. The more sophisticated re-coding of physicochemical properties of amino acids as the PseAAC feature performs better, but for all genes its performance is worse than the best-performing k-mer (**Supplementary Figure S2** and **Supplementary Table S10**).

For several genes, the highest value of WAS was obtained with multiple combinations of features and parameter values (**Supplementary Table S10**). Based on the above-described observations of the performance of individual features, we preferred parameter sets that did not include PseAAC and physicochemical composition features, and selected k-mer size that on average performed better than other k-mer sizes (see **Materials and Methods** for the full description of the parameter selection procedure).

For individual genes, the WAS of the selected parameter set ranges from 95.6 to 100% (**Table 2**), with 5 out of 11 genes reaching the WAS of 100%. The two genes with the highest WAS below 99% (g6 and g12) have the smallest number of viral homologs

available for training (**Table 2**). Additionally, several viral homologs in the training datasets for g6 and g12 genes have smaller phylogenetic distances to "true GTA" homologs than to other "true virus" homologs (**Supplementary Table S11**). As a result, the SVM classifier erroneously categorizes some of the RcGTA-like g6 and g12 genes as "viral", resulting in the reduced classifier efficacy (**Supplementary Table S10**).

Assessment of accuracy using the Matthews correlation coefficient (MCC) generally agrees with the results based on WAS (**Table 2** and **Supplementary Table 10**). For 10 out of 11 genes, the set of parameters with the highest WAS also has the highest MCC. For gene g6, there are sets of parameters with higher MCC than the MCC for set of parameters with the highest WAS, but the differences among the MCC values are small (**Supplementary Table S10**). Therefore, the combinations of features and parameters chosen using the WAS scheme (**Table 2**) were selected to classify homologs of the RcGTA genes in the 1,423 alphaproteobacterial genomes (**Supplementary Table S8**).

GTA-Hunter predicts abundance of RcGTA-like head-tail clusters in *Alphaproteobacteria*

The 1,423 examined alphaproteobacterial genomes contain 7,717 homologs of the 11 RcGTA genes. The GTA-Hunter classified 6,045 of these homologs as "GTA" genes (**Supplementary Table S12**). However, many genomes are known to contain regions of decaying viruses that may be too divergent to be recognizably "viral" and there is at least one known case of horizontal gene transfer of several GTA genes into a viral genome (Zhan et al., 2016), raising a possibility that some of the predicted "GTA" genes may not be part of "true GTA" genomic regions. To minimize such false positives, we imposed an extra requirement of multiple predicted RcGTA-like genes to be in proximity on a chromosome. Specifically, we called a genomic region the putative RcGTA-like cluster only if it consisted of at least 6 genes classified as "GTA". We found that the RcGTA-like clusters defined that way are present in one (and only one) copy in 818 of the 1,423 (~57.5%) examined alphaproteobacterial genomes (**Supplementary Table S13** and **Table 3**). Uneven taxonomic representation of *Alphaproteobacteria* among the analyzed genomes may inflate this estimation of the abundance of the GTA-harboring genomes within the class. To correct for this potential bias, 1,423 genomes were grouped into 797

Operational Taxonomic Units (OTUs) based on the average nucleotide identity (ANI) of their genomes (**Supplementary Table S14**). Although indeed some taxonomic groups are overrepresented in the set of 1,423 genomes, in 450 of the 797 OTUs (56.4%) all OTU members contain the putative RcGTA-like clusters (**Supplementary Table S14**).

Table 3. Distribution of prophages and RcGTA-like elements across different orders within class *Alphaproteobacteria*.

Order	Number of genomes	Number of prophages	Number of RcGTA-like clusters	Number of OTUs	Corrected abundance of RcGTA-like clusters[1]	Percentage of OTUs that have RcGTA-like clusters
Acetobacterales	62	34	0	34	0	0
Azospirillales	13	10	0	12	0	0
Caedibacterales	1	0	0	1	0	0
Caulobacterales	50	30	39	45	35	78
Elsterales	1	0	0	1	0	0
Kiloniellales	5	1	0	3	0	0
Oceanibaculales	2	1	0	2	0	0
Paracaedibacterales	1	2	0	1	0	0
Parvibaculales	5	5	2	5	2	40
Pelagibacterales	5	0	0	5	0	0
Rhizobiales	730	763	435	300	155	52
Rhodobacterales	241	318	208	174	150	86
Rhodospirillales	24	10	0	15	0	0
Rickettsiales	70	18	0	24	0	0
Sneathiellales	2	1	0	2	0	0
Sphingomonadales	207	115	132	169	110	65
Thalassobaculales	1	0	0	1	0	0
Unclassified order 1	1	0	0	1	0	0
Unclassified order 2	1	2	1	1	1	100
Unclassified order 3	1	2	1	1	1	100

[1] See "**Detection of RcGTA-like genes and head-tail clusters in *Alphaproteobacteria*"** subsection of the **Materials and Methods** for explanation about the correction.

RcGTA-like clusters are widely distributed within a large sub-clade of *Alphaproteobacteria*

The 818 genomes with the RcGTA-like gene clusters detected in this study are not evenly distributed across the class (**Table 3**), but are found only in a clade that includes

seven orders (clade 4 in **Figure 3**). Overall, 66% of the examined OTUs within the clade 4 are predicted to have an RcGTA-like cluster (**Table 3**). RcGTA-like clusters are most abundant in clade 6 (**Figure 3**), a group that consists of the orders *Rhodobacterales* and *Caulobacterales* (**Table 3**). Although the two unclassified orders that contain RcGTA-like clusters are represented by only two genomes (clades 2 and 3 in **Figure 3**), their position on the phylogenetic tree of *Alphaproteobacteria* suggests that the RcGTA-like element may have originated earlier than was proposed by Shakya et. al. (Shakya et al., 2017) (clade 5 on **Figure 3**). Given that RcGTA-like head-tail cluster genes are readily detectable in viral genomes, it is unlikely that the RcGTA-like clusters remained completely undetectable in the examined genomes outside of the clade 4 due to the sequence divergence. Therefore, an RcGTA-like element was unlikely to be present in the last common ancestor of all *Alphaproteobacteria* (clade 7 on **Figure 3**), which was suggested when only a limited number of genomic data was available (Lang & Beatty, 2007).

Most of the detected RcGTA-like clusters can be mistaken for prophages

Among the 818 detected RcGTA-like clusters, the functional annotations of the 11 examined genes were similar to the prophages and none of them refer to a "gene transfer agent" (data not shown). Since at least 11 of the 17 RcGTA head-tail cluster genes have detectable sequence similarity to viral genes (**Supplementary Table S3**), it is likely that, if not recognized as GTAs, many of the putative RcGTA-like clusters will be designated as "prophages" in genome-wide searches of prophage-like regions. To evaluate this hypothesis, we predicted prophages in the set of 1,423 alphaproteobacterial genomes, and limited our analyses to the predicted prophage regions that are more likely to be functional integrated viruses ('intact' prophages; see **Materials and Methods** for the criteria). Indeed, of the 1,235 'intact' prophage regions predicted in the clade 4 genomes, 664 (54%) coincide with the RcGTA-like clusters (**Figure 4**). Conversely, 664 out of 818 of the predicted RcGTA-like clusters (81%) are classified as intact prophages. Of the 351 RcGTA-like clusters that contain all 11 examined genes, 323 (92%) are classified as intact prophages.

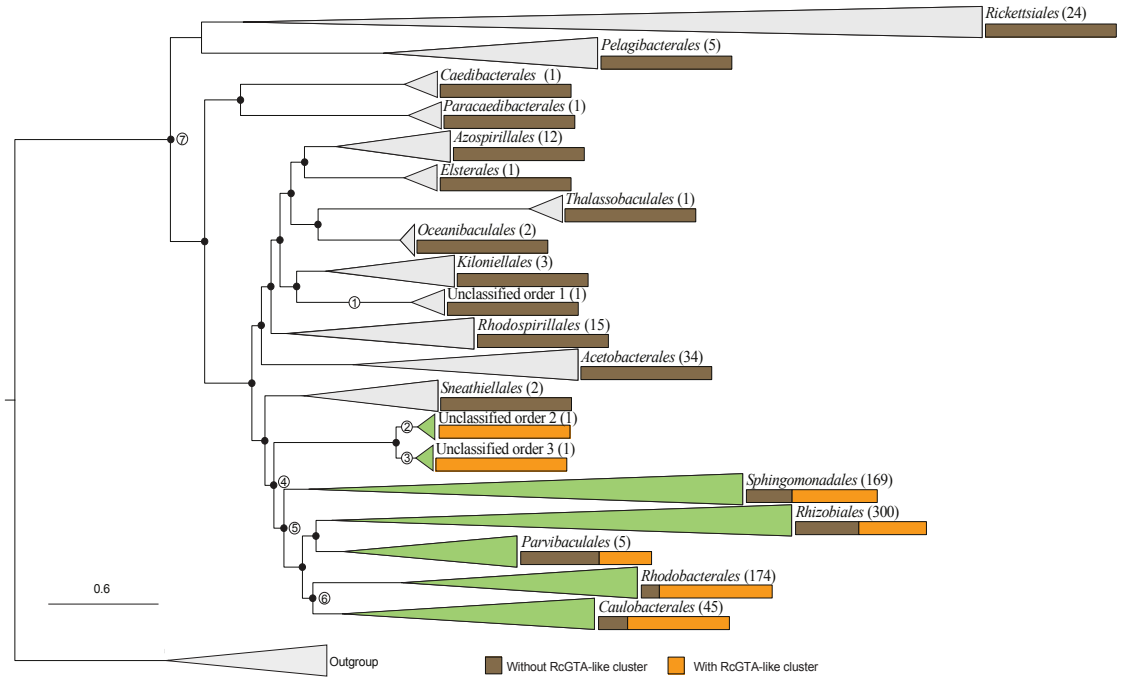

Figure 3. Distribution of the detected RcGTA-like clusters across the class Alphaproteobacteria. The presence of RcGTA-like clusters is mapped to a reference phylogenetic tree that was reconstructed from a concatenated alignment of 83 marker genes (See **Materials and Methods** and **Supplementary Table S9**). The branches of the reference tree are collapsed at the taxonomic rank of "order", and the number of OTUs within the collapsed clade is shown in parentheses next to the order name. Orange and brown bars depict the proportion of OTUs with and without the predicted RcGTA-like clusters, respectively. The orders that contain at least one OTU with an RcGTA-like cluster are colored in green. Nodes 1, 2 and 3 mark the last common ancestors of the unclassified orders. Node 4 marks the lineage where, based on this study, the RcGTA-like element should have already been present. Nodes 5 and 7 mark the lineages that were previously inferred to represent last common ancestor of the RcGTA-like element by Shakya et al. (2017) and Lang and Beatty (2007), respectively. Node 6 marks the clade where RcGTA-like elements are the most abundant. The tree is rooted using homologs from *Escherichia coli* str. K12 substr. DH10B and *Pseudomonas aeruginosa* PAO1 genomes. Branches with ultrafast bootstrap values >= 95% are marked with black circles. The scale bar shows the number of substitutions per site. The full reference tree is provided in the **FigShare** repository.

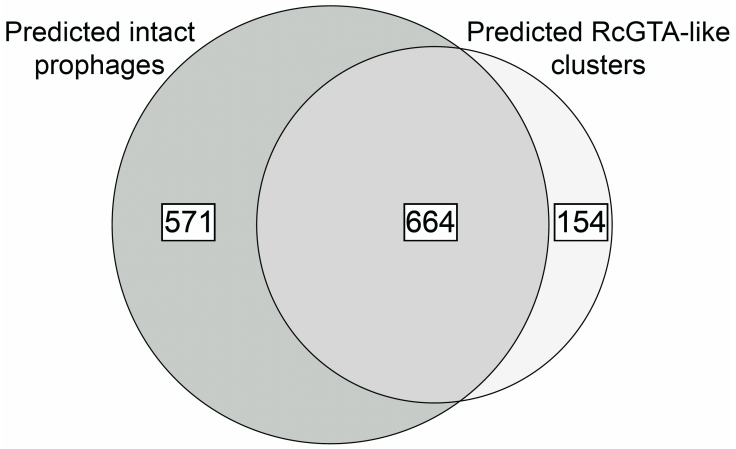

Figure 4. An overlap between prophage and GTA predictions. The "predicted RcGTA-like clusters" set refers to the GTA-Hunter predictions, while the "predicted intact prophages" set denotes predictions made by the PHASTER program (Arndt et al., 2016) on the subset of the genomes that are found within clade 4 (**Figure 3**).

Interestingly, within 818 genomes that contain RcGTA-like clusters, the average number of predicted intact prophages is 1.23 per genome (**Figure 5**), which is significantly higher than 0.51 prophages per genome in genomes not predicted to contain RcGTA-like clusters (p-value < $0.22 * 10^{-17}$; Mann-Whitney U test). If the 664 RcGTA-like regions classified as intact prophages are removed from the genomes that contain them, the average number of predicted 'intact' prophages per genome drops to 0.42 (**Figure 5**) and the difference becomes insignificant (p-value = 0.1492; Mann-Whitney U test). This analysis suggests that an elevated number of the observed predicted prophage-like regions in some alphaproteobacterial genomes may be due to the presence of unrecognized RcGTA-like elements.

Discussion

Our study demonstrates that RcGTA-like and bona fide viral homologs can be clearly separated from each other using a machine learning approach. The highest accuracy of the classifier is achieved when it primarily relies on short amino acid k-mers present in the examined genes. This suggests that the distinct primary amino acid

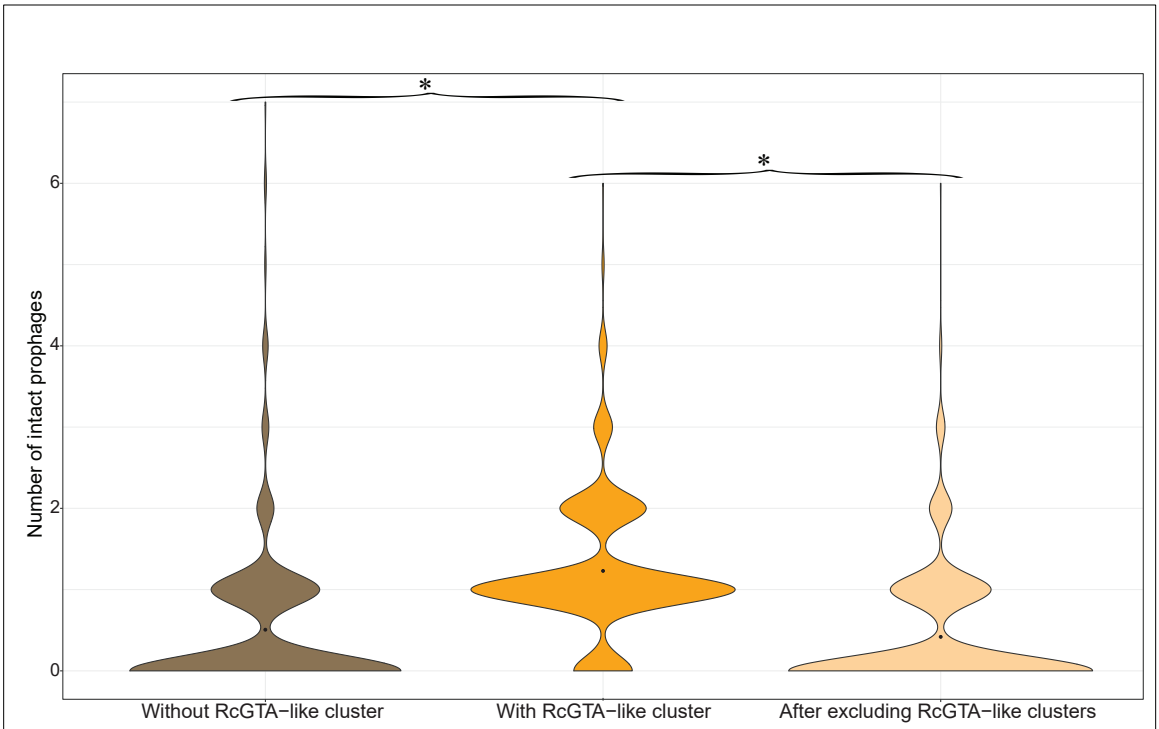

Figure 5. The number of predicted 'intact' prophages in alphaproteobacterial genomes. The 1,423 genomes were divided into two groups: those without GTA-Hunter-predicted RcGTA-like clusters (in brown) and those with these RcGTA-like clusters (in dark orange). For the latter group, the number of prophages was re-calculated after the RcGTA-like clusters that were designated as prophages were removed (in light orange). The distribution of the number of predicted intact prophages within each dataset is shown as a violin plot with the black point denoting the average value. The datasets with significantly different average values are denoted by asterisks ($p < 0.001$; Mann-Whitney U test).

composition of the RcGTA-like and truly viral proteins is what allows the separation of the two classes of elements (**Figure 1**). However, the cause of the amino acid preferences of the RcGTA-like genes, and especially enrichment of the encoded proteins in alanine and glycine amino acids (**Figure 1**), remains unknown. Given the structure of the genetic code, the skewed amino acid composition may be the driving force behind the earlier described significantly higher %G+C of the genomic region encoding the RcGTA-like head-tail cluster than the average %G+C in the host genome (Shakya et al., 2017). Regardless of the cause of the skewed amino acid composition, the successful

identification of the putative RcGTA-like elements in alphaproteobacterial taxa only distantly related to *Rhodobacter capsulatus* (clade 4 in **Figure 3**) suggests that the selection to maintain these elements likely extends beyond the *Rhodobacterales* order. Nevertheless, whether these putative elements indeed encode GTAs, as we currently understand them, remains to be experimentally validated.

The benefits associated with the GTA production that would underlie the selection to maintain them remain unknown. In a recently published high-throughput screen for phenotypes associated with specific genes (Price et al., 2018), knockout of the RcGTA-like genes in the three genomes that encode the RcGTA-like elements resulted in decreased fitness of the mutants (in comparison to the wild type) under some of the tested conditions (**Supplementary Table S15**). Interestingly, the conditions associated with the most statistically significant decreases in fitness correspond to the growth on non-glucose sugars, such as D-Raffinose, β-Lactose, D-Xylose and m-Inositol. Overall, carbon source utilization is the most common condition that elicits statistically significant fitness decreases in the mutants. The RcGTA production was also experimentally demonstrated to be stimulated by carbon depletion (Westbye, O'Neill, et al., 2017). Further experimental work is needed to identify the link between the RcGTA-like genes expression and carbon utilization. Conversely, absence of the RcGTA-like elements in some of the clade 4 genomes (**Figure 3**) indicates that in some ecological settings RcGTA-like elements are either deleterious or "useless" and thus their genes were either purged from the host genomes (if RcGTA-like element evolution is dominated by vertical inheritance) or not acquired (if horizontal gene transfer plays a role in the RcGTA-like element dissemination).

Previous analyses inferred that RcGTA-like elements had evolved primarily vertically, with few horizontal gene exchanges between closely related taxa (Hynes et al., 2016; Lang & Beatty, 2007; Shakya et al., 2017). Under this hypothesis, the distribution of the RcGTA-like head-tail clusters in alphaproteobacterial genomes suggests that RcGTA-like element originated prior to the last common ancestor of the taxa in clade 4 (**Figure 3**). This places the origin of the RcGTA-like element to even earlier timepoint than the one proposed in Shakya et al. (2017) (Shakya et al., 2017). However, it should be noted that our inference is sensitive to the correctness of the inferred relationships of taxa

within the alphaproteobacterial class, which remain to be disputed due to compositional biases and unequal rates of evolution of some alphaproteobacterial lineages (Munoz-Gomez et al., 2019). The most recent phylogenetic inference that takes into account these heterogeneities (Munoz-Gomez et al., 2019) is different from the reference phylogeny shown in **Figure 3**. Relevant to the evolution of RcGTA-like elements, on the phylogeny in Munoz-Gomez et al. (2019) the order *Pelagibacterales* is located within the clade 4 instead of being one of the early-branching alphaproteobacterial orders (**Figure 3**). No RcGTA-like clusters were detected in *Pelagibacterales*, although in our analyses the order is represented by only five genomes. Better sampling of genomes within this order would be needed either to show a loss of the RcGTA-like element in this order or to re-assess the hypothesis about origin and transmission of the RcGTA-like elements within *Alphaproteobacteria*.

Genes in the detected RcGTA-like head-tail clusters remain mainly unannotated as "gene transfer agents" in GenBank records, and therefore they can be easily confused with prophages. For example, recently described "conserved prophage" in *Sphingomonadales* (Viswanathan et al., 2017) is predicted to be an RcGTA-like element by GTA-Hunter. Incorporation of a GTA-Hunter-like machine learning classification into an automated genome annotation pipeline will help improve quality of the gene annotations in GenBank records and facilitate discovery of GTA-like elements in other taxa. Moreover, application of the presented GTA-Hunter program is not limited to the detection of the RcGTA-like elements. With appropriate training datasets, the program can be applied to the detection of GTAs that do not share evolutionary history with the RcGTA (Lang et al., 2017) and of other elements that are homologous to viruses or viral sub-structures, such as type VI secretion system (Leiman et al., 2009) and encapsulins (Giessen & Silver, 2017).

Acknowledgements

This work was supported by the National Science Foundation [NSF-DEB 1551674 to O.Z.]; the Simons Foundation Investigator in Mathematical Modeling of Living Systems [327936 to O.Z.]; Dartmouth Dean of Faculty start-up funds to O.Z.; and Dartmouth James O. Freedman Presidential Scholarship to T.N.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Andersen, M., Dahl, J., Liu, Z., & Vandenberghe, L. (2012). Interior-point methods for large-scale cone programming. In S. Sra, S. Nowozin, & S. J. Wright (Eds.), *Optimization for Machine Learning* (pp. 55–83). MIT Press. https://cvxopt.org/

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res*, *44*(W1), W16-21. https://doi.org/10.1093/nar/gkw387

Bhardwaj, N., Langlois, R. E., Zhao, G., & Lu, H. (2005). Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res*, *33*(20), 6486-6493. https://doi.org/10.1093/nar/gki949

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol*, *65*(6), 997-1008. https://doi.org/10.1093/sysbio/syw037

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, *43*(3), 246-255. https://doi.org/10.1002/prot.1035

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

de Sousa, A. L., Maues, D., Lobato, A., Franco, E. F., Pinheiro, K., Araujo, F., Pantoja, Y., da Costa da Silva, A. L., Morais, J., & Ramos, R. T. J. (2018). PhageWeb - web interface for rapid identification and characterization of prophages in bacterial genomes. *Front Genet*, *9*, 644. https://doi.org/10.3389/fgene.2018.00644

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, *32*(5), 1792-1797. https://doi.org/10.1093/nar/gkh340

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. kdd,

Fu, Y., MacLeod, D. M., Rivkin, R. B., Chen, F., Buchan, A., & Lang, A. S. (2010). High diversity of *Rhodobacterales* in the subarctic North Atlantic Ocean and gene transfer agent protein expression in isolated strains. *Aquatic Microbial Ecology*, *59*(3), 283-293. https://doi.org/10.3354/ame01398

Giessen, T. W., & Silver, P. A. (2017). Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nat Microbiol*, *2*, 17029. https://doi.org/10.1038/nmicrobiol.2017.29

Grull, M. P., Mulligan, M. E., & Lang, A. S. (2018). Small extracellular particles with big potential for horizontal gene transfer: membrane vesicles and gene transfer agents. *FEMS Microbiol Lett*, *365*(19). https://doi.org/10.1093/femsle/fny192

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*, *35*(2), 518-522. https://doi.org/10.1093/molbev/msx281

Hynes, A. P., Mercer, R. G., Watton, D. E., Buckley, C. B., & Lang, A. S. (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*(2), 314-325. https://doi.org/10.1111/j.1365-2958.2012.08113.x

Hynes, A. P., Shakya, M., Mercer, R. G., Grull, M. P., Bown, L., Davidson, F., Steffen, E., Matchem, H., Peach, M. E., Berger, T., Grebe, K., Zhaxybayeva, O., & Lang, A. S. (2016). Functional and evolutionary characterization of a gene transfer agent's

multilocus “genome". *Mol Biol Evol*, *33*(10), 2530-2543. https://doi.org/10.1093/molbev/msw125

Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Karchin, R., Karplus, K., & Haussler, D. (2002). Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, *18*(1), 147-159. https://doi.org/10.1093/bioinformatics/18.1.147

Kaundal, R., Sahu, S. S., Verma, R., & Weirick, T. (2013). Identification and characterization of plastid-type proteins from sequence-attributed features using machine learning. *BMC Bioinformatics*, *14 Suppl 14*(Suppl 14), S7. https://doi.org/10.1186/1471-2105-14-S14-S7

Keen, E. C. (2015). A century of phage research: bacteriophages and the shaping of modern biology. *Bioessays*, *37*(1), 6-9. https://doi.org/10.1002/bies.201400152

Koonin, E. V., & Krupovic, M. (2018). The depths of virus exaptation. *Curr Opin Virol*, *31*, 1-8. https://doi.org/10.1016/j.coviro.2018.07.011

Lang, A. S., & Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*, *15*(2), 54-62. https://doi.org/10.1016/j.tim.2006.12.001

Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol*, *4*(1), 87-104. https://doi.org/10.1146/annurev-virology-101416-041624

Lang, A. S., Zhaxybayeva, O., & Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol*, *10*(7), 472-482. https://doi.org/10.1038/nrmicro2802

Leiman, P. G., Basler, M., Ramagopal, U. A., Bonanno, J. B., Sauder, J. M., Pukatzki, S., Burley, S. K., Almo, S. C., & Mekalanos, J. J. (2009). Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. *Proc Natl Acad Sci U S A*, *106*(11), 4154-4159. https://doi.org/10.1073/pnas.0813360106

Marrs, B. (1974). Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad Sci U S A*, *71*(3), 971-973. https://doi.org/10.1073/pnas.71.3.971

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*, 538. https://doi.org/10.1186/1471-2105-11-538

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, *405*(2), 442-451. https://doi.org/10.1016/0005-2795(75)90109-9

Meher, P. K., Sahu, T. K., Raghunandan, K., Gahoi, S., Choudhury, N. K., & Rao, A. R. (2019). HRGPred: Prediction of herbicide resistant genes with k-mer nucleotide compositional features and support vector machine. *Sci Rep*, *9*(1), 778. https://doi.org/10.1038/s41598-018-37309-9

Minh, B. Q., Nguyen, M. A., & von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*, *30*(5), 1188-1195. https://doi.org/10.1093/molbev/mst024

Munoz-Gomez, S. A., Hess, S., Burger, G., Lang, B. F., Susko, E., Slamovits, C. H., & Roger, A. J. (2019). An updated phylogeny of the *Alphaproteobacteria* reveals that the parasitic *Rickettsiales* and *Holosporales* have independent origins. *eLife*, *8*. https://doi.org/10.7554/eLife.42535

Nagao, N., Yamamoto, J., Komatsu, H., Suzuki, H., Hirose, Y., Umekage, S., Ohyama, T., & Kikuchi, Y. (2015). The gene transfer agent-like particle of the marine

phototrophic bacterium *Rhodovulum sulfidophilum*. *Biochem Biophys Rep*, *4*, 369-374. https://doi.org/10.1016/j.bbrep.2015.11.002

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*, *36*(10), 996-1004. https://doi.org/10.1038/nbt.4229

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, *2*(11), 1533-1542. https://doi.org/10.1038/s41564-017-0012-7

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., . . . Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503-509. https://doi.org/10.1038/s41586-018-0124-0

Shakya, M., Soucy, S. M., & Zhaxybayeva, O. (2017). Insights into origin and evolution of alpha-proteobacterial gene transfer agents. *Virus Evol*, *3*(2), vex036. https://doi.org/10.1093/ve/vex036

Song, W., Sun, H. X., Zhang, C., Cheng, L., Peng, Y., Deng, Z., Wang, D., Wang, Y., Hu, M., Liu, W., Yang, H., Shen, Y., Li, J., You, L., & Xiao, M. (2019). Prophage Hunter: an integrative hunting tool for active prophages. *Nucleic Acids Res*, *47*(W1), W74-W80. https://doi.org/10.1093/nar/gkz380

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Tomasch, J., Wang, H., Hall, A. T. K., Patzelt, D., Preusse, M., Petersen, J., Brinkmann, H., Bunk, B., Bhuju, S., Jarek, M., Geffers, R., Lang, A. S., & Wagner-Dobler, I. (2018). Packaging of *Dinoroseobacter shibae* DNA into gene transfer agent particles is not random. *Genome Biol Evol*, *10*(1), 359-369. https://doi.org/10.1093/gbe/evy005

Touchon, M., Bernheim, A., & Rocha, E. P. (2016). Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J*, *10*(11), 2744-2754. https://doi.org/10.1038/ismej.2016.47

Viswanathan, V., Narjala, A., Ravichandran, A., Jayaprasad, S., & Siddaramappa, S. (2017). Evolutionary genomics of an ancient prophage of the order *Sphingomonadales*. *Genome Biol Evol*, *9*(3), 646-658. https://doi.org/10.1093/gbe/evx024

Westbye, A. B., O'Neill, Z., Schellenberg-Beaver, T., & Beatty, J. T. (2017). The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology (Reading)*, *163*(9), 1355-1363. https://doi.org/10.1099/mic.0.000519

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, *28*(7), 1033-1034. https://doi.org/10.1093/bioinformatics/bts079

Xu, B., Tan, Z., Li, K., Jiang, T., & Peng, Y. (2017). Predicting the host of influenza viruses based on the word vector. *PeerJ*, *5*, e3579. https://doi.org/10.7717/peerj.3579

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*, *39*(3), 306-314. https://doi.org/10.1007/BF00160154

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. https://doi.org/10.1038/srep30372

Chapter 3

Selection for reducing energy cost of protein production drives the GC content and amino acid composition bias in gene transfer agents

Roman Kogay[1], Yuri I. Wolf[2], Eugene V. Koonin[2], and Olga Zhaxybayeva[1,3]

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

[2]National Center of Biotechnology Information, National Institutes of Health, Bethesda, MD, USA

[3]Department of Computer Science, Dartmouth College, Hanover, NH, USA

Published in *mBio* on 14 July 2020
(DOI: 10.1128/mbio.01206-20)

Supplementary Material is available online (DOI: 10.1128/mbio.01206-20)

Author contributions

All authors designed the study. RK collected data and performed analyses. All authors interpreted results. RK and OZ wrote the initial draft. All authors revised and finalized the manuscript.

Abstract

Gene transfer agents (GTAs) are virus-like elements integrated into bacterial genomes, particularly, those of *Alphaproteobacteria*. The GTAs can be induced under nutritional stress, incorporate random fragments of bacterial DNA into mini-phage particles, lyse the host cells and infect neighboring bacteria, thus enhancing horizontal gene transfer. We show that the GTA genes evolve under pronounced positive selection for the reduction of the energy cost of protein production as shown by comparison of the amino acid compositions with both homologous viral genes and host genes. The energy saving in GTA genes is comparable to or even more pronounced than that in the genes encoding the most abundant, essential bacterial proteins. In cases when viruses acquire genes from GTAs, the bias in amino acid composition disappears in the course of evolution, showing that reduction of the energy cost of protein is an important factor of evolution of GTAs but not bacterial viruses. These findings strongly suggest that GTAs are bacterial adaptations rather than selfish, virus-like elements. Because GTA production kills the host cell and does not propagate the GTA genome, it appears likely that the GTAs are retained in the course of evolution via kin or group selection. Therefore, we hypothesize that GTA facilitate the survival of bacterial populations under energy-limiting conditions through the spread of metabolic and transport capabilities via horizontal gene transfer and increase of nutrient availability resulting from the altruistic suicide of GTA-producing cells.

Importance

Kin and group selection remain controversial topics in evolutionary biology. We argue that these types of selection are likely to operate in bacterial populations by showing that bacterial Gene Transfer Agents (GTAs), but not related viruses, evolve under positive selection for the reduction of the energy cost of a GTA particle production. We hypothesize that GTAs are dedicated devices for the survival of bacteria under the conditions of nutrient limitation. The benefits conferred by GTAs under nutritional stress appear to include horizontal dissemination of genes that could provide bacteria with

enhanced capabilities for nutrient utilization and the increase of nutrient availability through the lysis of GTA-producing bacteria.

Introduction

Gene transfer agents (GTAs) are phage-like entities that are known to be produced by several groups of bacteria and archaea (Lang et al., 2017; Lang et al., 2012). Unlike phages, GTAs do not package genes encoding their own structural proteins, and instead package pieces of DNA of the cell that produces them. The biological functions of the GTAs are not well understood, but the leading hypothesis is that GTAs are dedicated vehicles for horizontal gene transfer (HGT) (Brimacombe et al., 2014; Brimacombe et al., 2015). The GTAs can be induced by stress (Westbye, O'Neill, et al., 2017) and, after packaging host DNA and lysing the host cell, can infect neighboring cells (Fogg, 2019; Lang et al., 2012). These cells can integrate the DNA contained within the GTAs, and thus can acquire new alleles, some of which could increase their fitness (McDaniel et al., 2010). GTAs are thought to have evolved from different viral ancestors on at least five independent occasions (Lang et al., 2017), and in *Alphaproteobacteria*, GTAs appear to have been maintained for many millions of years (Shakya et al., 2017). Such convergent acquisition, long-term persistence and sequence conservation of these elements suggests that GTAs provide a selective advantage for their host populations (Lang et al., 2017).

The best-studied GTA (RcGTA) comes from the alphaproteobacterium *Rhodobacter capsulatus* (Marrs, 1974). Its production is directed by at least five loci that are scattered across the *R. capsulatus* genome, with 17 genes that encode most of the proteins necessary for the production of the RcGTA particles located in one locus (**Table S1**) (Hynes et al., 2016). This locus, also known as the 'head-tail' cluster (Lang et al., 2017), is detectable in many alphaproteobacterial genomes (Kogay et al., 2019; Shakya et al., 2017). Across *Alphaproteobacteria*, the RcGTA-like 'head-tail' clusters appear to evolve relatively slowly (Lang et al., 2012), have an elevated GC-content relative to the host genome (Shakya et al., 2017), and have skewed amino acid composition when compared to their viral homologs (Kogay et al., 2019).

Because bacteria and archaea occupy diverse ecological niches, they face different levels and directions of selective pressures and have different mutation rates, skewed GC-content and amino acid composition that emerged from multiple, intertwined processes. As a result, the genomic GC-content of bacterial and archaeal species varies in the wide range from less than 20% to more than 75% (Hildebrand et al., 2010) and cannot be explained solely by the universal mutational AT-bias (Hershberg & Petrov, 2010). Several studies have shown that the availability of different nutrients in the environment can act as a selective force and is involved in shaping the GC content of genomes and amino acid content of the encoded proteins. For example, inhabitants of nitrogen-poor environments tend to have a low content of G and C nucleotides and of amino acids containing nitrogen in their side chains (Grzymski & Dussaq, 2012; Luo et al., 2015). Because A and T each contain one nitrogen atom less than G and C, respectively, the reduced usage of the G and C allows an organism to minimize the demand for the limiting nitrogen during replication and transcription. By contrast, carbon limitation could drive long-term elevation of the genomic GC-content (Hellweger et al., 2018; Mende et al., 2017), likely, because small (carbon-poor) amino acids are preferentially encoded by GC-rich codons (Bragg & Hyder, 2004).

In addition to the GC-content fluctuation between species, there is also a considerable GC-content heterogeneity within single bacterial and archaeal genomes. For example, bacterial genomes can be subject to GC-biased gene conversion and thus recombination hotspots within a genome can have elevated GC-content compared to the rest of the genome (Lassalle et al., 2015). (19). Also, highly expressed genes tend to have an elevated GC-content and, accordingly, their highly abundant protein products have a skewed amino acid composition (Chen et al., 2016). Because highly abundant proteins appear to be optimized for low cost of production (Raiford et al., 2012; Swire, 2007), the elevated GC-content of highly expressed genes can be explained by selection for GC-rich codons that tend to encode small, energetically cheap amino acids. Generally, molecular composition of genes and proteins appears to reflect various selection pressures, among which those associated with energy savings are prominent.

Thus, there are two possible explanations for the observed skew in both the GC-content and amino acid composition of the RcGTA-like genes and proteins. Under one

scenario, selection and mutational biases act on the base composition, so that the amino acid bias is a byproduct of the skewed GC-content. Under the second scenario, selection could favor the skewed amino acid composition, resulting in a biased GC-content due to the structure of the genetic code. Here, we present evidence for the second scenario and show that the observed amino acid bias is driven by selection to reduce carbon utilization and biosynthetic cost of production of the RcGTA-like proteins. We show that the energy expense of the production of RcGTA-like proteins is comparable to that of the highly expressed housekeeping genes. For some of the amino acid changes, we identify clear signatures of positive selection towards amino acids with a smaller number of carbons in their side chains. We hypothesize that evolution of RcGTA-like elements was affected by selection to minimize cellular energy investment into their production under nutrient-poor conditions.

Results

Elevated GC-content in RcGTA-like regions is due to the higher GC-content in the first and second codon positions of the coding genes.

Because of the degeneracy of the genetic code, GC3-content is known to track the overall GC-content of genomic regions (Palidwor et al., 2010). Hence, if the GC-content of RcGTA-like 'head-tail' clusters is elevated because they reside in GC-rich genomic regions, the GC-content in the third, primarily synonymous codon positions (GC3-content) of the RcGTA-like genes is expected to be higher compared to the genomic average of the GC3-content. Moreover, the elevated GC3-content would not be limited to the genes in the RcGTA-like region but would be apparent in the adjacent genes as well. To test this hypothesis, we examined homologs of one RcGTA locus ('head-tail' cluster) in 212 alphaproteobacterial genomes (see **Materials and Methods**) (Kogay et al., 2019; Shakya et al., 2017). Although we analyzed homologs of only one locus from one GTA only, for brevity, we hereafter refer to these regions simply as "GTA regions", and to genes and encoded proteins in these regions as "GTA genes" and "GTA proteins". Contrary to the aforementioned expectation, we found no significant difference between the GC3-content of GTA genes of the 212 alphaproteobacterial genomes, their

neighboring genes and all genes in the genome (Kruskal-Wallis H test, p-value = 0.62; **Figure 1**). By contrast, the GC1- and GC2-content of GTA genes are significantly higher than the corresponding values for both the neighboring genes (Dunn's test, p-value < 0.0001) and the genes across the entire genome (Dunn's test, p-value < 0.0001) (**Figure 1**). Furthermore, the genes adjacent to the GTA regions do not have elevated GC1- and GC2-content when compared to the genes in the entire genome (Dunn's test, p-value = 1), indicating that the elevated GC1- and GC2-content is limited to the GTA genes. Due to the relationship between codons and amino acids in the genetic code, the elevated GC1- and GC2- content of an open reading frame (ORF) translates into a biased amino acid composition of the encoded protein. Indeed, a significant amino acid composition bias in the GTA proteins has been demonstrated previously (Kogay et al., 2019). Specifically, the

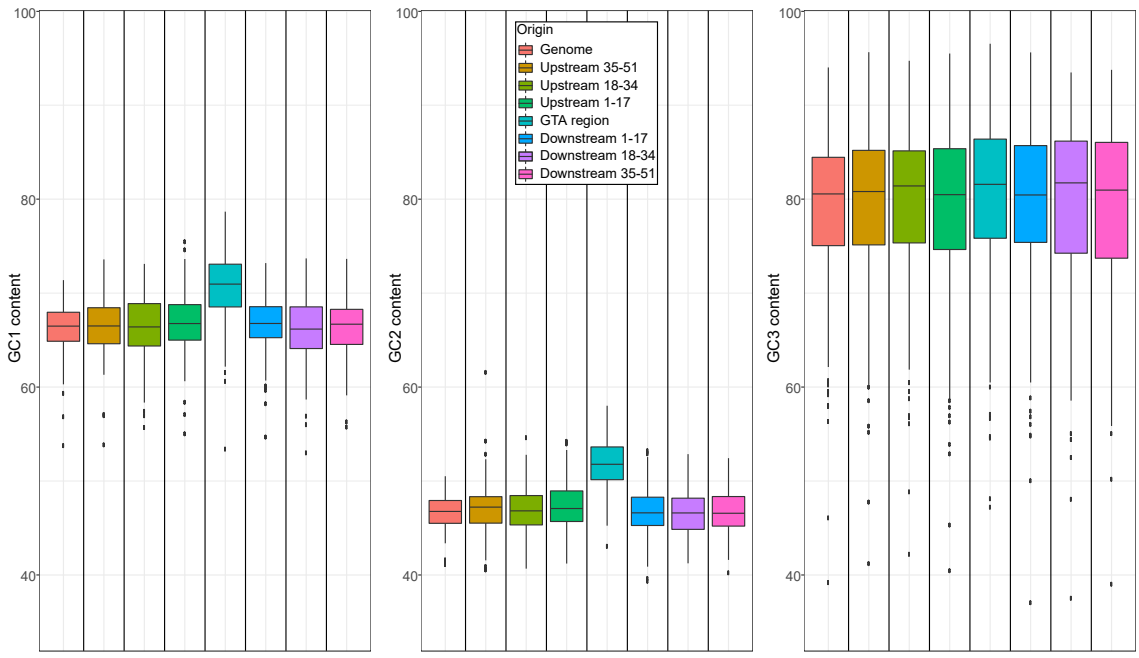

Figure 1. The GC1-, GC2- and GC3-content of GTA regions, their immediate neighborhoods and all protein-coding genes in 212 alphaproteobacterial genomes. The neighborhoods immediately upstream and downstream of a GTA region consists of 17 genes each. Boxplots represent median values bounded by the first and third quartiles. Whiskers show the values that lie in the range of 1.5*interquartile rule. Dots outside of the whiskers are the outliers.

relative abundance of amino acids encoded by GC-rich codons is significantly higher in the GTA genes than the genomic average (**Figure S1;** Student's t-test, p-value < 0.0001; see **Materials and Methods** for definition of GC-rich codons). Taken together, these findings suggest that the GC-content of GTA regions in *Alphaproteobacteria* is driven by selection for a specific amino acid composition of the encoded proteins.

Proteins encoded in GTA regions contain smaller number of carbons and are energetically cheaper than their viral homologs.

The RcGTA production has been experimentally demonstrated to be stimulated by carbon depletion (Westbye, O'Neill, et al., 2017). Furthermore, knockout of the RcGTA-like genes in three alphaproteobacterial strains (Price et al., 2018) resulted in a significant decrease in fitness of the mutants under growth conditions with alternative carbon sources that might not be utilized by these strains (Kogay et al., 2019). If GTAs are indeed produced under conditions of limited carbon availability, the observed amino acid bias in the GTA genes might represent an adaptation in the GTA-containing lineages to utilize energetically cheaper amino acids for GTA particle production. To test this hypothesis, we compared the number of carbons in amino acid side chains and costs of amino acid biosynthesis (measured as the number of high-energy phosphate bonds) in GTA proteins and by their viral homologs. We assumed that (a) all amino acids are produced by bacteria de novo, as at least 174 of the analyzed genomes can produce 19 or all 20 amino acids (**Figure S2**), and (b) viral infections are not specifically associated with the carbon-limited conditions, and therefore, viral homologs of RcGTA genes should not be subject to selection for energy saving. Consistent with the proposed hypothesis, for all of the 12 genes with sufficient number of viral homologs to estimate statistical significance (**Table S1**), GTA proteins have both a significantly smaller number of carbons (Mann-Whitney U test, all 12 Bonferroni-corrected p-values < 0.01; **Figure 2A**) and a significantly reduced cost of amino acid biosynthesis than their viral homologs (Mann-Whitney U test, all 12 Bonferroni-corrected p-values < 0.01; **Figure 2B**).

To demonstrate that the observed differences in the carbon content of the GTA and viral proteins are not simply due to the compositional bias present in the ancestor of the alphaproteobacterial GTA elements (Shakya et al., 2017), we sought to examine only

a subset of viral homologs that are presumed to be horizontally acquired from the GTA regions. Genes with significant sequence similarity to GTA genes have been previously found in viruses and inferred to be horizontally acquired from GTAs on the basis of phylogenetic reconstruction (Hynes et al., 2016; Zhan et al., 2016). In our phylogenetic analyses, we examined several viral genes of this apparent origin (**Table 1, Figure S3**; also see **Materials and Methods** for details). Under the assumption of no selection for energy saving in viruses, we expect the carbon content of the GTA genes acquired by viruses to increase after their relocation to the virus genomes. Indeed, in all cases, the carbon content of the now-viral homologs consistently (and, overall, significantly) increased compared to the inferred ancestral state at the time of acquisition (**Table 1, SI Figure S3**).

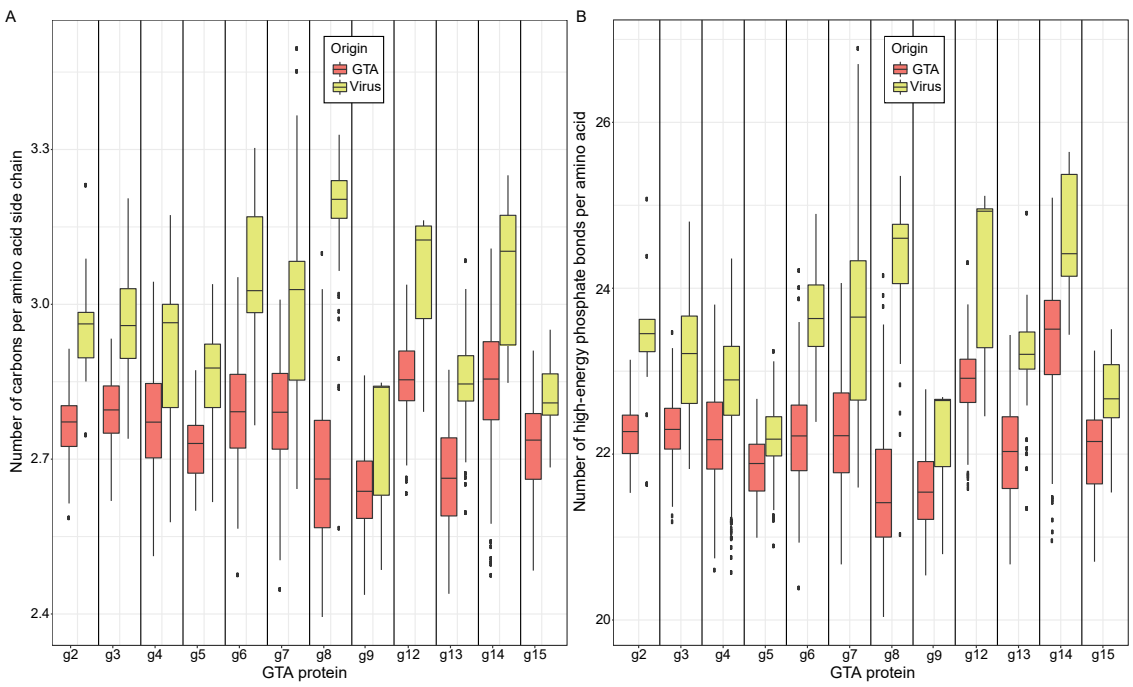

Figure 2. Carbon content (A) and biosynthetic cost (B) of proteins encoded by GTA genes in 212 alphaproteobacterial genomes and their viral homologs. Boxplots represent median values that are bounded by the first and third quartiles. Whiskers show the values that lie in the range of 1.5*interquartile rule. Dots outside of the whiskers are the outliers. The number of data points in each boxplot is listed in **Table S1**.

Table 1. Change in the carbon content between viral homologs of the GTA proteins and their closest GTA ancestral node.

GTA gene	Virus name	Change in the number of carbons per side chain of an amino acid	p-value	Alignment length
g6	Cellulophaga phage phi10 1	+0.605	<0.001	193
g7	Cellulophaga phage phi18 1	+0.394	0.001	147
g7	Streptomyces phage phiSASD1	+0.167	0.179	147
g7	Salmonella phage ST64B	+0.222	0.048	147
g7	Salmonella phage 118970 sal3	+0.229	0.042	147
g7	Shigella phage SfIV	+0.184	0.115	147
g7	Enterobacteria phage SfV	+0.244	0.083	147
g7	Shigella phage SfII	+0.191	0.107	147
g10	Rhizobium phage 16-3	+0.105	0.271	123
g12	Rhodobacter phage RcCronus	+0.123	0.081	228
g13	Paracoccus phage vB PmaS R3	+0.048	0.226	304
g13	Dinoroseobacter phage vB DshS R5C	+0.027	0.383	304
g13	Roseobacter phage RDJL Phi 1	+0.005	0.447	304
g13	Roseobacter phage RDJL Phi 2	+0.019	0.388	304
g14	Rhodobacter phage RcRhea	+0.191	0.108	166
g15	Rhodobacter phage RcRhea	+0.147	<0.001	1369
g15	Rhodobacter phage RcCronus	+0.143	<0.001	1369
Cumulative across 7 genes		+0.163	<0.001	2530

Energetic cost of the GTA proteins is as low as that of essential bacterial proteins.

Highly expressed genes have been demonstrated to evolve under selection to decrease the energetic cost of the encoded protein production (Chen et al., 2016). Indeed, 20 single-copy housekeeping genes involved in translation ([J] COG category; (Galperin et al., 2019)) (**Table S2**), and therefore presumed to be expressed at relatively high levels under any conditions, collectively, have a significantly lower energetic cost than the average of all proteins encoded in a genome, as measured by both side chain carbon utilization and biosynthetic cost of production per amino acid (**Figure 3;** Mann-Whitney

U test, p-values < 0.0001). The biosynthetic cost per amino acid of the GTA proteins was found to be statistically indistinguishable from that of the products of the 20 highly expressed genes (Mann-Whitney U test, p-value = 0.3372), and remarkably, utilize even less carbon (Mann-Whitney U test , p-value < 0.0001) (**Figure 3**).

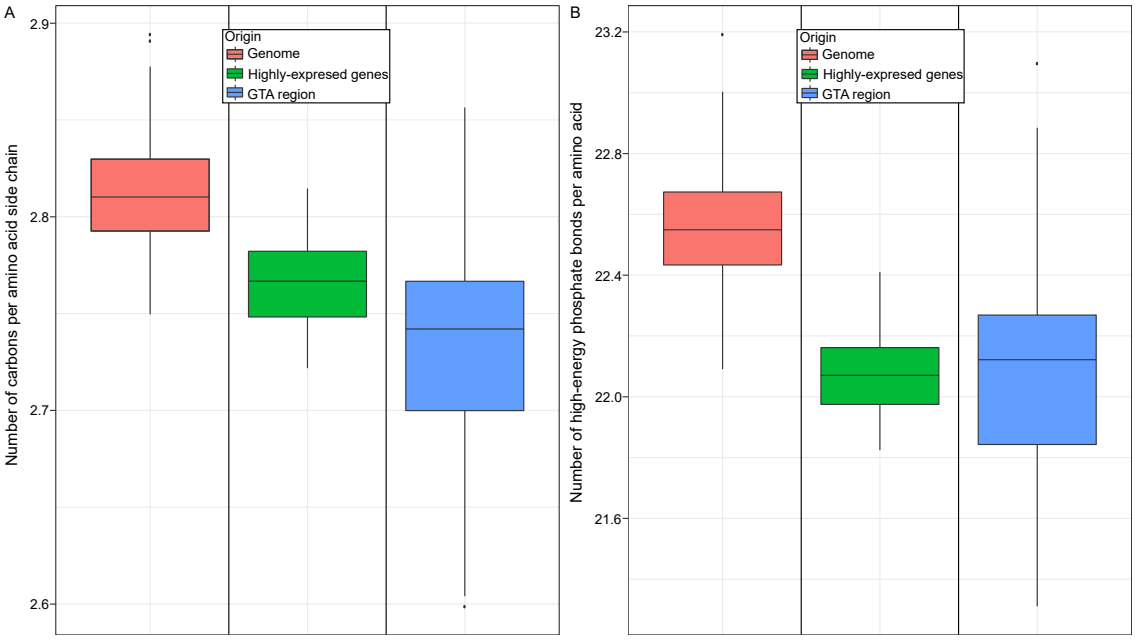

Figure 3. The number of carbons (A) and number of high-energy phosphates (B) in proteins encoded by all protein-coding genes in 212 genomes, highly expressed genes, and GTA genes. Boxplots represent median values that are bounded by the first and third quartiles. Whiskers show the values that lie in the range of 1.5*interquartile rule. Dots outside of the whiskers are outliers.

Reduction in carbon utilization varies among GTA genes and across bacterial taxa.

To investigate how reduction of carbon content evolved from the common ancestor of the examined GTA genes to the extant forms, we reconstructed the number of carbons per amino acid at the ancestral nodes of individual evolutionary trees of 14 GTA genes (those with at least one detectable viral homolog; **Table S1**). To correct for differences in the GC-content across taxa (which affects the carbon content of the encoded proteins), for each taxon we normalized the number of carbons per amino acid of GTA proteins by that of 26 housekeeping proteins (**Table S2**). No unifying pattern of directional selection

towards the lower carbon content was detected across all genes and all taxa (**Figure S4**). This lack of an overall signal was not surprising because GTA genes can be horizontally transferred across taxa (Shakya et al., 2017), have different evolutionary rates among and within taxa (Shakya et al., 2017), and are likely to reach unequal translation levels during GTA production (Chen et al., 2009). These differences would make the carbon content optimization gene- and taxon- specific, blurring the net effect. However, members of the order *Sphingomonadales* show the most pronounced reduction in carbon utilization for the GTA regions overall, as well as for the majority of individual genes (**Figure 4**). Notably, many *Sphingomonadales* species can live under nutrient-depleted conditions (Balkwill et al., 2006).

In *Sphingomonadales*, the decrease in carbon content of GTA proteins is driven by positive selection.

To evaluate whether diversifying (positive) selection plays a role in the observed reduction of carbon utilization in the GTA genes in *Sphingomonadales*, we tested for evidence of positive selection in individual sites on the branch leading to this clade. For 9 of the 14 evaluated genes, the model of positive selection on the branch was a significantly better fit than the neutral null model (**Table S3**). For 8 of these 9 genes, members of the *Spingomonadales* clade showed significant decrease in the carbon utilization relative to three other orders (Mann-Whitney U Test; α of 0.01, p -values < 0.01; **Table S4**; **Figure 4**). Conversely, for 4 of the 5 genes that did not show evidence of positive selection, there was no significant decrease in the carbon content of proteins in the *Sphingomonadales* genomes (**Figure 4**).

To assess how the specific sites that are inferred to be subject to positive selection contribute to the carbon content of the *Sphingomonadales'* GTA genes, we examined carbon content of amino acids in the sites with >0.95 posterior probability of being subject to positive selection. For 8 of the 9 positively selected genes, these sites substantially contributed to the decrease in carbon utilization in *Sphingomonadales* (**Table 2, Table S5**). This trend is manifested, in particular, by the observed replacements of aromatic amino acids, which contain relatively high numbers of carbons and have excessive biosynthetic costs, with non-aromatic amino acids (**Figure S5**). The observed

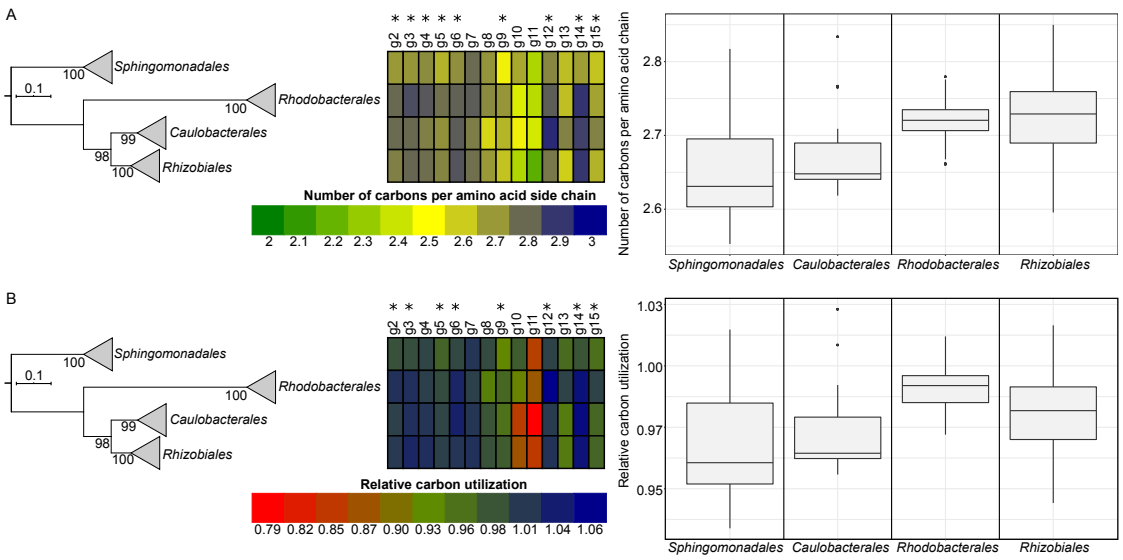

Figure 4. Carbon content of GTA proteins for four orders of the class *Alphaproteobacteria*. For each GTA protein, the heatmap visualizes the number of carbons per the side chain in amino acid averaged across taxonomic order. The numbers are shown either as raw values (**panel A**), or as values normalized by the carbon content of proteins encoded by 26 single-copy genes (**panel B**). The asterisks mark GTA proteins with significantly lower numbers of carbons per amino acid in the *Sphingomonadales* order than in the other three orders combined (α of 0.01; Mann-Whitney U test, all p-values < 0.01). Boxplots summarize the distribution of carbon content within each alphaproteobacterial order averaged across the examined GTA genes. Median values are bounded by the first and third quartiles. Whiskers show the values that lie in the range of 1.5*interquartile rule and dots outside of the whiskers are the outliers. The phylogenetic tree is the reference alphaproteobacterial phylogeny (see **Materials and Methods** for details), in which branches are collapsed at the taxonomic rank of order. Numbers at the tree nodes represent bootstrap support values. Scale bar, number of substitutions per site.

replacements of tryptophan with phenylalanine indicate that, under a constraint of maintaining an amino acid with similar physicochemical properties, there is selection for utilization of a cheaper amino acid (**Figure S5**). Mapping of the positively selected sites in the *Sphingomonadales*' g5 homolog onto a structural model of the T5 bacteriophage

major capsid protein shows that these sites tend to be located on the surface of the protein (**Supplementary Movie** in **FigShare Repository**). This example suggests that carbon-saving replacements preferentially occur in sites that are not involved in the folding of GTA proteins, allowing the GTAs to preserve the functionality of their proteins at reduced production costs.

Table 2. Contribution of positively selected sites to the reduction of carbon utilization in GTA proteins of *Sphingomonadales*.

GTA protein	Number of sites under the positive selection	Average change in number of carbons by the contribution of all sites under the positive selection	Number of sites that contribute to the decrease in number of carbons
g2	13	-0.22	6
g3	33	-0.72	22
g4	29	-0.42	13
g5	12	-0.39	8
g6	11	+0.16	5
g9	29	-0.68	16
g12	23	-0.52	13
g13	31	-0.44	16
g15	27	-0.55	15

Discussion

We show here that the elevated GC-content of GTA regions is driven by selection towards encoding proteins with energetically cheaper amino acids. Although GC-rich genes have an increased cost of mRNA expression, cells spend much more energy on the synthesis of amino acids than on the synthesis of ribonucleotides (Chen et al., 2016; Lynch & Marinov, 2015). Hence, the elevation of GC-content in non-synonymous codon positions (GC1 and GC2) reduces the energetic expenses on the production of the respective proteins. Consistent with this notion, energy savings for GTA proteins are as pronounced or even greater than those for highly expressed housekeeping genes that are known to utilize cheaper and smaller amino acids (Chen et al., 2016). Given that production of RcGTA-like particles in *Alphaproteobacteria* occurs in the stationary phase (Lang et al., 2017; Solioz et al., 1975) and is associated with carbon depletion (Kogay et

al., 2019; Westbye, O'Neill, et al., 2017), the shift in GC-content of GTA genes and amino acid composition of their products likely reflects the adaptation for their efficient expression under such conditions.

The change in the amino acid composition of GTA proteins was not uniform across the examined alphaproteobacterial lineages. These differences are not unexpected because GTA-carrying bacteria live in different environments and under different selection pressures. We demonstrated that, on the branch leading to *Sphingomonadales*, the decrease in carbon content of the GTA proteins is driven by positive selection for the use of cheaper amino acids. We hypothesize that the last common ancestor of *Sphingomonadales* evolved in a nutrient-depleted environment that selected for the reduction in the use of energetically expensive amino acids in the GTA proteins.

Although bacterial viruses also spend disproportionate amounts of energy on translation (Mahmoudabadi et al., 2017), our analysis of viral genes that apparently were acquired by viruses from bacterial GTAs shows a decrease in GC1 and GC2 content, with the concomitant increase in protein production energy cost. Thus, positive selection for cost saving, probably ceases to substantially affect the evolution of these genes once they are transferred to virus genomes. Lytic bacteriophages reproduce rapidly, with a typical burst size of about 200 virions that hijacks about 30% of the host energy budget (Mahmoudabadi et al., 2017). Under the conditions of such brief, explosive growth, energy saving might not be an important selective factor. Differences in the viral burst sizes imply that selection for energy saving could play some role. However, such selection is expected to be weak due to other constraints affecting the lytic viruses, such as fluctuations in the host energy budget, often error-prone viral replication machinery, and the main evolutionary pressure being evasion of host defense systems (Paez-Espino et al., 2015; Paterson et al., 2010). Thus, our observations provide additional evidence that GTAs are not selfish, virus-like agents but rather microbial adaptations.

Taken together, our findings, and in particular, the evidence of positive selection for energy saving in *Sphingomonadales*, are in line with the previous suggestions that maintenance of GTAs and production of GTA particles confers some advantage to the bacterial hosts (Lang et al., 2017; McDaniel et al., 2010). Because GTA-producing cell

lyses and GTA genes are not transferred to the recipient cell, the reduction of energy utilization for the production of GTA particles has to be beneficial at the population or community level, that is, it needs to involve some form of kin or group selection (Smith, 1964; West et al., 2006). The nature of such benefit(s) is not entirely clear, but it appears likely that the GTAs, effectively, are devices for survival under energy- or nutrient-limited conditions that are common in bacterial ecology. More specifically, GTAs could provide two types of adaptations. Previous studies suggest that oligotrophic conditions do not interfere with the capacity of bacteria to engage in genetic exchange (Goodman et al., 1994). Moreover, the nutrient limitation can upregulate horizontal gene transfer via transformation (Meibom et al., 2005), suggesting potential benefits of gene exchange under adverse conditions of energy or nutrient limitations. Conceivably, HGT mediated by the GTAs can confer additional metabolic or transport capacities to the recipient bacteria. Additionally, GTAs could be perceived as a mechanism of bacterial programmed cell death (Engelberg-Kulka et al., 2006; Peeters & de Jonge, 2018). Under this type of adaptation, the GTA-mediated lysis of a fraction of the bacterial community would decrease the population density and increase the nutrient availability per cell, by supplying additional nutrients released from the lysed cells.

Materials and Methods

Reduction in carbon utilization varies among GTA genes and across bacterial taxa.

The initial dataset of 422 GTA regions in 419 alphaproteobacterial genomes consisted of 88 regions identified by Shakya et al. (Shakya et al., 2017) and 334 regions in complete alphaproteobacterial genomes predicted by Kogay et al. (Kogay et al., 2019). Four GTA regions from the Methylobacterium nodulans ORS2060 genome were removed due to their questionable assignment as GTAs (Shakya et al., 2017). Because our previous GTA prediction procedure (Kogay et al., 2019) screened for the presence of only 11 of the 17 homologs of the RcGTA head-tail cluster (Lang et al., 2012), the remaining 6 homologs were identified using BLASTP (Altschul et al., 1997) (version 2.6.0, e-value = 0.1, manually curated homologs from Kogay et al. (Kogay et al., 2019) as queries), with subsequent restriction of the hits to the regions with previously identified GTA genes. To

reduce the computational cost of the downstream analyses, highly similar GTA regions were excluded. To this end, genomes that contained the 418 GTA regions were clustered into OTUs using furthest neighbor clustering and Average Nucleotide Identity (ANI) cutoff of 95%. The ANI values were calculated using fastANI v.1.1 (Jain et al., 2018). From each of the identified 215 OTUs, only the GTA region with the largest number of the relevant genes was retained. Further removal of the regions that contained less than 9 genes resulted in the final dataset of 212 GTA regions.

To obtain viral homologs of the GTA genes, genes from the 212 GTA regions were used as queries in BLASTP searches (Altschul et al., 1997) (version 2.6.0, e-value = 0.001, query and subject coverage of at least 60%) against the viral RefSeq database (release 96, accessed on October 2019) (Brister et al., 2015).

The numbers of identified alphaproteobacterial and viral homologs for the 17 RcGTA genes are shown in **Table S1.**

Calculation of GC-content for the 212 alphaproteobacterial genomes.

The GTA region's neighborhood was defined as 51 genes upstream and 51 genes downstream of the region. Each neighborhood was divided into 6 non-overlapping regions with 17 genes each. For each neighborhood region, the GTA region, and all annotated genes in the genome, GC1-, GC2-, and GC3-content values were calculated using an in-house script. The significance of the GC-content differences among the obtained 8 groups was assessed using the Kruskal-Wallis H test followed by the Dunn's test (Dunn, 1964). The p-values were adjusted for multiple testing using the Bonferroni correction.

Calculation of the relative abundance of amino acids encoded by GC-rich codons for 212 alphaproteobacterial genomes.

The amino acids that are encoded by GC-rich codons were defined as those that have G or C in the first and second codon positions (alanine, arginine, glycine and proline). For each genome, the amino acid frequencies were calculated for the pooled set of proteins encoded by genes in the GTA region, as well as for the pooled set of proteins encoded by all genes in a genome. The significance of the difference in relative

abundances of the 4 amino acids encoded by GC-rich codons in the two sets was assessed using the Student's t-test.

Calculation of carbon content and biosynthetic cost of amino acids in the encoded proteins.

Because differences in the carbon-content of amino acids are determined solely by the composition of their side chains, for each amino acid sequence encoded by a GTA gene (or its viral homolog), the number of carbons in the side chains of the amino acids was counted and normalized by the length of the encoded polypeptide. Additionally, for each amino acid sequence encoded by a GTA gene (or its viral homolog), the average biosynthetic cost of protein production per amino acid, defined as the number of high-energy phosphate bonds needed to produce a particular amino acid, was calculated. Because almost all of the 212 alphaproteobacteria containing the GTA regions are either obligate or facultative aerobes, the individual costs of amino acid production already computed for Escherichia coli by Akashi and Gojobori (Akashi & Gojobori, 2002) were used. The significance of the differences in the carbon utilization and biosynthetic cost between GTA proteins and viral homologs was assessed using the Mann-Whitney U test, followed by the Bonferroni correction of p-values to account for multiple testing.

Verification of amino acid biosynthesis pathways in the alphaproteobacterial genomes.

Presence of the amino acid biosynthesis pathways in the genomes was evaluated using the KEGG database release 92 (Kanehisa & Goto, 2000). For 189 of the 212 alphaproteobacteria, either its own genome (186 genomes) or the genome of a close relative (ANI > 95%; 3 genomes) were examined. For the remaining 23 genomes, no information from the closely related genomes was available in KEGG. For each of the 189 genomes, the map of amino acid biosynthesis (map number = 01230) was examined for completeness. If key enzymes were missing, additional maps (map number = 00250 – 00400) were evaluated to identify alternative enzymes that could catalyze the same reactions. If alternative enzymes were not found, Escherichia coli homologs that catalyze the missing steps were used as queries for a BLASTP search of the genome (version

2.6.0, e-value 0.001, query coverage of at least 50%) and the RefSeq annotations of the obtained matches were examined. If a complete biosynthetic pathway of an amino acid could not be reconstructed, the genome was designated as “auxotrophic” for the biosynthesis of the given amino acid.

Exclusion of divergent viral homologs.

To minimize possible misplacement of viral homologs due to long branch attraction, we have identified and excluded divergent viral homologs using the following procedure. Amino acid sequences of GTA genes and their viral homologs were aligned using MAFFT v 7.305 with the ‘auto’ setting (Katoh & Standley, 2013). Phylogenetic trees from individual gene alignments were reconstructed in the IQ-TREE v 1.6.7 (Nguyen et al., 2015) using the best substitution model detected by ModelFinder (Kalyaanamoorthy et al., 2017). The obtained trees were used as guides for the reconstruction of more accurate trees, using the profile mixture model “LG+C60+F+G” and the site-specific frequency models that were approximated by the posterior mean site frequency (Wang et al., 2018) as implemented in IQ-TREE.

To exclude viral homologs that are not closely related to GTA genes, only viral homologs nested within the taxonomic rank of alphaproteobacterial order with ultrafast bootstrap support >= 60% (1,000 pseudoreplicates; (Hoang et al., 2018)) were retained. Because, for genes g3, g4 and g8, large numbers of viral homologs were retained, only top 5 non-identical viral proteins most closely related to the alphaproteobacterial homologs were kept. The retained viral homologs were realigned with the GTA genes, and the phylogenetic trees were reconstructed and examined as described above. The process was repeated until all retained viral homologs grouped within alphaproteobacterial orders.

Reconstruction of ancestral amino acid sequences.

Amino acid sequences of the ancestral nodes of the reconstructed phylogenetic trees were reconstructed using FastML v 3.11 (Ashkenazy et al., 2012). Indels in the ancestral sequences were inferred using the maximum likelihood and probability cutoff of 0.5. Ancestral amino acid states of non-gapped states were determined using marginal

reconstruction under LG substitution matrix (Le & Gascuel, 2008), with heterogeneity in substitution rates among sites modeled using Gamma distribution (Yang, 1994).

Reconstruction of the alphaproteobacterial reference phylogeny.

In each of the 212 genomes containing GTA regions, 31 phylogenetic markers were detected and retrieved using AMPHORA2 (Wu & Scott, 2012). Amino acid sequences of these markers were aligned using MAFFT v 7.305 with the 'auto' setting (Katoh & Standley, 2013). The best substitution matrix for each gene was determined using the ProteinModelSelection.pl script obtained from https://github.com/stamatak/standard-RAxML/tree/master/usefulScripts (last accessed November 2019). The individual gene alignments were concatenated, and each gene was treated as a separate partition (Chernomor et al., 2016) in the subsequent phylogenetic reconstruction. The maximum likelihood tree was reconstructed by the IQ-TREE v 1.6.7 (Nguyen et al., 2015), and Gamma distribution with four categories was used to account for heterogeneity in substitution rates among sites (Yang, 1994). Although no outgroup sequences were included into the alignment, for presentation purposes, the tree was rooted to reflect the branching of *Alphaproteobacteria* as previously observed (Kogay et al., 2019). Phylogenetic tree was visualized using iTOL (Letunic & Bork, 2019).

Retrieval of selected single-copy and highly-expressed genes.

Twenty-six of the 120 phylogenetically informative genes (Parks et al., 2017) were found to be present in a single copy in all 212 genomes (**SI Table S2**). The 26 genes were extracted from each genome using hmmersearch v 3.1b2 via modified scripts from AMPHORA2 (Wu & Scott, 2012). The functional annotations of the 26 genes were examined using the eggNOG-mapper (Huerta-Cepas et al., 2017) based on the eggNOG orthology database v. 4.5 (Huerta-Cepas, Szklarczyk, et al., 2016). Twenty of the 26 genes belong to the [J] COG category ("Translation, ribosomal structure and biogenesis") and therefore were designated "highly-expressed" genes.

Calculation of carbon utilization in extant and ancestral GTA genes.

The relative carbon utilization of each extant protein encoded by a GTA gene was defined as the ratio of the average number of carbon atoms per site to that of the 26

single-copy genes. To calculate carbon utilization for the ancestral states, amino acid sequences of 14 GTA proteins with at least one viral homolog were aligned by MAFFT v 7.305 with the "auto" setting (Katoh & Standley, 2013), and phylogenetic trees were reconstructed using IQ-TREE v 1.6.7 (Nguyen et al., 2015) using the best substitution model detected with ModelFinder (Kalyaanamoorthy et al., 2017). Using reconstructed phylogenies and carbon utilization data for extant proteins, carbon utilization at the internal nodes was inferred using the marginal maximum likelihood reconstruction, as implemented in the phytools package (Revell, 2012). The change of carbon utilization along the tree branches was deduced via equation 2 of Felsenstein (Felsenstein, 1985), also as implemented in the phytools package (Revell, 2012).

To assess the significance in the increase of carbon content of the selected viral proteins in comparison to their inferred ancestral protein, for each of the seven GTA genes with such viral homologs, amino acid sequences of these extant viruses and their closest inferred ancestral sequence were retrieved and aligned via MAFFT using "linsi" settings (Katoh & Standley, 2013). For each gene alignment, 1000 bootstrap replicates were generated in RAxML v 8.2.11 (Stamatakis, 2014). For each bootstrap replicate, the net change in the number of carbons per amino acid between the viral protein and the ancestral protein was calculated. The p-value was defined as the proportion of bootstrap replicates with a zero or negative net change in the number of carbons per amino acid. Additionally, the cumulative net change in the number of carbons per amino acid across all 7 GTA proteins (**Table 1**) was calculated by adding up the net changes across individual genes. For genes with more than one viral homolog, the viral homolog with the smallest difference in the number of carbons per amino acid was selected to obtain a conservative estimate. The p-values were calculated as they were for individual comparisons.

Detection of positive selection on the branch leading to *Sphingomonadales*.

Using the phylogenetic trees and amino acid sequence alignments of the GTA proteins (see "**Calculation of carbon utilization states in contemporary and ancestral GTA genes"** section), evidence of episodic events of positive selection in the *Sphingomonadales* clade was inferred under the branch site A model, as implemented in

the codeml package of PAML version 4 (Yang, 2007). Codon alignments of nucleotide sequences were obtained using pal2nal (Suyama et al., 2006). The branch lengths in the corresponding phylogenetic trees were re-estimated in PAML. Because g12 and g15 genes vary in length between *Sphingomonadales* and other alphaproteobacterial orders, codons that were present in less than 50% and 80% of sequences in g12 and g15 datasets, respectively, were removed. For the null model (no positive selection), ω2a and ω2b were fixed to 1, and the significance for the alternative model (positive selection) was tested using the likelihood ratio test with one degree of freedom and α of 0.01. P-values were adjusted for multiple testing using the Bonferroni correction. A site was classified as being "under positive selection" if it had the probability of at least 0.95 in the Bayes Empirical Bayes estimation (Yang et al., 2005), and was present in at least of 50% of the *Sphingomonadales* branches and 50% of the remaining branches.

Visualization of positively selected sites on the 3D model of capsomer.

The amino acid sequences of the RcGTA genes were used in a BLASTP search (e-value < 0.01, low-complexity masking, and query coverage of at least 50%) against the PDB database (Berman et al., 2000) (last accessed November 2019). Only the g5 gene query returned significant matches to the PDB database. The amino acid sequence of the top-scoring match (PDB ID – 5TJT) was retrieved and aligned with the representative g5 homolog from *Sphingomonadales* (*Sphingobium amiense* DSM 16289) using the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970). Of the 12 sites classified as being under positive selection in the *Sphingobium amiense* DSM 16289 homolog, 2 sites did not have homologous positions in the 5TJT sequence. The remaining 10 sites were mapped onto the 5TJT PDB structure using PyMol version 2.3 (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.)

Data availability

List of accession numbers of 212 alphaproteobacterial genomes with GTA regions, amino acid sequences of identified GTA proteins in alphaproteobacteria and viruses, as well as sequence alignments and phylogenetic trees used in the described analyses have been deposited to FigShare under the doi: 10.6084/m9.figshare.12071223.

Acknowledgements

This work was supported in part by the National Science Foundation [NSF-DEB 1551674 to O.Z.]; by the Simons Foundation Investigator in Mathematical Modeling of Living Systems [327936 to O.Z.]; and by Intramural Research Program of the National Institutes of Health of the USA (National Library of Medicine), to YIW and EVK.

References

Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A*, *99*(6), 3695-3700. https://doi.org/10.1073/pnas.062526999

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., & Pupko, T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res*, *40*(Web Server issue), W580-584. https://doi.org/10.1093/nar/gks498

Balkwill, D. L., Fredrickson, J. K., & Romine, M. F. (2006). *Sphingomonas* and related genera. In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, & E. Stackebrandt (Eds.), *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass* (pp. 605-629). Springer New York. https://doi.org/10.1007/0-387-30747-8_23

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, *28*(1), 235-242. https://doi.org/10.1093/nar/28.1.235

Bragg, J. G., & Hyder, C. L. (2004). Nitrogen versus carbon use in prokaryotic genomes and proteomes. *Proc Biol Sci*, *271 Suppl 5*(Suppl 5), S374-377. https://doi.org/10.1098/rsbl.2004.0193

Brimacombe, C. A., Ding, H., & Beatty, J. T. (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*(6), 1260-1278. https://doi.org/10.1111/mmi.12628

Brimacombe, C. A., Ding, H., Johnson, J. A., & Beatty, J. T. (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*(16), 2653-2663. https://doi.org/10.1128/JB.00332-15

Brister, J. R., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015). NCBI viral genomes resource. *Nucleic Acids Res*, *43*(Database issue), D571-577. https://doi.org/10.1093/nar/gku1207

Chen, F., Spano, A., Goodman, B. E., Blasier, K. R., Sabat, A., Jeffery, E., Norris, A., Shabanowitz, J., Hunt, D. F., & Lebedev, N. (2009). Proteomic analysis and identification of the structural and regulatory proteins of the Rhodobacter capsulatus gene transfer agent. *J Proteome Res*, *8*(2), 967-973. https://doi.org/10.1021/pr8006045

Chen, W. H., Lu, G., Bork, P., Hu, S., & Lercher, M. J. (2016). Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat Commun*, *7*, 11334. https://doi.org/10.1038/ncomms11334

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol*, *65*(6), 997-1008. https://doi.org/10.1093/sysbio/syw037

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241-252. https://doi.org/10.1080/00401706.1964.10490181

Engelberg-Kulka, H., Amitai, S., Kolodkin-Gal, I., & Hazan, R. (2006). Bacterial programmed cell death and multicellular behavior in bacteria. *PLoS Genet*, *2*(10), e135. https://doi.org/10.1371/journal.pgen.0020135

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, *125*(1), 1-15. https://doi.org/10.1086/284325

Fogg, P. C. M. (2019). Identification and characterization of a direct activator of a gene transfer agent. *Nat Commun*, *10*(1), 595. https://doi.org/10.1038/s41467-019-08526-1

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2019). Microbial genome analysis: the COG approach. *Brief Bioinform*, *20*(4), 1063-1070. https://doi.org/10.1093/bib/bbx117

Goodman, A. E., Marshall, K. C., & Hermansson, M. (1994). Gene transfer among bacteria under conditions of nutrient depletion in simulated and natural aquatic environments. *FEMS Microbiology Ecology*, *15*(1-2), 55-60. https://doi.org/10.1111/j.1574-6941.1994.tb00229.x

Grzymski, J. J., & Dussaq, A. M. (2012). The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J*, *6*(1), 71-80. https://doi.org/10.1038/ismej.2011.72

Hellweger, F. L., Huang, Y., & Luo, H. (2018). Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. *ISME J*, *12*(5), 1180-1187. https://doi.org/10.1038/s41396-017-0023-7

Hershberg, R., & Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*, *6*(9), e1001115. https://doi.org/10.1371/journal.pgen.1001115

Hildebrand, F., Meyer, A., & Eyre-Walker, A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet*, *6*(9), e1001107. https://doi.org/10.1371/journal.pgen.1001107

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*, *35*(2), 518-522. https://doi.org/10.1093/molbev/msx281

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol*, *34*(8), 2115-2122. https://doi.org/10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., & Bork, P. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*, *44*(D1), D286-293. https://doi.org/10.1093/nar/gkv1248

Hynes, A. P., Shakya, M., Mercer, R. G., Grull, M. P., Bown, L., Davidson, F., Steffen, E., Matchem, H., Peach, M. E., Berger, T., Grebe, K., Zhaxybayeva, O., & Lang, A. S. (2016). Functional and evolutionary characterization of a gene transfer agent's multilocus “genome". *Mol Biol Evol*, *33*(10), 2530-2543. https://doi.org/10.1093/molbev/msw125

Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*, *14*(6), 587-589. https://doi.org/10.1038/nmeth.4285

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, *28*(1), 27-30. https://doi.org/10.1093/nar/28.1.27

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kogay, R., Neely, T. B., Birnbaum, D. P., Hankel, C. R., Shakya, M., & Zhaxybayeva, O. (2019). Machine-learning classification suggests that many alphaproteobacterial prophages may instead be gene transfer agents. *Genome Biol Evol*, *11*(10), 2941-2953. https://doi.org/10.1093/gbe/evz206

Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol*, *4*(1), 87-104. https://doi.org/10.1146/annurev-virology-101416-041624

Lang, A. S., Zhaxybayeva, O., & Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol*, *10*(7), 472-482. https://doi.org/10.1038/nrmicro2802

Lassalle, F., Perian, S., Bataillon, T., Nesme, X., Duret, L., & Daubin, V. (2015). GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet*, *11*(2), e1004941. https://doi.org/10.1371/journal.pgen.1004941

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol*, *25*(7), 1307-1320. https://doi.org/10.1093/molbev/msn067

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*, *47*(W1), W256-W259. https://doi.org/10.1093/nar/gkz239

Luo, H., Thompson, L. R., Stingl, U., & Hughes, A. L. (2015). Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol*, *32*(10), 2738-2748. https://doi.org/10.1093/molbev/msv149

Lynch, M., & Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proc Natl Acad Sci U S A*, *112*(51), 15690-15695. https://doi.org/10.1073/pnas.1514974112

Mahmoudabadi, G., Milo, R., & Phillips, R. (2017). Energetic cost of building a virus. *Proc Natl Acad Sci U S A*, *114*(22), E4324-E4333. https://doi.org/10.1073/pnas.1701670114

Marrs, B. (1974). Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad Sci U S A*, *71*(3), 971-973. https://doi.org/10.1073/pnas.71.3.971

McDaniel, L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., & Paul, J. H. (2010). High frequency of horizontal gene transfer in the oceans. *Science*, *330*(6000), 50. https://doi.org/10.1126/science.1192243

Meibom, K. L., Blokesch, M., Dolganov, N. A., Wu, C. Y., & Schoolnik, G. K. (2005). Chitin induces natural competence in *Vibrio cholerae*. *Science*, *310*(5755), 1824-1827. https://doi.org/10.1126/science.1120096

Mende, D. R., Bryant, J. A., Aylward, F. O., Eppley, J. M., Nielsen, T., Karl, D. M., & DeLong, E. F. (2017). Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat Microbiol*, *2*(10), 1367-1373. https://doi.org/10.1038/s41564-017-0008-3

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, *48*(3), 443-453. https://doi.org/10.1016/0022-2836(70)90057-4

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B. C., Barrangou, R., & Banfield, J. F. (2015). CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *mBio*, *6*(2). https://doi.org/10.1128/mBio.00262-15

Palidwor, G. A., Perkins, T. J., & Xia, X. (2010). A general model of codon bias due to GC mutational bias. *PLoS One*, *5*(10), e13431. https://doi.org/10.1371/journal.pone.0013431

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, *2*(11), 1533-1542. https://doi.org/10.1038/s41564-017-0012-7

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., Quail, M., Smith, F., Walker, D., Libberton, B., Fenton, A., Hall, N., & Brockhurst, M. A. (2010). Antagonistic coevolution accelerates molecular evolution. *Nature*, *464*(7286), 275-278. https://doi.org/10.1038/nature08798

Peeters, S. H., & de Jonge, M. I. (2018). For the greater good: Programmed cell death in bacterial communities. *Microbiol Res*, *207*, 161-169. https://doi.org/10.1016/j.micres.2017.11.016

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., . . . Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503-509. https://doi.org/10.1038/s41586-018-0124-0

Raiford, D. W., Heizer, E. M., Jr., Miller, R. V., Doom, T. E., Raymer, M. L., & Krane, D. E. (2012). Metabolic and translational efficiency in microbial organisms. *J Mol Evol*, *74*(3-4), 206-216. https://doi.org/10.1007/s00239-012-9500-9

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*(2), 217-223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Shakya, M., Soucy, S. M., & Zhaxybayeva, O. (2017). Insights into origin and evolution of alpha-proteobacterial gene transfer agents. *Virus Evol*, *3*(2), vex036. https://doi.org/10.1093/ve/vex036

Smith, J. M. (1964). Group selection and kin selection. *Nature*, *201*(4924), 1145-1147. https://doi.org/10.1038/2011145a0

Solioz, M., Yen, H. C., & Marris, B. (1975). Release and uptake of gene transfer agent by *Rhodopseudomonas capsulata*. *J Bacteriol*, *123*(2), 651-657. https://doi.org/10.1128/jb.123.2.651-657.1975

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*, *34*(Web Server issue), W609-612. https://doi.org/10.1093/nar/gkl315

Swire, J. (2007). Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J Mol Evol*, *64*(5), 558-571. https://doi.org/10.1007/s00239-006-0206-8

Wang, H. C., Minh, B. Q., Susko, E., & Roger, A. J. (2018). Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol*, *67*(2), 216-235. https://doi.org/10.1093/sysbio/syx068

West, S. A., Griffin, A. S., Gardner, A., & Diggle, S. P. (2006). Social evolution theory for microorganisms. *Nat Rev Microbiol*, *4*(8), 597-607. https://doi.org/10.1038/nrmicro1461

Westbye, A. B., O'Neill, Z., Schellenberg-Beaver, T., & Beatty, J. T. (2017). The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology (Reading)*, *163*(9), 1355-1363. https://doi.org/10.1099/mic.0.000519

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, *28*(7), 1033-1034. https://doi.org/10.1093/bioinformatics/bts079

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*, *39*(3), 306-314. https://doi.org/10.1007/BF00160154

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, *24*(8), 1586-1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., Wong, W. S., & Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*, *22*(4), 1107-1118. https://doi.org/10.1093/molbev/msi097

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. https://doi.org/10.1038/srep30372

Chapter 4

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

Published in *mSystems* on 14 November 2022
(DOI: 10.1128/msystems.00892-22)

Supplementary Material is available online (DOI: 10.1128/msystems.00892-22)

Author contributions

RK and OZ designed the study. RK collected data and performed analyses. RK and OZ interpreted results. RK and OZ wrote the initial draft. RK and OZ revised and finalized the manuscript.

Abstract

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

Importance

Horizontal gene transfer (HGT) is a fundamental process that drives evolution of microorganisms. HGT can result a rapid dissemination of beneficial genes within and among microbial communities, and can be achieved via multiple mechanisms. One peculiar HGT mechanism involves viruses “domesticated” by some bacteria and archaea (their hosts). These so-called gene transfer agents (GTAs) are encoded in hosts’ genomes, produced under starvation conditions, and cannot propagate themselves as viruses. We show that GTA genes are under selection to improve efficiency of their translation when the host activates GTA production. The selection is especially pronounced in bacteria that

occupy nutrient-depleted environments. Intriguingly, several genes involved in DNA incorporation into a genome are under similar selection pressure, suggesting that they may facilitate integration of GTA-delivered DNA into the host genome. Our findings underscore the potential importance of GTAs as a mechanism of HGT under nutrient-limited conditions, which are widespread in microbial habitats.

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 (Lang et al., 2012; Marrs, 1974). Released GTA particles can deliver the packaged genetic material to other cells (Brimacombe et al., 2014), impacting exchange of genetic material in prokaryotic populations (Brimacombe et al., 2015; McDaniel et al., 2010; Québatte 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), enable population-level exchange of traits needed under the conditions of a nutritional stress via horizontal gene transfer (HGT) (McDaniel et al., 2010) or decrease population density during the carbon starvation periods (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 (Kogay et al., 2019; Lang & Beatty, 2007; Shakya et al., 2017). Additionally, multiple cellular genes regulate RcGTA production, release and reception (Fogg, 2019; Hynes et al., 2016). 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 nutrient depletion (Westbye, O'Neill, et al., 2017), under which a small fraction of the *R. capsulatus* population becomes dedicated to GTA production (P. C. Fogg et al., 2012; Hynes et al., 2012). As a result, RcGTA-producing cells likely express GTA genes at high levels. By extension, RcGTA-like GTA genes in other alphaproteobacteria (hereafter referred to as "GTA genes" for brevity) also likely to be highly expressed in GTA-producing cells of alphaproteobacterial populations.

Highly expressed genes that are involved in core biological processes, such as translational machinery, are known to exhibit a 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 (Quax et al., 2015; Rocha, 2004; 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. This type of selection is known as the "selection for translational efficiency" and is ubiquitous among bacteria (Supek et al., 2010).

Besides constitutively highly expressed genes, selection for translational efficiency also acts on genes that are highly expressed under specific environmental conditions that microorganisms experience (LaBella et al., 2021; Roller et al., 2013; Supek et al., 2010). For instance, genes that utilize galactose have higher codon usage biases in budding yeasts that live in dairy-associated habitats than in yeasts that occupy alcohol-associated habitats (LaBella et al., 2021). Additionally, genes that encode interacting proteins and genes involved in the same pathway often exhibit similar codon usage biases (Fraser et al., 2004; LaBella et al., 2021).

In earlier work, we have discovered that alphaproteobacterial GTA genes have a striking bias towards GC-rich codons in comparison to the rest of the genome (Kogay et al., 2020; Shakya et al., 2017). However, this bias is different from the codon usage bias: skewed composition of the encoded proteins towards containing energetically cheaper

amino acids caused the first two positions of the codons to be enriched in Gs and Cs, due to the structure of the genetic code (Kogay et al., 2020). In this study, we examined GTA genes in 208 alphaproteobacterial genomes and assessed if there is additionally a codon usage bias due to the genes being under selection for the translation efficiency. For this purpose, we used two well-established metrics for assessment of codon usage 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). ENC quantifies how equally synonymous codons are used in a gene, and varies from 20 (when only one codon is used per each amino acid; strong bias) to 61 (when all codons are used equally; no bias). The tAI measures how optimally the codon usage of each gene fits the available tRNA pool by correlating frequency of each codon in the gene with the abundance of its cognate tRNA. The degree of adaptation of a gene is gauged by comparing its tAI value to the tAI values of all other genes in a genome. 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 usage bias of GTA genes and its match to available tRNAs varies across GTA genes and GTA-containing genomes

To assess the presence of codon usage 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 usage bias, corrected for the genomic GC content (dos Reis et al., 2004). Indeed, we found that 1,543 out of 2,308 (66.8%) reference GTA 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 usage 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 the strength of selection for translational efficiency should be examined in individual GTA genes and in specific taxonomic groups, which we investigate in the next two sections.

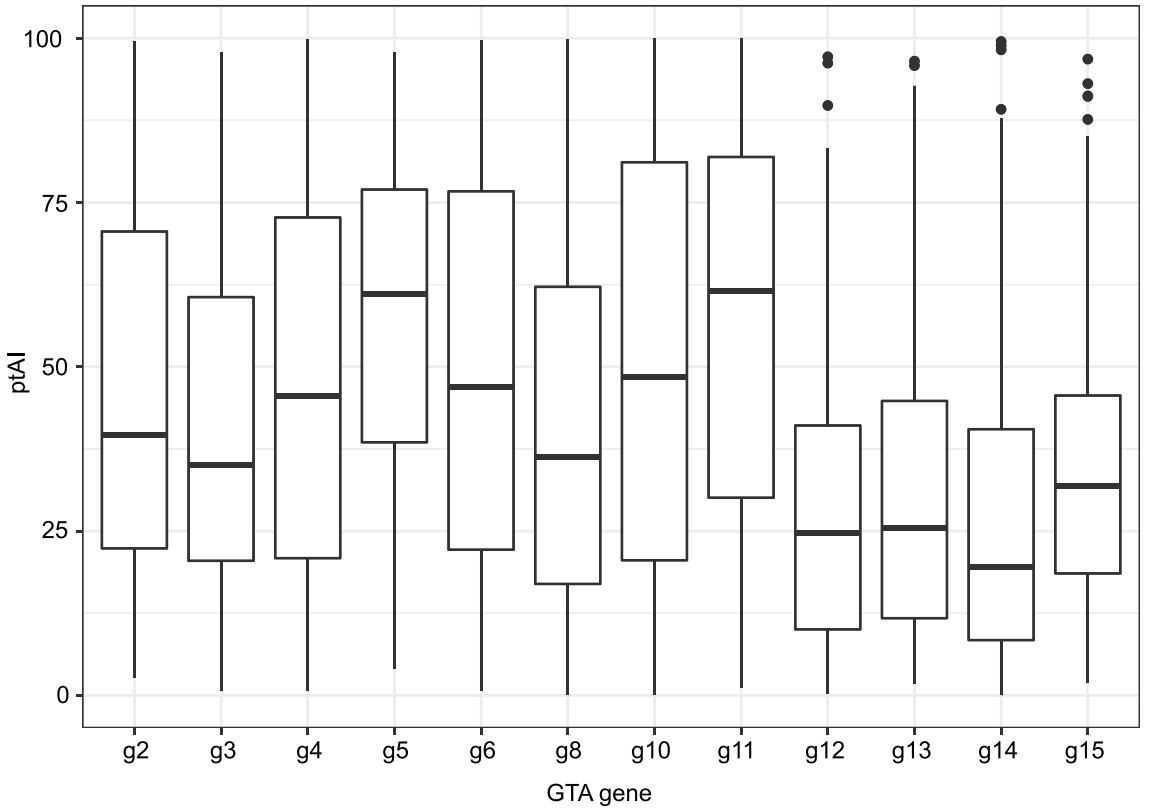

Figure 1. Distribution of ptAI values among reference GTA genes from GTA head-tail clusters in 208 alphaproteobacterial genomes. Line within a box displays the median ptAI value for a GTA gene across all genomes, in which the gene was detected. The boxes are bounded by first and third quartiles. Whiskers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

Selection for translational efficiency is uneven among GTA genes

The differences of ptAI values among the reference GTA genes 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 higher 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 of copies (1-6) 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 & 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 (Fogg, 2019; Lang & Beatty, 2000). 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 variability in ptAI values across different GTA genes, ptAI values of individual GTA genes vary substantially 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 fourteen 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 especially beneficial 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 them (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.

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

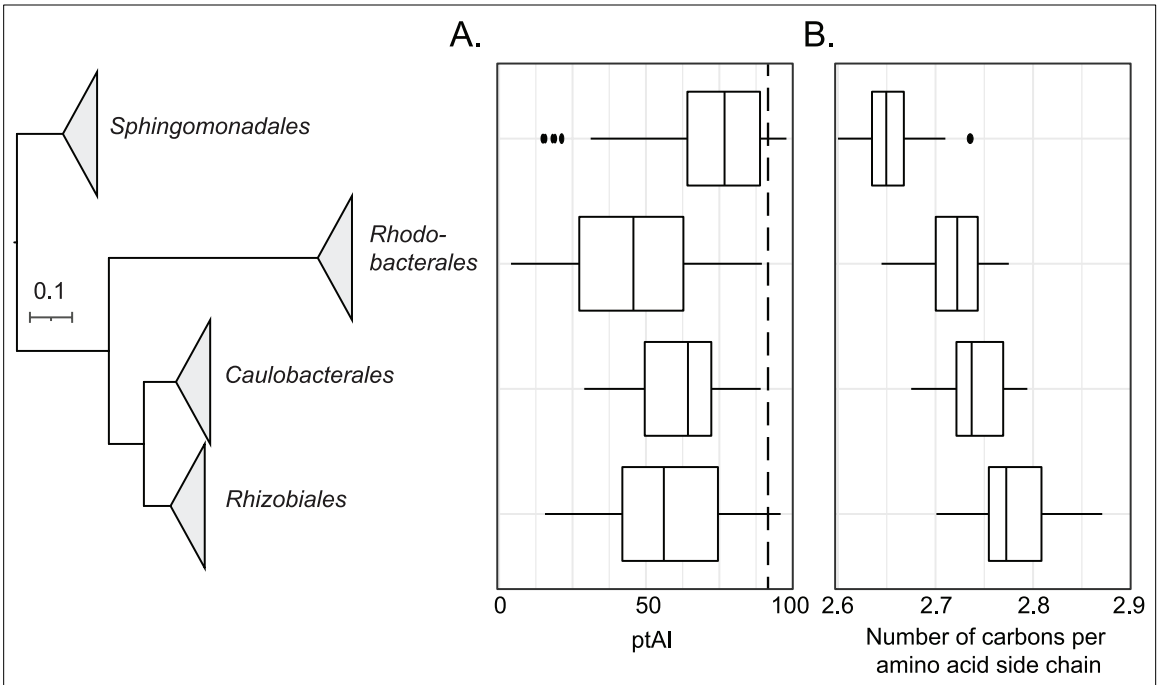

Figure 2. Distributions of (A) ptAI values in the major capsid protein-encoding gene (g5) and (B) carbon content of amino acids in the g5 protein across four orders of the class Alphaproteobacteria. On both panels A and B, line within a box displays the median ptAI value for g5 representatives within 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 phylogenomic tree (see **Methods** for details), in which branches are collapsed at the taxonomic order level. Dashed line in Panel A marks ptAI value of 90.

While the strength of selection for translational efficiency acting on GTA genes varies across gene and genomes, we found that the combinations of ptAI values across all reference GTA genes in a genome have similar trends across the 208 genomes. The similarity is significant in all pairwise reference GTA gene comparisons (**Supplemental Figure S4)**, as determined using phylogenetic generalized least squares (PGLS) method. We conjecture that ptAI values of genes in the other loci of a GTA "genome", as well as the host genes involved in GTA lifecycle, would exhibit similar trends to the ptAI values of reference GTA genes, 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)**.

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

One of 14 identified gene families is a homolog of gafA, which encodes a crucial transcription activator of GTA particles production in *Rhodobacter capsulatus* (Fogg, 2019; Hynes et al., 2016). This gene is located outside of the RcGTA's head-tail cluster, and therefore was not included in the set of reference GTA genes, 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* (Fogg, 2019; Hynes et al., 2016; Shakya et al., 2017). However, with different criteria in the OrthoFinder-based similarity searches, we were able to identify this regulator in 196 of the 208 genomes (94.7%), spanning all GTA-containing alphaproteobacterial orders. The evolutionary history of the gafA homologs is largely congruent with the reference phylogenomic tree (normalized quartet score of 0.87) and even more so with the phylogeny of the concatenated GTA reference genes (normalized quartet score of 0.92) (tree topologies are available at https://doi.org/10.6084/m9.figshare.20082749), suggesting that the gafA gene had co-evolved with the GTA 'head-tail' cluster since the last common ancestor of RcGTA-like GTAs.

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

Protein products of the addA and addB genes 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 the 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 exodeoxyribonuclease VII large subunit (xseA), which is encoded by the xseA gene, 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 usage bias between the xseB gene and GTA genes, we speculate that instead of involvement in processing of GTA DNA, xseA gene product facilitates lysis of GTA producing cells and release of GTA particles.

The tonB gene encodes tonB energy transducer. 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). Similar to the quorum-sensing regulated expression of the gene encoding GTA receptor in the non-GTA-producing cells of a *Rhodobacter capsulatus* population (Brimacombe et al., 2014), the tonB gene could also be regulated to be expressed in the non-GTA-producing cells to aid the uptake of the nutrients released from the lysed cells via TonB-dependent transporters. The tonB gene is currently detected only in members of *Sphingomonadales* order, suggesting that such nutrient uptake is most relevant in the nutrient-limited environments.

Three genes from the ‘molecular chaperones and related functions’ COG category are less likely to be directly involved in GTA lifecycle, because GTAs already encode their own chaperones that assist GTA protein folding (Bardy et al., 2020). However, 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 the reference GTA genes is due to their expression being triggered by the similar environmental conditions.

To evaluate if the 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

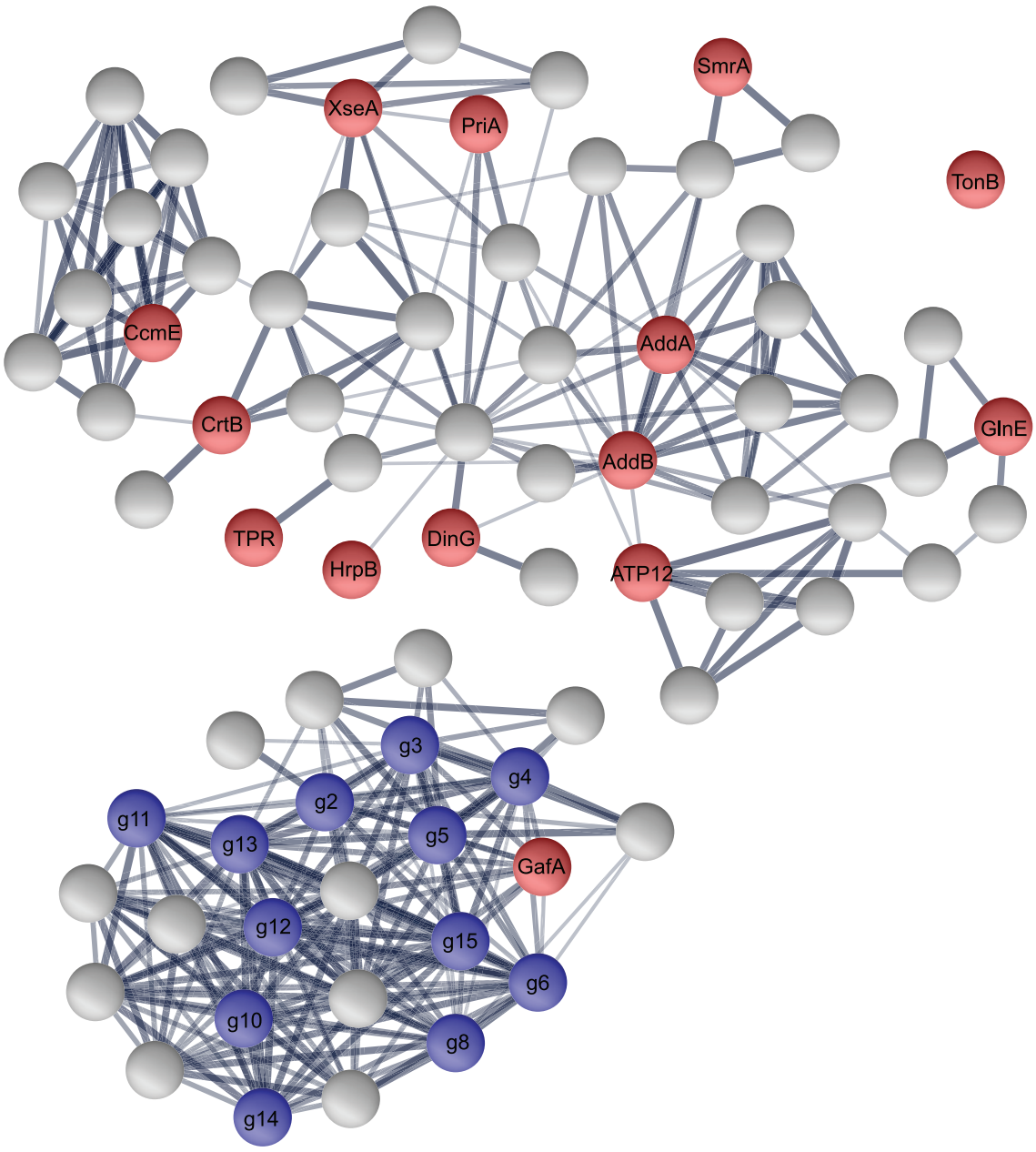

Figure 3. The protein-protein interactions among 12 GTA reference proteins and 14 proteins putatively co-expressed with GTAs. Nodes represent individual proteins. Blue-colored nodes correspond to GTA reference proteins and red-colored nodes correspond to 14 putatively co-expressed proteins. Gray-colored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

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 of the 13 gene families being assigned to the "replication, recombination, and repair" COG category, two of the KEGG pathways are 'homologous recombination' and 'mismatch repair', further corroborating 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 & Dudareva, 2007; Tyc et al., 2017), and carbon starvation leads to the upregulation of carotenoid biosynthesis pathway (Ram et al., 2020; Yang et al., 2015). Similar to the above-described genes encoding chaperones, we hypothesize thatexpression 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 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 (Kogay et al., 2019; Shakya et al., 2017). 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* (GenBank accessions are available at https://doi.org/10.6084/m9.figshare.20082749), 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

(**Supplemental Figure S5**). 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 usage biases suggest that alphaproteobacterial GTA systems are under selection for an optimal translation of GTA proteins from GTA genes achieved by using codons with more readily available tRNAs. The strength of such selection for translational efficiency 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. Additionally, the strength of the selection for translational efficiency varies across taxonomic groups, but is particularly prominent in *Sphingomonadales* order, whose members typically inhabit nutrient-limited conditions. We hypothesize that the observed variation in the selection strength depends on severity and duration of the nutrient scarcity experienced by a population capable of producing GTAs. On the one hand, a long-term exposure to nutrient-depleted conditions would trigger a more efficient and/or more frequent production of GTA particles, which would lead to a greater survival of communities with a better translational efficiency of GTA systems and thus a higher codon usage bias in the GTA genes. On the other hand, if GTAs are needed only in rare occasions due to the stable and abundant nutrient supplies, the selection for translational efficiency would be weak and would result in a lower codon usage bias. Combined with an observation that production of GTAs is triggered by

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 lengths and secondary structures. The secondary structures of *R. capsulatus* g2 (PDB ID 6TUI_8), Enterobacteria phage lambda gpW (1HYW), and Bacillus phage SPP1 (2KCA) proteins were retrieved from the PDB database. The secondary 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 names.

the nutritional stress (Westbye, O'Neill, et al., 2017), our findings that the selection is the strongest in alphaproteobacteria that inhabit nutrient-limited environments further underscore the earlier hypothesized importance of GTA systems in situations of nutrient scarcity (Kogay et al., 2020).

Additionally, the stronger selection for translation efficiency in GTA genes is associated with a 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.

We hypothesized that genes that are located outside of the GTA head-tail cluster, but are involved in GTA lifecycle, including processing and integration of the GTA-delivered DNA, would have signatures of selection for translational efficiency similar to those of 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. Several of these genes are involved in recombination and mismatch repair, providing bioinformatic evidence for the hypothesis that GTAs facilitate HGT by distributing genetic fragments that become incorporated into recipients' genomes via homologous recombination (Brimacombe et al., 2014). Involvement of other genes with similar selection pressures in GTA lifecycle is speculative and needs to be investigated experimentally. But the putative co-expression of xseA and tonB genes with GTA genes raises an intriguing possibility that, in addition to HGT, GTA production may provide an extra benefit in a nutrient-depleted environment: scavenging of scarce organic matter from GTA-producing cells. The lysis of the GTA-producing cells could be mediated by XseA and their debris could be imported as nutrients by the surviving cells via TonB-dependent transporters.

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 & Beatty, 2007; Shakya et al., 2017). However,

similar to the HGT influence onto many other regions of a typical bacterial genome (Soucy et al., 2015), it is very likely that over time GTAs experienced gene replacements via HGT (Shakya et al., 2017). Instances of HGT 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).

Materials and 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 (Altschul et al., 1997) searches with the e-value cutoff of 0.1 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 (see **"Evaluation of the adaptiveness of protein-coding genes to the tRNA pool"** section below for details) did not converge. As a result, only 208 genomes were retained in the reported analyses (GenBank accessions are available at https://doi.org/10.6084/m9.figshare.20082749).

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 & Kelly, 2019) with default parameters and DIAMOND (Buchfink et al., 2015) for the amino acid 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 & Kelly, 2019) with default parameters and 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 S3**), which we designate throughout the manuscript as "GTA reference genes".

Amino acid sequences of GTA reference genes were aligned individually using MAFFT-linsi v7.455 (Katoh & Standley, 2013) and then concatenated into a single alignment. Each gene was treated as a separate partition in the alignment and the best substitution model for each gene was determined by ModelFinder (Kalyaanamoorthy et al., 2017). The maximum likelihood tree was reconstructed using IQ-TREE v1.6.7 (Nguyen et al., 2015) and the support values were calculated via 1,000 ultrafast bootstrap replicates (Hoang et al., 2018).

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 & Scott, 2012). Amino acid sequences within each of the 29 marker families were aligned using MAFFT-linsi v7.455 (Katoh & 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 v1.6.7 (Nguyen et al., 2015) and the support values were calculated using 1,000 ultrafast bootstrap replicates (Hoang et al., 2018).

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 (available in the FigShare repository; see below). For every gene, the deviation of its ENC from the null model was calculated using the in-house script. Genes that have observed ENC higher than the expected 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 (Chan et al., 2021; Lowe & Eddy, 1997) and the Infernal mode without HMM filter to improve the sensitivity of the search (Nawrocki & Eddy, 2013). tRNA gene copy number was used as the proxy for tRNA abundance, following the previously reported observation that the two correlate strongly (dos Reis et al., 2004; Duret, 2000). 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 S6**), 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 & 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 S4**), 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 & 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. For gene families found in at least 40 genomes, a gene family was designated to be associated with a GTA, if obtained fit of the model was statistically significant across all reference GTA genes. Because gene families found in less than 40 genomes were kept only if the genes are located within the genomic regions encoding GTA head-tail clusters (see the "Identification of gene families in 208 alphaproteobacterial genomes" section), a more relaxed criterion was adopted for such gene families: a gene family was designated to be associated with a GTA if the fit of the model was statistically significant across at least 40% of 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 < 1,000) or across random 1,000 combinations of paralogs (if the total number of combinations was > 1,000). 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

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 listed in **Table 1**. 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, the secondary structure of its proteins were predicted using Porter v5.0 (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 & 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 & 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 (tree topology is available at https://doi.org/10.6084/m9.figshare.20082749).

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. (Kogay et al., 2020). The total number of carbons in each protein was normalized by the protein length.

Retrieval and phylogenetic analyses of gafA homologs

Amino acid sequences of gafA homologs in 196 alphaproteobacterial genomes were detected via Orthofinder (gene family OG0001218). Only homologs found in single copy in a genome (194 in total) were retained for phylogenetic analysis. These homologs were aligned using MAFFT-linsi v7.455 (Katoh & Standley, 2013). The best substitution model (LG+F+R6) was determined by ModelFinder (Kalyaanamoorthy et al., 2017) and the maximum-likelihood tree was reconstructed by IQ-TREE v1.6.7 (Nguyen et al., 2015) with the number of iterations to stop set to 500. The support values were calculated using 1,000 ultrafast bootstrap replicates (Hoang et al., 2018). Both GTA reference tree and reference phylogenomic tree were pruned to match the taxa in gafA phylogeny. The normalized quartet scores were calculated using ASTRAL v5.7.8 (Zhang et al., 2018).

Data availability

The following data are available in the FigShare repository under DOI 10.6084/m9.figshare.20082749 (https://doi.org/10.6084/m9.figshare.20082749): accession numbers of 208 analyzed alphaproteobacterial genomes; 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; an in-house script for ENC calculations; slopes and p-values of associations detected in PGLS analyses; accession numbers of the putative g7 proteins in *Sphingomonadales* genomes; multiple sequence alignments and phylogenetic trees of tonB and gafA gene families, concatenated phylogenomic markers, and concatenated GTA reference genes.

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, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Balkwill, D. L., Fredrickson, J. K., & Romine, M. F. (2006). *Sphingomonas* and related genera. In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, & E. Stackebrandt (Eds.), *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass* (pp. 605-629). Springer New York. https://doi.org/10.1007/0-387-30747-8_23

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*(1), 3034. https://doi.org/10.1038/s41467-020-16669-9

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, *28*(1), 235-242. https://doi.org/10.1093/nar/28.1.235

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*(2), e224. https://doi.org/10.1371/journal.pone.0000224

Brimacombe, C. A., Ding, H., & Beatty, J. T. (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*(6), 1260-1278. https://doi.org/10.1111/mmi.12628

Brimacombe, C. A., Ding, H., Johnson, J. A., & Beatty, J. T. (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*(16), 2653-2663. https://doi.org/10.1128/JB.00332-15

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, *12*(1), 59-60. https://doi.org/10.1038/nmeth.3176

Cantalapiedra, C. P., 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*(12), 5825-5829. https://doi.org/10.1093/molbev/msab293

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res*, *49*(16), 9077-9096. https://doi.org/10.1093/nar/gkab688

dos Reis, M., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res*, *32*(17), 5036-5044. https://doi.org/10.1093/nar/gkh834

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*(23), 6976-6985. https://doi.org/10.1093/nar/gkg897

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*(7), 287-289. https://doi.org/10.1016/s0168-9525(00)02041-2

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*(15), 5230-5238. https://doi.org/10.1128/JB.00194-08

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Fogg, P. C., Westbye, A. B., & Beatty, J. T. (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*(8), e43772. https://doi.org/10.1371/journal.pone.0043772

Fogg, P. C. M. (2019). Identification and characterization of a direct activator of a gene transfer agent. *Nat Commun*, *10*(1), 595. https://doi.org/10.1038/s41467-019-08526-1

Fraser, H. B., Hirsh, A. E., Wall, D. P., & Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A*, *101*(24), 9033-9038. https://doi.org/10.1073/pnas.0402591101

Genest, O., Wickner, S., & Doyle, S. M. (2019). Hsp90 and Hsp70 chaperones: Collaborators in protein remodeling. *J Biol Chem*, *294*(6), 2109-2120. https://doi.org/10.1074/jbc.REV118.002806

Gershenzon, J., & Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nat Chem Biol*, *3*(7), 408-414. https://doi.org/10.1038/nchembio.2007.5

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*, *35*(2), 518-522. https://doi.org/10.1093/molbev/msx281

Hynes, A. P., Mercer, R. G., Watton, D. E., Buckley, C. B., & Lang, A. S. (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*(2), 314-325. https://doi.org/10.1111/j.1365-2958.2012.08113.x

Hynes, A. P., Shakya, M., Mercer, R. G., Grull, M. P., Bown, L., Davidson, F., Steffen, E., Matchem, H., Peach, M. E., Berger, T., Grebe, K., Zhaxybayeva, O., & Lang, A. S. (2016). Functional and evolutionary characterization of a gene transfer agent's multilocus “genome". *Mol Biol Evol*, *33*(10), 2530-2543. https://doi.org/10.1093/molbev/msw125

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*(12), 820-828. https://doi.org/10.1007/s12275-015-5304-0

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*, *14*(6), 587-589. https://doi.org/10.1038/nmeth.4285

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*, *49*(D1), D545-D551. https://doi.org/10.1093/nar/gkaa970

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kogay, R., Koppenhofer, S., Beatty, J. T., Kuhn, J. H., Lang, A. S., & Zhaxybayeva, O. (2022). Formal recognition and classification of gene transfer agents as viriforms. *Virus Evol*, *8*(2), veac100. https://doi.org/10.1093/ve/veac100

Kogay, R., Neely, T. B., Birnbaum, D. P., Hankel, C. R., Shakya, M., & Zhaxybayeva, O. (2019). Machine-learning classification suggests that many alphaproteobacterial

prophages may instead be gene transfer agents. *Genome Biol Evol*, *11*(10), 2941-2953. https://doi.org/10.1093/gbe/evz206

Kogay, R., Wolf, Y. I., Koonin, E. V., & 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*(4), e01206-01220. https://doi.org/10.1128/mBio.01206-20

Kooistra, J., Haijema, B. J., & Venema, G. (1993). The *Bacillus subtilis* addAB genes are fully functional in *Escherichia coli*. *Mol Microbiol*, *7*(6), 915-923. https://doi.org/10.1111/j.1365-2958.1993.tb01182.x

LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T., & Rokas, A. (2019). Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet*, *15*(7), e1008304. https://doi.org/10.1371/journal.pgen.1008304

LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T., & Rokas, A. (2021). Signatures of optimal codon usage in metabolic genes inform budding yeast ecology. *PLoS Biol*, *19*(4), e3001185. https://doi.org/10.1371/journal.pbio.3001185

Lang, A. S., & Beatty, J. T. (2000). Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. *Proc Natl Acad Sci U S A*, *97*(2), 859-864. https://doi.org/10.1073/pnas.97.2.859

Lang, A. S., & Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*, *15*(2), 54-62. https://doi.org/10.1016/j.tim.2006.12.001

Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol*, *4*(1), 87-104. https://doi.org/10.1146/annurev-virology-101416-041624

Lang, A. S., Zhaxybayeva, O., & Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol*, *10*(7), 472-482. https://doi.org/10.1038/nrmicro2802

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*, *49*(W1), W293-W296. https://doi.org/10.1093/nar/gkab301

Lim, H. N., Lee, Y., & Hussein, R. (2011). Fundamental relationship between operon organization and gene expression. *Proc Natl Acad Sci U S A*, *108*(26), 10626-10631. https://doi.org/10.1073/pnas.1105692108

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, *25*(5), 955-964. https://doi.org/10.1093/nar/25.5.955

Marrs, B. (1974). Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad Sci U S A*, *71*(3), 971-973. https://doi.org/10.1073/pnas.71.3.971

Marrs, B., Wall, J. D., & Gest, H. (1977). Emergence of the biochemical genetics and molecular biology of photosynthetic bacteria. *Trends Biochem Sci*, *2*(5), 105-108. https://doi.org/10.1016/0968-0004(77)90173-6

Marsin, S., Lopes, A., Mathieu, A., Dizet, E., Orillard, E., Guerois, R., & Radicella, J. P. (2010). Genetic dissection of *Helicobacter pylori* AddAB role in homologous recombination. *FEMS Microbiol Lett*, *311*(1), 44-50. https://doi.org/10.1111/j.1574-6968.2010.02077.x

Martins, E. P., & Hansen, T. F. (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, L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., & Paul, J. H. (2010). High frequency of horizontal gene transfer in the oceans. *Science*, *330*(6000), 50. https://doi.org/10.1126/science.1192243

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933-2935. https://doi.org/10.1093/bioinformatics/btt509

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

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*(4), 1355-1361. https://doi.org/10.1128/AEM.02268-06

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

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

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, *12*(1), 32-42. https://doi.org/10.1038/nrg2899

Quax, T. E., Claassens, N. J., Soll, D., & van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Mol Cell*, *59*(2), 149-161. https://doi.org/10.1016/j.molcel.2015.05.035

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*(6), 611-621 e616. https://doi.org/10.1016/j.cels.2017.05.011

Ram, S., Mitra, M., Shah, F., Tirkey, S. R., & 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. https://doi.org/https://doi.org/10.1016/j.jff.2020.103867

Rocha, E. P. (2004). Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res*, *14*(11), 2279-2286. https://doi.org/10.1101/gr.2896904

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*(4), 846-854. https://doi.org/10.1128/JB.180.4.846-854.1998

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*(19), 8842-8852. https://doi.org/10.1093/nar/gkt673

Sabi, R., Volvovitch Daniel, R., & Tuller, T. (2017). stAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics*, *33*(4), 589-591. https://doi.org/10.1093/bioinformatics/btw647

Shakya, M., Soucy, S. M., & Zhaxybayeva, O. (2017). Insights into origin and evolution of alpha-proteobacterial gene transfer agents. *Virus Evol*, *3*(2), vex036. https://doi.org/10.1093/ve/vex036

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nat Rev Genet*, *16*(8), 472-482. https://doi.org/10.1038/nrg3962

Supek, F., Skunca, N., Repar, J., Vlahovicek, K., & Smuc, T. (2010). Translational selection is ubiquitous in prokaryotes. *PLoS Genet*, *6*(6), e1001004. https://doi.org/10.1371/journal.pgen.1001004

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*, *49*(D1), D605-D612. https://doi.org/10.1093/nar/gkaa1074

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*(7), e41204. https://doi.org/10.1371/journal.pone.0041204

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*(1), 12374. https://doi.org/10.1038/s41598-019-48786-x

Tyc, O., Song, C., Dickschat, J. S., Vos, M., & Garbeva, P. (2017). The ecological role of volatile and soluble secondary metabolites produced by soil bacteria. *Trends Microbiol*, *25*(4), 280-292. https://doi.org/10.1016/j.tim.2016.12.002

UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, *49*(D1), D480-D489. https://doi.org/10.1093/nar/gkaa1100

Westbye, A. B., O'Neill, Z., Schellenberg-Beaver, T., & Beatty, J. T. (2017). The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology (Reading)*, *163*(9), 1355-1363. https://doi.org/10.1099/mic.0.000519

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, *87*(1), 23-29. https://doi.org/10.1016/0378-1119(90)90491-9

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, *28*(7), 1033-1034. https://doi.org/10.1093/bioinformatics/bts079

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. https://doi.org/10.1038/srep08331

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. https://doi.org/10.1038/srep30372

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(Suppl 6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C. H., 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*(41), E6117-E6125. https://doi.org/10.1073/pnas.1606724113

Chapter 5

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

Available in *bioRxiv* (Submitted to *Journal of Bacteriology*)
(DOI: 10.1101/2023.08.11.553018)

Supplementary Material is available online (DOI: 10.1101/2023.08.11.553018)

Author contributions

RK and OZ designed the study. RK collected data and performed analyses. RK and OZ interpreted results. RK and OZ wrote the manuscript.

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.

Importance

Gene transfer agents (GTAs) are viruses domesticated by some archaea and bacteria as vehicles for carrying pieces of the host genome. Produced under certain environmental conditions, GTA particles can deliver DNA to neighboring, closely related cells. Function of GTAs remains uncertain. While making GTAs is suicidal for a cell, GTA-encoding genes are widespread in genomes of alphaproteobacteria. Such GTA persistence implies functional benefits but raises question about how selection maintains this lethal trait. By showing that GTA genes co-evolve with genes involved in stress

response, DNA repair, and biofilm formation, we provide support for the hypothesis that GTAs facilitate DNA exchange during the stress conditions and construct a model for how GTAs persist in biofilm-forming bacterial populations despite being lethal.

Introduction

Multiple bacteria and archaea produce Gene Transfer Agents (GTAs) – the viriforms whose function and mode of selection to maintain them remain unsolved (Kogay et al., 2022; Kuhn & Koonin, 2023; Lang et al., 2017). 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 (Kogay et al., 2022; Lang & Beatty, 2001). Experimentally, GTAs are most studied and characterized in the alphaproteobacteria *Rhodobacter capsulatus* (RcGTA) and *Caulobacter crescentus* (Gozzi et al., 2022; Lang & Beatty, 2000; Marrs, 1974), but they are also produced by several additional bacterial and archaeal species (Lang et al., 2012). Many more prokaryotes encode GTA-like genes (Fallon & Carroll, 2023; George et al., 2022; Lang & Beatty, 2007; Lang et al., 2002; Québatte et al., 2017; Shakya et al., 2017; Tamarit et al., 2018), 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 & 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 (Farrera-Calderon et al., 2021; Mercer et al., 2012). Serine acetyltransferase (cysE1), which is required for biofilm formation, plays a critical role for the optimal receipt of RcGTAs (Sherlock & 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 (Brunette et al., 2019; Kim et al., 2004). The co-evolution among genes can be effectively identified via Evolutionary Rate Covariation (ERC) approach (Clark et al., 2012; Goh et al., 2000; Sato et al., 2005). 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 & 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 (Fogg, 2019; Lang & Beatty, 2000) 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 (Bardy et al., 2020; Hynes et al., 2016; Lang et al., 2017) (**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

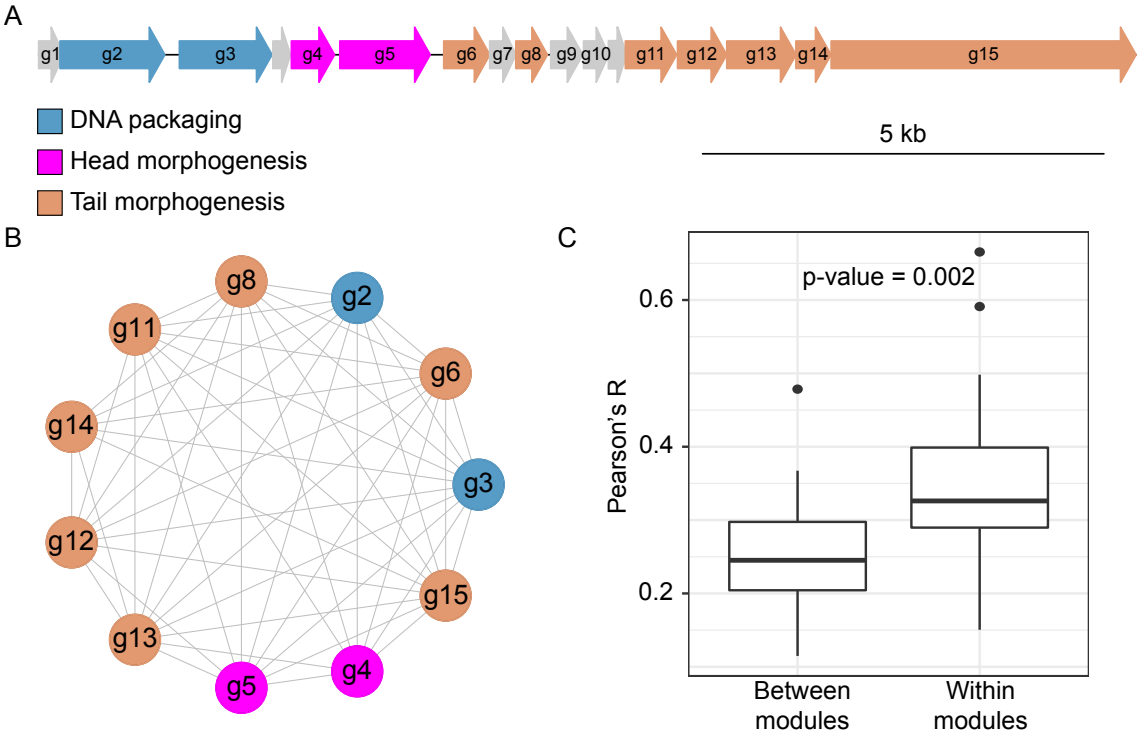

Figure 1. Co-evolution of 11 reference GTA genes from head-tail clusters. (A) The head-tail cluster of the RcGTA, as encoded in *R. capsulatus* strain SB 1003 genome (GenBank accession CP001312.1; spanning locus tags RCAP_rcc01682 - RCAP_rcc01698). Each gene is represented by an arrow and is drawn proportional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

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

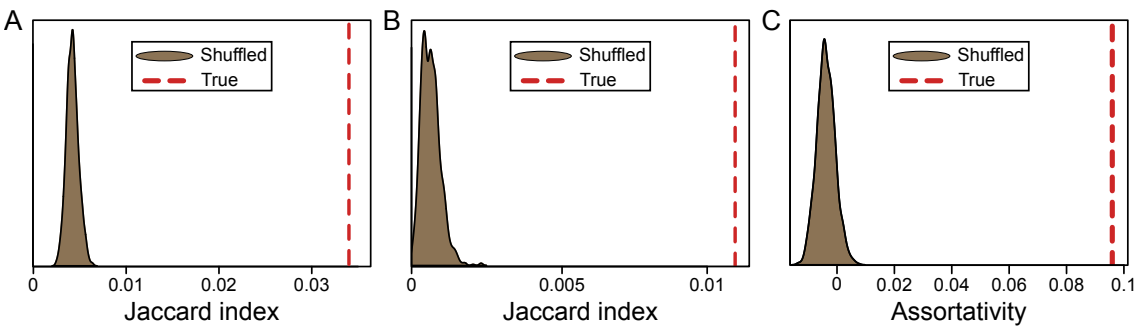

Figure 2. Strengths of correlations between the covariation evolutionary rate of genes and function of the proteins the genes encode, as measured by permutation tests. (A) Comparison of the co-evolution network and the PPI network of *Phaeobacter inhibens*. The distribution in brown corresponds to distances 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.

STRING. Indeed, ERC analysis has been used to uncover novel protein-protein interactions, especially between hypothetical proteins (Brunette et al., 2019; Forsythe et al., 2021; Raza et al., 2019). 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 (Han et al., 2016; Narita et al., 2013).

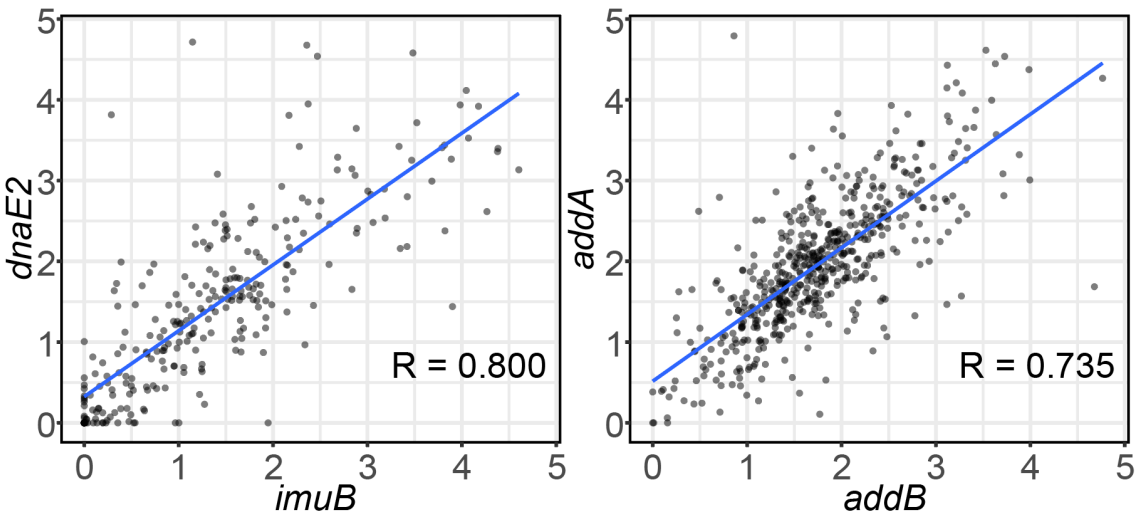

Figure 3. Covariation of evolutionary rates for two gene pairs with the largest Pearson's coefficients. Each dot represents the normalized branch length of the proteins shown on x- and y-axes. Blue line depicts the linear regression line. The value of Pearson's R is shown on each panel.

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-

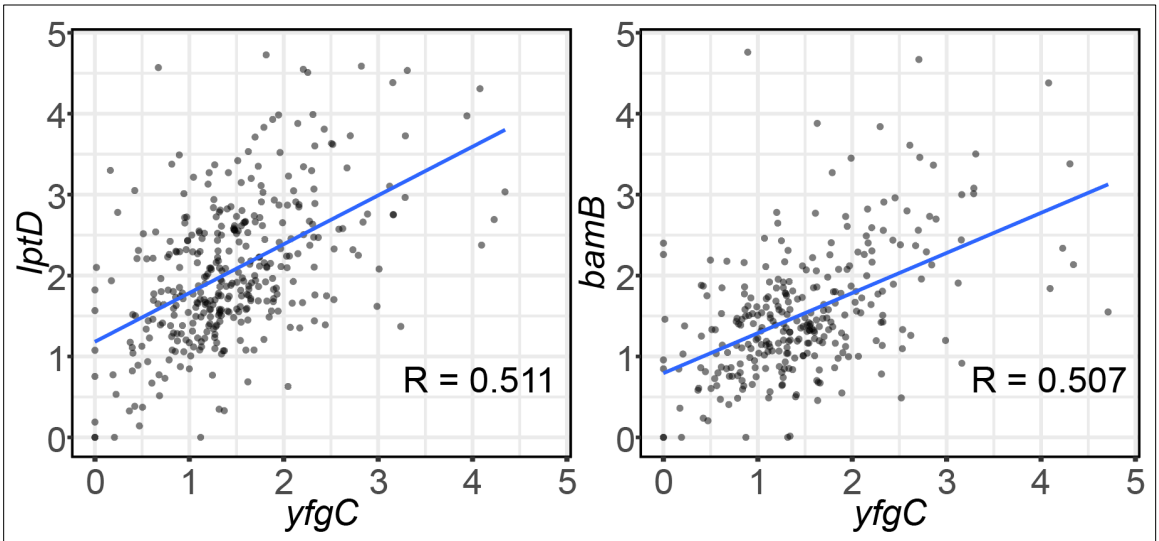

Figure 4. Covariation of evolutionary rates for two gene pairs with no evidence of interactions for their protein products in the STRING database. Each dot represents the normalized branch length of the proteins shown on x- and y-axes. Blue line depicts the linear regression line. The value of Pearson's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

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 (Chen et al., 2014; Kim & Sundin, 2001; Krokan et al., 1997). 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 (Gozzi et al., 2022; Marrs, 1974). Photolyase and tatD exonuclease do not directly participate in homologous recombination, but are involved in DNA repair process (Chen et al., 2014; Gozzi et al., 2022; Kim & Sundin, 2001) and thus are likely to be under the similar selection pressure as GTA genes.

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

Gene name	Representative GenBank accession No.	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

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 (Martin-Rodriguez, 2022; Schinner et al., 2020). 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 (Gozzi et al., 2022; Westbye, O'Neill, et al., 2017), GTAs being involved in DNA repair in recipient cell (Gozzi et al., 2022; Kogay & Zhaxybayeva, 2022), and, most recently, of GTA production occurring in biofilms (Sherlock & 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; Sherlock & Fogg, 2022; Zhang et al., 2019). Biofilms provide a microbial community with benefits that cannot be achieved by the individual cells, such as protection against antibiotics (Stewart & Costerton, 2001) and viral infections (Simmons et al., 2020; Vidakovic et al., 2018). Despite either shown or hypothesized benefits of GTA-disseminated DNA to the recipient cells (Gozzi et al., 2022; Lang et al., 2017; Westbye, Beatty, et al., 2017), 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 (Vos, 2009; Vulic et al., 1997), 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 &

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 (Kogay et al., 2019; Lang & Beatty, 2007; Shakya et al., 2017). 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 & 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 (Lehtonen, 2016; Price, 1970). 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 & 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.

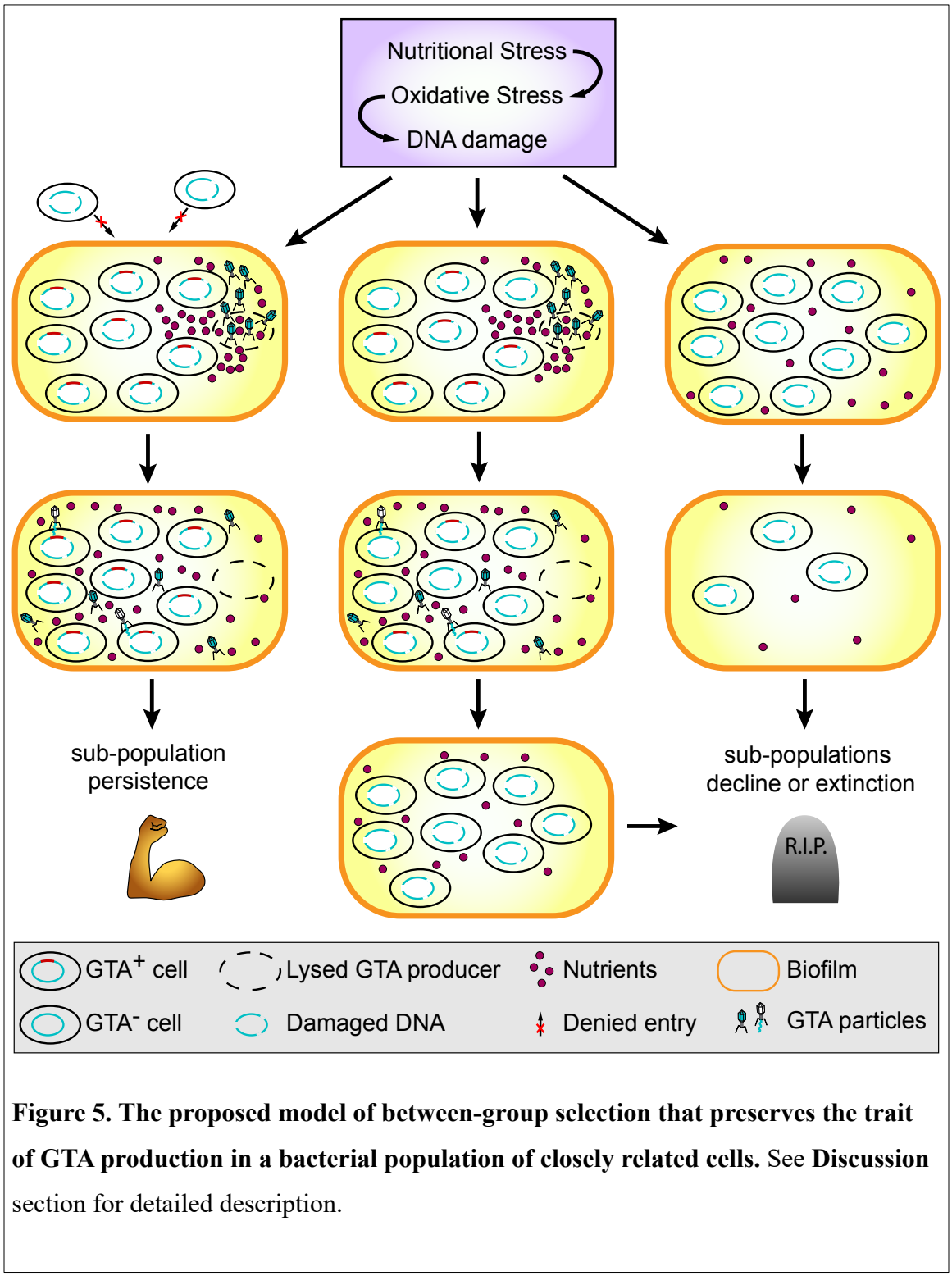

Figure 5. The proposed model of between-group selection that preserves the trait of GTA production in a bacterial population of closely related cells. See **Discussion** section for detailed description.

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 & Scott, 2012). Amino acid sequences of each of these 84 gene families were aligned using MAFFT v7.505 with the 'linsi' option (Katoh & 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 (Kogay et al., 2019; Shakya et al., 2017).

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 & 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 & 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, Serra, 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 & 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 & 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*(1), 212-222. https://doi.org/10.1111/j.1365-2958.2004.04260.x

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*(1), e1609. https://doi.org/10.1038/cddis.2014.570

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Andreae, C. A., Titball, R. W., & Butler, C. S. (2014). Influence of the molybdenum cofactor biosynthesis on anaerobic respiration, biofilm formation and motility in *Burkholderia thailandensis*. *Res Microbiol*, *165*(1), 41-49. https://doi.org/10.1016/j.resmic.2013.10.009

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*(1), 3034. https://doi.org/10.1038/s41467-020-16669-9

Bayles, K. W. (2007). The biological role of death and lysis in biofilm development. *Nat Rev Microbiol*, *5*(9), 721-726. https://doi.org/10.1038/nrmicro1743

Brimacombe, C. A., Ding, H., & Beatty, J. T. (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*(6), 1260-1278. https://doi.org/10.1111/mmi.12628

Brimacombe, C. A., Ding, H., Johnson, J. A., & Beatty, J. T. (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*(16), 2653-2663. https://doi.org/10.1128/JB.00332-15

Brimacombe, C. A., Stevens, A., Jun, D., Mercer, R., Lang, A. S., & Beatty, J. T. (2013). Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). *Mol Microbiol*, *87*(4), 802-817. https://doi.org/10.1111/mmi.12132

Brunette, G. J., Jamalruddin, M. A., Baldock, R. A., Clark, N. L., & Bernstein, K. A. (2019). Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes. *Proc Natl Acad Sci U S A*, *116*(39), 19593-19599. https://doi.org/10.1073/pnas.1906559116

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, *12*(1), 59-60. https://doi.org/10.1038/nmeth.3176

Cantalapiedra, C. P., 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*(12), 5825-5829. https://doi.org/10.1093/molbev/msab293

Chen, Y. C., Li, C. L., Hsiao, Y. Y., Duh, Y., & Yuan, H. S. (2014). Structure and function of TatD exonuclease in DNA repair. *Nucleic Acids Res*, *42*(16), 10776-10785. https://doi.org/10.1093/nar/gku732

Clark, N. L., Alani, E., & Aquadro, C. F. (2012). Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res*, *22*(4), 714-720. https://doi.org/10.1101/gr.132647.111

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, complex systems*, *1695*(5), 1-9.

Derelle, R., Philippe, H., & Colbourne, J. K. (2020). Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol Biol Evol*, *37*(11), 3389-3396. https://doi.org/10.1093/molbev/msaa159

Devaraj, A., Buzzo, J. R., Mashburn-Warren, L., Gloag, E. S., Novotny, L. A., Stoodley, P., Bakaletz, L. O., & Goodman, S. D. (2019). The extracellular DNA lattice of bacterial biofilms is structurally related to Holliday junction recombination intermediates. *Proc Natl Acad Sci U S A*, *116*(50), 25068-25077. https://doi.org/10.1073/pnas.1909017116

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*(1), 66-77. https://doi.org/10.1093/nar/gkj412

Fallon, A. M., & Carroll, E. M. (2023). Virus-like particles from *Wolbachia*-infected cells may include a gene transfer agent. *Insects*, *14*(6), 516. https://doi.org/10.3390/insects14060516

Farrera-Calderon, R. G., Pallegar, P., Westbye, A. B., Wiesmann, C., Lang, A. S., & Beatty, J. T. (2021). The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer agent production Is bidirectional and regulated by cyclic di-GMP. *J Bacteriol*, *203*(5), e00525-00520. https://doi.org/10.1128/JB.00525-20

Fogg, P. C. M. (2019). Identification and characterization of a direct activator of a gene transfer agent. *Nat Commun*, *10*(1), 595. https://doi.org/10.1038/s41467-019-08526-1

Forsythe, E. S., Williams, A. M., & Sloan, D. B. (2021). Genome-wide signatures of plastid-nuclear coevolution point to repeated perturbations of plastid proteostasis systems across angiosperms. *Plant Cell*, *33*(4), 980-997. https://doi.org/10.1093/plcell/koab021

George, E. E., Tashyreva, D., Kwong, W. K., Okamoto, N., Horak, A., Husnik, F., Lukes, J., & Keeling, P. J. (2022). Gene transfer agents in bacterial endosymbionts of microbial eukaryotes. *Genome Biol Evol*, *14*(7), evac099. https://doi.org/10.1093/gbe/evac099

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J Mol Biol*, *299*(2), 283-293. https://doi.org/10.1006/jmbi.2000.3732

Gozzi, K., Tran, N. T., Modell, J. W., Le, T. B. K., & Laub, M. T. (2022). Prophage-like gene transfer agents promote *Caulobacter crescentus* survival and DNA repair during stationary phase. *PLoS Biol*, *20*(11), e3001790. https://doi.org/10.1371/journal.pbio.3001790

Granato, L. M., Picchi, S. C., Andrade Mde, O., Takita, M. A., de Souza, A. A., Wang, N., & Machado, M. A. (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. https://doi.org/10.1186/s12866-016-0655-1

Gross, M., Marianovsky, I., & Glaser, G. (2006). MazG -- a regulator of programmed cell death in *Escherichia coli*. *Mol Microbiol*, *59*(2), 590-601. https://doi.org/10.1111/j.1365-2958.2005.04956.x

Han, L., Zheng, J., Wang, Y., Yang, X., Liu, Y., Sun, C., Cao, B., Zhou, H., Ni, D., Lou, J., Zhao, Y., & Huang, Y. (2016). Structure of the BAM complex and its implications for biogenesis of outer-membrane proteins. *Nat Struct Mol Biol*, *23*(3), 192-196. https://doi.org/10.1038/nsmb.3181

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*, *33*(6), 1635-1638. https://doi.org/10.1093/molbev/msw046

Hynes, A. P., Shakya, M., Mercer, R. G., Grull, M. P., Bown, L., Davidson, F., Steffen, E., Matchem, H., Peach, M. E., Berger, T., Grebe, K., Zhaxybayeva, O., & Lang, A. S. (2016). Functional and evolutionary characterization of a gene transfer agent's multilocus “genome". *Mol Biol Evol*, *33*(10), 2530-2543. https://doi.org/10.1093/molbev/msw125

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, *11*(2), 37-50.

Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*, *14*(6), 587-589. https://doi.org/10.1038/nmeth.4285

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol*, *428*(4), 726-731. https://doi.org/10.1016/j.jmb.2015.11.006

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kim, J. J., & Sundin, G. W. (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*(4), 1405-1411. https://doi.org/10.1128/AEM.67.4.1405-1411.2001

Kim, W. K., Bolser, D. M., & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, *20*(7), 1138-1150. https://doi.org/10.1093/bioinformatics/bth053

Kogay, R., Koppenhofer, S., Beatty, J. T., Kuhn, J. H., Lang, A. S., & Zhaxybayeva, O. (2022). Formal recognition and classification of gene transfer agents as viriforms. *Virus Evol*, *8*(2), veac100. https://doi.org/10.1093/ve/veac100

Kogay, R., Neely, T. B., Birnbaum, D. P., Hankel, C. R., Shakya, M., & Zhaxybayeva, O. (2019). Machine-learning classification suggests that many alphaproteobacterial prophages may instead be gene transfer agents. *Genome Biol Evol*, *11*(10), 2941-2953. https://doi.org/10.1093/gbe/evz206

Kogay, R., Wolf, Y. I., Koonin, E. V., & 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*(4), e01206-01220. https://doi.org/10.1128/mBio.01206-20

Kogay, R., & Zhaxybayeva, O. (2022). Selection for translational efficiency in genes associated with alphaproteobacterial gene transfer agents. *mSystems*, *7*(6), e0089222. https://doi.org/10.1128/msystems.00892-22

Kooistra, J., Haijema, B. J., & Venema, G. (1993). The *Bacillus subtilis* addAB genes are fully functional in *Escherichia coli*. *Mol Microbiol*, *7*(6), 915-923. https://doi.org/10.1111/j.1365-2958.1993.tb01182.x

Krokan, H. E., Standal, R., & Slupphaug, G. (1997). DNA glycosylases in the base excision repair of DNA. *Biochem J*, *325 (Pt 1)*(Pt 1), 1-16. https://doi.org/10.1042/bj3250001

Kuhn, J. H., & Koonin, E. V. (2023). Viriforms-a new category of classifiable virus-derived genetic elements. *Biomolecules*, *13*(2), 289. https://doi.org/10.3390/biom13020289

Lang, A. S., & Beatty, J. T. (2000). Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. *Proc Natl Acad Sci U S A*, *97*(2), 859-864. https://doi.org/10.1073/pnas.97.2.859

Lang, A. S., & Beatty, J. T. (2001). The gene transfer agent of *Rhodobacter capsulatus* and "constitutive transduction" in prokaryotes. *Arch Microbiol*, *175*(4), 241-249. https://doi.org/10.1007/s002030100260

Lang, A. S., & Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*, *15*(2), 54-62. https://doi.org/10.1016/j.tim.2006.12.001

Lang, A. S., Taylor, T. A., & Beatty, J. T. (2002). Evolutionary implications of phylogenetic analyses of the gene transfer agent (GTA) of *Rhodobacter capsulatus*. *J Mol Evol*, *55*(5), 534-543. https://doi.org/10.1007/s00239-002-2348-7

Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol*, *4*(1), 87-104. https://doi.org/10.1146/annurev-virology-101416-041624

Lang, A. S., Zhaxybayeva, O., & Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol*, *10*(7), 472-482. https://doi.org/10.1038/nrmicro2802

Lehtonen, J. (2016). Multilevel selection in kin selection language. *Trends Ecol Evol*, *31*(10), 752-762. https://doi.org/10.1016/j.tree.2016.07.006

Leimkuhler, S., Buhning, M., & Beilschmidt, L. (2017). Shared sulfur mobilization routes for tRNA thiolation and molybdenum cofactor biosynthesis in prokaryotes and eukaryotes. *Biomolecules*, *7*(1), 5. https://doi.org/10.3390/biom7010005

Leung, M. M., Brimacombe, C. A., Spiegelman, G. B., & Beatty, J. T. (2012). The GtaR protein negatively regulates transcription of the gtaRI operon and modulates gene transfer agent (RcGTA) expression in *Rhodobacter capsulatus*. *Mol Microbiol*, *83*(4), 759-774. https://doi.org/10.1111/j.1365-2958.2011.07963.x

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*, *48*(D1), D265-D268. https://doi.org/10.1093/nar/gkz991

Marrs, B. (1974). Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad Sci U S A*, *71*(3), 971-973. https://doi.org/10.1073/pnas.71.3.971

Marrs, B., Wall, J. D., & Gest, H. (1977). Emergence of the biochemical genetics and molecular biology of photosynthetic bacteria. *Trends Biochem Sci*, *2*(5), 105-108. https://doi.org/10.1016/0968-0004(77)90173-6

Martin-Rodriguez, A. J. (2022). Respiration-induced biofilm formation as a driver for bacterial niche colonization. *Trends Microbiol*, 120-134. https://doi.org/10.1016/j.tim.2022.08.007

McBee, M. E., Chionh, Y. H., Sharaf, M. L., Ho, P., Cai, M. W., & Dedon, P. C. (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. https://doi.org/10.3389/fmicb.2017.00459

McDaniel, L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., & Paul, J. H. (2010). High frequency of horizontal gene transfer in the oceans. *Science*, *330*(6000), 50. https://doi.org/10.1126/science.1192243

Mercer, R. G., Quinlan, M., Rose, A. R., Noll, S., Beatty, J. T., & Lang, A. S. (2012). Regulatory systems controlling motility and gene transfer agent production and

release in *Rhodobacter capsulatus*. *FEMS Microbiol Lett*, *331*(1), 53-62. https://doi.org/10.1111/j.1574-6968.2012.02553.x

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., 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*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet*, *17*(10), 589-596. https://doi.org/10.1016/s0168-9525(01)02447-7

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*(38), E3612-3621. https://doi.org/10.1073/pnas.1312012110

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Oliveira, N. M., Martinez-Garcia, E., Xavier, J., Durham, W. M., Kolter, R., Kim, W., & Foster, K. R. (2015). Biofilm formation as a response to ecological competition. *PLoS Biol*, *13*(7), e1002191. https://doi.org/10.1371/journal.pbio.1002191

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*, *36*(10), 996-1004. https://doi.org/10.1038/nbt.4229

Price, G. R. (1970). Selection and covariance. *Nature*, *227*(5257), 520-521. https://doi.org/10.1038/227520a0

Price, M. N., & Arkin, A. P. (2017). PaperBLAST: text mining papers for information about homologs. *mSystems*, *2*(4), e00039-00017. https://doi.org/10.1128/mSystems.00039-17

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., . . . Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503-509. https://doi.org/10.1038/s41586-018-0124-0

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*(6), 611-621 e616. https://doi.org/10.1016/j.cels.2017.05.011

Rajagopala, S. V., Casjens, S., & Uetz, P. (2011). The protein interaction map of bacteriophage lambda. *BMC Microbiol*, *11*, 213. https://doi.org/10.1186/1471-2180-11-213

Raza, Q., Choi, J. Y., Li, Y., O'Dowd, R. M., Watkins, S. C., Chikina, M., Hong, Y., Clark, N. L., & Kwiatkowski, A. V. (2019). Evolutionary rate covariation analysis of E-cadherin identifies Raskol as a regulator of cell adhesion and actin dynamics in Drosophila. *PLoS Genet*, *15*(2), e1007720. https://doi.org/10.1371/journal.pgen.1007720

Redfield, R. J., & Soucy, S. M. (2018). Evolution of bacterial gene transfer agents. *Front Microbiol*, *9*, 2527. https://doi.org/10.3389/fmicb.2018.02527

Rossy, T., Nadell, C. D., & Persat, A. (2019). Cellular advective-diffusion drives the emergence of bacterial surface colonization patterns and heterogeneity. *Nat Commun*, *10*(1), 2471. https://doi.org/10.1038/s41467-019-10469-6

Saikrishnan, K., Yeeles, J. T., Gilhooly, N. S., Krajewski, W. W., Dillingham, M. S., & Wigley, D. B. (2012). Insights into Chi recognition from the structure of an AddAB-type helicase-nuclease complex. *EMBO J*, *31*(6), 1568-1578. https://doi.org/10.1038/emboj.2012.9

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*(17), 3482-3489. https://doi.org/10.1093/bioinformatics/bti564

Schinner, S., Engelhardt, F., Preusse, M., Thoming, J. G., Tomasch, J., & Haussler, S. (2020). Genetic determinants of *Pseudomonas aeruginosa* fitness during biofilm growth. *Biofilm*, *2*, 100023. https://doi.org/10.1016/j.bioflm.2020.100023

Shakya, M., Soucy, S. M., & Zhaxybayeva, O. (2017). Insights into origin and evolution of alpha-proteobacterial gene transfer agents. *Virus Evol*, *3*(2), vex036. https://doi.org/10.1093/ve/vex036

Sherlock, D., & Fogg, P. C. M. (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*(19), e0094422. https://doi.org/10.1128/aem.00944-22

Simmons, E. L., Bond, M. C., Koskella, B., Drescher, K., Bucci, V., & Nadell, C. D. (2020). Biofilm structure promotes coexistence of phage-resistant and phage-susceptible bacteria. *mSystems*, *5*(3). https://doi.org/10.1128/mSystems.00877-19

Smith, P., & Schuster, M. (2019). Public goods and cheating in microbes. *Curr Biol*, *29*(11), R442-R447. https://doi.org/10.1016/j.cub.2019.03.001

Spek, E. J., Vuong, L. N., Matsuguchi, T., Marinus, M. G., & Engelward, B. P. (2002). Nitric oxide-induced homologous recombination in Escherichia coli is promoted by DNA glycosylases. *J Bacteriol*, *184*(13), 3501-3507. https://doi.org/10.1128/JB.184.13.3501-3507.2002

Steenwyk, J. L., Buida, T. J., Labella, A. L., Li, Y., Shen, X. X., & Rokas, A. (2021). PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics*, *37*(16), 2325–2331. https://doi.org/10.1093/bioinformatics/btab096

Steenwyk, J. L., Phillips, M. A., Yang, F., Date, S. S., Graham, T. R., Berman, J., Hittinger, C. T., & Rokas, A. (2022). An orthologous gene coevolution network provides insight into eukaryotic cellular and genomic structure and function. *Sci Adv*, *8*(18), eabn0105. https://doi.org/10.1126/sciadv.abn0105

Stewart, P. S., & Costerton, J. W. (2001). Antibiotic resistance of bacteria in biofilms. *Lancet*, *358*(9276), 135-138. https://doi.org/10.1016/s0140-6736(01)05321-1

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*, *49*(D1), D605-D612. https://doi.org/10.1093/nar/gkaa1074

Tamarit, D., Neuvonen, M. M., Engel, P., Guy, L., & Andersson, S. G. E. (2018). Origin and evolution of the *Bartonella* gene transfer agent. *Mol Biol Evol*, *35*(2), 451-464. https://doi.org/10.1093/molbev/msx299

Vidakovic, L., Singh, P. K., Hartmann, R., Nadell, C. D., & Drescher, K. (2018). Dynamic biofilm architecture confers individual and collective mechanisms of viral protection. *Nat Microbiol*, *3*(1), 26-31. https://doi.org/10.1038/s41564-017-0050-1

Vos, M. (2009). Why do bacteria engage in homologous recombination? *Trends Microbiol*, *17*(6), 226-232. https://doi.org/10.1016/j.tim.2009.03.001

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*(18), 9763-9767. https://doi.org/10.1073/pnas.94.18.9763

Westbye, A. B., Beatty, J. T., & Lang, A. S. (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. https://doi.org/10.1016/j.mib.2017.05.003

Westbye, A. B., O'Neill, Z., Schellenberg-Beaver, T., & Beatty, J. T. (2017). The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology (Reading)*, *163*(9), 1355-1363. https://doi.org/10.1099/mic.0.000519

Wilson, D. S. (1975). A theory of group selection. *Proc Natl Acad Sci U S A*, *72*(1), 143-146. https://doi.org/10.1073/pnas.72.1.143

Wong, E. H. J., Ng, C. G., Goh, K. L., Vadivelu, J., Ho, B., & Loke, M. F. (2018). Metabolomic analysis of low and high biofilm-forming *Helicobacter pylori* strains. *Sci Rep*, *8*(1), 1409. https://doi.org/10.1038/s41598-018-19697-0

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, *28*(7), 1033-1034. https://doi.org/10.1093/bioinformatics/bts079

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Yang, Y., Yue, Y., Song, N., Li, C., Yuan, Z., Wang, Y., Ma, Y., Li, H., Zhang, F., Wang, W., Jia, H., Li, P., Li, X., Wang, Q., Ding, Z., Dong, H., Gu, L., & Li, B. (2020).

The YdiU domain modulates bacterial stress signaling through Mn(2+)-dependent UMPylation. *Cell Rep*, *32*(12), 108161. https://doi.org/10.1016/j.celrep.2020.108161

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(Suppl 6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhang, W., Ding, W., Li, Y. X., Tam, C., Bougouffa, S., Wang, R., Pei, B., Chiang, H., Leung, P., Lu, Y., Sun, J., Fu, H., Bajic, V. B., Liu, H., Webster, N. S., & Qian, P. Y. (2019). Marine biofilms constitute a bank of hidden microbial diversity and functional potential. *Nat Commun*, *10*(1), 517. https://doi.org/10.1038/s41467-019-08463-z

Chapter 6

Formal recognition and classification of gene transfer agents as viriforms

Roman Kogay[1], Sonja Koppenhöfer[2], J. Thomas Beatty[3], Jens H. Kuhn[4], Andrew S. Lang[2], and Olga Zhaxybayeva[1]

[1]Department of Biological Sciences, Dartmouth College, Hanover, NH, USA
[2]Department of Biology, Memorial University of Newfoundland, St. John's, NL, Canada
[3]Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada
[4]Integrated Research Facility at Fort Detrick, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, MD, USA

Published in *Virus Evolution* on 15 October 2022
(DOI: 10.1093/ve/veac100)

Supplementary Material is available online (DOI: 10.1093/ve/veac100)

Author contributions

JTB, JHK, ASL and OZ designed the study. RK, and SK collected data and performed phylogenetic analyses. All authors contributed to the initial draft. RK, JTB, JHK, ASL, and OZ revised the manuscript. All authors finalized the manuscript.

Abstract

Morphological and genetic features strongly suggest that gene transfer agents (GTAs) are caudoviricete-derived entities that have evolved in concert with cellular genomes to such a degree that they should not be considered viruses. Indeed, GTA particles resemble caudoviricete virions but, in contrast to caudoviricetes (or any viruses), GTAs can encapsidate at best only part of their own genomes, are induced solely in small subpopulations of prokaryotic host cells and are transmitted vertically as part of cellular genomes during replication and division. Therefore, the lifecycles of GTAs are analogous to virus-derived entities found in parasitoid wasps, which have recently been recognized as non-virus entities and therefore reclassified as viriforms. We evaluated three distinct, independently exapted GTA groups for which the genetic basis for GTA particle production has been established. Based on the evidence, we outline a classification scheme for these viriforms.

Introduction

In 2021, the International Committee on Taxonomy of Viruses (ICTV) ratified a taxonomic proposal to formally accept a new operational definition of the term "virus" (Koonin et al., 2021; Kuhn et al., 2020; Walker et al., 2021). Consequently, the most current version of the International Code of Virus Classification and Nomenclature (ICVCN) states that viruses are

> " … a type of MGEs [mobile genetic elements] that encode at least one protein that is a major component of the virion encasing the nucleic acid of the respective MGE and therefore the gene encoding the major virion protein itself; or MGEs that are clearly demonstrable to be members of a line of evolutionary descent of such major virion protein-encoding entities" (ICVCN Rule 3.3) (International Committee on Taxonomy of Viruses, 2022; Kuhn et al., 2020).

This definition also formalized the postulate that some MGEs, long understood by the general virology community to be distinct from viruses, are indeed distinct. At the

time, the ICTV had already classified viroids and satellite nucleic acids in taxa separate from viral taxa (in families/genera with names that end with suffixes -viroidae/-viroid and -satellitidae/-satellite, respectively, as opposed to -viridae/-virus) (International Committee on Taxonomy of Viruses, 2022), and these elements were logically placed into the perivirosphere rather than the orthovirosphere (Koonin et al., 2021; Kuhn et al., 2020).

The adoption of the new virus definition brought into question the taxonomic standing of one official virus family, Polydnaviridae. Indeed, entities classified into this polyphyletic family fundamentally deviate from MGEs fulfilling the virus definition because "polydna" particles encapsidate multiple segments of circular double-stranded DNAs that, however, do not encode the entire "polydna" genomes. Instead, the genomes are permanently endogenized into the "polydna" host (i.e., parasitoid wasp) genomes and inherited vertically. The resultant non-mobile nonviral entities are used by the wasps to deliver immunomodulatory genes into insects that serve as prey for the wasps (Drezen et al., 2017; Herniou et al., 2013). "Polydna" entities are likely evolutionarily derived from various groups of insect viruses, including nudivirids (Darboux et al., 2019; Drezen et al., 2017; Gauthier et al., 2018; Petersen et al., 2022; Strand & Burke, 2020; Theze et al., 2011), but, because they have lost the ability to replicate and instead have been fully exapted by their wasp hosts, they have left the virosphere altogether (Koonin et al., 2021; Koonin & Krupovic, 2018). Consequently, in 2021, the ICTV recognized "polydna" entities as representatives of a new MGE category distinct from viruses called "viriforms" (Koonin et al., 2021; Kuhn et al., 2020; Walker et al., 2021), and reclassified *Polydnaviridae* as (still polyphyletic) *Polydnaviriformidae* (Kuhn et al., 2021; Walker et al., 2022). In the ICVCN, viriforms are defined operationally as

> " … a type of virus-derived MGEs that have been exapted by their organismal (cellular) hosts to fulfill functions important for the host life cycle; or MGEs that are derived from such entities in the course of evolution" (ICVCN Rule 3.3) (International Committee on Taxonomy of Viruses, 2022; Kuhn et al., 2020).

Importantly, the following comment was added to the Rule 3.3:

> "Gene transfer agents (GTAs) and the MGEs previously classified in the family *Polydnaviridae* are considered to be viriforms in classification and nomenclature" (International Committee on Taxonomy of Viruses, 2022; Kuhn et al., 2020).

Notably, there are no discernible evolutionary relationships between GTAs and polydnaviriformids. The term "viriform", similar to the term "virus", is an umbrella term for certain MGEs with comparable lifecycles and properties; it is currently applied to six realms of MGEs that are not evolutionary related to each other.

Based on the properties of entities referred to as "GTAs" in the literature (reviewed in (Lang et al., 2017; Lang et al., 2012) we define GTAs as viriforms with the following features:

1. GTAs use caudoviricete ancestor-derived proteins (established either via significant similarity of at least some GTA proteins to caudeviricete proteins or by image-based evidence of caudovirion-like particles) to form caudovirion-like particles;
2. GTAs encapsidate mostly random pieces of host DNA (established experimentally);
3. GTA genomes are fully endogenized in host genomes, often across multiple loci (established experimentally and via genomic examination);
4. GTA genomes are not/cannot be fully packaged into particles due to limited particle head size (established via comparison of the packaged DNA length and size of GTA loci);
5. GTA genomes are mostly vertically inherited and GTAs co-diversify with their hosts (established via congruence between phylogenies of host and GTA genes); and
6. DNA encapsidated in GTA particles is delivered to other cells (established experimentally).

Having these attributes, GTAs have lost the ability to replicate and have become fully exapted by their cellular hosts. They are produced under specific conditions (e.g., nutrient

depletion (Westbye, O'Neill, et al., 2017)) and mediate horizontal gene transfer (HGT), typically among cells of the same species.

The first GTA discovered, of the alphaproteobacterium *Rhodobacter capsulatus,* was described in 1974 by Barry Marrs (Marrs, 1974). Since that time, distinct functional GTAs have been described in other alphaproteobacteria, a sulfate-reducing deltaproteobacterium, a methanogenic archaeon, and a spirochete that infects domestic pigs (Bertani, 1999; Guy et al., 2013; Humphrey et al., 1997; Rapp & Wall, 1987). Additionally, clusters of genes homologous to those encoding the *R. capsulatus* GTA are found in many alphaproteobacterial genomes, suggesting a wider prevalence of GTA production than presently appreciated (Kogay et al., 2019; Lang & Beatty, 2007; Lang et al., 2002; Shakya et al., 2017). Indeed, some of these bacteria produce functional GTAs (Biers et al., 2008; Nagao et al., 2015; Tomasch et al., 2018).

The recent ICTV recognition of viriforms and the formal establishment of *Polydnaviriformidae* provides an opportunity to initiate a systematic classification of GTAs. Here we outline initial steps to establish such a formal taxonomic scheme for GTA viriforms, focusing specifically on GTAs experimentally documented as being produced by cells and performing gene transfer—and for which the genetic basis of particle production has been established. Simultaneously, we have also officially proposed this taxonomic scheme to the ICTV for the 2022–2023 proposal cycle.

Nomenclature of Gene Transfer Agents and Associated Taxa

Per ICTV rules, virus names are written in lower case (except if a name component is a proper noun), without italics in any part of the name (even if a host species name is part of the name), and ending in the term “virus”, which in virus name abbreviations is “V”. Examples are measles virus (MeV) and Ebola virus (EBOV). The nomenclature of already classified viriforms (polydnaviriformids) follows these rules, with “virus” being replaced by “viriform” and the abbreviation “V” being replaced with “Vf” (e.g., “Glyptapanteles liparidis bracoviriform” is abbreviated “GlBVf”). We suggest applying these general rules to GTAs, but with “viriform” being replaced by “gene transfer agent” due to the long-established use of this phrase and “Vf” being replaced

with “GTA”. Therefore, the gene transfer agent produced by *Rhodobacter capsulatus* would be called “*Rhodobacter capsulatus* gene transfer agent” and abbreviated as “RcGTA”, consistent with the established use of this abbreviation in the literature.

Rules for viriform taxon naming have been established by the ICVCN. Specifically,

> “[t]he formal endings for taxon names of viriforms are the suffixes "-*viriformia*" for realms, "-*viriforma*" for subrealms, "-*viriformae*" for kingdoms, "-*viriformites*" for subkingdoms, "-*viriformicota*" for phyla, "-*viriformicotina*" for subphyla, "-*viriformicetes*" for classes, "-*viriformicetidae*" for subclasses, "-*viriformales*" for orders, "-*viriformineae*" for suborders, "-*viriformidae*" for families, "-*viriforminae*" for subfamilies, and "-*viriform*" for genera and subgenera” (ICVCN Rule 3.26) (International Committee on Taxonomy of Viruses, 2022; Kuhn et al., 2020)

and

> “[a] species name shall consist of only two distinct word components separated by a space. The first word component shall begin with a capital letter and be identical in spelling to the name of the genus to which the species belongs. The second word component shall not contain any suffixes specific for taxa of higher ranks. The entire species name (both word components) shall be italicized” (International Committee on Taxonomy of Viruses, 2022).

We suggest adding the infix -gta- prior to the taxon-specific suffixes for immediate recognition of GTA-specific taxa (e.g., -*gtaviriform*).

Gene Transfer Agents can be Assigned to at Least Three Major Clades

Based on functionally and genetically characterized GTAs, at least three major GTA clades can be delineated.

Alphaproteobacterial type I GTAs

The best characterized GTA of this clade is RcGTA, produced by *R. capsulatus* (*Pseudomonadota*: *Alphaproteobacteria*: *Rhodobacterales*: *Rhodobacteraceae*). We designate RcGTA here as the founding member of one major GTA clade, the alphaproteobacterial type I GTAs. For many years since its discovery (Marrs, 1974), RcGTA was the only known GTA. Now we know that homologous GTAs are produced by other bacteria from the order *Rhodobacterales*: *Dinoroseobacter shibae* (*Dinoroseobacter shibae* gene transfer agent [DsGTA]) (Tomasch et al., 2018), *Ruegeria pomeroyi* (*Ruegeria pomeroyi* gene transfer agent [RpGTA]) (Biers et al., 2008), and *Rhodovulum sulfidophilum* (*Rhodovulum sulfidophilum* gene transfer agent [RsGTA]) (Nagao et al., 2015). Additionally, genes encoding RcGTA-like GTAs are conserved in most genomes in the order *Rhodobacterales* and in many genomes of the alphaproteobacterial orders *Caulobacterales*, *Sphingomonadales*, *Parvibaculales*, and *Hyphomicrobiales* (formerly *Rhizobiales*) (Kogay et al., 2019; Lang & Beatty, 2007; Lang et al., 2002; Shakya et al., 2017).

RcGTA and RcGTA-like GTA genes are similar in sequence to those of viruses classified in the uroviricot class *Caudoviricetes* (*Duplodnaviria*: *Heunggongvirae*) (Shakya et al., 2017). These GTAs are transmitted vertically from a bacterial parent to progeny during cell division (Lang & Beatty, 2007; Shakya et al., 2017), similar to propagation of temperate viruses ("prophages"). However, in contrast to temperate virus genomes, the set of genes required for production of the GTA particle (the GTA "genome") is not necessarily localized in one region of the host genome. In the case of RcGTA, known structural and regulatory genes are scattered across five loci in the *R. capsulatus* genome (Hynes et al., 2016), cumulatively spanning approximately 20 kilobases (kb) (**Figure 1** and **Supplementary Table S1**). Moreover, cellular regulatory genes are involved in controlling GTA particle production (Westbye, Beatty, et al., 2017), adding another factor that makes the GTA genome difficult to differentiate from its host's genome.

RcGTA particles resemble virions of caudoviricetes (Yen et al., 1979) and have been structurally characterized at high resolution (Bardy et al., 2020). RcGTA particles

have head diameters of 38 nm and tail lengths of 49 nm. A small percentage of RcGTA particles have T = 3 quasi-icosahedral heads, but the capsid shape of most particles is oblate, as they lack the five hexamers of capsid protein needed to form genuine icosahedral heads. Because of the small head size, RcGTA particles can only package double-stranded DNA of approximately 4 kb in length (Yen et al., 1979). The DNA is also encapsidated at 10–25% lower density than typical caudoviricetes (Bardy et al., 2020). Both RcGTA particle production and acquisition of the GTA-packaged DNA by other host cells in the population are controlled by the same cellular regulatory systems (Westbye, Beatty, et al., 2017). Only 0.1–3.0% of cells produce GTA particles (P. C. M. Fogg et al., 2012; Hynes et al., 2012), whereas the remaining cells produce a GTA receptor (Brimacombe et al., 2013).

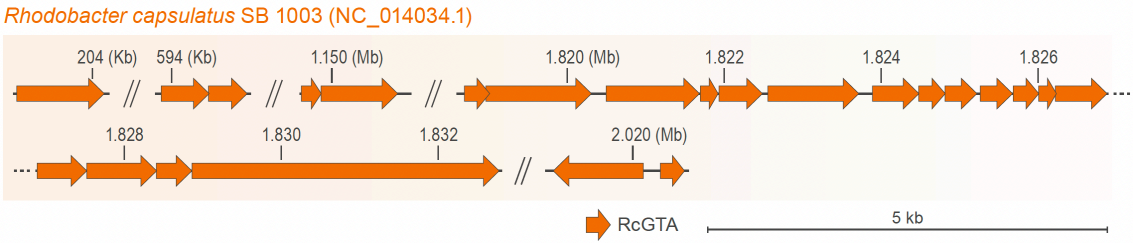

Figure 1. Genome of *Rhodobacter capsulatus* gene transfer agent (RcGTA). Genes (arrows) are depicted to scale, in their locations in the host genome (*R. capsulatus*). Exact coordinates of the RcGTA genes, their locus tags, and their functional annotations are listed in **Supplementary Table S1**.

Compositionally, structural proteins encoded by RcGTA and RcGTA-like GTAs are biased towards amino acids that are energetically cheaper to produce (Kogay et al., 2020). To date, such a bias has not yet been associated with viruses. Based on this difference in amino-acid composition, GTA proteins can be distinguished from their viral homologs using a machine-learning approach, which is implemented in the publicly available GTA-Hunter program (Kogay et al., 2019).

In a comprehensive evolutionary analysis of homologs of the large subunit of the DNA packaging terminase enzyme (TerL, encoded by the g2 gene in the RcGTA genome), RcGTA and RcGTA-like GTAs form a clade closely related to, but distinct

from, duplodnavirians (Esterman et al., 2021). To illustrate the relationships of alphaproteobacterial type I GTAs to each other and to their closest viral homologs, we reconstructed evolutionary histories of their TerL proteins and the HK97-like major capsid proteins (HK97-MCP, encoded by the g5 gene in the GTA genome, is the hallmark protein that defines the virus realm *Duplodnaviria* (Koonin et al., 2020)). Consistent with an earlier analysis (Esterman et al., 2021), RcGTA and RcGTA-like GTAs formed a clade closely related to, but distinct from, caudoviricetes (**Figure 2**), with a few exceptions that are likely artefacts of phylogenetic reconstruction.

Specifically, in the TerL phylogeny (**Figure 2A**), all viral homologs except one (Caulobacter virus Sansa) are separated from GTA proteins (with a solid bootstrap support of 81%). Caulobacter virus Sansa groups with one GTA sequence from a bacterium of the order *Sphingomonadales* (with a low bootstrap support of 50%), whereas all other GTAs of *Sphingomonadales* bacteria group together (with a strong bootstrap support of 96%). We hypothesize that the phylogenetic placement of the Caulobacter virus Sansa TerL is due to the long-branch attraction artefact (Felsenstein, 1978). We searched for a maximum-likelihood tree in which caudoviricete- and GTA-derived TerLs were required to group separately from each other and compared that tree to the tree depicted in **Figure 2A**. We found that the likelihoods of the two trees are not significantly different (approximately unbiased [AU] test; p-value = 0.555), confirming that the placement of the Caulobacter virus Sansa sequence within the GTA sequences is unreliable.

In the HK97-MCP phylogeny (**Figure 2B**), GTAs and most caudoviricetes are separated by a branch with 63% bootstrap support. Several caudoviricetes that group within GTAs are located on long branches, are situated outside of well-supported groups of GTAs from several alphaproteobacterial orders and have very low bootstrap support for their placements. It is therefore likely that the positions of these viral homologs are unreliable. To test this hypothesis, we identified a maximum-likelihood phylogeny among trees in which GTAs and caudoviricetes were required to be separated by a branch. The likelihoods of this tree and the phylogeny shown in **Figure 2B** are not significantly different (AU test; p-value = 0.534). Therefore, these viruses are likely positioned in different places in trees reconstructed from different bootstrap replicates, which would

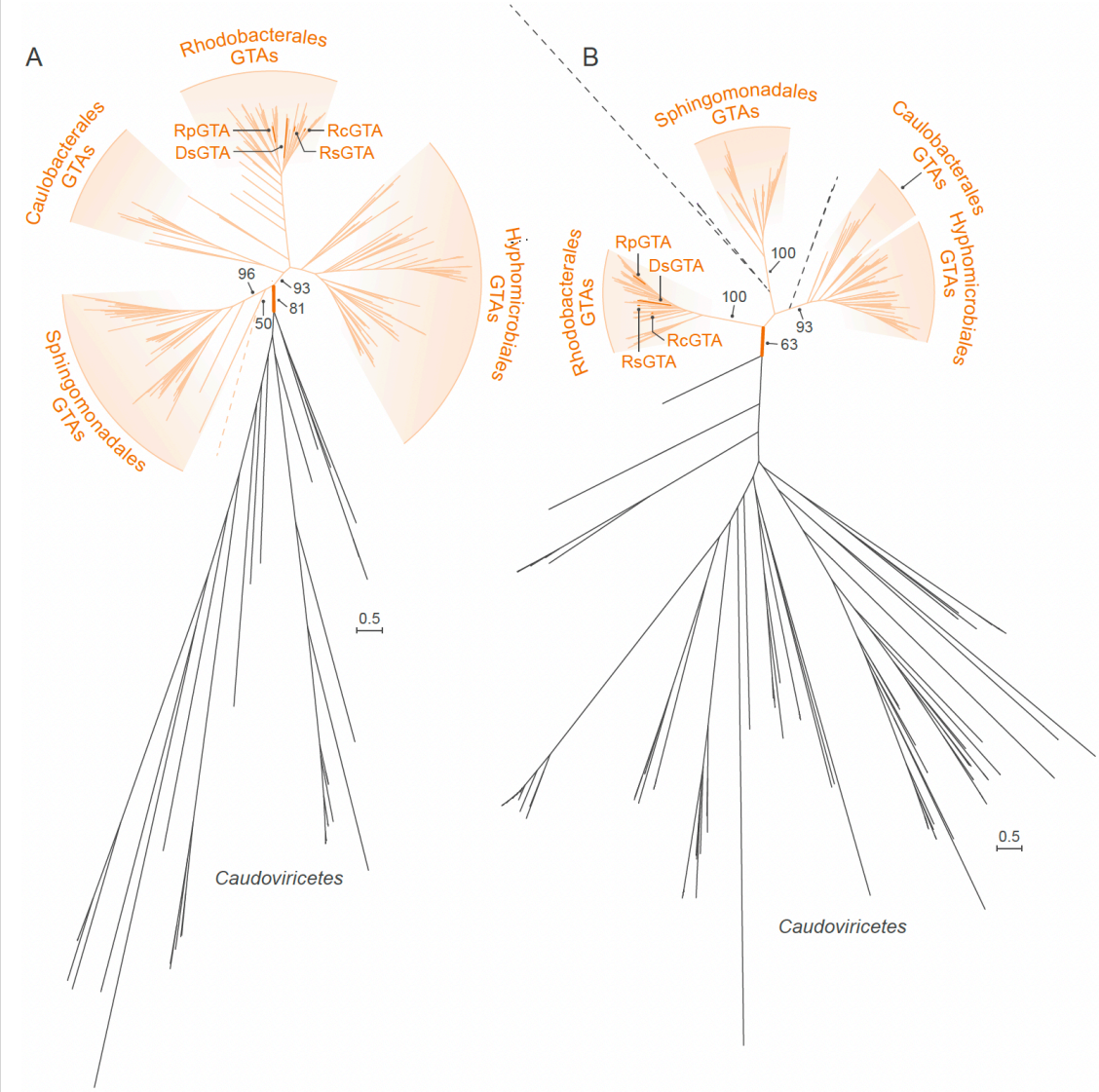

Figure 2. Maximum Likelihood phylogenies of (A) large terminase (TerL) subunits and (B) HK97 major capsid protein (HK97-MCP) sequences of rhodogtaviriformids and their closest known caudoviricete homologs.

Alphaproteobacterial type I gene transfer agent (GTA) (rhodogtaviriformid) lineages are shown in orange. Caudoviricete lineages that are nested within GTA lineages are shown in dashed black lines. Other caudoviricete lineages are shown in solid black lines. Bootstrap support values are shown only for selected branches. Scale bars represent substitutions per site. DsGTA, *Dinoroseobacter shibae* gene transfer agent; GTA, gene transfer agent; RcGTA, *Rhodobacter capsulatus* gene transfer agent; RpGTA, *Ruegeria pomeroyi* gene transfer agent; RsGTA, *Rhodovolum sulfidophilum*, gene transfer agent.

lead to their artificial (and poorly supported) basal positions with the GTA homologs on the tree shown in **Figure 2B.**

In the **Figure 2** trees, GTA branches have shorter lengths than their caudoviricete counterparts, conforming with the reported slower evolutionary rate of GTAs compared to viruses (Shakya et al., 2017). Additionally, on both phylogenetic trees, GTAs from alphaproteobacteria of different orders form separate groups with very high support, corroborating vertical inheritance of most GTA genes (Lang & Beatty, 2007; Lang et al., 2002; Shakya et al., 2017).

Together, these results justify the classification of RcGTA and three RcGTA-like GTAs in a common viriform taxon: family *Rhodogtaviriformidae* (from *Rhodobacterales*, infix -*gta*-, and family-specific suffix -*viriformidae*). Given limited dataset size (i.e., just four GTAs), it is challenging to establish quantifiable criteria for demarcating taxonomic relationships among the four GTAs. In the future, when more GTA sequences become available for analyses, a criterion based on percent sequence similarity among shared genes should be considered. For now, based on the evidence of co-evolution of these GTAs and their specific hosts, we argue that at least four rhodogtaviriformid genera, each for GTAs of bacteria classified in distinct genera included in *Rhodobacterales*, ought to be established:

- *Dinogtaviriform* (named after DsGTA host genus *Dinoroseobacter*, infix -gta-, and genus-specific suffix -*viriform*) to include one new species, *Dinogtaviriform tomaschi* (species epithet to honor GTA researcher Jürgen Tomasch, who was instrumental in the discovery of DsGTA) for DsGTA (**Supplementary Table S2**);
- *Rhodobactegtaviriform* (named after RcGTA host genus *Rhodobacter*, infix -*gta*-, and genus-specific suffix -*viriform*) to include one new species, *Rhodobactegtaviriform marrsi* (species epithet to honor GTA researcher Barry Marrs, who first discovered GTAs and coined the term “gene transfer agent”) for RcGTA (**Supplementary Table S1**);
- *Rhodovulugtaviriform* (named after RsGTA host genus *Rhodovulum*, infix -*gta*-, and genus-specific suffix -*viriform*) to include one new species, *Rhodovulugtaviriform kikuchii* (species epithet to honor GTA researcher Yo

Kikuchi, who was instrumental in the discovery of RsGTA) for RsGTA (**Supplementary Table S3**); and

- *Ruegerigtaviriform* (named after RpGTA host genus *Ruegeria*, infix -*gta*-, and genus-specific suffix -*viriform*) to include one new species, *Ruegerigtaviriform cheni* (species epithet to honor GTA researcher Feng Chen, who was instrumental in the discovery of RpGTA) for RpGTA (**Supplementary Table S4**).

Alphaproteobacterial type II GTAs

There was a lag between discovery of these elements and their recognitions as bona fide GTAs. Phage-like particles, originally referred to as bacteriophage-like particles (BLPs), that contained heterogenous DNA from *Bartonella* host genomes were first characterized in *Bartonella henselae* (Anderson et al., 1994), and noted to be similar in structure to particles produced by B. bacilliformis (Umemori et al., 1992). These B. bacilliformis particles were subsequently shown to also contain heterogeneous genomic DNA fragments, but attempts to demonstrate their gene transfer ability were not successful (Barbian & Minnick, 2000). Functionality of the particles produced by *Bartonella* for gene transfer (*Bartonella* gene transfer agent [BaGTA]) was eventually demonstrated by work on *B. henselae* (*Pseudomonadota*: *Alphaproteobacteria*: *Hyphomicrobiales*: *Bartonellaceae*) (Guy et al., 2013). BaGTA genes were initially proposed to be located within a single cluster of 11–13 genes spanning approximately 14 kb (Guy et al., 2013). However, a subsequent screen for genes essential for BaGTA functionality identified a total of 29 genes located within a larger (approximately 79-kb-long) region (Québatte et al., 2017) (**Figure 3** and **Supplementary Table S5**). Homologs of BaGTA genes (BaGTA-like GTAs) were found in the genomes of multiple species of *Bartonella* (Berglund et al., 2009; Guy et al., 2013; Tamarit et al., 2018). BaGTA genes are located near an active virus-derived origin of replication and next to genes encoding secretion systems (Guy et al., 2013). As a result, the region of the genome containing BaGTA and these secretion-system genes are amplified and packaged more often than other genomic regions (Guy et al., 2013; Québatte et al., 2017). These findings led to the hypothesis that BaGTA and BaGTA-like GTAs have been maintained due to their mediation of HGT of secretion-system and toxin genes, thereby enabling *Bartonella*

bacteria to adapt to diverse hosts (Guy et al., 2013). However, actual GTA-mediated DNA transfer among bacterial cells has only been demonstrated for *B. henselae* (Guy et al., 2013). There, BaGTA production is restricted to a distinct subpopulation of fast-growing cells, which comprise about 6% of the total population (Québatte et al., 2017), and the uptake of BaGTA-packaged DNA was proposed to be limited to cells undergoing division (Québatte et al., 2017).

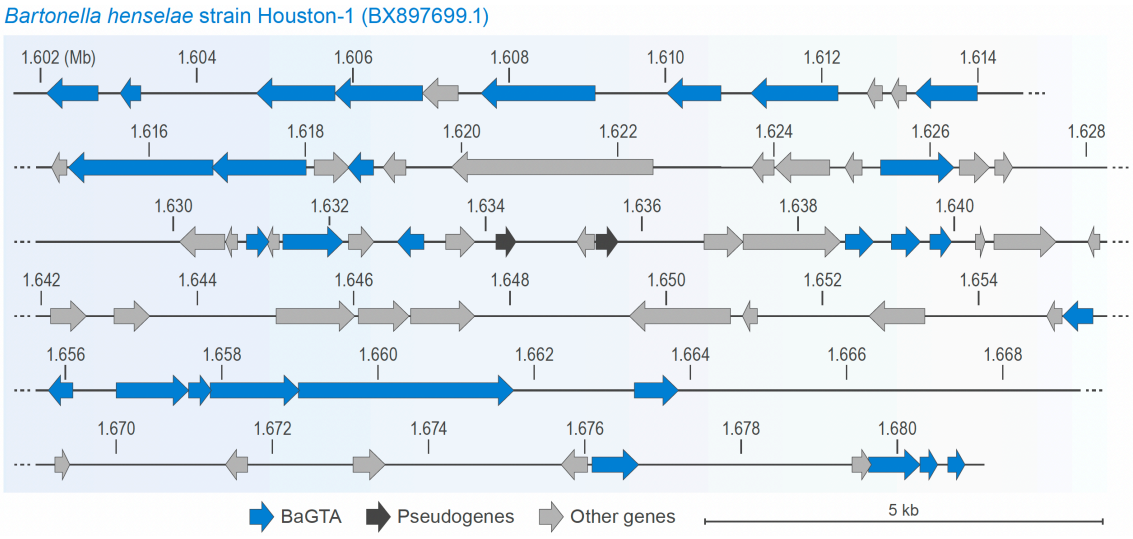

Figure 3. Genome of *Bartonella* gene transfer agent (BaGTA). Genes (arrows) are depicted to scale, in their locations in the host genome (*B. henselae*). Exact coordinates of the BaGTA genes, their locus tags, and their functional annotations are listed in **Supplementary Table S5**.

There are some discrepancies in the literature regarding the structure of BaGTA particles, suggesting some bacteria might release additional phage-like particles. The *B. henselae* particles were originally reported as particles without tails or with short non-contractile tails with a head diameter of 40 nm (Anderson et al., 1994). The head diameter of the B. bacilliformis particles was originally measured at 40 nm (Umemori et al., 1992) and subsequently 80 nm (Barbian & Minnick, 2000). Those of *B. grahamii* were reported as possessing long non-contractile tails and icosahedral heads of 50–70 nm or 80 nm and tails of 100 nm (Berglund et al., 2009). Although BaGTA particles are potentially able to package the entire main structural gene cluster of 11–13 genes, they

cannot package all 29 genes required for BaGTA production due to a capacity of 14 kb (Anderson et al., 1994; Guy et al., 2013; Lang et al., 2017).

In the TerL phylogeny, BaGTA-like homologs are separated from almost all caudoviricetes by longer branches (with 100% bootstrap support; **Figure 4A**). Two caudoviricete homologs (Sulfitobacter phage pCB2047-C and Sulfitobacter phage NYA-2014a) group together and are nested within the BaGTA-like group (with 84% bootstrap support). We hypothesize that the terL gene was horizontally transferred from GTAs to these caudoviricetes, with similar HGT events documented between RcGTA-like GTAs and caudoviricetes infecting bacteria of the *Rhodobacterales* (Zhan et al., 2016). In the HK97-MCP phylogeny, BaGTA homologs are located on shorter branches than their caudoviricete counterparts and are separated from caudoviricete homologs with 100% bootstrap support (**Figure 4B**). Phylogenomic analyses suggest that *Bartonella* GTAs have co-evolved with their hosts (Tamarit et al., 2018).

Together, these results justify the classification of BaGTA and BaGTA-like GTAs in a common viriform taxon, family *Bartogtaviriformidae* (from *Bartonella*, infix -*gta*-, and family-specific suffix -*viriformidae*). For now, we argue that at least one bartogtaviriformid genus ought to be established: *Bartonegtaviriform* (named after BaGTA host genus *Bartonella*, infix -*gta*-, and genus-specific suffix -*viriform*) including one new species, *Bartonegtaviriform andersoni* (species epithet to honor GTA researcher Burt Anderson, who first discovered BaGTA particles (Anderson et al., 1994)) for BaGTA.

GTAs of spirochaetes

A GTA originally called virus of *Serpulina hyodysenteriae* 1 (VSH-1) was identified in *Brachyspira* (formerly *Serpulina*) *hyodysenteriae* (*Spirochaetota*: *Spirochaetia*: *Brachyspirales*: *Brachyspiraceae*) (Humphrey et al., 1997). In accordance with the nomenclature rules established here, we suggest renaming this GTA to *Brachyspira hyodysenteriae* gene transfer agent (BhGTA). The structural gene cluster responsible for production of BhGTA particles—i.e., the BhGTA "genome"—is 16.3 kb in length (Matson et al., 2005) (**Figure 5** and **Supplementary Table S6**)

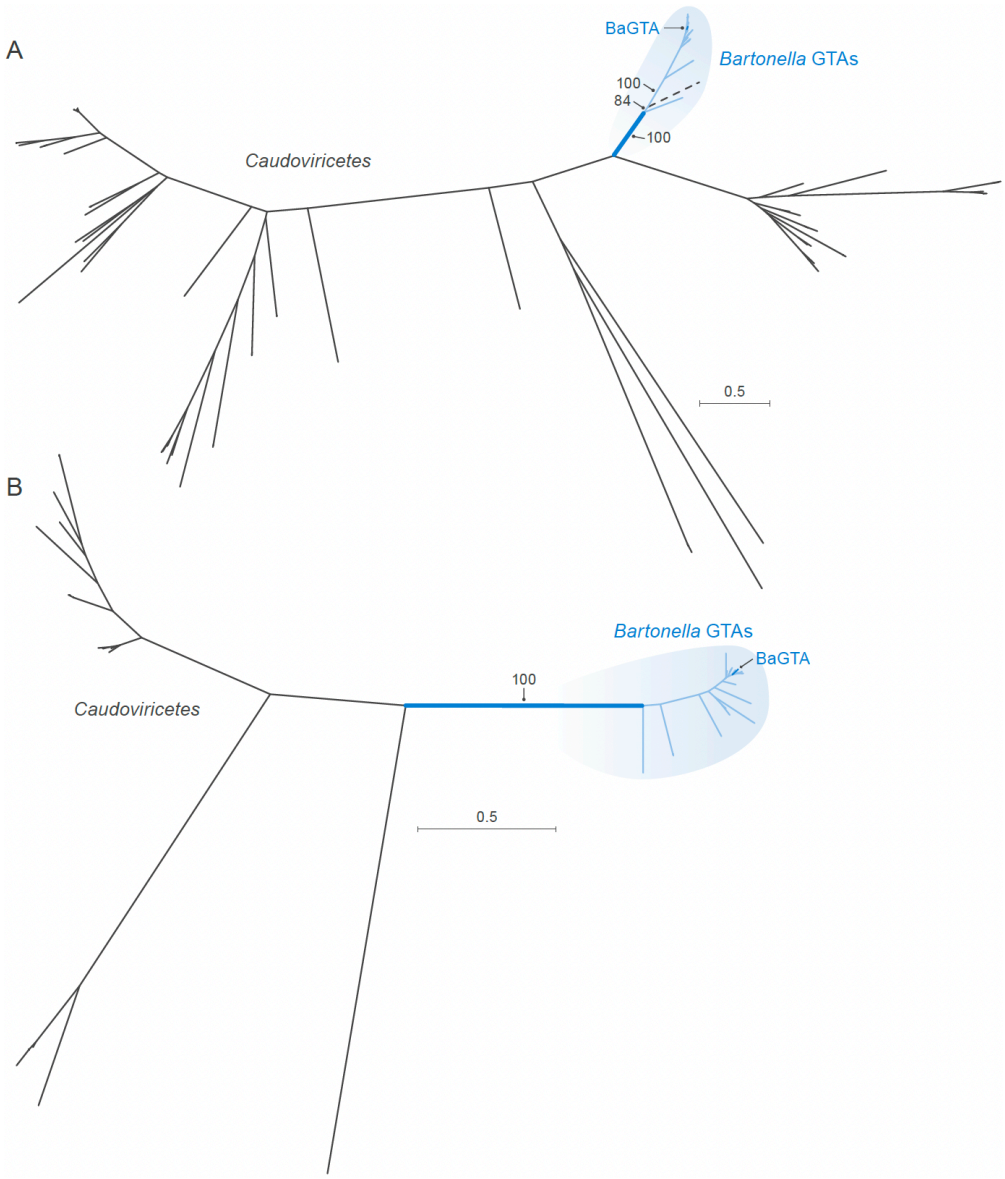

Figure 4. Maximum Likelihood phylogenies of (A) large terminase (TerL) subunits and (B) HK97 major capsid protein (HK97-MCP) sequences of bartogtaviriformids and their closest known caudoviricete homologs.
Alphaproteobacterial type II gene transfer agent (GTA) (bartogtaviriformid) lineages are shown in blue. Caudoviricete lineages are shown in black. Two nearly identical caudoviricete lineages that are nested within GTA lineages are shown in dashed black lines. A bootstrap support value is shown only for the branch separating GTA and caudoviricete sequences. Scale bars indicate substitutions per site. BaGTA, *Bartonella* gene transfer agent; GTA, gene transfer agent.

BhGTA particles have a head diameter of 45 nm and a flexible non-contractile tail of 65 nm (Humphrey et al., 1997). Like other GTAs, BhGTA is unable to package and transfer its entire genome, given the limiting capacity of 7.5 kb (Humphrey et al., 1997; Matson et al., 2005). Restriction enzyme digests of the packaged DNA and the range of marker genes that can be transferred by BhGTA particles suggest that they package any region of the B. hyodysenteriae genome (Humphrey et al., 1997) without an obvious bias for the genomic region that encodes BhGTA. The induction of BhGTA particle production by DNA-damaging agents, such as mitomycin C and antibiotics, results in large-scale lysis of cells (Stanton et al., 2008). However, the proportion of *B. hyodysenteriae* cells in a population that naturally produce and release BhGTA particles has not been quantified. BhGTA particles are capable of transferring antimicrobial resistance genes within the bacterial population (Stanton et al., 2008), pointing at possible selective advantages of maintaining the capability of BhGTA particle production.

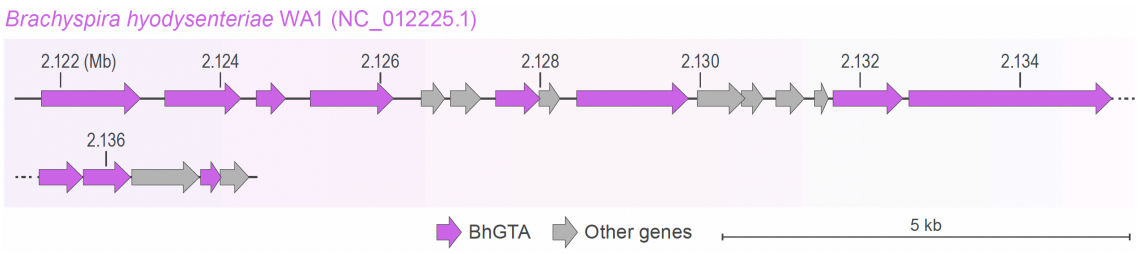

Figure 5. Genome of *Brachyspira hyodysenteriae* gene transfer agent (BhGTA). Genes (arrows) are depicted to scale, in their locations in the host genome (*B. hyodysenteriae*). Exact coordinates of the BhGTA genes, their locus tags, and their functional annotations are listed in **Supplementary Table S6**.

Homologs of genes in the BhGTA genome were found in the genomes of other members of the genus *Brachyspira*, but there is no gene synteny in their organization (Motro et al., 2009). Unlike in rhodogtaviriformids and bartogtaviriformids, an endolysin-encoding gene is the only gene in the BhGTA genome that has a significant sequence similarity to caudoviricete genes in the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database (accessed in May 2022). Some genes encoding the BhGTA particle proteins were experimentally validated

(including endolysin), and the particles structurally resemble those of caudoviricetes (Matson et al., 2005). Therefore, the absence of their homologs in the viral RefSeq database is likely due to the limited sampling of the virosphere.

In the endolysin phylogeny, the *Brachyspira* homologs group together and are separated from all caudoviricetes by a long branch (with 100% bootstrap support; **Figure 6A**). Additionally, the *B. hyodysenteriae* genome encodes a single copy of an identifiable terL gene, which is located outside of the currently delineated BhGTA genome. Homologs of this terL gene are also present in a single copy in genomes of other *Brachyspira* bacteria that encode BhGTA-like MCPs. These homologs are highly conserved, with pairwise amino-acid identities of 81–100%. In a phylogenetic tree, the *Brachyspira* TerLs are separated from all caudoviricete TerLs by a longer branch (with 100% bootstrap support; **Figure 6B**). Although the role of this TerL homolog in the BhGTA lifecycle has not been experimentally validated, the presence of the encoding gene as the only identifiable terL in the *Brachyspira* genomes, its high degree of conservation within the Brachyspira genus and its divergence from the related caudoviricete sequences support its potential involvement in the packaging of DNA into the BhGTA particles. Based on comparison of *Brachyspira* GTA and host genes, GTAs have co-diversified with *Brachyspira* (Motro et al., 2009). Together, these results justify the classification of BhGTA and BhGTA-like GTAs in a common viriform taxon, family *Brachygtaviriformidae* (form *Brachyspira*, infix -*gta*-, and family-specific suffix -*viriformidae*). For now, we argue that at least one brachygtaviriformid genus ought to be established: *Brachyspigtaviriform* (named after BhGTA host genus Brachyspira, infix -*gta*-, and genus-specific suffix -*viriform*) to include one new species, *Brachyspigtaviriform stantoni* (species epithet to honor GTA researcher Thaddeus Stanton, who first discovered BhGTA particles (Humphrey et al., 1997)) for BhGTA.

Independent origins of the three GTAs

Genes from the genomes of these three GTAs are either not homologous or too divergent to have significant sequence similarity in BLASTP searches of the encoded proteins. For example, pairwise amino-acid identity of TerLs, which is one of the most conserved GTA and caudoviricete proteins, is 14–20% among RcGTA, BaGTA, and

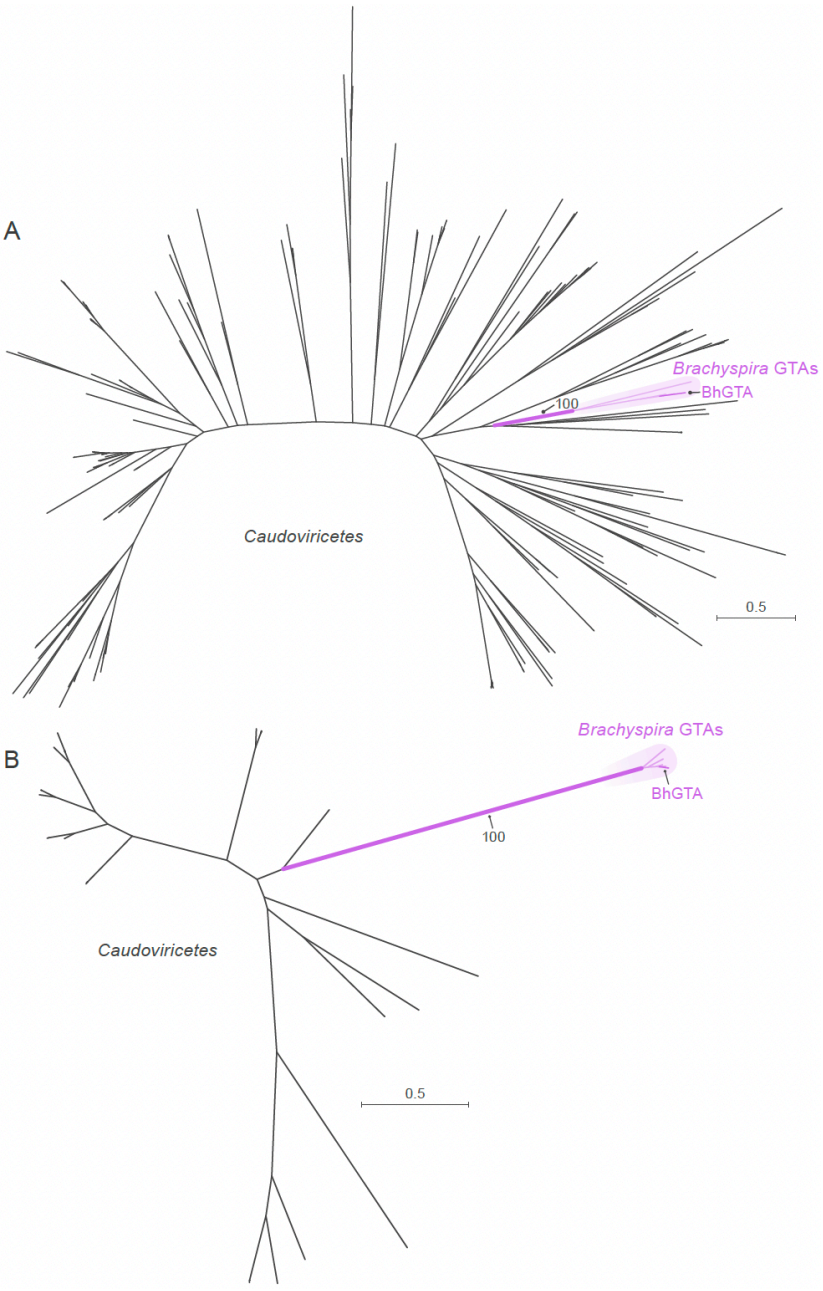

Figure 6. Maximum Likelihood phylogenies of (A) endolysin and (B) the putative large terminase (TerL) subunits of brachygtaviriformids and their closest known caudoviricete homologs. *Brachyspira* gene transfer agent (GTA) lineages are shown in purple. Caudoviricete lineages are shown in black. A bootstrap support value is shown only for the branch separating GTA and caudoviricete sequences. Scale bar indicates substitutions per site. BhGTA, *Brachyspira hyodysenteriae* gene transfer agent; GTA, gene transfer agent.

BhGTA. Nevertheless, an iterative clustering-alignment-phylogeny procedure (Wolf et al., 2018) established the homology among known TerL proteins that include RcGTA, BaGTA, and putative BhGTA TerLs (Esterman et al., 2021). The evolutionary history of RcGTA-like, BaGTA-like, putative BhGTA-like TerLs, and their closest known caudoviricete homologs (**Figure 7**) demonstrates that GTA-like TerLs appear in three distinct clades within viral TerLs. Based on this phylogenetic evidence, we propose that these three GTA clades are a result of three independent exaptation events. Therefore, just like viruses (which are classified in at least six unrelated realms), GTA viriforms are polyphyletic.

Discussion

Based on the evolutionary differences between GTA and caudoviricete genes encoding well-conserved proteins and on morphological differences of GTA particles, we propose three families for these GTAs. The greatest number of functionally confirmed and putative GTAs are in the alphaproteobacterial type I clade, which, for now, is proposed to be a family *Rhodogtaviriformidae* that includes at least four genera. The members of this family are currently restricted to a single cellular order (*Rhodobacterales*). The TerLs and MCPs of these RcGTA-like GTAs and alphaproteobacterial type II GTAs (*Bartogtaviriformidae*) are clearly distinguishable from each other and their caudoviricete homologs and evolve at a slower rate (**Figure 2** and **Figure 4**) (Esterman et al., 2021; Shakya et al., 2017). The spirochaete GTAs (*Brachygtaviriformidae*) are more difficult to distinguish from caudoviricetes due to a lack of available viral representatives in GenBank for all but one experimentally validated BhGTA gene. Nevertheless, both the experimentally validated BhGTA endolysin and the putative BhGTA TerL and their *Brachyspira* homologs also form a well-supported cluster distinct from caudoviricete lineages; moreover, brachygtaviriformid TerLs evolve at a slower rate than their spirochete homologs (**Figure 6B**). As in the case with the experimentally validated RcGTA, the “genome” of BhGTA is also likely dispersed across multiple loci.

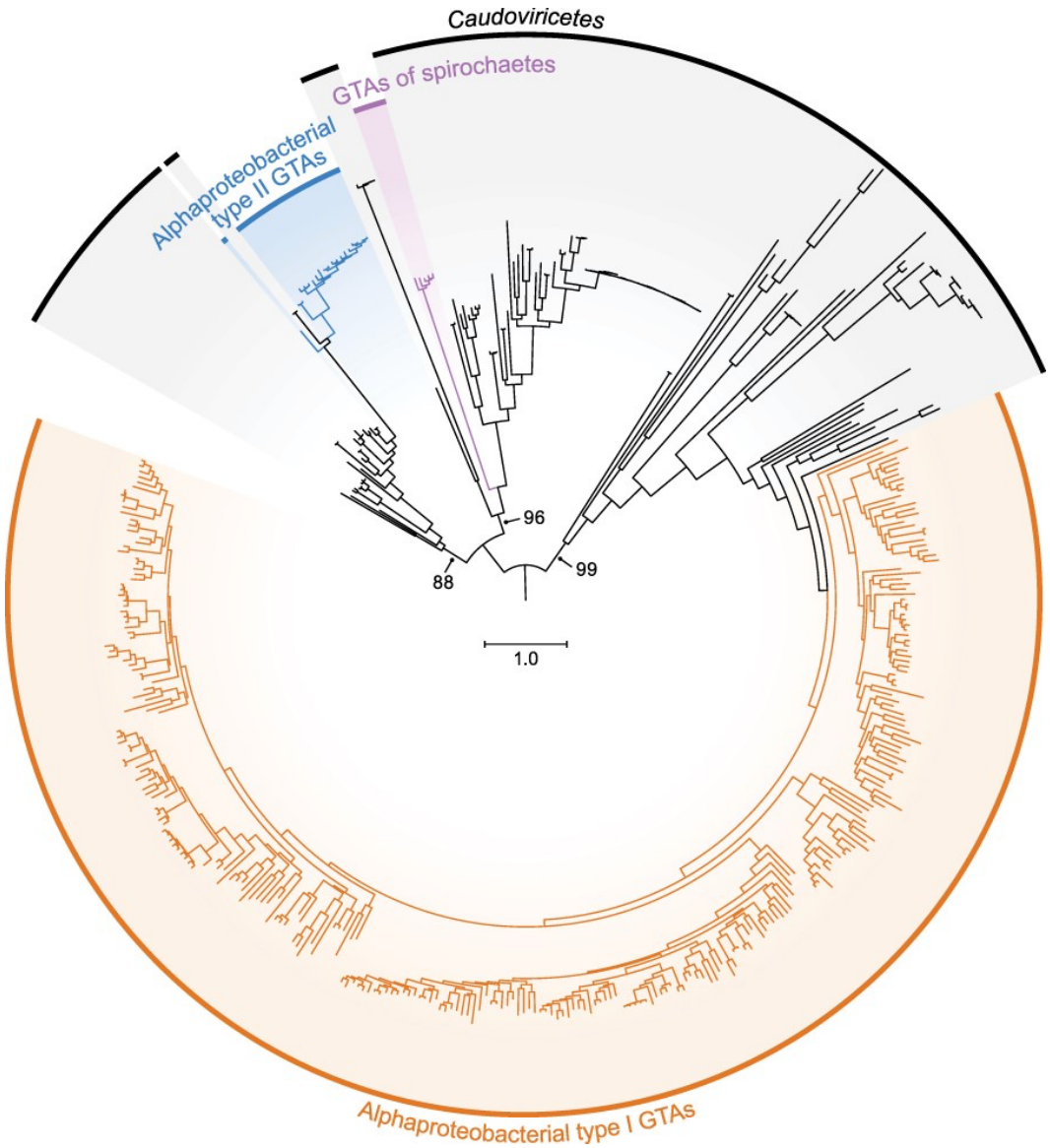

Figure 7. Maximum Likelihood phylogeny of the large terminase (TerL) subunits of three major clades of GTAs and their closest known caudoviricete homologs. This tree includes all TerL homologs from Figures 2A, 4A, and 6A, using their color coding. Bootstrap support values are shown only for the branches separating three GTA clades and their closest caudoviricete sequences. Scale bar indicates substitutions per site. GTA, gene transfer agent.

Analyses of environmental samples and genome sequences suggest the existence of a large number of GTAs, especially those related to the rhodogtaviriformids (Biers et al., 2008; Yunyun Fu et al., 2010; McDaniel et al., 2010; Zhao et al., 2009). In a genome-

wide screen of 1,423 alphaproteobacterial genomes, 57.5% were found to encode RcGTA-like “genomes”, which are often annotated as either intact or incomplete prophages (Kogay et al., 2019). The great majority of RcGTA-like genes in alphaproteobacterial genomes are associated with bacteria for which a GTA-based gene-transfer activity has not been documented, and it is possible that some of these RcGTA-like genes may not be expressed to produce functional particles. Therefore, we have restricted our proposal to those GTAs that have been shown to be functional. However, we speculate that at least some (and perhaps many) of these GTA-like gene clusters will be shown to produce functional GTAs that will need to be classified.

Based on the evolutionary history of TerL proteins (**Figure 7**), it is likely that the proposed three GTA families had distinct caudoviricete progenitors. Eventual deduction of the relatives of these progenitors may make it possible (or necessary) to include these GTA families in the virus class Caudoviricetes, thereby creating an overarching taxon for distinct MGEs (viruses and viriforms). Since the exaptation events, however, the three families have evolved as part of the host genomes (Esterman et al., 2021; Lang & Beatty, 2007; Lang et al., 2002; Shakya et al., 2017), in the case of the rhodogtaviriformids for hundreds of millions of years (Shakya et al., 2017). As a result, GTAs effectively became a component of cellular genomes, integrated into cellular regulatory circuits that also control processes such as motility, quorum sensing, extracellular polysaccharide synthesis, and biofilm formation (Lang et al., 2017; Pallegar et al., 2020; Shimizu et al., 2022). There is also mounting evidence that GTA genes experience selective pressures to be maintained in their host genomes (Kogay et al., 2020; Lang et al., 2012). Although the fitness benefits associated with GTA production remain to be elucidated, the time is now ripe to have the known GTAs officially recognized and classified as specific viriforms. We recognize this step as the initiation of a taxonomic framework that undoubtedly will rapidly expand and change in the future.

Materials and Methods

To identify alphaproteobacterial type I GTAs, we searched for RcGTA-like sequences in 1,248 complete alphaproteobacterial genomes extracted from the NCBI RefSeq database (accessed in October 2020) using GTA-Hunter (Kogay et al., 2019). We

identified 503 genomes that contained at least six RcGTA homologs in the same genetic neighborhood and had both g2 (encoding TerL) and g5 (encoding HK97-MCP) genes. To remove redundancy, we clustered genomes into the operational taxonomic units (OTUs) using an average nucleotide identity threshold of 95%. From all genomes within an OTU, we selected one genome with the largest number of GTA genes. This strategy resulted in 290 representative GTAs selected for further analysis. We identified the closest viral homologs of the TerL and HK97-MCP proteins from these GTAs by conducting a BLASTP search (Altschul et al., 1997) of the RefSeq database (accessed in March 2021) (O'Leary et al., 2016), using TerL and HK97-MCP proteins from representative GTAs as queries, an e-value cutoff of 0.001, and query coverage of at least 50%. Retrieved viral homologs with identical amino-acid sequences were removed from further analyses. For both proteins, we aligned amino-acid sequences of GTA and virus homologs using MAFFT v7.455 with -linsi option (Katoh & Standley, 2013). We reconstructed phylogenetic trees using IQ-TREE v2 (Minh et al., 2020), identifying the best substitution models using the built-in ModelFinder (Kalyaanamoorthy et al., 2017). The selected models were LG+F+R9 and LG+F+R7 for TerL and HK97-MCP datasets, respectively. Branch support values were assessed using 1,000 ultrafast bootstrap replicates and a hill-climbing nearest-neighbor interchange search for optimal trees (Hoang et al., 2018). Additionally, for both protein phylogenies, we reconstructed a phylogenetic tree in IQ-TREE v2 (Minh et al., 2020) using a tree search that was constrained by requiring all GTAs and all viruses to be separated by a branch. We compared the resultant trees in unconstrained and constrained searches using the AU test (Shimodaira, 2002), as implemented in the IQ-TREE v2 program.

To identify alphaproteobacterial type II GTAs, we used the BaGTA TerL and HK97-MCP sequences (accession numbers WP_034448260.1 and WP_011181178.1, respectively) as queries in a BLASTP search against the 57 complete Bartonella genomes extracted from the RefSeq database (accessed in May 2022). We restricted our search only to matches for which BaGTA TerL and HK97-MCP homologs are in the same genomic neighborhood (defined as being within 5 kb of each other). In genomes with multiple matches to the query protein, we retained only the homolog with the highest BLASTP bit score. We clustered 57 genomes using a 95% average nucleotide identity

(ANI) threshold and randomly selected one TerL and HK97-MCP representative from each cluster for phylogenetic analysis. We identified caudoviricete homologs by conducting a BLASTP search (e-value cutoff of 0.001, and query coverage of at least 50%) against viral RefSeq database (accessed in May 2022). We performed phylogenetic reconstructions as described above for alphaproteobacterial type I GTAs. The selected best substitution models were LG+R6 and LG+G4 for TerL and HK97-MCP datasets, respectively.

To identify GTAs of spirochaetes, we used BhGTA's MCP sequence (GenBank accession number WP_012671344.1) as a query in a BLASTP search (with an e-value cutoff of 0.001 and query coverage of at least 50%) against the 13 complete Brachyspira genomes extracted from the RefSeq database (accessed in May 2022). We used TerL of B. hyodysenteriae (GenBank accession number WP_012671469.1) and endolysin protein of B. hyodysenteriae (GenBank accession number WP_012671356.1) as queries in a BLASTP search (with an e-value cutoff of 0.001 and query coverage of at least 50%) against the same set of 13 genomes. For endolysins, we only retained matches that co-localized within the BhGTA region on the chromosome. We clustered 13 genomes using a 95% ANI threshold and randomly chose one TerL and endolysin representative from each cluster for phylogenetic analyses. We identified caudoviricete homologs by doing BLASTP searches (e-value cutoff of 0.001 and query coverage of at least 50%) against the viral RefSeq database (accessed in May 2022). We performed phylogenetic reconstructions as described above for the alphaproteobacterial type I GTAs. The selected best substitution models were VT+F+R3 and WAG+R6 for TerL and endolysin datasets, respectively.

To reconstruct the phylogeny that includes all three clades of GTAs, we combined all TerL homologs extracted in the above-described procedures into one dataset. We aligned the TerL sequences using MAFFT v7.455 with -dash option (Rozewicki et al., 2019) and trimmed the obtained alignment using ClipKIT with -gappy option (Steenwyk et al., 2020). We computed the phylogenetic tree using IQ-TREE v2 (Minh et al., 2020) as described above with the LG+F+R10 substitution model selected by ModelFinder. We rooted the tree using a larger TerL phylogeny presented in (Esterman et al., 2021).

We visualized all phylogenetic trees in iTOL v6 (Letunic & Bork, 2021).

Data availability

All data used in this manuscript were retrieved from publicly available GenBank databases, as described in the Methods. The accession numbers of database records used in the phylogenetic analyses can be found in alignments that are included in the **Supplementary Data**.

Acknowledgements

We thank Anya Crane and Jiro Wada (Integrated Research Facility at Fort Detrick, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health) for editing the manuscript and figures, respectively.

This work was supported in part through Laulima Government Solutions, LLC, prime contract with the U.S. National Institute of Allergy and Infectious Diseases (NIAID) under Contract No. HHSN272201800013C; J.H.K. performed this work as an employee of Tunnell Government Services (TGS), a subcontractor of Laulima Government Solutions, LLC, under Contract No. HHSN272201800013C. This work was also supported by the Simons Foundation Investigator in Mathematical Modeling of Living Systems award #327936 (O.Z.), the Canadian Natural Sciences and Engineering Research Council (NSERC) award RGPIN 2018-03898 (J.T.B.), NSERC award RGPIN-2017-04636 (A.S.L.). S.K. was partially supported by funding from the Memorial University of Newfoundland School of Graduate Studies.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Health and Human Services or of the institutions and companies affiliated with the authors, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Anderson, B., Goldsmith, C., Johnson, A., Padmalayam, I., & Baumstark, B. (1994). Bacteriophage-like particle of *Rochalimaea henselae*. *Mol Microbiol*, *13*(1), 67-73. https://doi.org/10.1111/j.1365-2958.1994.tb00402.x

Barbian, K. D., & Minnick, M. F. (2000). A bacteriophage-like particle from *Bartonella bacilliformis*. *Microbiology (Reading)*, *146 (Pt 3)*, 599-609. https://doi.org/10.1099/00221287-146-3-599

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*(1), 3034. https://doi.org/10.1038/s41467-020-16669-9

Berglund, E. C., Frank, A. C., Calteau, A., Vinnere Pettersson, O., Granberg, F., Eriksson, A. S., Näslund, K., Holmberg, M., Lindroos, H., & Andersson, S. G. (2009). Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting Bartonella grahamii. *PLoS Genet*, *5*(7), e1000546. https://doi.org/10.1371/journal.pgen.1000546

Bertani, G. (1999). Transduction-like gene transfer in the methanogen *Methanococcus voltae*. *J Bacteriol*, *181*(10), 2992-3002. https://doi.org/10.1128/JB.181.10.2992-3002.1999

Biers, E. J., Wang, K., Pennington, C., Belas, R., Chen, F., & Moran, M. A. (2008). Occurrence and expression of gene transfer agent genes in marine bacterioplankton. *Appl Environ Microbiol*, *74*(10), 2933-2939. https://doi.org/10.1128/AEM.02129-07

Brimacombe, C. A., Stevens, A., Jun, D., Mercer, R., Lang, A. S., & Beatty, J. T. (2013). Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). *Mol Microbiol*, *87*(4), 802-817. https://doi.org/10.1111/mmi.12132

Darboux, I., Cusson, M., & Volkoff, A. N. (2019). The dual life of ichnoviruses. *Curr Opin Insect Sci*, *32*, 47-53. https://doi.org/10.1016/j.cois.2018.10.007

Drezen, J.-M., Leobold, M., Bézier, A., Huguet, E., Volkoff, A.-N., & Herniou, E. A. (2017). Endogenous viruses of parasitic wasps: variations on a common theme. *Curr Opin Virol*, *25*, 41-48. https://doi.org/10.1016/j.coviro.2017.07.002

Esterman, E. S., Wolf, Y. I., Kogay, R., Koonin, E. V., & Zhaxybayeva, O. (2021). Evolution of DNA packaging in gene transfer agents. *Virus Evol*, *7*(1), veab015. https://doi.org/10.1093/ve/veab015

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst Biol*, *27*(4), 401-410. https://doi.org/10.1093/sysbio/27.4.401

Fogg, P. C. M., Westbye, A. B., & Beatty, J. T. (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*(8), e43772. https://doi.org/10.1371/journal.pone.0043772

Fu, Y., MacLeod, D. M., Rivkin, R. B., Chen, F., Buchan, A., & Lang, A. S. (2010). High diversity of *Rhodobacterales* in the subarctic North Atlantic Ocean and gene transfer agent protein expression in isolated strains. *Aquat Microb Ecol*, *59*, 283–293.

Gauthier, J., Drezen, J.-M., & Herniou, E. A. (2018). The recurrent domestication of viruses: major evolutionary transitions in parasitic wasps. *Parasitology*, *145*(6), 713-723. https://doi.org/10.1017/s0031182017000725

Guy, L., Nystedt, B., Toft, C., Zaremba-Niedzwiedzka, K., Berglund, E. C., Granberg, F., Näslund, K., Eriksson, A. S., & Andersson, S. G. (2013). A gene transfer agent and a dynamic repertoire of secretion systems hold the keys to the explosive radiation of the emerging pathogen *Bartonella*. *PLoS Genet*, *9*(3), e1003393. https://doi.org/10.1371/journal.pgen.1003393

Herniou, E. A., Huguet, E., Thézé, J., Bézier, A., Periquet, G., & Drezen, J.-M. (2013). When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. *Philos Trans R Soc Lond B Biol Sci*, *368*(1626), 20130051. https://doi.org/10.1098/rstb.2013.0051

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*, *35*(2), 518-522. https://doi.org/10.1093/molbev/msx281

Humphrey, S. B., Stanton, T. B., Jensen, N. S., & Zuerner, R. L. (1997). Purification and characterization of VSH-1, a generalized transducing bacteriophage of *Serpulina hyodysenteriae*. *J Bacteriol*, *179*(2), 323-329. https://doi.org/10.1128/jb.179.2.323-329.1997

Hynes, A. P., Mercer, R. G., Watton, D. E., Buckley, C. B., & Lang, A. S. (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*(2), 314-325. https://doi.org/10.1111/j.1365-2958.2012.08113.x

Hynes, A. P., Shakya, M., Mercer, R. G., Grull, M. P., Bown, L., Davidson, F., Steffen, E., Matchem, H., Peach, M. E., Berger, T., Grebe, K., Zhaxybayeva, O., & Lang, A. S. (2016). Functional and evolutionary characterization of a gene transfer agent's multilocus “genome". *Mol Biol Evol*, *33*(10), 2530-2543. https://doi.org/10.1093/molbev/msw125

International Committee on Taxonomy of Viruses. (2022). ICTV Code. The International Code of Virus Classification and Nomenclature (ICVCN). March 2021. https://talk.ictvonline.org/information/w/ictv-information/383/ictv-code.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*, *14*(6), 587-589. https://doi.org/10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kogay, R., Neely, T. B., Birnbaum, D. P., Hankel, C. R., Shakya, M., & Zhaxybayeva, O. (2019). Machine-learning classification suggests that many alphaproteobacterial prophages may instead be gene transfer agents. *Genome Biol Evol*, *11*(10), 2941-2953. https://doi.org/10.1093/gbe/evz206

Kogay, R., Wolf, Y. I., Koonin, E. V., & 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*(4), e01206-01220. https://doi.org/10.1128/mBio.01206-20

Koonin, E. V., Dolja, V. V., Krupovic, M., & Kuhn, J. H. (2021). Viruses defined by the position of the virosphere within the replicator space. *Microbiol Mol Biol Rev*, *85*(4), e0019320. https://doi.org/10.1128/MMBR.00193-20

Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global organization and proposed megataxonomy of the virus world. *Microbiol Mol Biol Rev*, *84*(2), e00061-00019. https://doi.org/10.1128/MMBR.00061-19

Koonin, E. V., & Krupovic, M. (2018). The depths of virus exaptation. *Curr Opin Virol*, *31*, 1-8. https://doi.org/10.1016/j.coviro.2018.07.011

Kuhn, J. H., Dolja, V. V., Krupovic, M., Adriaenssens, E. M., Di Serio, F., Dutilh, B. E., Flores, R., Harrach, B., Mushegian, A., Owens, B., Randles, J., Rubino, L., Sabanadzovic, S., Simmonds, P., Varsani, A., Zerbini, M., & Koonin, E. (2020). Expand, amend, and emend the International Code of Virus Classification and Nomenclature (ICVCN; "the Code") and the Statutes to clearly define the remit of the ICTV. International Committee on Taxonomy of Viruses (ICTV) TaxoProp 2020.005G.R.Code_and_Statute_Change. https://talk.ictvonline.org/files/ictv_official_taxonomy_updates_since_the_8th_report/m/general-2008/11061.

Kuhn, J. H., Postler, T., Dolja, V., Krupovic, M., Adriaenssens, E., Di Serio, F., Dutilh, B., Flores, R., Harrach, B., Mushegian, A., Owens, B., Randles, J., Rubino, L., Sabanadzovic, S., Simmonds, P., Varsani, A., Zerbini, M., & Koonin, E. (2021). Rename the family Polydnaviridae (as Polydnaviriformidae), rename the genus Bracovirus (as Bracoviriform) and rename all polydnaviriformid species to comply with the newly ICTV-mandated binomial format. International Committee on Taxonomy of Viruses (ICTV) TaxoProp 2021.006D.R.Polydnaviriformidae_1renfam_3rensp. https://talk.ictvonline.org/files/ictv_official_taxonomy_updates_since_the_8th_report/m/animal-dna-viruses-and-retroviruses/13248.

Lang, A. S., & Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*, *15*(2), 54-62. https://doi.org/10.1016/j.tim.2006.12.001

Lang, A. S., Taylor, T. A., & Beatty, J. T. (2002). Evolutionary implications of phylogenetic analyses of the gene transfer agent (GTA) of *Rhodobacter capsulatus*. *J Mol Evol*, *55*(5), 534-543. https://doi.org/10.1007/s00239-002-2348-7

Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol*, *4*(1), 87-104. https://doi.org/10.1146/annurev-virology-101416-041624

Lang, A. S., Zhaxybayeva, O., & Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol*, *10*(7), 472-482. https://doi.org/10.1038/nrmicro2802

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*, *49*(W1), W293-W296. https://doi.org/10.1093/nar/gkab301

Marrs, B. (1974). Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad Sci U S A*, *71*(3), 971-973. https://doi.org/10.1073/pnas.71.3.971

Matson, E. G., Thompson, M. G., Humphrey, S. B., Zuerner, R. L., & Stanton, T. B. (2005). Identification of genes of VSH-1, a prophage-like gene transfer agent of *Brachyspira hyodysenteriae*. *J Bacteriol*, *187*(17), 5885-5892. https://doi.org/10.1128/JB.187.17.5885-5892.2005

McDaniel, L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., & Paul, J. H. (2010). High frequency of horizontal gene transfer in the oceans. *Science*, *330*(6000), 50. https://doi.org/10.1126/science.1192243

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., 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*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015

Motro, Y., La, T., Bellgard, M. I., Dunn, D. S., Phillips, N. D., & Hampson, D. J. (2009). Identification of genes associated with prophage-like gene transfer agents in the pathogenic intestinal spirochaetes *Brachyspira hyodysenteriae*, *Brachyspira pilosicoli* and *Brachyspira intermedia*. *Vet Microbiol*, *134*(3-4), 340-345. https://doi.org/10.1016/j.vetmic.2008.09.051

Nagao, N., Yamamoto, J., Komatsu, H., Suzuki, H., Hirose, Y., Umekage, S., Ohyama, T., & Kikuchi, Y. (2015). The gene transfer agent-like particle of the marine

phototrophic bacterium *Rhodovulum sulfidophilum*. *Biochem Biophys Rep*, *4*, 369-374. https://doi.org/10.1016/j.bbrep.2015.11.002

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Pallegar, P., Pena-Castillo, L., Langille, E., Gomelsky, M., & Lang, A. S. (2020). Cyclic di-GMP-mediated regulation of gene transfer and motility in *Rhodobacter capsulatus*. *J Bacteriol*, *202*(2). https://doi.org/10.1128/JB.00554-19

Petersen, J. M., Bézier, A., Drezen, J.-M., & van Oers, M. M. (2022). The naked truth: An updated review on nudiviruses and their relationship to bracoviruses and baculoviruses. *J Invertebr Pathol*, *189*, 107718. https://doi.org/10.1016/j.jip.2022.107718

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*(6), 611-621 e616. https://doi.org/10.1016/j.cels.2017.05.011

Rapp, B. J., & Wall, J. D. (1987). Genetic transfer in *Desulfovibrio desulfuricans*. *Proc Natl Acad Sci U S A*, *84*(24), 9128-9130. https://doi.org/10.1073/pnas.84.24.9128

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res*, *47*(W1), W5-W10. https://doi.org/10.1093/nar/gkz342

Shakya, M., Soucy, S. M., & Zhaxybayeva, O. (2017). Insights into origin and evolution of alpha-proteobacterial gene transfer agents. *Virus Evol*, *3*(2), vex036. https://doi.org/10.1093/ve/vex036

Shimizu, T., Aritoshi, T., Beatty, J. T., & Masuda, T. (2022). Persulfide-responsive transcription factor SqrR regulates gene transfer and biofilm formation via the metabolic modulation of cyclic di-GMP in *Rhodobacter capsulatus*. *Microorganisms*, *10*(5), 908. https://www.mdpi.com/2076-2607/10/5/908

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst Biol*, *51*(3), 492-508. https://doi.org/10.1080/10635150290069913

Stanton, T. B., Humphrey, S. B., Sharma, V. K., & Zuerner, R. L. (2008). Collateral effects of antibiotics: carbadox and metronidazole induce VSH-1 and facilitate gene transfer among *Brachyspira hyodysenteriae* strains. *Appl Environ Microbiol*, *74*(10), 2950-2956. https://doi.org/10.1128/AEM.00189-08

Steenwyk, J. L., Buida, T. J., III, Li, Y., Shen, X.-X., & Rokas, A. (2020). ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol*, *18*(12), e3001007. https://doi.org/10.1371/journal.pbio.3001007

Strand, M. R., & Burke, G. R. (2020). Polydnaviruses: evolution and function. *Curr Issues Mol Biol*, *34*, 163-182. https://doi.org/10.21775/cimb.034.163

Tamarit, D., Neuvonen, M. M., Engel, P., Guy, L., & Andersson, S. G. E. (2018). Origin and evolution of the *Bartonella* gene transfer agent. *Mol Biol Evol*, *35*(2), 451-464. https://doi.org/10.1093/molbev/msx299

Theze, J., Bezier, A., Periquet, G., Drezen, J. M., & Herniou, E. A. (2011). Paleozoic origin of insect large dsDNA viruses. *Proc Natl Acad Sci U S A*, *108*(38), 15931-15935. https://doi.org/10.1073/pnas.1105580108

Tomasch, J., Wang, H., Hall, A. T. K., Patzelt, D., Preusse, M., Petersen, J., Brinkmann, H., Bunk, B., Bhuju, S., Jarek, M., Geffers, R., Lang, A. S., & Wagner-Dobler, I. (2018). Packaging of *Dinoroseobacter shibae* DNA into gene transfer agent particles is not random. *Genome Biol Evol*, *10*(1), 359-369. https://doi.org/10.1093/gbe/evy005

Umemori, E., Sasaki, Y., Amano, K., & Amano, Y. (1992). A phage in *Bartonella bacilliformis*. *Microbiol Immunol*, *36*(7), 731-736. https://doi.org/10.1111/j.1348-0421.1992.tb02075.x

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Davison, A. J., Dempsey, D. M., Dutilh, B. E., Garcia, M. L., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Krupovic, M., Kuhn, J. H., Lambert, A. J., Lobocka, M., . . . Zerbini, F. M. (2021). Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch Virol*, *166*(9), 2633-2648. https://doi.org/10.1007/s00705-021-05156-1

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., Garcia, M. L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Lobocka, M., Oksanen, H. M., Orton, R. J., Robertson, D. L., Rubino, L., . . . Zerbini, F. M. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch Virol*, *167*(11), 2429-2440. https://doi.org/10.1007/s00705-022-05516-5

Westbye, A. B., Beatty, J. T., & Lang, A. S. (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. https://doi.org/10.1016/j.mib.2017.05.003

Westbye, A. B., O'Neill, Z., Schellenberg-Beaver, T., & Beatty, J. T. (2017). The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology (Reading)*, *163*(9), 1355-1363. https://doi.org/10.1099/mic.0.000519

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucia-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja, V. V., & Koonin, E. V. (2018). Origins and evolution of the global RNA virome. *mBio*, *9*(6), e02329-02318. https://doi.org/10.1128/mBio.02329-18

Yen, H. C., Hu, N. T., & Marrs, B. L. (1979). Characterization of the gene transfer agent made by an overproducer mutant of *Rhodopseudomonas capsulata*. *J Mol Biol*, *131*(2), 157-168. https://doi.org/10.1016/0022-2836(79)90071-8

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. https://doi.org/10.1038/srep30372

Zhao, Y., Wang, K., Budinoff, C., Buchan, A., Lang, A., Jiao, N., & Chen, F. (2009). Gene transfer agent (GTA) genes reveal diverse and dynamic *Roseobacter* and *Rhodobacter* populations in the Chesapeake Bay. *Isme j*, *3*(3), 364-373. https://doi.org/10.1038/ismej.2008.115

Chapter 7

Outlook and Conclusions

Roman Kogay[1]

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

Despite that GTA genes share ancestry with viruses, they can be clearly differentiated from each other by phylogenetics and comparative genomics analyses. Reconstructing phylogenetic trees and examining genetic neighborhoods are generally computationally demanding tasks, and the development of GTA-Hunter software, introduced in chapter 2, substantially improves the efficiency of detecting GTA clusters. By examining 1,423 alphaproteobacterial genomes using the GTA-Hunter, we found that a majority of these genomes contain GTA clusters. These findings are consistent with results from the earlier studies, confirming that GTAs are widely encoded by alphaproteobacterial species (Lang & Beatty, 2007; Lang et al., 2002; Shakya et al., 2017). Do all these predicted GTAs maintain their functional capabilities? More detailed examination of these GTA clusters reveals a common occurrence of pseudogenization and the loss events of GTA genes that are essential for the proper production of GTAs. This suggests that at least some GTA clusters are very unlikely to be functional. Interestingly, even closely related genomes contain heterogeneous GTA clusters. For example, despite that *Brucella* species have high genomic identity with each other (more than 99% of Average Nucleotide Identity (Jain et al., 2018; Konstantinidis & Tiedje, 2005)), they exhibit different pseudogenization/loss patterns in GTA genes (**Figure 1**). It raises an intriguing possibility that strains with incomplete or partially pseudogenized GTA clusters represent emergent lineages of 'cheaters' that do not produce GTA particles due to the inactivation of GTA genes, as we discussed in the chapter 5. Additional analyses are needed to examine the pseudogenization rate of GTA genes relatively to other gene families within various closely related species. It will allow to better understand whether selection pressure favors pseudogenization of GTA genes in different taxonomic groups.

Furthermore, multiple clades have convergently lost GTA genes, suggesting that under some ecological conditions GTAs are useless or deleterious and can get purged from the genomes. In fact, a strong mutational bias toward deletions in bacteria profoundly shapes their genomic architecture, eliminating junk DNA in relatively short periods of time (Mira et al., 2001). Interestingly, we did not detect presence of GTAs in endosymbiont genomes (Kogay et al., 2019). Indeed, endosymbionts generally undergo

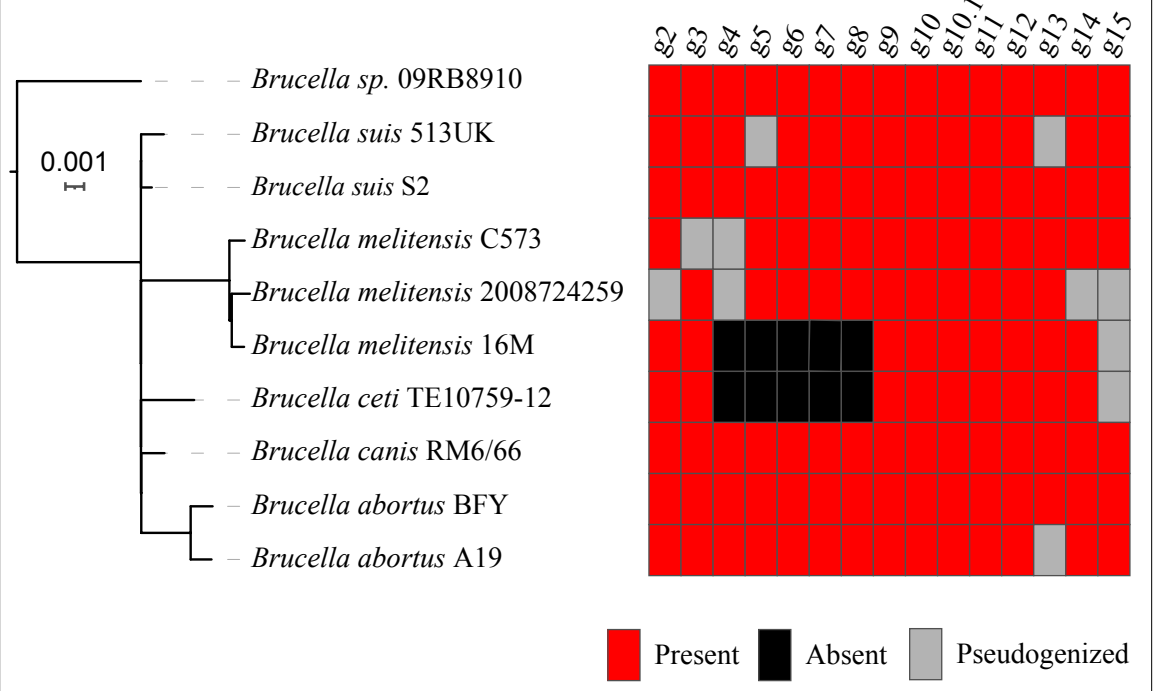

Figure 1. Distribution of GTA genes in ten representative *Brucella* species. The patterns of presence, absence and pseudogenization of GTA genes are mapped to a phylogenomic tree that was reconstructed from a concatenated alignment of 2,061 marker genes. Genes *g1* and *g3.5* were excluded, as they are generally not identified outside of *Rhodobacterales* order due to their short length and high substitution rate. The unit of the scale bar is the number of substitutions per site.

the process of Muller's ratchet by accumulating deleterious mutations due to the genetic drift, losing even beneficial phenotypic traits in the process (Moran, 1996). However, some of the latest studies suggest that endosymbionts may actually produce GTA particles (Fallon & Carroll, 2023; George et al., 2022). This inconsistency might stem from the fact that GTA-Hunter relies on the notable bias towards amino acids that are predominantly encoded by GC-rich codons, whereas genomic composition of endosymbiotic genomes is highly AT-biased (Clark et al., 1999). The main benefit of producing GTAs within populations of endosymbionts might be to escape the Muller's ratchet by increasing the recombination rate. This hypothesis can be computationally evaluated by examining the recombination rate in potentially GTA-producing endosymbiotic species relative to those that do not encode them.

The considerable progress toward better understanding of GTAs was made over the last five decades. Although only dozens of GTAs were experimentally confirmed, it is clear that more GTAs will be discovered and validated in the future. However, the lingering question that largely remains unanswered is the extent to which GTAs are produced in the natural environments. This gap in knowledge can be potentially addressed by studying metaviromes that were collected and sequenced from different environmental habitats. Interestingly, the collected metavirome samples generally contain prokaryotic DNA (Hurwitz & Sullivan, 2013; Kristensen et al., 2010). Partially this could be attributed to contamination and limitations of the virome collection protocols. However, some of the microbial DNA in metaviromes could be due to GTAs, as they can pass through the filters that retain microbial cells and can protect DNA from nuclease treatment. Indeed, the study of the soil virome suggests that ~25% of the assembled reads could be coming from GTAs (Trubl et al., 2018). Thus, the systematic analysis of properly collected metavirome samples is a promising approach to advance our knowledge about GTA production in Nature.

References

Clark, M. A., Moran, N. A., & Baumann, P. (1999). Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol Biol Evol*, *16*(11), 1586-1598. https://doi.org/10.1093/oxfordjournals.molbev.a026071

Fallon, A. M., & Carroll, E. M. (2023). Virus-like particles from *Wolbachia*-infected cells may include a gene transfer agent. *Insects*, *14*(6), 516. https://doi.org/10.3390/insects14060516

George, E. E., Tashyreva, D., Kwong, W. K., Okamoto, N., Horak, A., Husnik, F., Lukes, J., & Keeling, P. J. (2022). Gene transfer agents in bacterial endosymbionts of microbial eukaryotes. *Genome Biol Evol*, *14*(7), evac099. https://doi.org/10.1093/gbe/evac099

Hurwitz, B. L., & Sullivan, M. B. (2013). The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One*, *8*(2), e57355. https://doi.org/10.1371/journal.pone.0057355

Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Kogay, R., Neely, T. B., Birnbaum, D. P., Hankel, C. R., Shakya, M., & Zhaxybayeva, O. (2019). Machine-learning classification suggests that many alphaproteobacterial prophages may instead be gene transfer agents. *Genome Biol Evol*, *11*(10), 2941-2953. https://doi.org/10.1093/gbe/evz206

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A*, *102*(7), 2567-2572. https://doi.org/10.1073/pnas.0409727102

Kristensen, D. M., Mushegian, A. R., Dolja, V. V., & Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends Microbiol*, *18*(1), 11-19. https://doi.org/10.1016/j.tim.2009.11.003

Lang, A. S., & Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*, *15*(2), 54-62. https://doi.org/10.1016/j.tim.2006.12.001

Lang, A. S., Taylor, T. A., & Beatty, J. T. (2002). Evolutionary implications of phylogenetic analyses of the gene transfer agent (GTA) of *Rhodobacter capsulatus*. *J Mol Evol*, *55*(5), 534-543. https://doi.org/10.1007/s00239-002-2348-7

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet*, *17*(10), 589-596. https://doi.org/10.1016/s0168-9525(01)02447-7

Moran, N. A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A*, *93*(7), 2873-2878. https://doi.org/10.1073/pnas.93.7.2873

Shakya, M., Soucy, S. M., & Zhaxybayeva, O. (2017). Insights into origin and evolution of alpha-proteobacterial gene transfer agents. *Virus Evol*, *3*(2), vex036. https://doi.org/10.1093/ve/vex036

Trubl, G., Jang, H. B., Roux, S., Emerson, J. B., Solonenko, N., Vik, D. R., Solden, L., Ellenbogen, J., Runyon, A. T., Bolduc, B., Woodcroft, B. J., Saleska, S. R., Tyson, G. W., Wrighton, K. C., Sullivan, M. B., & Rich, V. I. (2018). Soil viruses are underexplored players in ecosystem carbon processing. *mSystems*, *3*(5). https://doi.org/10.1128/mSystems.00076-18