New insights into the regulation and evolution of gene transfer agents in Alphaproteobacteria through comparative genomics and transcriptomics

by Sonja Elena Koppenhöfer

A Thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biology

Memorial University of Newfoundland

Mai 2022

St. John’s, Newfoundland and Labrador

Abstract

Gene transfer agents (GTA) are phage-like particles that can transfer fragments of the host genome between cells. GTAs are particularly widely distributed in the Alphaproteobacteria where their expression is controlled by their hosts gene regulatory network in the model organisms. In this thesis, I used computational comparative analyses of publicly available datasets to further extend our knowledge about the regulatory networks controlling GTA production and to identify genomic traits that might help explain the strong conservation of GTAs in these bacteria. Reanalysis of transcriptomic datasets revealed a regulatory interplay between nitrogen- and oxygen-sensing regulators with the CtrA phosphorelay, which is an established regulator of GTA gene expression. I evaluated the chromosomal locations of genes encoding GTA regulators, components of the CtrA phosphorelay, and enzymes associated with the synthesis and degradation of the GTA-affecting second messenger c-di-GMP. These showed some conserved distribution patterns along the chromosome, which might indicate that GTA production is coordinated with the cell cycle. Finally, I showed that GTA genes share many properties with core genes, such as the preferred location on the leading strand, in low plasticity regions and in GC skew peaks. Overall, the results of my dissertation provide novel insights for understanding the regulation of GTA gene expression and the evolution of the GTA gene cluster in Alphaproteobacteria and revealed several potentially important aspects that can be followed up on in the future with experimental studies to test their relevance to GTA biology.

Acknowledgements

First, I would like to thank Dr. Andrew Lang for his guidance and mentoring during my time here at Memorial University. His support and encouragement helped me through the difficult phases of the doctoral process and encouraged me to pursue own interests.

I would also like to thank my committee members, Dr. Suzanne Dufour and Dr. Dawn Bignell for helping me to stay focused and not to lose the thread even though the project got so bioinformatic. I would especially like to thank Suzanne for her kindness and always open door.

My sincere thanks to the post-doctoral fellows Dr. Marta Canuti and Dr. Joost Verhoeven for being so motivating, always trying to help and not losing patience, even when I tried to understand phylogenetic analysis or to learn Linux.

I want to thank Dr. Jürgen Tomasch, who has contributed so comprehensively through his insights and knowledge of the matter and helped making a very confusing story less confusing.

My former and current laboratory colleagues, thank you for accompanying me on this trip. I can't even imagine what it must be like to do a PhD without the friendship, support, and motivation that I got from you.

My doctoral thesis was supported by graduate scholarships from the Memorial University School of Graduate Studies (SGS) and the Natural Science and Engineering Research Council (NSERC), for what I am very grateful.

Finally, I would like to thank my family and especially my father, who always encouraged and supported me to think outside the box and to follow my dreams.

Co-Authorship Statement

I am the primary author of chapters 2, 4 and 5 of this thesis, and co-first author of chapter 3.

Chapter 2: The conceptualization, methodology, data analysis, visualization and original draft preparation of this research were done by me. Writing, review, editing, and supervision were performed by A. S. Lang.

Koppenhöfer, S.; Lang, A.S. Interactions among Redox Regulators and the CtrA Phosphorelay in *Dinoroseobacter shibae* and *Rhodobacter capsulatus*. *Microorganisms* **2020**, *8*. https://doi.org/10.3390/microorganisms8040562.

Chapter 3: I did all the data analysis of this research. Together with J. Tomasch I visualized the data and together with J. Tomasch and A. S. Lang I wrote the manuscript. J. Tomasch had the original idea for the study.

Tomasch, J.; **Koppenhöfer, S.**; Lang, A.S. Connection between chromosomal location and function of CtrA phosphorelay genes in Alphaproteobacteria. *Frontiers in Microbiology* **2021**, *12*, 662907. https://doi.org/10.3389/fmicb.2021.662907.

Chapter 4: The conceptualization, methodology, data analysis, visualization and original draft preparation of this research were done by me. A. S. Lang supervised the work. Writing, review and editing were done by J. Tomasch and A. S. Lang.

This manuscript is currently under review in the journal *Microbial Genomics*. **Koppenhöfer, S.**; Tomasch, J; Lang, A.S. Shared properties of gene transfer agent and core genes revealed by comparative genomics of Alphaproteobacteria.

Chapter 5: I did the conceptualization, methodology and original draft preparation of this research. Data analysis and visualization were done by me and A. S. Lang. Writing, review, editing, and supervision were performed by A. S. Lang.

This manuscript is currently under review in the journal *Microbial Genomics*. **Koppenhöfer, S.**; Lang, A.S. Patterns in the abundance, chromosomal localizations, and domain organizations of c-di-GMP-associated genes revealed by comparative genomics of five alphaproteobacterial orders.

Table of contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Meaning
%	Percent
µ	Micro
Δ	Delta
AAP	Aerobic anoxygenic photosynthesis
AHL	Acyl-homoserine lactone (N-acyl homoserine lactone)
AI	Autoinducer
ATP	Adenosine tri phosphate
BaGTA	*Bartonella* gene transfer agent
Bchl-a	Bacteriochlorophyll-a
VSH-1	*Brachyspira hyodysenteriae*
Bp	Base pair
ChIP-Seq	Chromatin immunoprecipitation sequencing
CO_2	Carbon dioxide
C-di-GMP	Cyclic dimeric guanosine monophosphate
Dd1	*Desulfovibrio desulfuricans* gene transfer agent
DGC	Diguanylate cyclase
Dif	Deletion induced filamentation
DNA	Deoxyribonucleic acid
DsGTA	*Dinoroseobacter shibae* gene transfer agent
g2	Large terminase protein
g5	Major capsid protein
g9	Major tail protein
GTA	Gene transfer agent

GC skew	Guanine cytosine skew
GDP	Guanosine diphosphate
GMP	Guanosine monophosphate
GTP	Guanosine triphosphate
HSL	N-acyl homoserine lactone (Acyl-homoserine lactone)
Mb	Mega base pair
NCBI	National Center for Biotechnology Information
PDE	Phosphodiesterase
(p)ppGpp	Guanosine (penta-) tetra-phosphate
QS	Quorum sensing
RcGTA	*Rhodobacter capsulatus* gene transfer agent
ssDNA	Single stranded deoxyribonucleic acid
Ori	Origin of replication
Tad	Tight adherence
Ter	Terminus of replication
VTA	*Methanococcus voltae* gene transfer agent

1 CHAPTER 1: Introduction

In my PhD thesis I use comparative computational analyses to study transcriptomic and genomic datasets and identify patterns related to the regulation and evolution of gene transfer agents within the class Alphaproteobacteria. The introduction is structured to provide relevant background information on the topics and concepts that are featured in the subsequent research chapters.

1.1 Alphaproteobacteria

The Alphaproteobacteria are a class of Gram-negative bacteria that can be found in diverse environments. They differ in terms of their geographic distributions and metabolic capabilities, and include free-living organisms in marine, freshwater or terrestrial environments, host-associated organisms, intracellular symbionts, and pathogens. They also show high genomic diversity, with varying degrees of AT and GC contents, genome sizes, numbers of protein-encoding genes, and evolutionary rates [1–3]. It was previously thought that a member of the Alphaproteobacteria gave rise to mitochondria, but this hypothesis was recently re-evaluated considering new phylogenetic and genomic analyses and it seems that the mitochondrial origin goes back to a proteobacterial lineage that split off before the Alphaproteobacteria [4,5]. Currently this subphylum contains eight orders, and the phylogeny shows that the Pelagibacterales first split from a common ancestor and represent the deepest branching lineage, followed by the Rickettsiales and Holosporales. Interestingly, the GC contents and sizes of genomes in the different lineages increased according to the split-off time. These events were followed by the separation of the Rhodospirillales, Sphingomonadales, Rhodobacterales, Caulobacterales and Rhizobiales [1,2].

The order Pelagibacterales includes the SAR11 clade, which is the most abundant bacterial group in the ocean and its members are major players in biogeochemical cycles [6]. On average the Pelagibacterales constitute over 25% total, and in some areas up to 60%, of marine microbial communities [6,7]. While the reason for their success is still unclarified, it has been hypothesized that streamlining their genomes has helped them survive in nutrient-poor oceanic regions [3,7]. This genomic simplification mainly involves the reduction of the numbers of mobile genetic elements, accessory genes,

pseudogenes and paralogs while most central metabolism pathways are maintained [3].

On the contrary, bacteria within the Rickettsiales and Holosporales, which are endosymbionts, rely on metabolic pathways of their hosts. The Rickettsiales are a group of obligate intracellular bacteria, some members of which can infect both vertebrate and invertebrate hosts. This lifestyle allowed the host's metabolism to be harnessed, saving energy, resulting in increased gene deletion rates and a reduction in genome size [8]. This order is primarily known for the pathogenicity of its members and as causative agents of serious diseases, such as typhus, ehrlichioses, and heartwater disease [9–11]. The Holosporales is a recently defined order and includes parasites that, despite being physiologically similar to bacteria within the Rickettsiales, are genetically different and are thought to have evolved independently from a free-living ancestor [1].

The Rhodospirillales appear to be a paraphyletic order since three genera of the Rhizobiales were nested in this order with high support [12]. The Rhodospirillales contains metabolically diverse bacteria, such as acetic acid, magnetotactic, and photosynthetic bacteria [12,13] and model organisms used to study their diverse light-harvesting antenna structures [14,15]. Rhodospirillales can be host-associated, such as *Rhodospirillales* bacterium strain TMPK1 that can be found within the intestines of *Drosophila* flies, or free-living [13,16]. The phylogenetic division of the families within the Sphingomonadales was recently put into question and reorganization of some genera into a new family was suggested [12]. Many members of the Caulobacterales are dimorphic, prosthecate bacteria whose cell division leads to two cells that are morphologically and behaviorally different from each other. They are often oligotrophs and are found in fresh, brackish, and marine waters [17–19]. Members of the Rhizobiales can also adapt to a multitude of different lifestyles, a feature that requires high genomic plasticity [17,20,21]. They can be found associated with specific hosts, for example as nitrogen-fixing and pathogenic symbionts of plants and animals, respectively, a fact that makes them interesting for agricultural and medical research [20,22].

The Rhodobacterales is one of the best-studied orders of the Alphaproteobacteria, mainly because it includes the roseobacters, one of the most abundant and cultivatable marine groups of bacteria [23]. However, examination of the phylogeny of marine and non-marine members showed that both groups

were inter-nested. This indicates that the switch from marine to non-marine habitats occurred several times independently and each led to characteristic changes in genome content and metabolism. The roseobacters were defined as paraphyletic in 2017 [24] and proposed to be relocated into a new, monophyletic family called "Roseobacteraceae" in 2021, on the basis of distinct environmental adaptions, genomic, phylogenetic, and *in silico*-predicted phenotypic data [25]. The Rhodobacterales are also of interest for this thesis work because of their high level of conservation of gene transfer agent genes [26].

1.2 Gene transfer agents (GTA) and their regulation

Gene transfer agents (GTA) are named for their ability to package their hosts' DNA and deliver it to other cells. GTA research began in 1974 when Barry Marrs discovered a deoxyribonuclease-resistant, cell-cell contact-independent DNA transfer between cells and concluded a new kind of "vehicle" was responsible [27]. At the time of their discovery, this was the first genetic recombination system observed in photosynthetic bacteria [28]. Currently, five GTA families (discovered in *Rhodobacter capsulatus* (RcGTA), *Bartonella* spp. (BaGTA), *Brachyspira hyodysenteriae* (VSH-1), *Desulfovibrio desulfuricans* (Dd1), and *Methanococcus voltae* (VTA)) have been identified. They have some shared and unique properties and have been found in both bacteria and archaea [29].

Rhodobacter capsulatus became the first model organism and namesake for its GTA family (RcGTA) [29], which also includes the GTA of *Dinoroseobacter shibae* (DsGTA) [31]. Since the 1960s *R. capsulatus*, originally known as *Rhodopseudomonas capsulata*, has been widely used as a model to investigate the metabolism of purple non-sulfur bacteria [28]. It has a genome size of 3.9 Mb, is associated with freshwater environments and, as a purple non-sulfur photoheterotrophic bacterium [32], is capable of aerobic and anaerobic respiration, fermentation and anaerobic phototrophy. Only recently *D. shibae* has also been added as a model organism for studying GTAs. It belongs to the roseobacter group and was discovered growing in association with the benthic dinoflagellate *Prorocentrum minimum* in 2005 [33]. Since then, it has been associated with a variety of other cosmopolitan marine toxic and non-toxic algae [34]. *D. shibae* has a genome size of 4.4 Mb and is an aerobic anoxygenic photoheterotroph (AAP) [33].

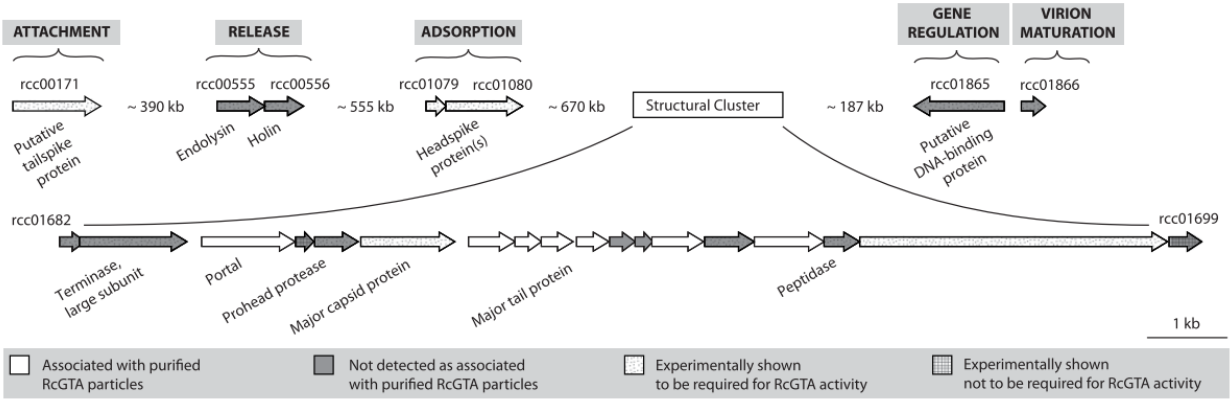

Figure 1.1: GTA genes of *Rhodobacter capsulatus*. The RcGTA gene cluster and its related loci are indicated with their putative roles in particle production and gene transfer and gene identification (prefix "RCAP_" missing for NCBI usage). Open reading frames are drawn to scale and shaded according to function/characterization. Taken from [36].

RcGTA-related genes are particularly widespread in the Alphaproteobacteria, with 57.5% of sequenced members containing at least some of the genes and almost all members of the order Rhodobacterales containing complete copies of the RcGTA-like gene cluster (~ 14 kb) as well as some of the additional isolated GTA genes located in other regions of the genome (Figure 1.1) [35]. The typical RcGTA gene cluster consists of 15 genes and encodes most of the structure of the particle. There are additional genes distributed across the chromosome that encode proteins involved in GTA-cell interactions, and particle maturation and release [36]. The highest levels of sequence conservation among different species is found for the open reading frames *g2*, *g5* and *g9*, which encode the large terminase, the major capsid, and the major tail proteins, respectively [35]. Due to the conservation of the GTA gene cluster and since many genes involved in GTA production share viral homologues, these are believed to derive from the structural genes of a prophage that integrated into an alphaproteobacterial ancestor and since coevolved with the host genome [35,37-39]. However, there are fundamental points that distinguish GTAs from phages. First of all, most GTAs can only transfer short fragments of linear dsDNA (~4-5 kb for RcGTA) [40]. This constraint does not allow the transfer of the entire RcGTA structural gene cluster within a single GTA particle. However, in other cases, such as BaGTA, the particle is encoded by a smaller genomic region and thus the whole cluster could theoretically be self-transmissible (although this

region is currently poorly defined) [29]. Furthermore, unlike phages, GTAs predominantly package host DNA and GTA-encoding genes are only encapsulated as they are part of the bacterial genome. In *Rhodobacter*, for example, transferred fragments are random genomic pieces with the GTA gene cluster showing lower packaging rates compared to the rest of the genome [31,35]. In *D. shibae*, certain genomic regions are packaged more frequently than others, possibly due to specific DNA properties, but this is still under investigation [31]. In *Bartonella*, the chromosomal region around the *ori* of a defective phage is amplified by run-off replication and the increased copy number of this DNA also leads to an increased packaging rate of this region in the GTAs [41,42].

Morphologically, GTAs resemble small, tailed phages, more precisely siphoviruses in the cases of RcGTA and DsGTA [31,40]. Although GTAs are encoded by all cells of a population, their production is restricted to a subpopulation, whose size varies among organisms, strains, and GTA families. For RcGTA, the proportion of GTA-producing cells is <3% [31,38,43,44]. Similar to some phages, GTA release results in cell death of the producing microbial subpopulation. From RcGTA it is known that, also similar to phages, these particles interact with recipient cells and inject their DNA through a tail. However, in contrast to phages which inject DNA into the cytosol, RcGTAs inject it into the periplasm, from where it is then transported into the cytosol by Com-family proteins, similar to transformation [45]. Thus, gene transfer by RcGTAs can be seen as a combination of transformation and transduction [46-49]. Lastly, it has been determined that RcGTA-family genes evolve significantly slower than their viral homologs [39].

1.2.1 The CtrA phosphorelay

GTAs are controlled by a variety of regulators and environmental signals. One of the most important regulatory systems for RcGTA is the CtrA histidyl-aspartyl phosphorelay that is widely conserved in the Alphaproteobacteria [50]. The CtrA phosphorelay has three members: the histidine kinase CckA, which autophosphorylates in response to a currently unknown stimulus, the phosphotransferase ChpT that accepts the phosphate residue from CckA, and the transcriptional regulator CtrA, which is phosphorylated by ChpT. A second histidine kinase, CcsA, has been identified in

Sphingomonas melonis that also feeds into the ChpT-CtrA relay, likely in response to blue light [51], and it is possible there are additional kinases that can affect this pathway in other species.

The high diversity in habitats and lifestyles that characterizes the Alphaproteobacteria is reflected by the fact that this regulatory system is broadly conserved across the class and operates in a similar way in different species but its functions and importance for bacterial survival differ. This phosphorelay is most well-known from its extensively studied role in cell cycle regulation in *Caulobacter crescentus* [52]. The original discovery of the CtrA phosphorelay in *R. capsulatus* was with respect to its regulation of GTA production and subsequently flagellar motility [53,54]. In *R. capsulatus*, loss of CtrA affects the expression of 6% of the bacterium's genes [55], including those relevant for GTA production, release and recipient capability and, even though direct regulation of the GTA gene cluster by CtrA does not appear to occur, these traits are differentially controlled by different CtrA phosphorylation states (Figure 1.3) [56]. Other members of the CtrA regulon that were identified for *D. shibae* are the Tad (tight adherence) pilus, c-di-GMP-associated enzymes, the master regulator of the stress response LexA, heterogeneous morphology and cell division as well as parts of the quorum sensing system (Figure 1.2) [57]. Therefore, CtrA in *D. shibae* seems to also be linked to cell-cell communication and differentiation. CtrA affects gene expression differently when it is in a phosphorylated (CtrA~P) or unphosphorylated status. For example, in *C. crescentus* increased concentrations of CtrA~P activate numerous targets but inhibit chromosome replication, which only progresses when CtrA is unphosphorylated [58,59]. The production of GTA occurs in two different steps, the maturation of phage-like particles and their lytic release, which depend on low and high concentrations of CtrA~P, respectively [60].

1.2.2 Integrated signals: Cell density, DNA stress and nutrient availability

All bacterial cells use signaling proteins and pathways to sense and respond to the environment by converting extracellular cues into intracellular signals. Three environmental signals are known to interact with the CtrA phosphorelay and GTA production in *R. capsulatus* and *D. shibae*. These involve cell-cell communication via quorum sensing (QS), the integration of the stress response via its master regulator LexA, and nutritional availability via the signaling molecule (p)ppGpp (guanosine-5',3'-penta/tetraphosphate) [46,61,62].

QS is a process of synthesizing, secreting, and sensing small diffusible molecules known as autoinducers (AIs). When a critical cell concentration threshold or quorum is reached, the AIs, which are detected by cell receptors, activate signal cascades that ultimately regulate the expression of genes for specific and coordinated behaviors, such as virulence, biofilm formation, cell cycle regulation and horizontal gene transfer [63]. Multiple AIs exist, although oligopeptides and N-acyl-L-homoserine lactones (AHLs) make up the majority of such molecules in Gram-positive and Gram-negative bacteria, respectively [64]. The QS system of *D. shibae* consists of three AHL synthases and each one is encoded within an operon with a putative LuxR-type transcriptional regulator [65]. *D. shibae* produces long chain AHLs with varying side chain lengths (C14–C18) and modifications [66,67]. The *D. shibae* QS system is organized hierarchically, as revealed by

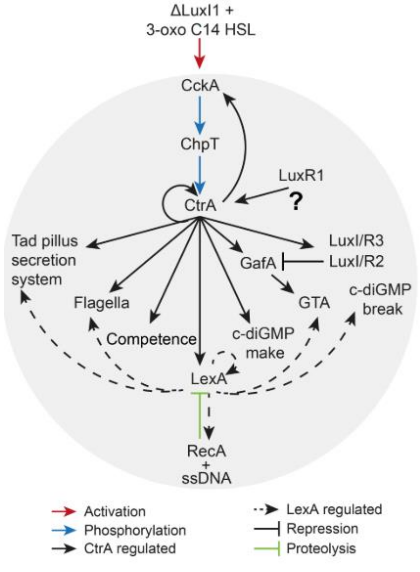

Figure 1.2: *D. shibae* CtrA phosphorelay regulatory network. Addition of 3-oxo C14 HSL to a knockout strain of *luxI*$_1$ is assumed to be detected by the histidine kinase CckA (red arrow) and results in a phosphorylation cascade activating CtrA (blue) and subsequently its target genes (continuous arrows). The concentration of GafA might be limited by the inhibition by LuxI$_2$ and defines the subpopulation size of GTA producers (repression symbol). Multiple traits are regulated by CtrA and the SOS stress response regulator LexA (dashed arrows). LexA inhibition on the genome is released by binding and proteolysis of RecA to ssDNA (green repression symbol arrow) [57].

gene deletion analyses: the deletion of *luxI*$_1$ results in a QS null mutant which does not produce detectable levels of AHLs and thus also abolishes AHL production by the other two synthases [65-67]. The product of the LuxI$_1$ synthase induces the expression of the CtrA phosphorelay genes, which induces the expression of the *luxI*$_2$ and *luxI*$_3$ operons as well as QS target genes, including those for biosynthesis of flagella and GTAs [67]. Additionally, QS activation of the CtrA phosphorelay has been shown to affect the chromosome content of cells and their morphological differentiation. The effects of the *luxI*$_1$ knockout could be complemented by external addition of a wide variety of long chain AHLs. While knockouts of *luxI*$_1$ and CtrA phosphorelay genes resulted in the downregulation of GTA genes, knockout of *luxI*$_2$ resulted in their overexpression [65,67].

R. capsulatus has one autoinducer synthase-receptor pair, GtaI and GtaR, respectively. In QS knockouts, GTA production is reduced but can be restored by exogenous addition of N-hexadecanoyl-homoserine lactones (HSL). GTA recipient capability is also influenced by QS [61], where genes involved in extracellular polysaccharide biosynthesis (e.g., RCAP_rcc01081 and RCAP_rcc01932), which are under QS control and necessary for the formation of the cell's capsule, are required for the adsorption of GTAs to the cell as the capsule acts as a receptor on the cell surface [47].

DNA strand stress is mediated by the SOS master regulator LexA. Single-stranded DNA (ssDNA), for example derived from DNA damage, interacts with the recombinase RecA. This complex then binds to the transcriptional regulator LexA and promotes its autoproteolysis protease activity. In *R. capsulatus* LexA autoproteolysis leads to its release from the *cckA* promoter region which can then be transcribed and CckA can activate GTA production via CtrA (Figure 1.3). Curiously, DNA damage induced by mitomycin C does not affect GTA production in *R. capsulatus* but DNA-damaging antibiotics induce the activation of VSH-1, the *Brachyspira hyodysenteriae* GTA [68]. In *D. shibae* determination of transcription factor binding sites revealed that all CtrA-regulated traits are also part of the LexA regulon, besides the CtrA phosphorelay itself and the GTA gene cluster (Figure 1.2). *Vice versa*, a CtrA binding site on the *lexA* promoter has been identified in *D. shibae*. Thus, the connection between the LexA and CtrA regulatory systems differs between *R. capsulatus* and *D. shibae* [69,70].

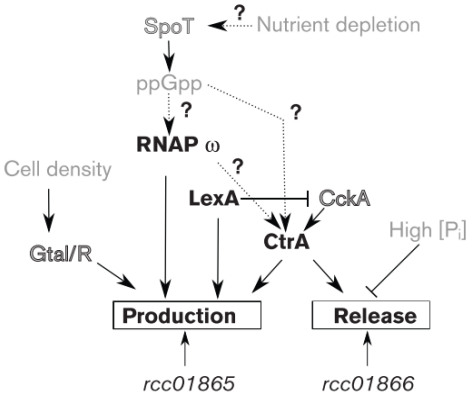

Figure 1.3: Possible regulatory network controlling GTA production in *R. capsulatus*. Regulators that might be essential (black) or are involved (grey) in GTA production and release are indicated. Connections found in other organisms are marked by questions marks (dashed lines). Taken from [62].

The combination of QS on a collective level and stress response on an individual level have been considered as a mechanism that allows a more fine-tuned adaptation to changing environments [71]. Similar regulation in other bacteria, such as *Pseudomonas aeruginosa* and *Bacillus subtilis*, has been found to control processes such as antibiotic resistance, horizontal gene transfer and virulence [69].

Finally, (p)ppGpp (guanosine 3‘, 5‘-bisdiphosphate) is a bacterial alarmone that is generated in response to a variety of environmental signals. It is hydrolyzed and synthesized by the bifunctional enzymes of the RelA/SpoT protein family from and to ATP, GDP or GTP [72]. In most Alphaproteobacteria RelA/SpoT, collectively referred to as SpoT, is expressed in response to stress, entry to stationary phase, or nutrient (e.g., nitrogen or carbon) starvation. In *C. crescentus* and other

Alphaproteobacteria, this system is known to control CtrA [72] with (p)ppGpp levels inducing a non-replicative motile cell state and other secondary messengers inducing reproduction, entry into S-phase, and adhesive structures for a sessile lifestyle [73]. In *R. capsulatus* (p)ppGpp levels decrease pigmentation, photosynthesis, and GTA production in response to amino acid, carbon and, to a lesser extent, phosphate depletion [62,72,74]. In *Bartonella* (p)ppGpp was shown to be synthesized upon nitrogen and carbon starvation and to repress GTA production in these conditions. In contrast, low levels of this alarmone were found to coincide with optimal cell growth and GTA production in this system. However, GTA DNA uptake is not activated by low levels of (p)ppGpp but it relies on a protein complex also required for outer membrane invagination during cell division [75].

1.2.3 Partner-switching system and GafA

As discussed above, the loss of CtrA in *R. capsulatus* has far-reaching implications for a variety of genes and systems. These include genes predicted to encode an anti-sigma factor, anti-anti-sigma factor, phosphatase partner-switching system (RsbW, RsbV, and RsbY) [55] with sequence homology to a system that has been intensively investigated for its role in stress response, motility, and spore formation in Gram-positive bacteria [76-78]. When the respective genes were knocked out in *R. capsulatus*, GTA production increased in Δ*rbaW* and decreased in Δ*rbaV* and Δ*rbaY* strains compared to the wild type [79]. This effect is at least partly due to a change in the size of the subpopulation and increased expression of GTA. These proteins were also shown to be involved in flagellar motility. It is presumed that an additional environmental signal is integrated via RbaY and transmitted to RbaV by dephosphorylation (Figure 1.4) [79]. However, the phenotypes of the different

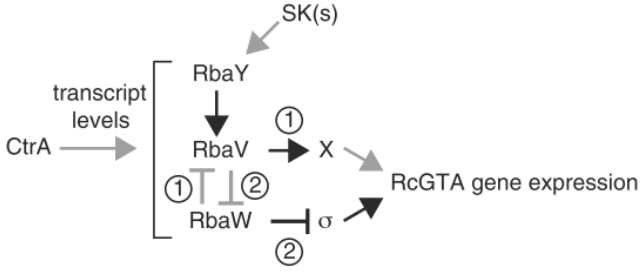

Figure 1.4: Regulatory interaction of the *R. capsulatus* partner switching system. CtrA activates the three components (RbaY, RbaV and RbaW) transcriptionally and an unknown signal integrates on protein level via phosphorylation of RbaY. RbaV is activated by RbaY. ① RbaV indirectly activates GTA synthesis and RbaW inhibits GTA production. ② RbaV binds to RbaW to release its inhibition of a sigma factor and thus to stimulate GTA production. Taken from [79].

mutant strains did not match with what would be expected of a sigma factor-regulating system and the exact functions of RbaV and RbaW and their targets remain unknown. However, this partner-switching system is located downstream of the CtrA phosphorelay and upstream of the GTA expression and thus represents an additional regulatory intermediate. Alternatively, those three systems could represent a triangulation where CtrA controls the partner-switching system as well as GTA gene expression, and the activated partner-switching system then integrates the additional environmental signal and regulates GTA expression to fine tune GTA synthesis.

GafA has been suggested as a possible sigma factor, possibly targeted by RbaW [56]. This transcriptional regulator is part of the CtrA and GtaR regulons. Protein binding assays have shown direct binding of CtrA to the *gafA* promoter and GafA binding to the GTA gene cluster promoter (Figure 1.5). Comparison of the transcriptomes of two *R. capsulatus* strains, the wild type strain SB1003 and a GTA-

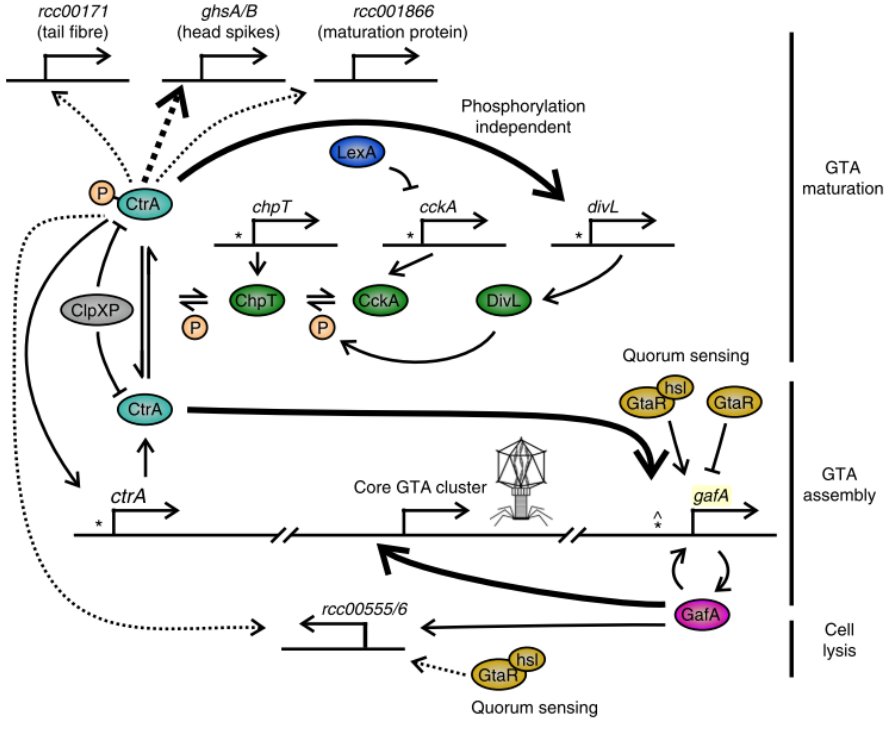

Figure 1.5: Regulation of GTA in *R. capsulatus*. Binding sites of CtrA (*) and GtaR (^) are shown. Proteins and promoter regions are colored and indicated by kinked arrows, respectively. Direct, indirect/unknown regulation and essential regulation are shown by solid, dashed, and bold arrows, respectively. Taken from [56].

overproducing strain DE442, characterized by subpopulations of GTA-producing cells corresponding to <3% and >30% [36], respectively, indicated that GafA may be involved in the determination of the size of the subpopulation that synthesizes GTA [56]. However, another gene, RCAP_rcc00280, was also identified to be relevant for determining the size of this subpopulation. This gene was also found to be mutated in the overproducing strain (DE442). Thus, it was suggested that this gene is additionally located upstream of GafA in the signaling pathway [80].

Transcriptomic and *in silico* binding site predictions showed that CtrA directly binds *gafA* in *D. shibae* and is thus indirectly controlled by the master QS synthase $LuxI_1$ [57]. A mutant lacking the autoinducer synthase $LuxI_2$ shows greatly increased production of GTAs, which is why it is reasonable to assume that *luxI*$_2$ and *gafA* have some genetic interaction [57].

1.2.4 Cyclic dimeric guanosine monophosphate

Bis-(3'-5')-cyclic dimeric guanosine monophosphate (c-di-GMP) is synthesized from two guanosine triphosphate (GTP) molecules by diguanylate cyclases (DGCs) and hydrolyzed to two molecules of guanosine monophosphate (GMP) by phosphodiesterases (PDEs). There are also bifunctional enzymes that can both synthesize and hydrolyze c-di-GMP and thus represent a “biochemical conundrum” [81]. C-di-GMP can be found in all domains of life and allows organisms to adapt to changing environments by controlling phenotypic heterogeneity. To do so, the cellular concentrations of c-di-GMP is modified in response to a variety of external and internal stimuli and thereby controls a variety of other features. This dinucleotide second messenger’s enzymes are associated with numerous sensory domains and most bacteria contain multiple enzymes for c-di-GMP production and hydrolysis.

Like (p)ppGpp, c-di-GMP is a second messenger that can regulate GTA production [82]. Among members of the Alphaproteobacteria, this molecule is best known for its role in the cell cycle of *C. crescentus*, where it is a driving force for cell development and differentiation into the flagellated and stalked cells through inhibition of CtrA. On the contrary, (p)ppGpp, which is made in response to starvation, indirectly stabilizes CtrA levels and inhibits cell cycle progression [83]. The levels of c-di-GMP vary as the bacterium progresses through the cell cycle. Increased concentrations of c-di-GMP are

able to change the function of CckA from a kinase to a phosphatase [84], causing a reduction of intracellular CtrA~P levels. This change induces cell division and the formation of the stalked cell pole. Increased levels of intracellular CtrA~P lead to the formation of the flagellum and hold the cell in the G1 phase [85].

Similarly in *R. capsulatus*, c-di-GMP stimulates the phosphatase activity of CckA and thus increased levels of unphosphorylated CtrA, which controls the maturation process of the GTA particle, while low c-di-GMP and high CtrA~P concentrations lead to cell lysis and release of GTA particles (Figure 1.6) [60]. *R. capsulatus* has two genes each with DGC and PDE domains and 16 genes with both DGC and PDE domains. Eight of the involved genes are part of the CtrA regulon and four of those (1 DGC and 3 DGC/PDE) affect GTA gene expression and production as well as motility [82]. At least part of these effects could be through effects on CckA, but there could also be other c-di-GMP-GTA connections that remain to be discovered [82].

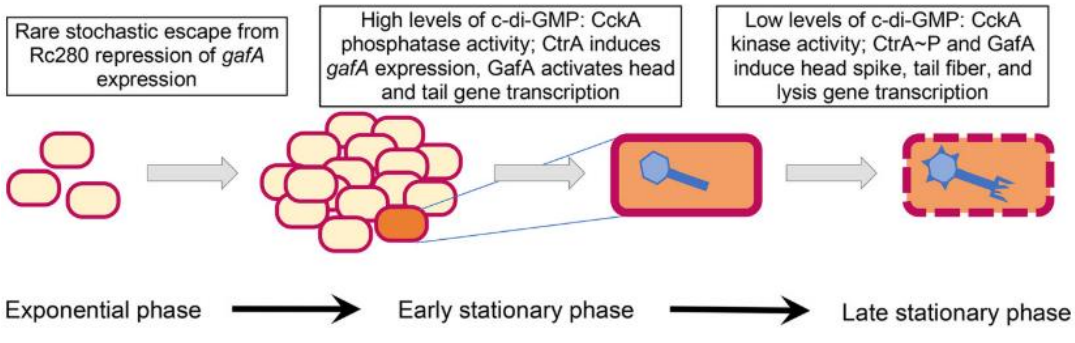

Figure 1.6: RcGTA regulation by c-di-GMP. At a low cell density during the exponential phase, GTA expression is repressed. Increased cell density at the early stationary phase releases the inhibition of *gafA* by RCAP_rcc00280. GafA activates the transcription of the RcGTA head and tail structural genes, as well as the *gafA* gene itself. Since c-di-GMP levels are high, CckA acts as phosphatase and leaves CtrA unphosphorylated, which induces *gafA* expression. This increases RcGTA head and tail synthesis. An unknown signal decreases c-di-GMP levels and CckA switches its function to a kinase. This increases the concentration of phosphorylated CtrA and activates transcription of head spike, tail fiber and lysis proteins. This allows the release of the mature RcGTA particle. Taken from [60].

D. shibae has two DGCs and two PDEs [86]. Interestingly, both DGC- and both PDE-encoding genes contain *in silico*-determined CtrA or LexA transcription factor binding sites, respectively (Figure 1.2) [57]. The regulation via CtrA by the QS synthase could also be observed *in vivo* when an external autoinducer was added to the synthase knockout. After the addition, time-resolved observations revealed a gradual increase of the c-di-GMP concentration in the cell, simultaneous with the increase of the protein level of the GTA major capsid protein, flagellar gene transcription and the activation of transcription of other CtrA-regulated genes. In addition, flow cytometric measurements revealed an interaction between GTA and replication, since fewer cells were dividing when GTA gene expression started [57].

1.3 Gene expression regulation through chromosomal gene location

Gene expression is regulated at multiple levels, e.g., directly by transcriptional regulators or indirectly because of the localization of the genes within the genome. However, the influence of genomic localization on gene expression has received much less attention, even though many examples exist from different domains of life for how the expression of specific genes is influenced by genomic localization, e.g. in the Gram-negative bacterium *Vibrio cholerae*, in the Gram-positive bacterium *Bacillus subtilis*, and due to chromosomal ploidies in eukaryotes [87].

1.3.1 Structured regions

The localization of genes in the genome is not random. On one hand, the distance between genes on the coiled DNA strand is important. In *E. coli*, for example, operons tend to be arranged to minimize the spatial distance between operons in the same regulon or involved in the same biological pathway. Genes that are spatially close on the coiled chromosome tend to be co-expressed and the resulting proteins can more easily interact with each other [88,89]. On the other hand, the gene order on the chromosome is also important. The *E. coli* chromosome, for example, is divided in six zones, four macro-domains and two less structured regions, which were defined on the basis of their recombination efficiency since the mobility of genes is constrained between the macro-domains and, to a lesser extent, in the less structured regions [90-92]. There are multiple examples of how genomic location is relevant for gene expression, e.g., essential genes are often found closer to the origin of replication (*ori*) and on the leading strand of DNA replication

and phage integration sites are primarily identified in the proximity of *ter* [93,94]. The preferential position on the leading strand is probably due to the fact that there are fewer clashes between the transcription and replication machineries and thus a lower mutation rate [95].

However, for many of the observed patterns, there are still no explanations. For example, it has been observed that in circular bacterial genomes there is an excess of guanines and cytosines on the leading and lagging strands, respectively. This observation is currently used to determine the positions of *ori* and *ter* because leading and lagging strands flip at *ori* and *ter*, respectively, due to the bidirectional replication of circular genomes. However, the significance of stronger differences in guanine and cytosine occurrence on leading and lagging strands is not known.

1.3.2 Regulatory impact of genomic arrangements

One reason for specific genomic arrangements of genes is for the purpose of regulation. For example, during sporulation of the Gram-positive bacterium *Bacillus subtilis*, the replicated chromosome progressively pushed into a spore, starting with the *ori*. As a result, genes near the *ori* and *ter* regions are expressed in the spore and mother cell, respectively. One sigma factor is encoded proximate to *ori*, and therefore the genes it regulates are expressed differentially during chromosomal transfer into the spore without being exposed to the effects of the sigma factor-inhibiting anti-sigma factor, which due to the location of its gene close location to *ter* is not yet expressed in the spore. The genes regulated by the sigma factor include those relevant to spore development, and thus their activation occurs during transfer of the chromosome into the spore and they are down-regulated after the complete chromosome is shifted and the inhibitory anti-sigma factor can act in the forespore [96]. A similar phenomenon can be observed in the differentiation of the two cell poles during cell division in *C. crescentus* [97]. Flow cytometric data of *D. shibae* have suggested that packaging of DNA into GTAs is timed based on the replication progress and thus might be dependent on genomic properties of the genes encoding its regulators [57].

In addition, DNA replication might also be linked to gene localization and regulation. Depending on the growth of the bacterium, multiple rounds of replication initiation can take place within the same cell, possibly leading to higher expression levels of genes located in proximity to *ori* due to their higher copy

number in the cell [98]. For example, *E. coli* cells can contain up to 8 copies of *ori* in one cell [98]. Consequently, the two following factors can influence gene regulation [87]. On the one hand, the periodic fluctuations in the copy number around the chromosome [99] and, on the other hand, the amplitude of the copy number, which results from the number of replications occurring simultaneously [100]. This can lead to a temporal imbalance between different components of a network depending on their locations on the chromosome. In *B. subtilis* this temporal imbalance of the translation of components of the same regulatory network has been shown to allow a well-timed onset of spore formation coupled to the cell cycle, allowing only full, non-damaged chromosomes to be transferred into each spore [99]. In the gammaproteobacterium *Vibrio cholerae* the translocation of the operon encoding the S10 ribosomal protein from the proximity of *ori* to the proximity of *ter* reduced the copy number during fast growth from 3 to 1.2, which abolished growth and infectivity. However, the addition of a second operon that led to copy numbers of 1.2 and 1.9, in total 3.1, during growth restored the phenotype [100]. Since there seems to be no imbalance of GTA-packaged DNA towards the *ori*, particularly in *D. shibae*, it is possible that similar to spore formation, DNA is packaged only at the end of the replication cycle [57].

1.4 Biological databases

In recent decades, the amount of biological data available has exploded, and its accumulation is accelerating. This is mainly due to the improving high-throughput technologies and their accessibility at decreasing costs. These data are generated for different purposes, resulting in different data types acquired with different methods. This amount and variability of data and the resulting diverse databases make a good management system necessary [101]. For example, there is a need for application of programming interfaces that enable automated combination and exchange between different databases. These data can then be optimally used to determine patterns, for example the preferred integration sites of phages or externally derived DNA in genomes, or their association with genome size and neighborhood core genes. Comparative analysis of these data can form the basis for future wet lab experiments that ensure well-founded decision-making and are therefore an important tool in biological research [102,103].

1.5 Research goals

In my thesis, I used comparative analyses of publicly available data to investigate the integration of GTAs in regulatory networks of their host cells and examined conserved properties of these networks. Furthermore, I aimed to identify genomic properties that could help to explain why GTAs are conserved in Alphaproteobacteria, and in particular in the order Rhodobacterales.

In chapter two, I aimed to further expand our understanding of the gene regulatory network controlling GTA gene expression. Therefore, I re-analyzed published datasets that documented the cellular response to oxygen and nitric oxide concentrations in the environment in *D. shibae* and *R. capsulatus*. I investigated the extent to which those environmental factors and the CtrA regulon interplay. When the data were originally analyzed the regulation of the conserved flagella was identified but the possible interaction with the CtrA phosphorelay and GTA gene expression was neglected [104,105]. I was able to show the incorporation of respiration and denitrification signals into the GTA regulatory network.

In chapter three I aimed to clarify whether the chromosomal locations of important regulators associated with the CtrA phosphorelay show noticeable, conserved patterns. Therefore, I compared the genomic contexts of these genes in five alphaproteobacterial orders. I was able to describe interesting chromosomal patterns that have the potential to affect GTA expression. In the Rhodobacterales the localization pattern suggests an uncoupling of the cell cycle and activity of CtrA targets which might ensure that GTA genes are only expressed once replication has finished.

In chapter four I aimed to determine genomic properties that might help to explain why GTA are so widely conserved in alphaproteobacteria. The study is an in-depth analysis of the GC skews, core gene locations, plasticity regions, and methylation patterns associated with GTA gene clusters. I found that GTA genes share multiple properties with core genes, particularly in the Rhodobacterales. This might help to explain their conservation, especially in this order.

In chapter five I aimed to investigate chromosomal patterns for the genes that control c-di-GMP concentrations in the cell in order to identify genomic commonalities between these highly diverse genes. Therefore, I performed a comparative analysis of the c-di-GMP synthesizing and hydrolyzing enzymes

that could have an important influence on the CtrA phosphorelay and GTAs [60,82,106]. Here, I found some interesting patterns, that suggest a coordination of c-di-GMP concentrations with the cell cycle or the establishment of a concentration gradient along the *ori-ter* axis in some of the studied bacteria.

1.6 References

1. Muñoz-Gómez, S.A.; Hess, S.; Burger, G.; Franz Lang, B.; Susko, E.; Slamovits, C.H.; Roger, A.J. An updated phylogeny of the alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. *Elife* **2019**, *8*, e42535.
2. Ettema, T.J.G.; Andersson, S.G.E. The α-proteobacteria: The Darwin finches of the bacterial world. *Biol. Lett.* **2009**, *5*, 429–432.
3. Giovannoni, S.J.; Tripp, H.J.; Givan, S.; Podar, M.; Vergin, K.L.; Baptista, D.; Bibbs, L.; Eads, J.; Richardson, T.H.; Noordewier, M.; et al. Genetics: Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **2005**, *309*, 1242–1245.
4. Martijn, J.; Vosseberg, J.; Guy, L.; Offre, P.; Ettema, T.J.G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **2018**, *557*, 101–105.
5. Wang, Z.; Wu, M. Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. *PLoS One* **2014**, *9*, e110685.
6. Dinasquet, J.; Landa, M.; Obernosterer, I. High contribution of Pelagibacterales to bacterial community composition and activity in spring blooms off Kerguelen Island (Southern Ocean). *bioRxiv* **2019**, 633925.
7. Ortmann, A.C.; Santos, T.T.L. Spatial and temporal patterns in the Pelagibacteraceae across an estuarine gradient. *FEMS Microbiol. Ecol.* **2016**, *92*, fiw133.
8. Lynch, M. Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.* **2006**, *60*, 327–349.
9. Blazejak, K.; Janecek, E.; Strube, C. A 10-year surveillance of Rickettsiales (*Rickettsia* spp. and *Anaplasma phagocytophilum*) in the city of Hanover, Germany, reveals *Rickettsia* spp. as emerging pathogens in ticks. *Parasites and Vectors* **2017**, *10*, 588.

10. Darby, A.C.; Cho, N.H.; Fuxelius, H.H.; Westberg, J.; Andersson, S.G.E. Intracellular pathogens go extreme: genome evolution in the Rickettsiales. *Trends Genet.* **2007**, *23*, 511–520.

11. Yu, XueJie, and Walker, D. H. "The order Rickettsiales." Prokaryotes 5 (2006): 493-528.

12. Hördt, A.; López, M.G.; Meier-Kolthoff, J.P.; Schleuning, M.; Weinhold, L.M.; Tindall, B.J.; Gronow, S.; Kyrpides, N.C.; Woyke, T.; Göker, M. Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of Alphaproteobacteria. *Front. Microbiol*. **2020**, *11*. 468.

13. Nakai, R., Kusada, Tmpk, B.S.; Soil, I. Draft genome sequence of novel filterable Rhodospirillales. *Microbiol. Resorce Announc*. **2021**, *10*. e0039321.

14. Tani, K.; Kanno, R.; Ji, X.C.; Hall, M.; Yu, L.J.; Kimura, Y.; Madigan, M.T.; Mizoguchi, A.; Humbel, B.M.; Wang-Otomo, Z.Y. Cryo-EM Structure of the photosynthetic LH1-RC complex from *Rhodospirillum rubrum*. *Biochemistry* **2021**, *60*, 2483–2491.

15. Vermeglio, R.A.; Lutz, M. More than two structurally distinct types of antenna in Rhodospirillales. In *In Advances in Photosynthesis Research*; Springer Netherlands; pp. 199–202.

16. Erkosar, B.; Storelli, G.; Defaye, A.; Leulier, F. Host-intestinal microbiota mutualism: "learning on the fly." *Cell Host and Microbe* **2013**, *13*, 8–14.

17. Mogollon-Pasapera, E.; Otvos, L.; Giordano, A.; Cassone, M. *Bartonella*: emerging pathogen or emerging awareness? *Int. J. Infect. Dis.* **2009**, *13*, 3–8.

18. Staley, J.T. *Prosthecomicrobium* and *Ancalomicrobium*: new prosthecate freshwater bacteria. *J. Bacteriol.* **1968**, *95*, 1921–1942.

19. Stahl, D.A.; Key, R.; Flesher, B.; Smit, J. The phylogeny of marine and freshwater Caulobacters reflects their habitat. *J. Bacteriol.* **1992**, *174*, 2193–2198.

20. Wang, S.; Meade, A.; Lam, H.-M.; Luo, H. Evolutionary timeline and genomic plasticity underlying the lifestyle diversity in Rhizobiales. *mSystems* **2020**, *5*. e00438-20.

21. Escobar, M.A.; Dandekar, A.M. *Agrobacterium tumefaciens* as an agent of disease. *Trends Plant. Sci.* **2003**, *8*, 380-386.

22. Carvalho, F.M.; Souza, R.C.; Barcellos, F.G.; Hungria, M.; Vasconcelos, A.T.R. Genomic and evolutionary comparisons of diazotrophic and pathogenic bacteria of the order Rhizobiales. *BMC Microbiol.* **2010**, *10*. 37.

23. Buchan, A.; González, J.M.; Moran, M.A. Overview of the marine Roseobacter lineage. *Appl. Environ. Microbiol.* **2005**, *71*, 5665–5677.

24. Simon, M.; Scheuner, C.; Meier-Kolthoff, J.P.; Brinkhoff, T.; Wagner-Döbler, I.; Ulbrich, M.; Klenk, H.-P.; Schomburg, D.; Petersen, J.; Göker, M. Phylogenomics of Rhodobacteraceae reveals evolutionary adaptation to marine and non-marine habitats. *ISME J.* **2017**, *11*, 1–17.

25. Liang, K.Y.H.; Orata, F.D.; Boucher, Y.F.; Case, R.J. Roseobacters in a sea of poly- and paraphyly: Whole genome-based taxonomy of the family Rhodobacteraceae and the proposal for the split of the "Roseobacter Clade" into a novel family, Roseobacteraceae fam. nov. *Front. Microbiol.* **2021**, *12*. 683109.

26. Kogay, R.; Neely, T.B.; Birnbaum, D.P.; Hankel, C.R.; Shakya, M.; Zhaxybayeva, O. Machine-learning classification suggests that many alphaproteobacterial prophages may instead be gene transfer agents. *Genome Biol. Evol.* **2019**, *11*, 2941–2953.

27. Marrs, B. Genetic Recombination in *Rhodopseudomonas capsulata*. *Proc. Natl. Acad. Sci. U S A* **1974**, *71*, 971–973.

28. Weaver, P.F.; Wall, J.D.; Gest, H. Characterization of *Rhodopseudomonas capsulata*. *Arch. Microbiol.* **1975**, *105*, 207–216.

29. Lang, A.S.; Westbye, A.B.; Beatty, J.T. The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu. Rev. Virol.* **2017**, *4*, 87–104.

30. Madigan and Jung, and *The Purple Phototrophic Bacteria*; 2009; Vol. 47; ISBN 9781402088148.

31. Tomasch, J.; Wang, H.; Hall, A.T.K.; Patzelt, D.; Preuße, M.; Brinkmann, H.; Bhuju, S.; Jarek, M.; Geffers, R.; Lang, A.S. Packaging of *Dinoroseobacter shibae* DNA into gene transfer agent particles is not random. *Genome Biol. Evol.* **2018**, *10*, 359–369.

32. Strnad, H.; Lapidus, A.; Paces, J.; Ulbrich, P.; Vlcek, C.; Paces, V.; Haselkorn, R. Complete

genome sequence of the photosynthetic purple nonsulfur bacterium *Rhodobacter capsulatus* SB 1003. *J. Bacteriol.* **2010**, *192*, 3545–3546.

33. Biebl, H.; Allgaier, M.; Tindall, B.J.; Koblizek, M.; Lünsdorf, H.; Pukall, R.; Wagner-Döbler, I. *Dinoroseobacter shibae* gen. nov., sp. nov., a new aerobic phototrophic bacterium isolated from dinoflagellates. *Int. J. Syst. Evol. Microbiol.* **2005**, *55*, 1089-1096.

34. Wagner-Döbler, I.; Ballhausen, B.; Berger, M.; Brinkhoff, T.; Buchholz, I.; Bunk, B.; Cypionka, H.; Daniel, R.; Drepper, T.; Gerdts, G.; et al. The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: a hitchhiker's guide to life in the sea. *ISME J.* **2010**, *4*, 61–77.

35. Lang, A.S.; Beatty, J.T. Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* **2007**, *15*, 54–62.

36. Hynes, A.P.; Shakya, M.; Mercer, R.G.; Grüll, M.P.; Bown, L.; Davidson, F.; Steffen, E.; Matchem, H.; Peach, M.E.; Berger, T.; et al. Functional and evolutionary characterization of a gene transfer agent's multilocus "Genome." *Mol. Biol. Evol.* **2016**, *33*, 2530–2543.

37. Lang, A.S.; Beatty, J.T. Genetic analysis of a bacterial genetic exchange element: The gene transfer agent of *Rhodobacter capsulatus*. *Proc. Natl. Acad. Sci.* **2000**, *97*, 859–864.

38. Hynes, A.P.; Mercer, R.G.; Watton, D.E.; Buckley, C.B.; Lang, A.S. 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.* **2012**, *85*, 314–325.

39. Shakya, M.; Soucy, S.M.; Zhaxybayeva, O. Insights into origin and evolution of α -proteobacterial gene transfer agents. Virus Evol. 2017, 3, vex036.

40. Lang, A.S.; Zhaxybayeva, O.; Beatty, J.T. Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* **2012**, *10*, 472–82.

41. Barbian, K.D.; Minnick, M.F. A bacteriophage-like particle from *Bartonella bacilliformis*. *Microbiology* **2000**, *146*, 599–609.

42. Berglund, E.C.; Frank, A.C.; Calteau, A.; Pettersson, O.V.; Granberg, F.; Eriksson, A.S.; Näslund,

K.; Holmberg, M.; Lindroos, H.; Andersson, S.G.E. Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting *Bartonella grahamii*. *PLoS Genet.* **2009**, *5*, e1000546.

43. Lang, A.S.; Beatty, J.T. Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. *Proc. Natl. Acad. Sci. U S A* **2000**, *97*, 859–864.

44. Fogg, P.C.M.; Westbye, A.B.; Beatty, J.T. 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* **2012**, *7*, e43772.

45. Seitz, P.; Blokesch, M. Cues and regulatory pathways involved in natural competence and transformation in pathogenic and environmental Gram-negative bacteria. *FEMS Microbiol. Rev.* **2013**, *37*, 336–363.

46. Kuchinski, K.S.; Brimacombe, C.A.; Westbye, A.B.; Ding, H.; Beatty, J.T. The SOS response master regulator LexA regulates the gene transfer agent of *Rhodobacter capsulatus* and represses transcription of the signal transduction protein CckA. *J. Bacteriol.* **2016**, *198*, 1137–1148.

47. Brimacombe, C.A.; Stevens, A.; Jun, D.; Mercer, R.; Lang, A.S.; Beatty, J.T. Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). *Mol. Microbiol.* **2013**, *87*, 802–817.

48. Westbye, A.B.; Kuchinski, K.; Yip, C.K.; Beatty, J.T. The gene transfer agent RcGTA contains head spikes needed for binding to the *Rhodobacter capsulatus* polysaccharide cell capsule. *J. Mol. Biol.* **2016**, *428*, 477–491.

49. Westbye, A.B.; Beatty, J.T.; Lang, A.S.; Rice, P. Guaranteeing a captive audience: coordinated regulation of gene transfer agent (GTA) production and recipient capability by cellular regulators This review comes from a themed issue on mobile genetic elements and HGT in prokaryotes. *Curr. Opin. Microbiol.* **2017**, *38*, 122–129.

50. Brilli, M.; Fondi, M.; Fani, R.; Mengoni, A.; Ferri, L.; Bazzicalupo, M.; Biondi, E.G. The diversity and evolution of cell cycle regulation in alphaproteobacteria: a comparative genomic

analysis. *BMC Systems Biology*; **2010**, 4, 52.

51. Francez-Charlot, A.; Kaczmarczyk, A.; Vorholt, J.A. The branched CcsA/CckA-ChpT-CtrA phosphorelay of *Sphingomonas melonis* controls motility and biofilm formation. *Mol. Microbiol.* **2015**, *97*, 47–63.

52. Laub, M.T.; Chen, S.L.; Shapiro, L.; Mcadams, H.H. Genes directly controlled by CtrA, a master regulator of the *Caulobacter* cell cycle. *PNAS* **2002**, *99*, 4632–4637.

53. Greene, S.E.; Brilli, M.; Biondi, E.G.; Komeili, A. Analysis of the CtrA pathway in *Magnetospirillum* reveals an ancestral role in motility in Alphaproteobacteria. *J. Bacteriol.* **2012**, *194*, 2973–2986.

54. Panis, G.L.; Murray, S.R.; Viollier, P.H. Versatility of global transcriptional regulators in Alphaproteobacteria: From essential cell cycle control to ancillary functions. *Microbiol. Rev.* **2015**, *39*, 120–133.

55. Mercer, R.G.; Callister, S.J.; Lipton, M.S.; Pasa-tolic, L.; Strnad, H.; Paces, V.; Beatty, J.T.; Lang, A.S.; Acteriol, J.B. Loss of the response regulator *ctrA* causes pleiotropic effects on gene expression but does not affect growth phase regulation in *Rhodobacter capsulatus*. *J. Bacteriol.* **2010**, *192*, 2701–2710.

56. Fogg, P.C.M. Identification and characterization of a direct activator of a gene transfer agent. *Nat. Commun.* **2019**, *10*, 595

57. Koppenhöfer, S.; Wang, H.; Scharfe, M.; Kaever, V.; Wagner-Döbler, I.; Tomasch, J. Integrated transcriptional regulatory network of quorum sensing, replication control, and SOS response in *Dinoroseobacter shibae*. *Front. Microbiol.* **2019**, *10*, 803.

58. Narayanan, S.; Kumar, L.; Kumar Radhakrishnan, S. Sensory domain of the cell cycle kinase CckA regulates the differential DNA binding of the master regulator CtrA in *Caulobacter crescentus. BBA - Gene Regul. Mech.* **2018**, 952–961.

59. Weston, B.R.; Tyson, J.J.; Cao, Y. Computational modeling of unphosphorylated CtrA: Cori binding in the *Caulobacter* cell cycle. *iScience* **2021**, *24*, 103413.

60. Farrera-Calderon, R.G.; Pallegar, P.; Westbye, A.B.; Wiesmann, C.; Lang, A.S.; Beatty, J.T. The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer agent (RcGTA) production is bi-directional and regulated by cyclic-di-GMP. *J. Bacteriol.* **2020.**

61. Leung, M.M.; Brimacombe, C.A.; Spiegelman, G.B.; Beatty, J.T. The GtaR protein negatively regulates transcription of the *gtaRI* operon and modulates gene transfer agent (RcGTA) expression in *Rhodobacter capsulatus. Mol. Microbiol.* **2012**, *83*, 759–774.

62. Westbye, A.B.; O'Neill, Z.; Schellenberg-Beaver, T.; Beatty, J.T. The *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology* **2017**, *163*, 1355–1363.

63. Bassler, B.L.; Losick, R. Bacterially Speaking. *Cell* **2006**, *125*. 237-246.

64. Papenfort, K.; Bassler, B.L. Quorum sensing signal–response systems in Gram-negative bacteria. *Nat. Rev. Microbiol.* **2016**, *14*, 576-588.

65. Patzelt, D.; Wang, H.; Buchholz, I.; Rohde, M.; Gröbe, L.; Pradella, S.; Neumann, A.; Schulz, S.; Heyber, S.; Münch, K.; et al. You are what you talk: quorum sensing induces individual morphologies and cell division modes in *Dinoroseobacter shibae. ISME J.* **2013**, *7*, 2274–86.

66. Neumann, A.; Patzelt, D.; Wagner-Döbler, I.; Schulz, S. Identification of new N-acylhomoserine lactone signalling compounds of *Dinoroseobacter shibae* DFL-12T by overexpression of luxI genes. *ChemBioChem* **2013**, *14*, 2355–2361.

67. Wang, H.; Ziesche, L.; Frank, O.; Michael, V.; Martin, M.; Petersen, J.; Schulz, S.; Wagner-Döbler, I.; Tomasch, J. The CtrA phosphorelay integrates differentiation and communication in the marine alphaproteobacterium *Dinoroseobacter shibae*. *BMC Genomics* **2014**, *15*, 130.

68. Stanton, T.B.; Humphrey, S.B.; Sharma, V.K.; Zuerner, R.L. Collateral effects of antibiotics: Carbadox and metronidazole induce VSH-1 and facilitate gene transfer among *Brachyspira hyodysenteriae* strains. *Appl. Environ. Microbiol.* **2008**, *74*, 2950–2956.

69. Kamenšek, S.; Podlesek, Z.; Gillor, O.; Žgur-Bertok, D. Genes regulated by the *Escherichia coli* SOS repressor LexA exhibit heterogenous expression. *BMC Microbiol.* **2010**, *10*, 1471–2180.

70. Mellies, J.; Haack, K.; Galligan, D. SOS regulation of the type III secretion system of enteropathogenic *Escherichia coli*. *J. Bacteriol.* **2007**, *189*, 2863–2872.

71. Singh, P.K.; Bartalomej, S.; Hartmann, R.; Jeckel, H.; Vidakovic, L.; Nadell, C.D.; Drescher, K. *Vibrio cholerae* combines individual and collective sensing to trigger biofilm dispersal. *Current Biology* **2017**, *27*, 3359–3366.

72. Hallez, R.; Delaby, M.; Sanselicio, S.; Viollier, P.H. Hit the right spots: Cell cycle control by phosphorylated guanosines in Alphaproteobacteria. *Nat. Rev. Microbiol.* **2017**, *15*, 137–148.

73. Wang, J. D.; Sanders, G. M.; Grossman, A. D. Nutritional control of elongation of DNA replication by (p)ppGpp. *Cell* **2007**, *128*, 865–875.

74. Masuda, S.; Bauer, C.E. Null mutation of *hvrA* compensates for loss of an essential *relA/spoT*-like gene in *Rhodobacter capsulatus*. *J. Bacteriol.* **2003**, *186*, 235–239.

75. Québatte, M.; Christen, M.; Harms, A.; Körner, J.; Christen, B.; Dehio, C. Gene transfer agent promotes evolvability within the fittest subpopulation of a bacterial pathogen. *Cell Syst.* **2017**, *4*, 611-621.e6.

76. Hecker, M.; Pané-Farré, J.; Völker, U. SigB-dependent general stress response in *Bacillus subtilis* and related gram-positive bacteria. *Annu. Rev. Microbiol.* **2007**, *61*, 215–236.

77. Bartolini, M.; Cogliati, S.; Vileta, D.; Bauman, C.; Rateni, L.; Leñini, C.; Argañaraz, F.; Francisco, M.; Villalba, J.M.; Steil, L.; et al. Regulation of biofilm aging and dispersal in *Bacillus subtilis* by the alternative sigma factor SigB. *J. Bacteriol.* **2019**, *201*, e00473-18.

78. Rodriguez Ayala, F.; Bartolini, M.; Grau, R. The Stress-responsive alternative sigma factor SigB of *Bacillus subtilis* and its relatives: An old friend with new functions. *Front. Microbiol*. **2020**, 11, 1761.

79. Mercer, R.G.; Lang, A.S. 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.* **2014**, *14*, 71.

80. Ding, H.; Grüll, M.P.; Mulligan, M.E.; Lang, A.S.; Beatty, J.T. Induction of *Rhodobacter*

capsulatus gene transfer agent gene expression is a bistable stochastic process repressed by an extracellular calcium-binding RTX protein homologue. *J. Bacteriol.* **2019**, *201*, e00430-19.

81. Chou, S. H., Guiliani, N., Lee, V.T., Römling, U. Microbial cyclic di-nucleotide signaling. **2020**, *822*. Cham, Switzerland: Springer, 2020.
82. Pallegar, P.; Peña-Castillo, L.; Langille, E.; Gomelsky, M.; Lang, A.S. Cyclic di-GMP-mediated regulation of gene transfer and motility in *Rhodobacter capsulatus*. *J. Bacteriol.* **2020**, *202*, e00554-19.
83. Xu, C.; Weston, B.R.; Tyson, J.J.; Cao, Y. Cell cycle control and environmental response by second messengers in *Caulobacter crescentus*. *BMC Bioinformatics* **2020**, 21, 408.
84. Lori, C.; Ozaki, S.; Steiner, S.; Böhm, R.; Abel, S.; Dubey, B.N.; Schirmer, T.; Hiller, S.; Jenal, U. Cyclic di-GMP acts as a cell cycle oscillator to drive chromosome replication. *Nature* **2015**, *523*, 236–239.
85. Jenal, U.; Reinders, A.; Lori, C. Cyclic di-GMP: second messenger extraordinaire. *Nature Rev. Microbiol*. **2017**, *15*, 271-284.
86. Bedrunka, P.; Olbrisch, F.; Rüger, M.; Zehner, S.; Frankenberg-Dinkel, N. Nitric oxide controls c-di-GMP turnover in *Dinoroseobacter shibae*. *Microbiol. (United Kingdom)* **2018**, *164*, 1405–1415.
87. Slager, J.; Veening, J.-W. Hard-Wired control of bacterial processes by chromosomal gene location. *Trends in Microbiology* **2016**, *24*, 788–800.
88. Xie, T.; Fu, L.Y.; Yang, Q.Y.; Xiong, H.; Xu, H.; Ma, B.G.; Zhang, H.Y. Spatial features for *Escherichia coli* genome organization. *BMC Genomics* **2015**, *16*, 37.
89. Feuerborn, A.; Cook, P.R. Why the activity of a gene depends on its neighbors. *Trends Genet.* **2015**, *31*, 483–490.
90. Espeli, O.; Mercier, R.; Boccard, F. DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. *Mol. Microbiol.* **2008**, *68*, 1418–1427.
91. Valens, M.; Penaud, S.; Rossignol, M.; Cornet, F.; Boccard, F. Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J.* **2004**, *23*, 4330–4341.

92. Lioy, V.S.; Cournac, A.; Marbouty, M.; Duigou, S.; Mozziconacci, J.; Espéli, O.; Boccard, F.; Koszul, R. Multiscale structuring of the *E. coli* chromosome by nucleoid-associated and condensin Proteins. *Cell* **2018**, *172*, 771-783.e18.

93. Rocha, E.P.C.; Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **2003**, *31*, 6570–6577.

94. Kopejtka, K.; Lin, Y.; Jakubovičová, M.; Koblížek, M.; Tomasch, J. Clustered core- and pan-genome content on Rhodobacteraceae chromosomes. *Genome Biol. Evol.* **2019,** *11*, 2208-2217.

95. Lobry, J.R.; Sueoka, N. Asymmetric directional mutation pressures in bacteria. *Genome Biol.* **2002**, *3*, RESEARCH0058.

96. Errington, J. Regulation of endospore formation in *Bacillus subtilis*. *Nat. Rev. Microbiol.* **2003**, *1*, 117–126.

97. Goley, E.D.; Iniesta, A.A.; Shapiro, L. Cell cycle regulation in *Caulobacter*: Location, location, location. *J. Cell Sci.* **2007**, *120*, 3501–7.

98. Fossum, S.; Crooke, E.; Skarstad, K. Organization of sister origins and replisomes during multifork DNA replication in *Escherichia coli*. *EMBO J.* **2007**, *26*, 4514–4522.

99. Narula, J.; Kuchina, A.; Lee, D.D.; Fujita, M.; Süel, G.M.; Igoshin, O.A. Chromosomal arrangement of phosphorelay genes couples sporulation and DNA replication. *Cell* **2015**, *162*, 328–337.

100. Soler-Bistué, A.; Mondotte, J.A.; Bland, M.J.; Val, M.-E.; Saleh, M.-C.; Mazel, D. Genomic location of the major ribosomal protein gene locus determines *Vibrio cholerae* global growth and infectivity. *PLOS Genet.* **2015**, *11*, e1005156.

101. Zou, D.; Ma, L.; Yu, J.; Zhang, Z. Biological databases for human research. *Genomics, Proteomics Bioinforma.* **2015**, *13*, 55–63.

102. Vignani, R.; Liò, P.; Scali, M. How to integrate wet lab and bioinformatics procedures for wine DNA admixture analysis and compositional profiling: Case studies and perspectives. *PLoS One* **2019**, *14*, e0211962.

103. Nobrega, M.A.; Pennacchio, L.A. Comparative genomic analysis as a tool for biological discovery. *J. Physiol.* **2004**, *554*, 31–39.

104. Ebert, M.; Laaß, S.; Thürmer, A.; Roselius, L.; Eckweiler, D.; Daniel, R.; Härtig, E. and; Jahn, D. FnrL and three Dnr regulators are used for the metabolic adaptation to low oxygen tension in *Dinoroseobacter shibae. Front. Microbiol.* **2017**, *8*, 642.

105. Kumka, J.E.; Schindel, H.; Fang, M.; Zappa, S.; Bauer, C.E. Transcriptomic analysis of aerobic respiratory and anaerobic photosynthetic states in *Rhodobacter capsulatus* and their modulation by global redox regulators RegA, FnrL and CrtJ. *Microb. Genomics* **2017**, *3*. e000125.

106. Pallegar, P.; Canuti, M.; Langille, E.; Peña-Castillo, L.; Lang, A.S. A Two-Component System Acquired by horizontal gene transfer modulates gene transfer and motility via cyclic dimeric GMP. *J. Mol. Biol.* **2020**, *432*, 4840–4855.

2 CHAPTER 2: Interactions among redox regulators and the CtrA phosphorelay in *Dinoroseobacter shibae* and *Rhodobacter capsulatus*

2.1 Abstract

Bacteria possess regulatory networks to detect environmental signals and respond appropriately, often by adjusting gene expression. Some regulatory networks influence many genes, and many genes are affected by multiple regulatory networks. Here, we investigate the extent to which regulatory systems controlling aerobic–anaerobic energetics overlap with the CtrA phosphorelay, an important system that controls a variety of behavioral processes, in two metabolically versatile alphaproteobacteria, *Dinoroseobacter shibae* and *Rhodobacter capsulatus*. We analyzed ten available transcriptomic datasets from relevant regulator deletion strains and environmental changes. We found that in *D. shibae*, the CtrA phosphorelay represses three of the four aerobic–anaerobic Crp/Fnr superfamily regulator-encoding genes (*fnrL*, *dnrD*, and especially *dnrF*). At the same time, all four Crp/Fnr regulators repress all three phosphorelay genes (*cckA*, *chpT*, *ctrA*). Loss of *dnrD* or *dnrF* resulted in activation of the entire examined CtrA regulon, regardless of oxygen tension. In *R. capsulatus* FnrL, in silico and ChIP-seq data, also suggested regulation of the CtrA regulon, but it was only with loss of the redox regulator RegA where an actual transcriptional effect on the CtrA regulon was observed. For the first time, we show that there are complex interactions between redox regulators and the CtrA phosphorelays in these bacteria and we present several models for how these interactions might occur.

2.2 Introduction

Bacteria sense and process environmental signals in order to adapt to changes in their surroundings. These signals are relayed through regulatory networks that adjust the cells' behavior, often through changes in gene expression. The alphaproteobacterium *Dinoroseobacter shibae* is a member of the marine roseobacter group and an aerobic anoxygenic photoheterotrophic bacterium, capable of both aerobic and anaerobic respiration [1]. It can be free-living or an algal symbiont [1] and is a metabolically versatile bacterium able to adapt to changes in its highly dynamic environment. For example, at the end of an algal bloom when the oxygen concentration drops, an alternative terminal electron acceptor such as nitrate can

be used for respiration [1,2]. The response to the change from aerobic to anaerobic conditions is controlled by four Crp/Fnr transcriptional regulators in *D. shibae* [3]. Crp/Fnr regulators are widely distributed among bacteria and form a superfamily consisting of 14 phylogenetic subgroups [4]. The versatility of this family is reflected by both the wide range of signals that are sensed, such as temperature [5], oxygen [6], and nitric oxide (NO) [7], and the range of metabolic processes regulated upon activation, which include respiration-related processes and especially the transition between aerobic and anaerobic lifestyles [3,8].

Two well-studied members of this family are the Dnr and Fnr proteins. Dnr proteins bind a heme cofactor that allows for sensing of NO [4,9], while Fnr proteins react to low oxygen tension [4,6]. In *D. shibae*, FnrL and DnrD regulate DnrE and DnrF in a cascade-type network that controls the transition from aerobic to anaerobic growth, heme and carotenoid synthesis, multiple other metabolic processes, and flagellar synthesis [3]. The importance of these regulators in *D. shibae* is well illustrated by the observation that loss of FnrL affects the transcript levels of over 400 genes [3]. Another important regulatory system in *D. shibae* is the CtrA phosphorelay [10]. Like the Crp/Fnr regulators, this phosphorelay integrates an environmental signal, in this case, the autoinducer concentration as an indicator of cell density, and adjusts gene expression in response [11]. This phosphorelay is conserved within the majority of alphaproteobacterial lineages and consists of the histidine kinase CckA, the phosphotransferase ChpT and the transcriptional regulator CtrA [10]. In *D. shibae*, the CtrA phosphorelay is activated by the quorum sensing (QS) signal of the main acyl-homoserine lactone (AHL) synthase ($LuxI_1$) with subsequent regulation of genes for flagellar motility, recombination and competence proteins, a tight adherence (*tad*) pilus involved in attachment to carbohydrates on the host cells [12], cell cycle control, gene transfer agent (GTA) production, bis-(3-5)-cyclic dimeric guanosine monophosphate (c-di-GMP) signaling, the NO-sensing heme-nitric oxide/oxygen binding domain (HNOX) protein, and the AHL synthases $LuxI_2$ and $LuxI_3$ [11,13,14]. Deletion of *cckA* has been found to abolish the mutualistic interaction between *D. shibae* and its algal host, demonstrating that the CtrA phosphorelay is essential for establishment of this symbiosis, at least partly due to the requirement for flagella [15]. The

Crp/Fnr and CtrA phosphorelay networks are connected by their shared regulation of flagellar gene expression and due to their involvement in symbiosis with the host dinoflagellate. There are three ways bacteria can be exposed to NO. Some bacteria generate NO during denitrification, and this is considered the activator for DnrD in *D. shibae* [3,16]. NO can be produced intracellularly through the oxidization of L-arginine to NO and L-citrulline [17] or via a nitric oxide synthase (NOS) [17,18]. NO released by some eukaryotic organisms can be a form of communication with their symbiotic bacteria and is then typically sensed by HNOX proteins [19]. The HNOX genes are often located adjacent to genes encoding c-di-GMP signaling proteins or histidine kinases. In the context of symbioses, only a few NO-detecting systems have been found that do not involve c-di-GMP signaling but instead directly integrate into QS systems [20–22]. In *D. shibae,* an HNOX protein detects NO and thereupon inhibits the c-di-GMP synthesizing enzyme Dgc1 [23]. The potential for overlap between Crp/Fnr-based regulation and the CtrA phosphorelay also exists in the purple non-sulfur alphaproteobacterium *Rhodobacter capsulatus*. Its CtrA phosphorelay was originally discovered due to its regulation of GTA production [24], but it also affects many other genes such as those associated with flagellar motility, gas vesicles, and c-di-GMP signaling [24,25]. Like *D. shibae, R. capsulatus* can switch between aerobic and anaerobic lifestyles, which involves Crp/Fnr regulation, the RegA/B two-component system, and CrtJ [26–28]. Loss of FnrL affects the transcript levels of 20% of *R. capsulatus* genes [29], including 42 that are directly regulated by FnrL (shown by Chip-seq binding sites) and encode c-di-GMP signaling, gas vesicle, and flagellar proteins, among others [29]. These initial surveys of the activities of redox regulators and the CtrA phosphorelays in *D. shibae* and *R. capsulatus* indicated a potential connection of the regulons. Therefore, we were interested in exploring in more detail the extent to which these regulatory systems interact. We re-analyzed ten available transcriptomic datasets for the two species. Deletion mutants, including those of redox regulators and the CtrA phosphorelay/QS networks, were analyzed to examine the regulon overlap of these systems and to evaluate their potential integration. We also included further analyses of available transcriptomic datasets of wild type strains undergoing physiological changes related to the environmental signals integrated by these regulatory systems.

2.3 Materials and Methods

2.3.1 Datasets analyzed in this study

Ten published and accessible microarray and RNA-seq transcriptomic datasets for chosen gene knockout strains and experiments monitoring responses to changes in environmental conditions were obtained from the NCBI GenBank database (Table 2.1).

Table 2.1. Description of the transcriptomic datasets analyzed in this study.

Species	Strains and culture conditions	Type of data	Accession number	Reference
D. shibae	Time-resolved response to addition of AHL to Δ*luxI*$_1$	RNA-seq	GSE122111	[13]
	Time resolved co-cultivation with *Prorocentrum minimum*	RNA-seq	GSE55371	[15]
	Knockouts of *ctrA*, *chpT* and *cckA* in exponential and stationary phases of growth	Agilent dual-color microarray	GSE47451	[11]
	Knockouts of *fnrL*, *dnrD*, *dnrE* and *dnrF* under aerobic conditions and 60 minutes after shift to anaerobic, denitrifying conditions	Agilent dual-color microarray	GSE93652	[3]
	Time-resolved growth of wild type and Δ*luxI*$_1$ strains from OD_{600} 0.1 to stationary phase	Agilent dual-color microarray	GSE42013	[14]
	Δ*luxI*$_2$ growth to OD_{600} of 0.4	RNA-seq	PRJEB20656	[30]
	Time-resolved shift of the wild type from aerobic to anaerobic growth conditions	Agilent single-color microarray	GSE47445	[31]
	Knockouts of *regA*, *crtJ* and *fnrL* in mid-exponential growth phase	RNA-seq	PRJNA357604	[32]

R. capsulatus	Knockouts of *ctrA* and *cckA* in mid-exponential growth phase	Affymetrix microarray	GSE53636	[33]
	Knockout of *ctrA* during exponential and stationary growth phases	Affymetrix microarray	GSE18149	[34]

2.3.2 Processing and analysis of datasets

This study includes four different types of transcriptomic data (Table 2.1) that could not be processed and analyzed as one dataset. We therefore used the changes in transcript levels (log2 fold change) compared to the controls used in the respective studies (e.g., wild type or time point before changes in the environmental conditions) for each dataset. RNA-seq data from *D. shibae* (reads per gene) and *R. capsulatus* (log2 fold change) were obtained from the respective publications (Table 2.1).

Agilent microarray datasets were processed using the LIMMA package in R [35]. Background correction was performed with the "normexp" method and an offset of 10. Two-color microarrays were normalized with the "loess" method before quantile normalization. Signals/intensities from spots were averaged.

Affymetrix microarray datasets were processed using the R packages LIMMA, makecdfenv, and affy [35–37]. The CDF environment for GSE18149 was generated using the corresponding CDF file downloaded from GEO (accession GPL9198). Data were normalized with the rma function. A linear fit model was generated for comparison.

In order to analyze the CckA and ChpT regulons, thresholds were set that allowed definition of regulated and non-regulated genes. These thresholds were applied to the log2 fold change in transcript level values in the *cckA* and *chpT* deletion mutants. A gene was not considered regulated when its log2 fold change was between 1 and −1 while a log2 fold change value above 1 or below −1 indicated an affected gene. The analyzed genes were grouped based on published, manually curated information about their functional categories as described (Table S1) [13].

2.4 Results

2.4.1 Overlap of the Crp/Fnr and CtrA regulons in *Dinorosebacter shibae*

The possible interaction between the Crp/Fnr regulator and CtrA phosphorelay networks was first assessed using transcriptomic datasets for regulator deletion mutants. The changes of transcript levels of known Crp/Fnr- and CtrA-controlled traits revealed an overlap of both regulons, with the regulator-encoding genes themselves affected by losses of the other regulators (Figure 2.1). Under both aerobic and anaerobic conditions, the loss of *dnrD* or *dnrF* resulted in increased transcript levels of the CtrA phosphorelay, QS, flagellar motility, *tad* pilus, competence and recombination, gene transfer agent (GTA), *divL* and c-di-GMP signaling genes (Figure 2.1A). In all datasets, the GTA genes showed comparatively small changes in transcript levels (Figure 2.1), probably as a result of a small subpopulation actually expressing these genes [13]. Only the loss of *dnrF* led to a change in gene expression between aerobic and anaerobic conditions, since a greater increase in the transcript levels could be observed under anaerobic conditions for most of its regulon (Figure 2.1C). The loss of *fnrL* or *dnrE* resulted in increased transcript levels of *ctrA*, *cckA*, *chpT*, $luxI_1$, $luxR_1$, and $luxR_2$ but had little to no effect on the downstream CtrA regulon (Figure 2.1B).

Almost all examined genes showed an opposite pattern in the CtrA phosphorelay and $luxI_1$ mutants (Figure 2.1D) compared to *dnrD* and *dnrF* (Figure 2.1A). Most of the genes showed decreased transcript levels in strains lacking any of the CtrA phosphorelay genes, with the exceptions of the Crp/Fnr regulators where the largest increase was found for *dnrF* (Figure 2.1D). Loss of $luxI_1$ resulted in increased transcript levels for *fnrL*, *dnrD*, and *dnrF*, but no changes were observed for *dnrE* (Figure 2.1D).

2.4.2 The role of ChpT in signal integration

In *D. shibae*, deletion of neither *ctrA* nor *cckA* had an influence on expression of *chpT*, whereas the loss of either *ctrA* or *chpT* resulted in decreased expression of *cckA* (Figure 2.1D) [11]. However, all three CtrA phosphorelay component genes showed reduced transcript levels in the absence of the AHL synthase $luxI_1$ (Figure 2.1D) [13], whereas loss of the Crp/Fnr regulators resulted in increased transcript levels of these genes (Figure 2.1A, B). Therefore, in contrast to *ctrA* and *ccka*, *chpT* is not regulated by

the CtrA phosphorelay itself, but by other factors that can thereby control the phosphorylation state of CtrA. These findings also suggest that *chpT* transcription is regulated oppositely by QS and the Crp/Fnr regulators.

This is supported by binding site predictions for FnrL [3] that suggest it binds at the promoter of *chpT* and *clpX*, which encodes a protease known to cleave CtrA [3,38,39]. Deletion of *fnrL* strongly increased the expression of *chpT* but only resulted in minimal changes for *clpX* (Figure 2.1B). Binding site prediction for the Dnr regulators did not find any binding sites near *clpX* or the CtrA phosphorelay genes [3].

It was previously found that more genes were affected by the loss of *chpT* than *cckA* [11], suggesting ChpT regulates some genes independent of CckA and that a different kinase might regulate its activity and thereby affect downstream gene expression. Among the genes affected by the loss of *chpT* but not *cckA*, *dnrF* was the most upregulated gene during exponential growth while *lexA* and *recA* were among those most downregulated genes in both exponential and stationary phases (Figure 2.2). Although there was a small increase in transcript levels of *dnrF* in the *cckA* deletion strain during exponential growth, it did not pass the threshold we defined (see Materials and Methods). These findings suggest a link between *dnrF* and *chpT*.

Additional discrepancies between CckA and ChpT are apparent from their opposing effects on the *nap* gene cluster during exponential growth (Figure 2.2A), although this is not maintained in stationary phase (Figure 2.2B). In exponential phase, loss of *cckA* led to decreased transcript levels of the nap gene cluster, while the loss of *ctrA* and *chpT* led to increased levels (Figure 2.2A). This cluster is the only denitrification cluster activated by FnrL but repressed by the three Dnr regulators [3]. Interestingly, transcript levels of all four denitrification gene clusters were increased in the AHL synthase knockout Δ*luxI*$_2$ but were unaffected in Δ*luxI*$_1$ (Figure 2.2C).

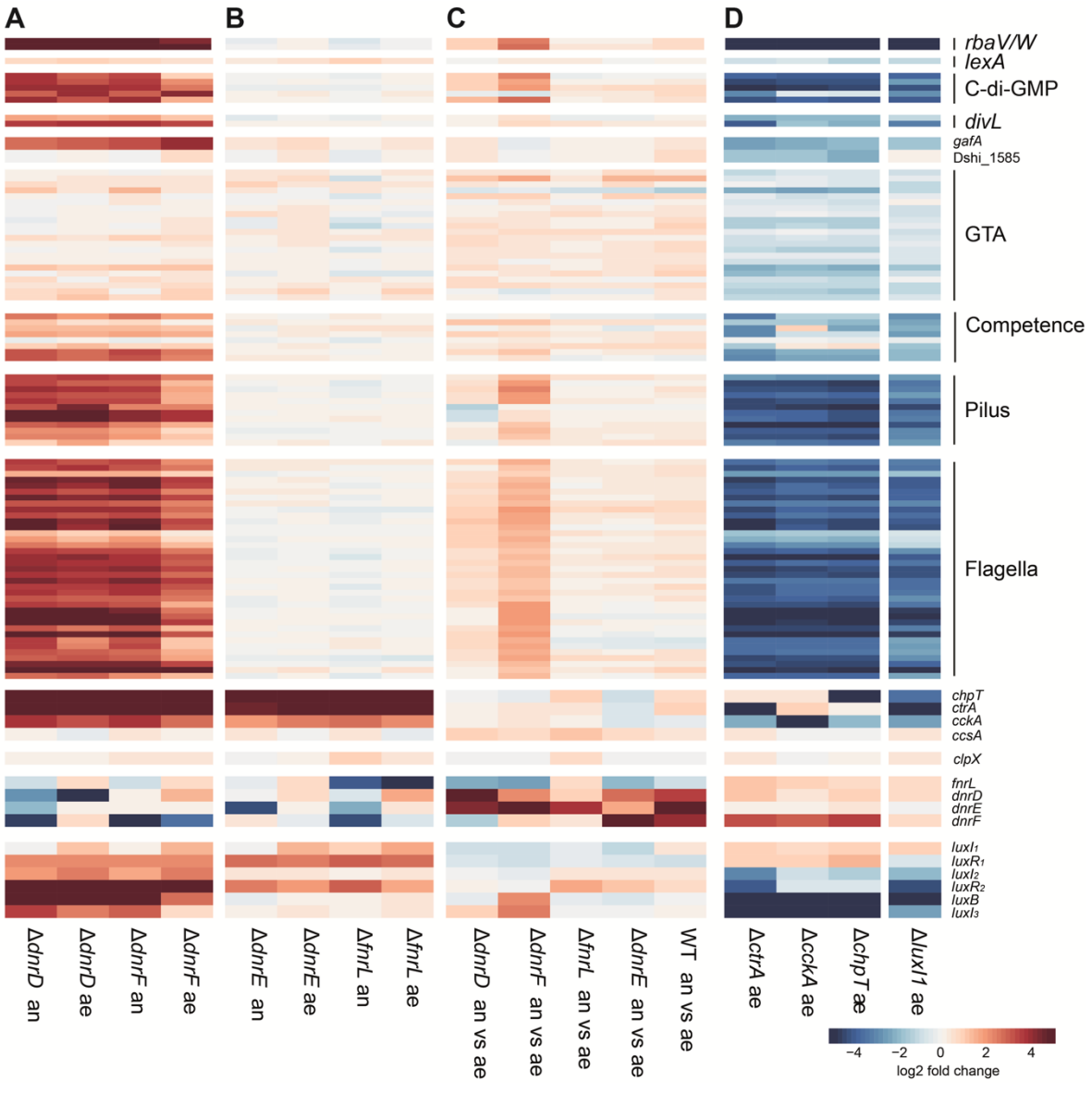

Figure 2.1: Transcriptomic data for genes in selected functional groups in different knockout strains. The four Crp/Fnr regulator knockouts were grown under aerobic (ae) or anaerobic (an) conditions. The $\log_2$ fold changes compared to the respective wild type (WT) (**A** and **B**) or against themselves grown at different conditions, are shown (**C**). The CtrA phosphorelay and quorum sensing system knockouts were grown aerobically and compared to the WT (**D**). The functional group assignments on the right are based on published information as described in Table S1.

2.4.3 Time-resolved evaluation of environmental changes and the regulation of c-di-GMP signaling genes

Interactions between the networks in *D. shibae* were further analyzed using time-resolved transcriptomic datasets. These were collected following the switch from aerobic to anaerobic conditions in wild type cells (Figures 2.3A and 2.4A) [31], following the external addition of AHL autoinducer to the AHL synthase mutant Δ*luxI*$_1$ (Figures 2.3B and 2.4B) [13], and through the culture growth phases for Δ*luxI*$_1$ in the absence of AHL (Figure 2.4C) [14].

Upon the shift to anaerobic conditions, all three *dnr* genes showed an immediate increase in transcript levels for 30 min and then stayed constant, whereas those of *fnrL* decreased (Figure 2.4A). These changes corresponded with increased transcript levels of the denitrification gene clusters, with the nap cluster showing a slightly different pattern than the *nir* and *nos* clusters (Figure 2.3A). Slight increases were observed for the c-di-GMP signaling, flagellar, tad pilus, and QS genes (Figure 2.3A). Four of the five c-di-GMP signaling genes showed increased transcript levels following the transfer to an anaerobic environment, whereas *dgc*$_2$ showed a slight decrease (Figure 2.4A).

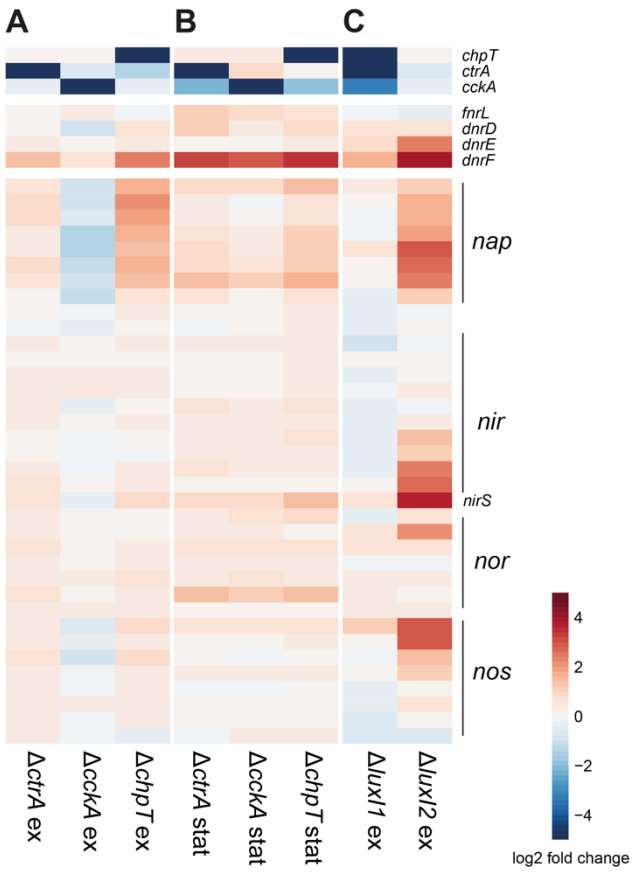

Figure 2.2: Comparison of CtrA phosphorelay, Crp/Fnr regulator and denitrification gene expression control by CtrA phosphorelay and LuxI$_{1/2}$ synthases during exponential and stationary growth phases. Samples for the *ctrA*, *cckA* and *chpT* knockout mutants were analyzed at mid-exponential (OD 0.4) (**A**) and stationary (six hours after onset of stationary phase) (**B**) phases of growth. The Δ*luxI*$_1$ data were obtained during stationary phase, six hours after the onset of stationary phase, and the D*luxI*$_2$ data were obtained during the mid-exponential growth phase (**C**).

The addition of AHL to the Δ*luxI*$_1$ strain led to increased transcript levels for all CtrA- and QS-controlled genes (Figure 2.3B). This included the CtrA phosphorelay and c-di-GMP signaling genes, with *dgc*$_2$ showing the largest increase (Figure 2.4B). No effect was visible for the Crp/Fnr regulator-encoding genes (Figure 2.4B) and only a minor increase of the *nap* gene cluster was observed among the denitrification genes (Figure 2.3B).

Due to the increased transcript levels observed for CtrA regulon genes in the *dnrD* and *dnrF* deletion strains, it was expected that the same genes would also be decreased under anaerobic conditions. Instead, it turned out that the change from aerobic to anaerobic conditions (Figure 2.3) resulted in

increased transcript levels for these genes. However, this increase was small, and effects were not observed for some genes that appeared to be controlled by the individual regulators based on the knockout transcriptomic data (Figure 2.1). This included the regulation of the CtrA phosphorelay genes by the Crp/Fnr regulators. Vice versa, loss of the CtrA phosphorelay genes indicated their repression of Crp/Fnr regulator gene expression (Figure 2.1D), but the contrary was observed in the respective physiological datasets where the Crp/Fnr regulators seem to be upregulated (Figure 2.3B). Notably however, in both physiological datasets, *dgc*$_2$ stands out as distinctly affected compared to other c-di-GMP signaling genes (Figure 2.4A, B). Also, in the non-induced ΔluxI1 culture, no influence of the QS null mutant on the Crp/Fnr regulators was observed, but the CtrA phosphorelay and c-di-GMP signaling genes were down-regulated (Figure 2.4C).

Interestingly, in contrast to *fnrL*, *dgc*$_2$, and *chpT*, the other Crp/Fnr regulators, c-di-GMP signaling, and CtrA phosphorelay genes all decreased at the onset of the stationary phase (Figure 2.4C). Moreover, analysis of the Crp/Fnr knockout data showed that the loss of *dnrF* or *dnrD* resulted in increased transcript levels of four of the c-di-GMP signaling genes under anaerobic growth conditions, with only *dgc*$_2$ being unaffected (Figure S1A). Loss of *luxI*$_1$ and the CtrA phosphorelay genes resulted in decreased transcripts for all five genes (Figure S1B, C), although the effects on *dgc*$_2$ were smaller than for the other genes in the stationary phase (Figure S1C).

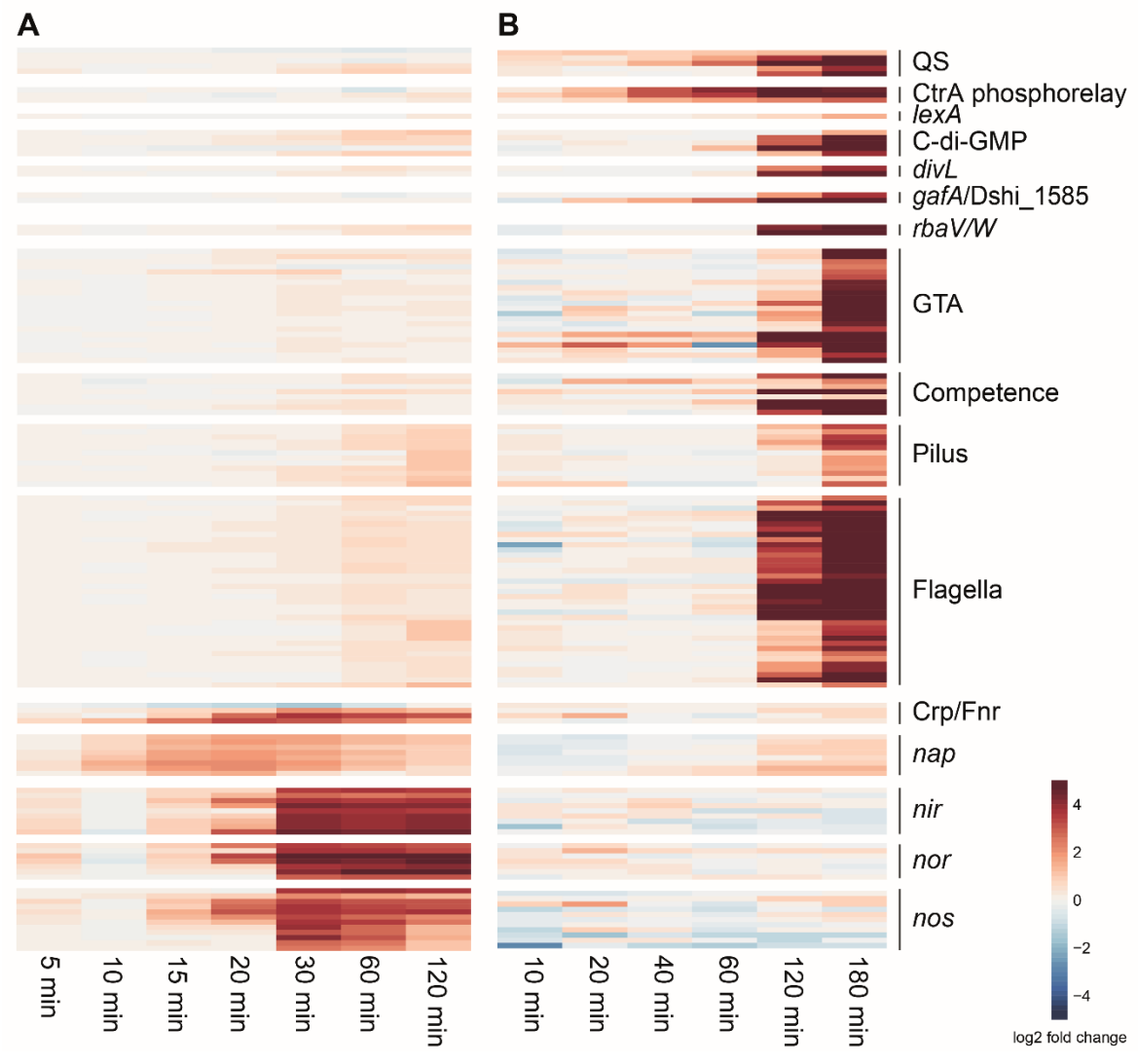

Figure 2.3: Transcript level changes for genes in selected groups in response to a shift from aerobic to anaerobic growth conditions or to external addition of autoinducer in a synthase null mutant. Transcriptomic data were obtained at seven time points post-shift between growth conditions the microarray sequences were plotted against growth under aerobic conditions (**A**). RNA-seq data were obtained at six time points post-addition of 3-oxo C14 HSL to a *luxI*$_1$ knockout mutant and plotted against time point t=0 (**B**).

2.4.4 Effects on the CtrA regulon during coculture of *Dinoroseobacter shibae* and its algal host

In the two-phase interaction of *D. shibae* with its dinoflagellate host *Prorocentrum minimum* [14,40], a mutualistic growth phase (0 to 21 days of cocultivation) is followed by a pathogenic growth phase (21 to 30 days of cocultivation) that leads to death of the algae [15]. Analysis of the transcriptomic

data of *D. shibae* during cocultivation showed an overall increase in the transcription for the CtrA regulon genes during the transition between the two phases (day 24 compared to day 18), followed by a decrease during the late-pathogenic phase, after 30 days (Figure S2).

Of the CtrA phosphorelay genes, only *chpT* remained upregulated during the pathogenic interaction. Evaluation of the denitrification gene clusters showed strong variation among these genes (Figure 2.2A), likely arising from overall low expression levels, and this made it difficult to draw any conclusions.

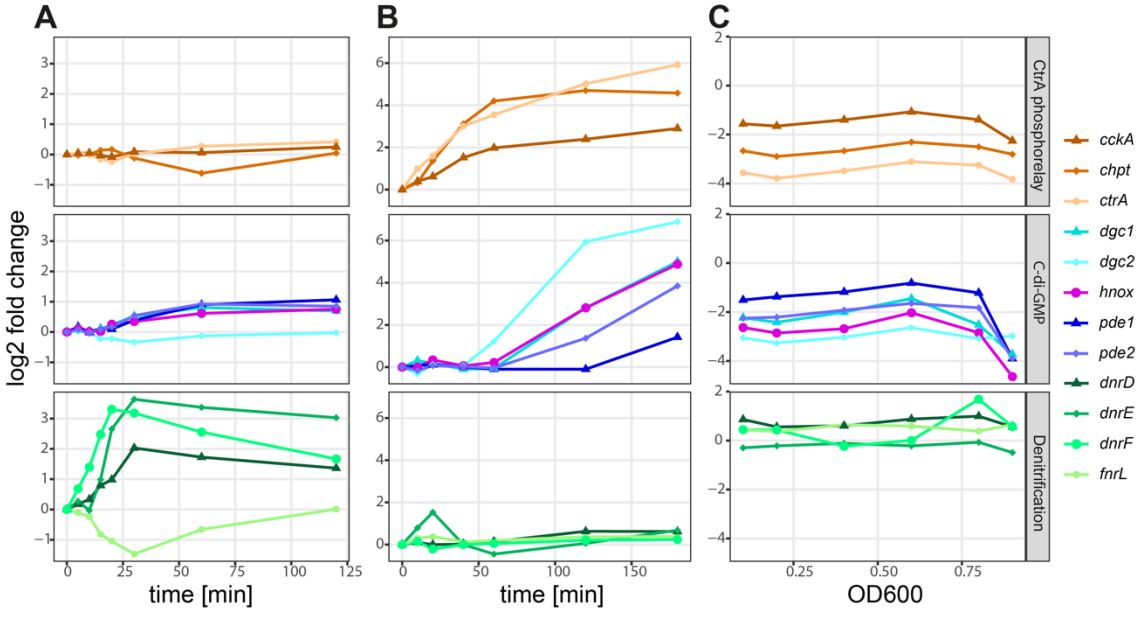

Figure 2.4: Time- and density-resolved transcript levels in three different conditions for three groups of regulators. The expression profiles of the CtrA phosphorelay genes (top), c-di-GMP signaling genes (middle) and four Crp/Fnr regulator-encoding genes (bottom) are plotted. The changes in transcript levels were monitored after the switch from aerobic to anaerobic growth over a time period of 120 min (**A**), after the external addition of autoinducer (3-oxo C14 HSL) to the synthase null mutant (D*luxI*$_1$) over a period of 180 min (**B**) and during logarithmic (samples 1-5) and stationary (sample 6) phases of growth as determined by optical density (**C**).

2.4.5 RegA activates the CtrA regulon in *Rhodobacter capsulatus*

Next, we asked if the observed overlap between redox regulators and the CtrA phosphorelay system is conserved in another member of the family Rhodobacteraceae. For *R. capsulatus*,

transcriptomic data were available for knockout mutants of *ctrA*, *cckA*, and the known redox regulator-encoding genes *fnrL*, *regA*, and *crtJ*. We identified three additional Crp/Fnr regulator-encoding genes in this bacterium based on blast searches (RCAP_rcp00107, RCAP_rcc01561, RCAP_rcc03255), but these genes showed no evidence of differential regulation in any of the analyzed datasets and we did not consider them further. A blast search also identified a homologue (RCAP_rcc02630) of the HNOX-encoding gene of *D. shibae* (Dshi_2815). This gene encodes a protein with a predicted heme nitric oxide/oxygen binding (HNOB) domain and is located adjacent to a c-di-GMP signaling gene (RCAP_rcc02629) that was recently demonstrated to affect GTA production and motility in *R. capsulatus* [41]. When bound to NO, the HNOX homologue in *D. shibae* inhibits the activity of the diguanylate cyclase Dgc1, which is encoded by the neighboring gene [23].

FnrL is the only Crp/Fnr regulator that has been studied in *R. capsulatus* [29]. Its loss did not result in any large changes in transcript levels for the examined traits under anaerobic phototrophic conditions (Figure 2.5), and the same was observed for the loss of *crtJ*, which encodes a transcription factor that controls numerous photosynthesis and cytochrome genes [32] (Figure 2.5). RegA is the response regulator of the RegB/A two-component system that controls photosynthesis, nitrogen and carbon fixation, denitrification, and respiration genes in response to oxygen availability [26]. In contrast to *fnrL* and *crtJ*, we found that the loss of *regA* resulted in a strong decrease in transcript levels of the CtrA regulon genes (Figure 2.5), indicating that RegA acts as a direct or indirect activator of these genes. Like the genes involved in regulation of photosynthesis and the change between aerobic/anaerobic lifestyle in *D. shibae*, loss of *regA* affected *chpT* the most among the CtrA phosphorelay genes in *R. capsulatus*. Loss of the CtrA phosphorelay genes had no influence on transcription of *fnrL*, *regA*, *regB*, or the other putative Crp/Fnr regulator-encoding genes (Figure S3). A comparison of photosynthetic anaerobic growth and aerobic cultivation in *R. capsulatus* showed the CtrA-regulated traits have reduced transcript levels under anaerobic conditions (Figure 2.5).

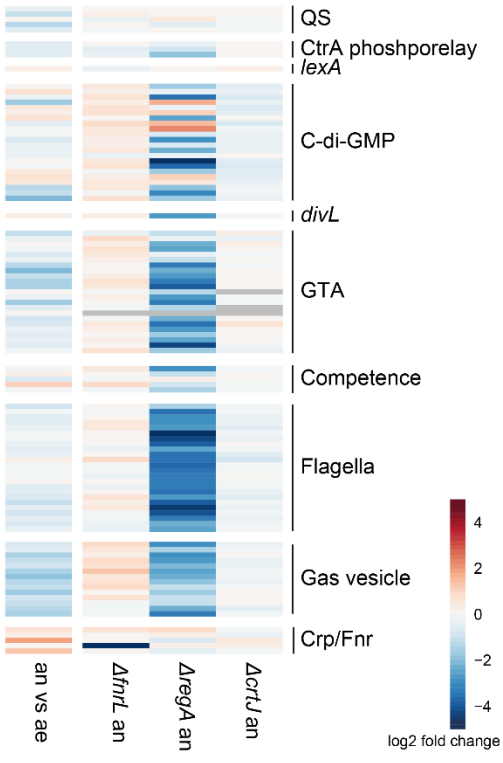

Figure 2.5: Effects of growth conditions and three regulator knockouts on the transcript levels of eight categorized groups of genes in *Rhodobacter capsulatus*. The microarray-based transcriptomic data for aerobic versus anaerobic growth in the wild type and for three mutants, *fnrL*, *regA* and *crtJ*, compared to the wild type are shown. Increased transcript levels under anaerobic relative to aerobic conditions are indicated by RED/BLUE in the heatmap. Negative effects on expression by the respective regulator are reflected by red in the heatmap, whereas positive effects are reflected by blue.

2.5 Discussion

2.5.1 The Crp/Fnr and CtrA/QS regulons overlap in *Dinoroseobacter shibae*

Our analysis revealed an inverse regulatory crosstalk between the Crp/Fnr and CtrA systems in *D. shibae*. We found the denitrification gene clusters and Crp/Fnr regulator genes, especially *dnrF*, to be part of the CtrA phosphorelay and $LuxI_2$ regulons. DnrE was affected exclusively by loss of $LuxI_2$, whereas loss of $LuxI_1$ only had minor effects on *fnrL*, *dnrD*, and *dnrF* and no effect on *dnrE*. In addition to their regulation by $LuxI_1$, which signals cell density, the Crp/Fnr regulators integrate oxygen and NO levels and affect all three CtrA phosphorelay genes.

Until now, overlapping regulation by the Crp/Fnr and CtrA systems has only been noted in *D.*

shibae for flagellar genes and *recA* [3,12,13], and to our knowledge this level of regulatory interaction has not been reported for alphaproteobacteria. However, a comparable connection between QS and Crp/Fnr regulators has been documented for the gammaproteobacterium *Pseudomonas aeruginosa* where the regulons of the FnrL homolog Anr and QS synthase LasR overlap. Here, denitrification genes are induced by Anr and inhibited by LasR. Additionally, in the absence of *lasR*, Anr regulates production of the QS molecule 4-hydroxy-2-alkylquinoline [42]. At the protein level, nitrite reductase (NirS), a flagellar protein (FliC), and the chaperone DnaK form a complex that influences flagellar formation and motility and thus creates a link between denitrification and motility [43]. In cystic fibrosis infections, *P. aeruginosa* is exposed to ambient conditions with low oxygen tension. The intracellular levels of c-di-GMP increase, which leads to biofilm formation. These conditions also lead to an increase in mutations in the QS transcriptional regulator-encoding gene *lasR*. As *lasR* deletion strains grow to higher cell densities and have higher denitrification rates, it has been suspected that these mutations increase the fitness of the population during infection [44–46].

Combined, these observations indicate that there may be a more widely conserved interaction of Crp/Fnr regulators and QS in proteobacteria. The CtrA phosphorelay is unique to alphaproteobacteria, indicating a potential independent evolution of this regulatory crosstalk in this lineage

2.5.2 Inverse control of the CtrA regulon by RegA and anaerobic photosynthetic growth conditions in *Rhodobacter capsulatus*

In *R. capsulatus*, the regulons of the redox-responsive two-component system RegA/B [47] and the CtrA phosphorelay overlap. Interestingly, *chpT* stands out because it is the only CtrA phosphorelay gene that is regulated by RegA. Similar to Dnr and Fnr in *D. shibae*, RegA controls the expression of photosynthesis and respiration genes [26]. ChIP-seq with RegA identified binding sites adjacent to several genes also targeted by CtrA: RCAP_rcc02857 (a c-di-GMP signaling gene involved in GTA production) and its divergently transcribed neighbor (RCAP_rcc02856), RCAP_rcc02683 (a chemotaxis gene), and *dksA* (a *dnaK* suppressor gene) [34].

As in *D. shibae*, transcriptomic data from a *fnrL* deletion strain showed no effects on the

CtrA-controlled traits outside of the CtrA phosphorelay genes themselves. However, ChIP-seq and in silico binding site predictions [29] suggest FnrL binding adjacent to *divL*, *dnaK*, *recA*, flagellar gene clusters, the RcGTA capsid protein-encoding gene, and c-di-GMP signaling genes (including those affecting RcGTA production [41]). Similarly, ChIP-seq with CrtJ [48], a regulator controlling expression of multiple genes involved in photosynthesis, also revealed a connection to the CtrA phosphorelay. Even though the observed transcript level changes in the *crtJ* mutant were small, binding was found adjacent to *ctrA*, *clpX*, a *luxR* family gene, *dnaA*, *spoT*, *ftsZ*, and the first gene in the GTA structural gene cluster (RCAP_rcc01682) under aerobic and anaerobic cultivation. Binding sites adjacent to *dnaK* and two flagellar genes (*flgB* and *flaA*) were identified under aerobic and anaerobic conditions, respectively.
In *D. shibae*, deletions of the Crp/Fnr regulator-encoding genes indicated an inhibition of the CtrA regulon, but the physiological changes detected by these regulators (switch from aerobic to anaerobic conditions) showed a tendency towards activation of the CtrA regulon. The same was observed for the deletion mutants of the CtrA phosphorelay components and their regulation of the Crp/Fnr regulator genes. In *R. capsulatus*, we could observe a similar pattern but in reverse for regulation of the CtrA regulon by RegA. While the *regA* knockout indicated activation of the CtrA regulon, the switch to anaerobic photosynthetic growth conditions showed an inhibition. This is probably indicative of a more complex interaction among these regulatory systems. However, the *regA* deletion transcriptomic data are supported by in vivo motility tests that showed reduced swimming ability of the Δ*regA* strain [26].

2.5.3 Integration of Crp/Fnr regulation into the CtrA phosphorelay and regulon

In *D. shibae*, CtrA binding site predictions and expression data for *ctrA* and *cckA* suggest that CtrA directly regulates its own expression and that of *cckA*, but not *chpT* [13]. Therefore, *chpT* transcription must be regulated from outside of the CtrA phosphorelay and upstream of CtrA. Both, regulatory control of *chpT* and signal integration upstream of CtrA is known for $LuxI_1$ [11]. A similar situation might be possible for Crp/Fnr signal integration due to their regulation of *chpT* (Figure 2.6A). Since *chpT* is the only RegA-regulated CtrA phosphorelay gene in *R. capsulatus* (and it has a RegA binding site), it seems to play a central role here, too. However, there are also RegA binding sites

associated with *clpX* and other genes of the CtrA regulon [26]. Interestingly, the Dnr/Fnr binding site in the *nosR2* promoter in *D. shibae* has the sequence 5-TTAAC-N4-GTCAA-3 [3], which shares a half-site binding motif with CtrA 5-TTAAC-N5-GTTAAC-3 [11]. Previously, comparison between transcriptional regulation and the presence of full and half-site motifs revealed the potential importance of half-site motifs for transcriptional control by CtrA in *R. capsulatus* [34]. Thus, CtrA and Fnr regulators might interact with some of the same/overlapping sequences (Figure 2.6B).

A distinct role for ChpT is supported by the observation that loss of *chpT* or *ctrA* but not *cckA* results in decreased transcript levels of *dnrF*. It is possible that ChpT integrates signals from more than one kinase into its regulation of CtrA. To our knowledge, the only other instance of a histidine kinase affecting phosphorylation of CtrA via ChpT is CcsA from *Sphingomonas melonis* [49]. Potential homologues of CcsA are encoded in *D. shibae* (Dshi_1893) and *R. capsulatus* (RCAP_rcc02545), but effects on transcript levels of these genes were not observed in any of the analyzed datasets. This does not exclude their involvement but also does not allow us to draw further conclusions (Figure 2.6C).

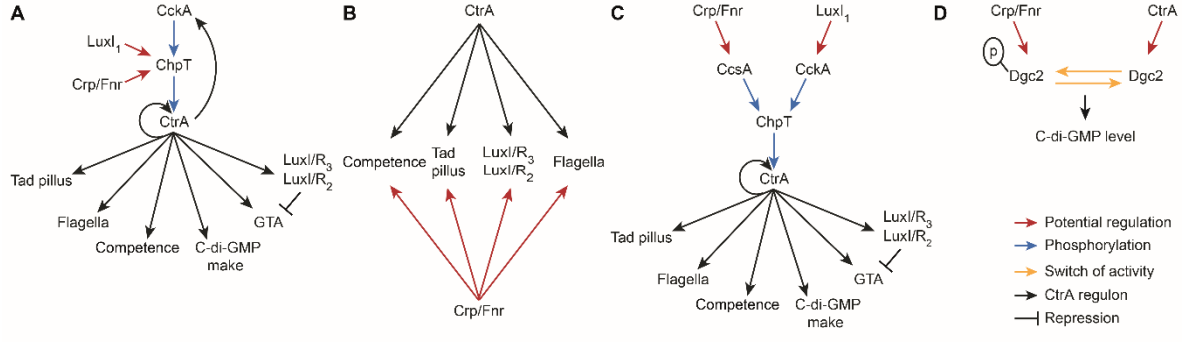

Figure 2.6: Possible mechanisms of integration of the Crp/Fnr and CtrA systems. (**A**) The $LuxI_1$ and Crp/Fnr signals could be integrated into the CtrA phosphorelay via *chpT* regulation, which does not happen via CckA or CtrA. (**B**) Shared binding site motifs for Crp/Fnr regulators and CtrA might allow direct integration of the NO/oxygen signal into the CtrA regulon. (**C**) An additional histidine kinase (CcsA) has been reported to phosphorylate ChpT in another bacterium, and this could integrate the Crp/Fnr signals and disconnect CckA from the integration. (**D**) Phosphorylation of the Dgc2 receiver domain likely regulates the enzyme's diguanylate cyclase activity and thereby alters the intracellular levels of c-di-GMP, which are known to affect the CtrA regulon.

2.5.4 Crp/Fnr regulation of the CtrA regulon is largely independent of oxygen tension

Among the Crp/Fnr regulators, only loss of the NO-sensing DnrF resulted in higher inhibition activity of the CtrA system under anaerobic conditions. In *P. aeruginosa*, swimming motility is controlled anaerobically and aerobically, and it was suggested that NirS promotes motility in multiple ways, at the protein level or via signaling pathways, depending on oxygen availability [50]. Regulation of QS traits by both NO and oxygen was also found in the interaction of *Vibrio fischeri* with the light organ of its squid host. Here, NO released by the host's immune system regulates the symbionts' settlement via biofilm production while the host's control of oxygen availability regulates bacterial bioluminescence in a circadian manner [51–53]. However, since the Crp/Fnr knockout and physiological change transcriptomic data have opposite effects on the CtrA phosphorelay (inhibition indicated by the knockouts and activation by the shift to anaerobic growth), it is difficult to determine the role of oxygen on the CtrA phosphorelay. In *R. capsulatus*, the knockout transcriptomic data were supported by in vivo experiments, so if the knockout transcriptomic data also reflect the actual CtrA regulon in *D. shibae*, the Crp/Fnr regulators have an inhibitory effect on the CtrA phosphorelay and its regulon. It is known that *Dinoroseobacter* establishes a mutualistic symbiosis with its dinoflagellate host via the CtrA phosphorelay and by means of flagella. It is possible this interaction is repressed towards the end of an algal bloom when oxygen concentrations change, resulting in downregulation of flagella (and other CtrA-regulated traits) via Crp/Fnr regulation.

2.5.5 The role of c-di-GMP

Multiple eukaryotic hosts are known to use NO for communication with microbial symbionts. In some of the characterized systems, NO is sensed by HNOX proteins, which then control c-di-GMP signaling proteins or histidine kinases encoded by genes adjacent to the HNOX-encoding gene. For example, in *Vibrio harveyi*, the HNOX-neighboring histidine kinase phosphorylates the QS transcription regulator LuxU [20], and in *D. shibae*, HNOX inhibits the c-di-GMP signaling enzyme Dgc1 [23]. However, *D. shibae* also has a second c-di-GMP synthesizing enzyme, Dgc_2. During adaptation to anaerobic cultivation and at the onset of stationary phase, dgc_2 transcriptional patterns were similar to

chpT and *fnrL*. The transcript levels of these three genes plateaued, whereas those of the other c-di-GMP signaling, CtrA phosphorelay and Crp/FnrL genes decreased. A unique regulation of *dgc*$_2$ was also observed in the *dnrF*, *dnrD*, *cckA*, and *chpT* knockout strains. Thus, both networks (Crp/Fnr and CtrA phosphorelay) regulate *dgc*$_2$ and affect its expression in a similar manner as a response to the onset of stationary phase.

The role of *dgc*$_2$ in the CtrA phosphorelay and FnrL networks and how it might connect both remain to be clarified. For example, it is possible that phosphorylation of the receiver domain of Dgc$_2$ regulates its c-di-GMP synthase activity. As a result, regulation by the Crp/Fnr or CtrA phosphorelay systems could have different effects on the shared traits (Figure 2.6D).

2.6 Conclusions

In this study we show that regulation of the CtrA regulon, including traits related to phenotypic heterogeneity, is additionally controlled by the aerobic–anerobic regulators Crp/Fnr in *D. shibae* and by FnrL/RegA in *R. capsulatus*. This finding is especially important for the understanding of the metabolically flexible lifestyles of these bacteria. The analysis of the available transcriptomic datasets revealed multiple possible integration sites of the Crp/Fnr signal into the CtrA phosphorelay, but a final explanation is still elusive based on these data. Nevertheless, this investigation provides the first insights into the integration of a second environmental signal into the CtrA phosphorelay and demonstrates a strong transcriptional connection between QS, CtrA-regulated traits, and Crp/Fnr regulators in alphaproteobacteria, which has an interesting parallel with QS and Crp/Fnr regulators in a second class of proteobacteria. To our knowledge, *D. shibae* and *R. capsulatus* are the first two organisms where both Dnr and HNOX NO-sensor proteins have been studied. Further investigation is necessary to clarify the interaction between the CtrA phosphorelay and the Crp/Fnr regulators. For example, it would be helpful to confirm if an additional kinase is indeed regulating ChpT in these bacteria.

2.7 Supplementary materials

This chapter includes two supplementary files that are available in digital format using this link:

The Word file contains the supplementary figures S2.1-S2.3. Table S2.1 is an Excel file, that contains the assignment of genes into functional categories for *D. shibae, D. shibae* denitrification genes and *R. capsulatus* in three tabs.

2.8 References

1. Wagner-Döbler, I.; Ballhausen, B.; Berger, M.; Brinkhoff, T.; Buchholz, I.; Bunk, B.; Cypionka, H.; Daniel, R.; Drepper, T.; Gerdts, G.; et al. The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: A hitchhiker's guide to life in the sea. *ISME J.* **2010**, *4*, 61–77.
2. Pitcher, G.C.; Probyn, T.A. Suffocating phytoplankton, suffocating waters-red tides and anoxia. *Front. Mar. Sci*. **2016**, *3*, 186.
3. Ebert, M.; Laaß, S.; Thürmer, A.; Roselius, L.; Eckweiler, D.; Daniel, R.; Härtig, E.; Jahn, D. FnrL and three Dnr regulators are used for the metabolic adaptation to low oxygen tension in *Dinoroseobacter shibae*. *Front. Microbiol*. **2017**, *8*, 642.
4. Körner, H.; Sofia, H.J.; Zumft,W.G. Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: Exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol. Rev*. **2003**, *27*, 559–592.
5. Leimeister-Wachter, M.; Domann, E.; Chakraborty, T. The expression of virulence genes in *Listeria monocytogenes* is thermoregulated. *J. Bacteriol.* **1992**, *174*, 947–952.
6. Volbeda, A.; Darnault, C.; Renoux, O.; Nicolet, Y.; Fontecilla-Camps, J.C. The crystal structure of the global anaerobic transcriptional regulator FNR explains its extremely fine-tuned monomer-dimer equilibrium. *Sci. Adv.* **2015**, *1*, e1501086.
7. Poole, R.K.; Anjum, M.F.; Membrillo-Hernandez, J.; Kim, S.O.; Hughes, M.N.; Stewart, V. Nitric oxide, nitrite, and Fnr regulation of *hmp* (flavohemoglobin) gene expression in *Escherichia coli* K-12, *J. Bacteriol*. **1996**, *178*, 5487–5492.
8. Beliaev, A.S.; Beliaev, A.S.; Thompson, D.K.; Thompson, D.K.; Fields, M.W.; Fields, M.W.; Wu, L.; Wu, L.; Lies, D.P.; Lies, D.P.; et al. Microarray transcription profiling of a *Shewanella oneidensis etrA* mutant. *J. Bacteriol*. **2002**, *184*, 4612–4616.

9. Ebert, M.; Schweyen, P.; Bröring, M.; Laass, S.; Härtig, E.; Jahn, D. Heme and nitric oxide binding by the transcriptional regulator DnrF from the marine bacterium *Dinoroseobacter shibae* increases *napD* promoter affinity. *J. Biol. Chem*. **2017**, *292*, 15468–15480.
10. Poncin, K.; Gillet, S.; De Bolle, X. Learning from the master: Targets and functions of the CtrA response regulator in *Brucella abortus* and other alpha-proteobacteria. *FEMS Microbiol. Rev.* **2018**, *019*, 500–513.
11. Wang, H.; Ziesche, L.; Frank, O.; Michael, V.; Martin, M.; Petersen, J.; Schulz, S.; Wagner-Döbler, I.; Tomasch, J. The CtrA phosphorelay integrates differentiation and communication in the marine alphaproteobacterium *Dinoroseobacter shibae*. *BMC Genomics* **2014**, *15*, 130.
12. Motherway, C.; Zomer, A.; Leahy, S.C.; Reunanen, J.; Bottacini, F.; Claesson, M.J.; Flynn, K.; Casey, P.G.; Antonio Moreno Munoz, J.; Kearney, B.; et al. Functional genome analysis of *Bifidobacterium breve* UCC2003 reveals type IVb tight adherence (Tad) pili as an essential and conserved host-colonization factor. *Proc. Natl. Acad. Sci.* USA **2011**, *108*, 11217–11222.
13. Koppenhöfer, S.; Wang, H.; Scharfe, M.; Kaever, V.; Wagner-Döbler, I.; Tomasch, J. Integrated transcriptional regulatory network of quorum sensing, replication control, and SOS response in *Dinoroseobacter shibae*. *Front. Microbiol.* **2019**, *10*, 803.
14. Patzelt, D.; Wang, H.; Buchholz, I.; Rohde, M.; Gröbe, L.; Pradella, S.; Neumann, A.; Schulz, S.; Heyber, S.; Münch, K.; et al. You are what you talk: Quorum sensing induces individual morphologies and cell division modes in *Dinoroseobacter shibae*. *ISME J.* **2013**, *7*, 2274–2286.
15. Wang, H.; Tomasch, J.; Michael, V.; Bhuju, S.; Jarek, M.; Petersen, J.; Wagner-Döbler, I. Identification of genetic modules mediating the Jekyll and Hyde interaction of *Dinoroseobacter shibae* with the dinoflagellate *Prorocentrum minimum*. *Front. Microbiol.* **2015**, *6*, 1262.
16. Zumft, W.G. Nitric oxide signaling and NO dependent transcriptional control in bacterial denitrification by members of the FNR-CRP regulator family, *J. Mol. Microbiol. Biotechnol.* **2002**, *4*, 277–286.
17. Crane, B.R.; Sudhamsu, J.; Patel, B.A. Bacterial nitric oxide synthases. *Annu. Rev. Biochem*.

2010, *79*, 445–470.

18. Rao, M.; Smith, B.C.; Marletta, M.A. Nitric oxide mediates biofilm formation and symbiosis in *Silicibacter* sp. strain TrichCH4B. *mBio* **2015**, *6*, e00206-15.
19. Wang, Y.; Ruby, E.G. The roles of NO in microbial symbioses. *Cell. Microbiol*. **2013**, *13*, 518–526.
20. Henares, B.M.; Higgins, K.E.; Boon, E.M. Discovery of a nitric oxide responsive quorum sensing circuit in *Vibrio harveyi*. *ACS Chem. Biol.* **2012**, *7*, 28.
21. Hossain, S.; Heckler, I.; Boon, E.M. Discovery of a nitric oxide responsive quorum sensing circuit in *Vibrio cholerae*. *ACS Chem. Biol.* **2018**, *13*, 56.
22. Nisbett, L.-M.; Boon, E.M. Nitric oxide regulation of H-NOX signaling pathways in bacteria. *Biochemistry* **2016**, *55*, 32.
23. Bedrunka, P.; Olbrisch, F.; Rüger, M.; Zehner, S.; Frankenberg-Dinkel, N. Nitric oxide controls c-di-GMP turnover in *Dinoroseobacter shibae*. *Microbiology* **2018**, *164*, 1405–1415.
24. Lang, A.S.; Beatty, J.T. Genetic analysis of a bacterial genetic exchange element: The gene transfer agent of *Rhodobacter capsulatus*. *Proc. Natl. Acad. Sci.* **2000**, *97*, 859–864.
25. Lang, A.S.; Beatty, J.T. A bacterial signal transduction system controls genetic exchange and motility a bacterial signal transduction system controls genetic exchange and motility. *J. Bacteriol.* **2002**, *184*, 913–918.
26. Schindel, H.S.; Bauer Biochemistry, C.E.; Bauer, C.E. The RegA regulon exhibits variability in response to altered growth conditions and differs markedly between *Rhodobacter* species. *Microb. Genomics* **2016**, *2*, e000081.
27. Smart, J.L.; Willett, J.W.; Bauer, C.E. Regulation of hem gene expression in *Rhodobacter capsulatus* by redox and photosystem regulators RegA, CrtJ, FnrL, and AerR. *J. Mol. Biol.* **2004**, *342*, 1171–1186.
28. Ponnampalam, S.N.; Bauer, C.E. DNA binding characteristics of RegA. J. Biol. Chem. **1998**, 273, 18509–18513. 29. Kumka, J.E.; Bauer, C.E. Analysis of the FnrL regulon in *Rhodobacter*

capsulatus reveals limited regulon overlap with orthologues from *Rhodobacter sphaeroides* and *Escherichia coli*. *BMC Genomics* **2015**, *16*.

29. Tomasch, J.; Wang, H.; Hall, A.T.K.; Patzelt, D.; Preuße, M.; Brinkmann, H.; Bhuju, S.; Jarek, M.; Geffers, R.; Lang, A.S. Packaging of *Dinoroseobacter shibae* DNA into Gene Transfer Agent particles is not random. *Genome Biol. Evol.* **2018**, *10*, 359–369.
30. Laass, S.; Kleist, S.; Bill, N.; Drüppel, K.; Kossmehl, S.; Wöhlbrand, L.; Rabus, R.; Klein, J.; Rohde, M.; Bartsch, A.; et al. Gene regulatory and metabolic adaptation processes of *Dinoroseobacter shibae* DFL12 T during oxygen depletion. *J. Biol. Chem.* **2014**, *289*, 13219–13231.
31. Kumka, J.E.; Schindel, H.; Fang, M.; Zappa, S.; Bauer, C.E. Transcriptomic analysis of aerobic respiratory and anaerobic photosynthetic states in *Rhodobacter capsulatus* and their modulation by global redox regulators RegA, FnrL and CrtJ. Microb. *Genomics* **2017**, *3*, e000125.
32. Peña-Castillo, L.; Mercer, R.G.; Gurinovich, A.; Callister, S.J.; Wright, A.T.; Westbye, A.B.; Beatty, J.T.; Lang, A.S. Gene co-expression network analysis in *Rhodobacter capsulatus* and application to comparative expression analysis of *Rhodobacter sphaeroides*. *BMC Genomics* **2014**, *15*, 730.
33. Mercer, R.G.; Callister, S.J.; Lipton, M.S.; Pasa-tolic, L.; Strnad, H.; Paces, V.; Beatty, J.T.; Lang, A.S.; Acteriol, J.B. Loss of the response regulator CtrA causes pleiotropic effects on gene expression but does not affect growth phase regulation in *Rhodobacter capsulatus*. *J. Bacteriol.* **2010**, *192*, 2701–2710.
34. Smyth, G.K. Limma: Linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor; Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S., Eds.; Springer: New York, NY, USA, **2005**; pp. 397–420.
35. Irizarry, R.A.; Gautier, L.; Huber, W.; Bolstad, B.M. makecdfenv: CDF Environment Maker. R Packag. Version 1.62.0. 2019. Available online: https://rdrr.io/bioc/makecdfenv/ (accessed on 1 March 2019).

36. Gautier, L.; Cope, L.; Bolstad, B.M.; Irizarry, R.A. Affy-analysis of affymetrix GeneChip data at the probe level. *Bioinformatics* **2004**, *20*, 307–315.

37. Westbye, A.; Kater, L.; Wiesmann, C.; Ding, H.; Yip, C.; Beatty, J. The protease ClpXP and the PAS-domain protein DivL regulate CtrA and gene transfer agent production in *Rhodobacter capsulatus*. *Appl. Environ. Microbiol.* **2018**, 84, e00275-18.

38. Brilli, M.; Fondi, M.; Fani, R.; Mengoni, A.; Ferri, L.; Bazzicalupo, M.; Biondi, E.G. The diversity and evolution of cell cycle regulation in alpha-proteobacteria: A comparative genomic analysis. *BMC systems biology* **2010**, 4, 52.

39. Wang, H.; Tomasch, J.; Jarek, M.; Wagner-Döbler, I. A dual-species co-cultivation system to study the interactions between Roseobacters and dinoflagellates. *Front. Microbiol.* **2014**, *5*, 311.

40. Pallegar, P.; Peña-Castillo, L.; Evan, L.; Mark, G.; Lang, A.S. Cyclic-di-GMP-mediated regulation of gene transfer and motility in *Rhodobacter capsulatus*. *J. Bacteriol.* **2020**, *202*, e00554-19.

41. Hammond, J.H.; Dolben, E.F.; Smith, T.J.; Bhuju, S.; Hogan, D.A. Links between Anr and quorum sensing in *Pseudomonas aeruginosa* biofilms. *J. Bacteriol.* **2015**, *197*, 2810–2820.

42. Heylen, K.; Gevers, D.; Vanparys, B.; Wittebolle, L.; Geets, J.; Boon, N.; De Vos, P. The incidence of *nirS* and *nirK* and their genetic heterogeneity in cultivated denitrifiers. *Environ. Microbiol.* **2006**, *8*, 2012–2021.

43. Wang, Y.; Gao, L.; Rao, X.; Wang, J.; Yu, H.; Jiang, J.; Zhou, W.; Wang, J.; Xiao, Y.; Li, M.; et al. Characterization of *lasR*-deficient clinical isolates of *Pseudomonas aeruginosa*. *Sci. Rep.* **2018**, *8*, 13344.

44. Barraud, N.; Hassett, D.J.; Hwang, S.-H.; Rice, S.A.; Kjelleberg, S.; Webb, J.S. Involvement of nitric oxide in biofilm dispersal of *Pseudomonas aeruginosa*. *J. Bacteriol.* **2006**, *188*, 7344–7353.

45. Toyofuku, M.; Nomura, N.; Fujii, T.; Takaya, N.; Maseda, H.; Sawada, I.; Nakajima, T.; Uchiyama, H. Quorum sensing regulates denitrification in *Pseudomonas aeruginosa PAO1*. *J. Bacteriol.* **2007**, *189*, 4969–4972.

46. Elsen, S.; Swem, L.R.; Swem, D.L.; Bauer, C.E. RegB/RegA, a highly conserved redox-responding global two-component regulatory system. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 263–279.

47. Cheng, Z.; Li, K.; Hammad, L.A.; Karty, J.A.; Bauer, C.E. Vitamin B12 regulates photosystem gene expression via the CrtJ antirepressor AerR in *Rhodobacter capsulatus*. *Mol. Microbiol.* **2014**, *91*, 649–664.

48. Francez-Charlot, A.; Kaczmarczyk, A.; Vorholt, J.A. The branched CcsA/CckA-ChpT-CtrA phosphorelay of *Sphingomonas melonis* controls motility and biofilm formation. *Mol. Microbiol.* **2015**, *97*, 47–63.

49. Cutruzzolà, F.; Frankenberg-Dinkel, N. Origin and impact of nitric oxide in *Pseudomonas aeruginosa* biofilms. *J. Bacteriol.* **2015**, *198*, 55–65.

50. Boettcher, K.J.; Ruby, E.G.; Mcfall-Ngai, M.J. Bioluminescence in the symbiotic squid *Euprymna scolopes* is controlled by a daily biological rhythm. *J. Comp. Physiol*. A **1996**, *179*, 65–73.

51. Wang, Y.; Dufour, Y.S.; Carlson, H.K.; Donohue, T.J.; Marletta, M.A.; Ruby, E.G. H-NOX-mediated nitric oxide sensing modulates symbiotic colonization by *Vibrio fischeri*. *Proc. Natl. Acad. Sci.* USA **2010**, *107*, 8375–8380.

52. Bouchard, J.N.; Yamasaki, H. Heat stress stimulates nitric oxide production in *Symbiodinium microadriaticum*: A possible linkage between nitric oxide and the coral bleaching phenomenon. *Plant. Cell Physiol.* **2008**, *49*, 641–652.

3 CHAPTER 3: Connection between chromosomal location and function of CtrA phosphorelay genes in Alphaproteobacteria

3.1 Abstract

Most bacterial chromosomes are circular, with replication starting at one origin (*ori*) and proceeding on both replichores toward the terminus (*ter*). Several studies have shown that the location of genes relative to *ori* and *ter* can have profound effects on regulatory networks and physiological processes. The CtrA phosphorelay is a gene regulatory system conserved in most alphaproteobacteria. It was first discovered in *Caulobacter crescentus* where it controls replication and division into a stalked and a motile cell in coordination with other factors. The locations of the *ctrA* gene and targets of this response regulator on the chromosome affect their expression through replication-induced DNA hemi-methylation and specific positioning along a CtrA activity gradient in the dividing cell, respectively. Here we asked to what extent the location of CtrA regulatory network genes might be conserved in the alphaproteobacteria. We determined the locations of the CtrA phosphorelay and associated genes in closed genomes with unambiguously identifiable *ori* from members of five alphaproteobacterial orders. The location of the phosphorelay genes was the least conserved in the Rhodospirillales followed by the Sphingomonadales. In the Rhizobiales a trend toward certain chromosomal positions could be observed. Compared to the other orders, the CtrA phosphorelay genes were conserved closer to *ori* in the Caulobacterales. In contrast, the genes were highly conserved closer to *ter* in the Rhodobacterales. Our data suggest selection pressure results in differential positioning of CtrA phosphorelay and associated genes in alphaproteobacteria, particularly in the orders Rhodobacterales, Caulobacterales and Rhizobiales that is worth deeper investigation.

3.2 Introduction

Most bacteria possess one circular chromosome. Replication is initiated through unwinding the two DNA strands at the origin of replication (*ori*) and proceeds on both replichores toward the terminus (*ter*). Here, the dimer of newly synthesized chromosomes is resolved, and cell division can be completed (reviewed by [1]). Close links between replication and organization of genes on the chromosome became

evident with the first complete bacterial genomes [2,3]. In many bacteria, genes are preferentially oriented co-directional to replication progression. This pattern probably evolved to avoid collisions between DNA and RNA polymerase complexes [4]. In recent years it also became apparent that the specific locations of genes can have a major influence on transcription levels and thereby control physiological processes [5]. For instance, chromosomal location results in differences in the copy- number of genes during replication. Therefore, genes that are more highly expressed, such as those encoding transcription and translation proteins, tend to be conserved near *ori* [6]. The importance of gene location has also been validated experimentally: Relocating *Vibrio cholerae* genes encoding ribosomal proteins to the *ter* region resulted in severe growth defects [7].

Positioning of genes on the chromosome might also be dictated by regulatory needs. In *Escherichia coli* and other gammaproteobacteria, genes coding for nucleoid-associated proteins and regulators are ordered according to their activities during the growth cycle [8]. For example, *rpoN*, expressed during exponential growth, is located closer to *ori* while *rpoS*, expressed in the stationary phase, is located closer to *ter*. The same trend was found for the targets of these sigma factors. One of the most fascinating examples is how replication- oriented location of regulatory genes is employed to control the timing of *Bacillus subtilis* spore formation [9]. Here, imbalance between the expression of *ori* and *ter* located members of a phosphorylation chain during replication inhibits activation of the sporulation-inducing transcription factor Spo0A. This ensures that spore formation is only induced in cells with two complete chromosomes.

Proper transcription of the important cell cycle regulatory gene *ctrA* of *Caulobacter crescentus* is dependent on its chromosomal location, too [10]. In alphaproteobacteria, the CtrA phosphorelay regulatory system is widely conserved [11,12]. We recently found that its key regulatory components are concentrated proximal to *ori* and *ter* in the Rhodobacterales *Dinoroseobacter shibae* and *Rhodobacter capsulatus* and, in contrast to *C. crescentus*, the *ctrA* gene itself is located close to *ter* in both organisms [13].

In this chapter, we will first provide a brief overview of the CtrA phosphorelay and its role in

controlling the cell cycle and other traits in different bacteria. We will focus on how changes in DNA methylation during replication and the formation of a phosphorylation gradient in predivisional cells influence the regulatory system. Then, we will show that chromosomal location of the regulatory genes is conserved to varying degrees within alphaproteobacterial orders and differs among them. We propose that one consequence of the differing gene locations might be altered timing of expression during the cell cycle. Understanding how their positioning shapes the functionality of the CtrA phosphorelay and associated genes might help to explain the evolution of distinct roles in different alphaproteobacterial orders.

3.3 CtrA and cell cycle control in Alphaproteobacteria

Replicating only once per cell division, the dimorphic bacterium *C. crescentus* displays a eukaryotic-like cell cycle (Figure 3.1A). The growth-arrested flagellated cell (G1 phase) can transform into a stalked cell that replicates (S phase) and prepares for division (G2 phase) into two physiologically different daughter cells. The old, stalked cell can directly undergo the next round of replication while the new, flagellated cell remains in a growth-arrested state [14]. Key to the replication-coupled differentiation is the orchestration of gene expression by an array of interconnected regulatory circuits (reviewed by [15]) among which the CtrA phosphorelay takes a leading role [16,17].

As part of the regulation of cell division events, autophosphorylation of the transmembrane histidine kinase CckA results in phosphorylation of the phosphotransferase ChpT, which in turn phosphorylates the response regulator CtrA [18]. Phosphorylated CtrA then activates differentiation-specific genes and inhibits replication initiation by blocking *ori* from binding by the replication initiator DnaA [19]. Replication rounds are controlled via the directed proteolysis of several transcriptional regulators, including fast degradation of CtrA in the stalked cell, mediated by ClpX, in order to enable initiation of replication (reviewed by [20]). The chromosomal position influences transcription of *ctrA* and CtrA targets through DNA methylation and establishment of a CtrA phosphorylation gradient, respectively (detailed in the following sections and Figure 3.1).

The *C. crescentus* chromosome is linearly ordered in the cell with the *ori* located at the stalked or

flagellated pole and *ter* oriented at the pole where the division plane will form [21]. The CcrM methyltransferase specifically methylates the adenosine in GANTC palindromic motifs. Expression of *ccrM* is restricted to the transition from S to G2 phase [22,23]. Thus, newly replicated DNA stays hemi-methylated until replication has finished (Figure 3.1A). Expression of *ori*-proximal *ctrA* is controlled by two different promoters (Figure 3.1B). Promoter P1 gets activated in the early S phase and CtrA then triggers its own expression from promoter P2 while inhibiting expression from P1 in pre-divisional cells and the flagellated daughter cell [24]. Activation of P1 requires hemi- methylation of an upstream GANTC site, and thus the replication fork has to pass the *ctrA* locus for this promoter to be active [10,25]. Activity of P1 is highest when the respective motif on the coding strand is methylated [111]. If *ctrA* is moved closer to *ter*, the P1-associated GANTC motif remains in the fully methylated state longer and the resulting delay of *ctrA* transcription leads to elongated flagellated daughter cells. The transcription factor responsible for P1 regulation is GcrA, which is active exclusively in S phase cells and oscillates with CtrA activity [26]. GcrA preferentially binds and activates promoters

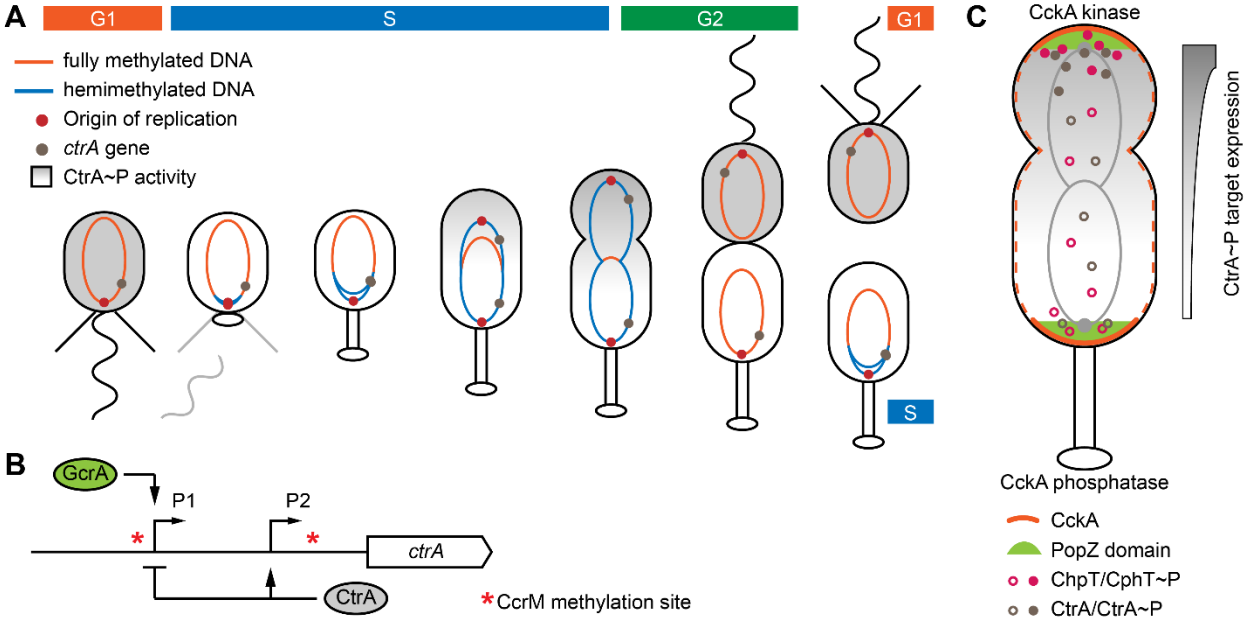

Figure 3.1: Mechanisms of *C. crescentus* differentiation for which chromosomal localization matters. (A) Changes of chromosome methylation state and CtrA activity during the *C. crescentus* cell cycle. Newly replicated DNA stays hemi-methylated during the S phase allowing *ctrA* transcription to be activated. CtrA activity is restricted to the late S phase and the

flagellated daughter cell. (B) Control of *ctrA* expression. Hemi-methylated P1 is activated by GcrA. Phosphorylated CtrA inhibits expression from P1 and activates expression from P2. (C) Establishment of a CtrA activity gradient through localized phosphorylation/dephosphorylation. The protein environment at the new pole triggers kinase functionality of CckA. The PopZ microdomain ensures proximity of phosphorelay components. Panel A inspired by Panis et al. (2015) and panel C inspired by Lasker et al. (2020).

carrying fully or hemi-methylated GANTC motifs [27,28]. Replication-controlled methylation also affects *ftsZ* expression, which encodes the divisome Z-ring protein. In this case the promoter is most active in the fully methylated state [29]. The regulatory function of CcrM is probably conserved broadly in alphaproteobacteria as GANTC motifs are enriched in intergenic regions on the vast majority of chromosomes [30].

CckA is dispersed throughout the inner membrane, but concentrates at the cell poles in pre-divisional cells [31]. It acts as a kinase at the new cell pole and as a phosphatase elsewhere. The switch in enzymatic activity is controlled by interaction with different sets of proteins [32]. Essential for triggering the kinase activity of CckA are its homo-oligomerization and direct interaction with the pseudo-kinase DivL, both concentrated at the cell poles [33]. Recently, [34] demonstrated the formation of diffusion-limiting microdomains at the cell poles that ensure close proximity of CckA, ChpT and CtrA in order to allow efficient phosphotransfer (Figure 3.1C). The polar localization of phosphorylating and dephosphorylating enzymatic chains ensures the formation of a CtrA activity gradient from the flagellated to the stalked pole in pre-divisional cells. When a promoter that is regulated exclusively by CtrA was repositioned on the chromosome, its expression decreased along the *ori-ter* axis, in accordance with the increasing distance from the flagellated cell pole [34].

The core components of the CtrA phosphorelay are highly conserved within the alphaproteobacteria and connected to accessory regulatory systems that are often restricted to specific orders [12]. In particular, most genes of the polarity module [35] essential for dimorphic development of *C. crescentus* are found only in members of the Caulobacterales and Rhizobiales orders, an exception being the more widely conserved *divL* gene. The CtrA regulon also differs among orders [11,12]. Flagellar genes are

controlled by CtrA in all studied orders, including the early branching Rhodospirillales, leading to the hypothesis that regulation of motility was the primordial role of CtrA and cell cycle control was acquired later [36]. Transcriptional activation of the DNA repair machinery, observed in several species, might also be a more ancient function of CtrA [37].

Construction of *ctrA* knockout mutants failed or they showed severe growth defects in the Caulobacterales *C. crescentus* [38-40] and *Hyphomonas neptunium* [41] as well as the Rhizobiales *Sinorhizobium meliloti* [42] and *Brucella abortus* [43], but has no negative effects on growth or viability in the Sphingomonadales *Sphingomonas melonis* [44], Rhodospirillales *Rhodospirillum centenum* [45,112] and *Magnetospirillum magneticum* [36], as well as the Rhodobacterales members studied so far [46-50]. While the influence of CtrA on replication has been demonstrated in *Sphingomonas melonis* and some Rhodobacterales, these bacteria lack a strict dimorphic lifestyle or polar growth that has been demonstrated for the Caulobacterales and Rhizobiales. These findings led to the hypothesis that essentiality of *ctrA* arose during the evolution of a lifestyle that couples differentiation to reproduction [41]. An intermediate step might have been a non-essential influence on replication and cell division.

3.4 Conserved location of CtrA-associated genes in alphaproteobacterial orders

Given that chromosomal localization influences expression of cell cycle-controlling genes in the model alphaproteobacterium *C. crescentus* and knowing that genes key to this process are conserved within this class, we asked if the location of these genes shows a pattern and if there are differences among orders in which the CtrA phosphorelay is an essential regulator of replication-coupled differentiation and those in which it is not. We used a dataset of 179 closed genomes from five alphaproteobacterial orders for which Ori-Finder [51] unambiguously identified *ori* and identified the orthologs of CtrA phosphorelay and associated genes using Proteinortho [52]. Only one representative strain was selected for each species to avoid species overrepresentation bias. Detailed analysis steps can be found in the Supplementary Material (Data sheet 1). The analyzed genomes from the Rhodospirillales, Sphingomonadales, Rhodobacterales, Caulobacterales and Rhizobiales are listed in Table S3.1, and the analyzed genes are listed in Tables S3.2, S3.3.

Figure 3.2A summarizes the localization analysis with the upper panel showing one representative genome and the lower panel showing the frequency of the respective genes within 20 segments on the *ori-ter* axis. The segments were obtained by normalizing all genomes to 100% and then dividing them into 20 sections of equal size. The orders are arranged phylogenetically with the earliest branching Rhodospirillales at the left and the latest branching Rhizobiales on the right [53]. In almost all analyzed genomes the *parAB* genes were located close to *ori* with some exceptions in the Rhodospirillales, Rhodobacterales, and Rhizobiales. This is in accordance with previous studies that found the par locus predominantly conserved close to *ori* [54]. The location of the other analyzed genes and the respective conservation differed among the orders.

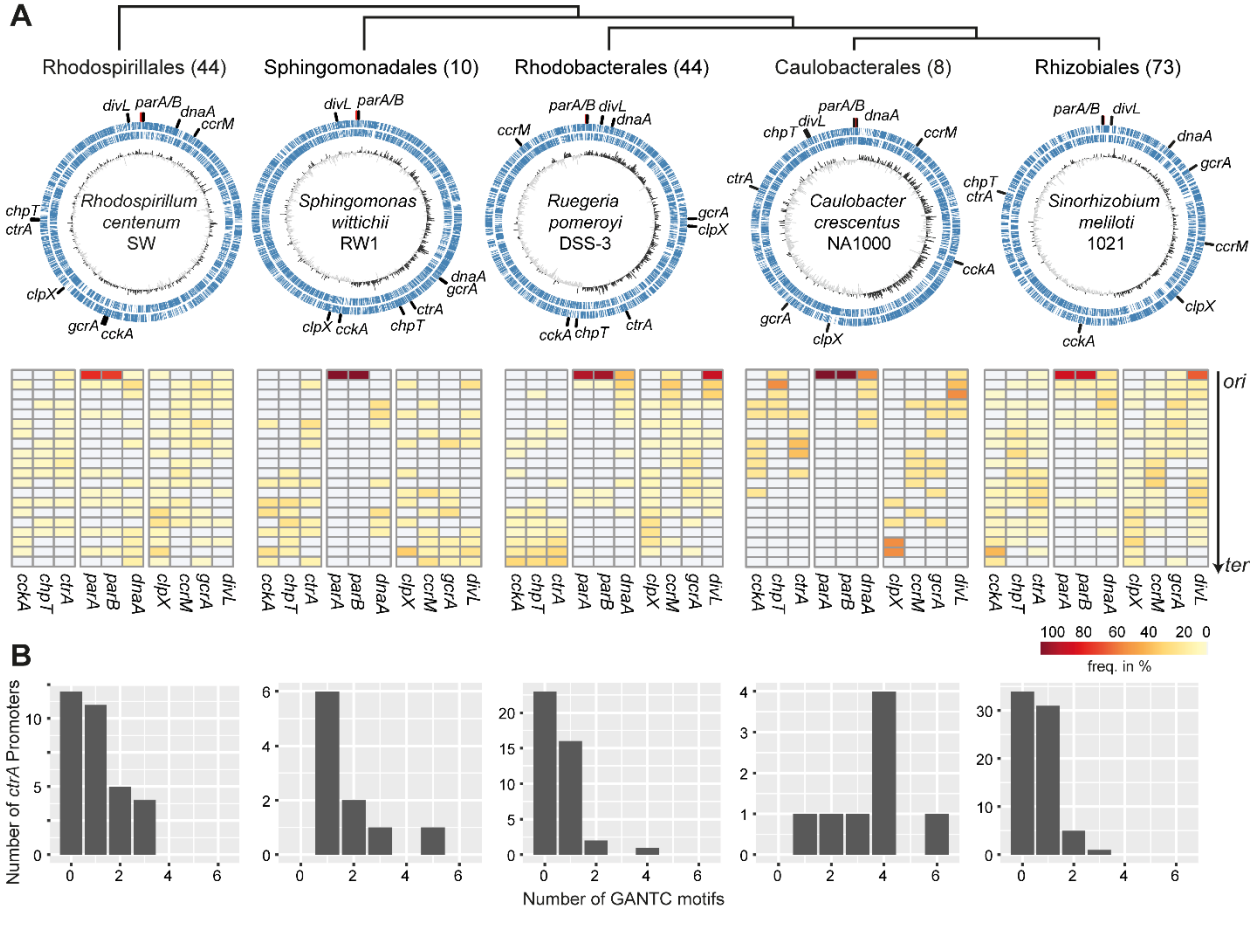

Figure 3.2: Chromosomal localization of CtrA phosphorelay component genes and methylation of the *ctrA* promoter region in alphaproteobacteria. (A) Alphaproteobacterial orders are arranged according to Muñoz-Gómez et al. (2019). The numbers of genomes per order are in brackets. Upper panel: representative chromosomes for each order with positions of regulators marked. Lower panel: percentage of regulator genes within 20 segments of the chromosomes of the particular order oriented along the *ori-*

ter axis. (B) Distribution of the number of CcrM methylation motifs in the *ctrA* promoter regions for each order.

We observed no conserved localization of the genes examined in the Rhodospirillales with the exceptions of *parAB* and *clpX*, which tended to be located closer to *ter* (Figure 3.2A). Surprisingly, and in contrast to the other orders, *ctrA*, *gcrA* and *ccrM* were identified in 72–82% of the genomes, whereas *chpT*, *cckA* and *divL* were identified in only 27–36% (Table S1). This might indicate that the selection pressure to maintain these genes is lower in Rhodospirillales than in the other orders. However, CckA and DivL are modular proteins, therefore their architecture might have evolved differently in this order and ChpT is a small protein that also shows greater divergence within orders [12,47], making definitive identification of homologs more difficult. Similarly, no clear distribution patterns of the genes analyzed were observed in the Sphingomonadales genomes. The only exceptions were *cckA* , which tended to be localized near *ter* (Figure 3.2A). Of note, in the Rhodospirillales and Sphingomonadales, the *dnaA* gene was not conserved close to *ori*, in contrast to the other three orders.

In the Caulobacterales the phosphorelay genes *ctrA*, *cckA* and *chpT* as well as *divL* showed conserved localization in the half of the chromosome closer to *ori*. The *clpX* gene was highly conserved in proximity to *ter*. By contrast, *ccrM* and *gcrA* were predominantly found midway between *ori* and *ter*. In the Rhizobiales, localization of *cckA* was conserved in the "lower half" of the chromosome with a peak close to *ter* while *chpT* and *ctrA* were preferentially located midway between *ori* and *ter*. The genes *clpX*, *ccrM* and *gcrA* showed similar trends as in the Caulobacterales. Interestingly, in many of the Rhizobiales genomes we identified two *divL* homologs, one of which was conserved near *ori* (Table S2).

In the Rhodobacterales a clear preference for *ter*-proximal localization was observed for *cckA*, *chpT* and *ctrA*, while *ccrM* was preferentially located close to *ori*. Like in the Rhizobiales, several genomes contained two paralogs of *divL* that were located mostly near *ori* (Table S2). We also analyzed the chromosomal position of other genes that are part of the CtrA regulon in this order and that regulate the gene transfer agent (GTA) gene cluster (Table S3). The direct activator of the GTA cluster *gafA* [55] and its neighbor (Dshi_1585 in *D. shibae*) were located in proximity to *ter* while the *rbaVW* genes that encode part of a partner-switching phosphorelay system [56] were preferentially found close to *ori*

(Figure S1). Interestingly, the CtrA-controlled genes [13] that are part of the DNA uptake and recombination machinery (*lexA*, *recA*, and *comECFM*) also showed a conserved location pattern.

As (hemi)-methylation is an important factor in the regulation of *ctrA* expression in *C. crescentus* we determined the number of GANTC motifs 300 bp upstream of the *ctrA* homologs in all orders (Figure 2B). This was determined by identifying the position of the *ctrA* gene, starting from the start codon, 300 bp upstream were used to look for the GANTC motif. All putative *ctrA* promotors contained at least one and up to five or six CcrM methylation sites in the Sphingomonadales and Caulobacterales, respectively. In 65% and 50% of all putative Rhodospirillales and Rhizobiales *ctrA* promoters, respectively, we identified the GANTC motif. The lowest number was found in Rhodobacterales where only 45% of all promoter regions contained this motif.

3.5 Discussion

Here, we evaluated whether key regulators associated with the CtrA phosphorelay have conserved chromosomal locations. The number of genomes available to analyze was small for the orders Caulobacterales and Sphingomonadales, leaving the possibility of a bias in our study. The employed Ori-Finder tool returned several possible *ori* positions for a considerable number of genomes that we excluded from further analysis. We found that the *parAB* genes might serve as a good anchor for manual curation of the *ori* position in Alphaproteobacteria. The locations of *ori* and *ter* can also be identified experimentally by sequencing DNA from growing cultures when there will be a coverage gradient decreasing from start toward end of replication [57,58]. This could be considered for all future genome sequencing projects.

Despite the limitations, we could identify localization patterns in all orders except for the early branching Rhodospirillales in which the conservation of the CtrA phosphorelay was also lower than in the other orders. Particularly striking was the strong conservation of the phosphorelay genes near *ter* on the Rhodobacterales chromosomes. This conserved localization is also remarkable because core genes in this order show very distinct location patterns among different species [59]. Localization near *ter* and the low occurrence of GANTC motifs in the *ctrA* promoter might indicate that replication-mediated changes of

the state of DNA methylation do not play a major role in regulation of gene expression in this order. On the other hand, establishment of a CtrA phosphorylation gradient might indeed also play a role in Rhodobacterales. The bifunctionality of CckA as a kinase and phosphatase has recently been demonstrated for *R. capsulatus* [60].

In some Rhodobacterales the CtrA phosphorelay is integrated into quorum sensing (QS) regulation [48,49,61]. CtrA-mediated QS communication induces subpopulation-specific responses, most notably the "decision" of a small number of cells to produce GTAs [13,62]. A loss of the CtrA phosphorelay genes is not lethal, the bacteria just resemble a "silent" population. The location of the phosphorelay genes close to *ter* might indicate that communication-induced differentiation is uncoupled from replication and cell division. It is also tempting to speculate that the location of genes controlling GTA expression at the opposite poles of the chromosome ensures repression of GTA production during replication, similar to spore formation in *B. subtilis* [9]. Indeed, no DNA packaging bias along the *ori-ter* axis has been observed for GTAs, which would be expected if they are produced in replicating cells [63,64]. In the Rhizobiales and Caulobacterales, however, most of the essential CtrA phosphorelay genes are located toward the upper half of the chromosome. This might result in their activation during replication as observed for *ctrA* of *C. crescentus* [10], leading to an interconnected essentiality of reproduction and physiological differentiation. Most essential *C. crescentus* genes are concentrated near *ori* or *ter* [39]. It would be interesting to see if this pattern is conserved in other species with a pronounced dimorphic lifestyle.

In conclusion, our analysis suggests selection pressure to fix the position of CtrA phosphorelay and associated genes in different chromosomal regions depending on their involvement in different cell physiological processes. This is particularly evident in the Rhodobacterales, Caulobacterales and Rhizobiales. Understanding the underlying evolutionary forces will require both comparative genomic analysis and experimental data beyond what is currently available for a limited number of established model organisms. Our analysis concentrated on the core components of the CtrA phosphorelay but could be expanded to include more accessory regulators and CtrA targets in the different orders. It would also

be interesting to identify highly related strains where recent chromosome rearrangements have led to different positions of genes of interest. In *Pseudomonas aeruginosa* a large-scale chromosome inversion resulted in large gene expression and physiological differences between two strains [65]. Similarly, analyzing the consequences of relocating genes, as has been done for *C. crescentus* and several other organisms, is a promising experimental approach for understanding the effects of chromosome positioning on gene regulation [7,10,29,34].

3.6 Supplementary materials

This chapter includes two supplementary files that are available in digital format using this link:

The PDF file contains the supplementary method section, supplementary figure S1 and the descriptions for the supplementary tables. These tables are a combined in three different tables in the Excel file, which contains the analyzed genomes from the Rhodospirillales, Sphingomonadales, Rhodobacterales, Caulobacterales and Rhizobiales, listed in Table S3.1, and the analyzed genes, listed in Tables S3.2, S3.3.

3.7 References

1. Reyes-Lamothe, R.; Nicolas, E.; Sherratt, D.J. Chromosome replication and segregation in bacteria. *Annu. Rev. Genet.* **2012**, *46*, 121–143.
2. Rocha, E.P.C. The replication-related organization of bacterial genomes. *Microbiology* **2004**, *150*, 1609–1627.
3. Touchon, M.; Rocha, E.P.C. Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harb. Perspect. Biol.* **2016**, *8*, 1–18.
4. Lang, K.S.; Merrikh, H. The Clash of Macromolecular Titans: Replication-Transcription Conflicts in Bacteria. *Annu. Rev. Microbiol.* **2018**, *72*, 71–88.
5. Slager, J.; Veening, J.-W. Hard-Wired Control of Bacterial Processes by Chromosomal Gene Location. **2016**, *24*, 788–800.
6. Couturier, E.; Rocha, E.P.C. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.* **2006**, *59*,

1506–1518.

7. Soler-Bistué, A.; Mondotte, J.A.; Bland, M.J.; Val, M.-E.; Saleh, M.-C.; Mazel, D. Genomic Location of the Major Ribosomal Protein Gene Locus Determines *Vibrio cholerae* Global Growth and Infectivity. *PLOS Genet.* **2015**, *11*, e1005156.
8. Sobetzko, P.; Travers, A.; Muskhelishvili, G. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. **2012**, *109*.
9. Narula, J.; Kuchina, A.; Lee, D.D.; Fujita, M.; Süel, G.M.; Igoshin, O.A. Chromosomal Arrangement of Phosphorelay Genes Couples Sporulation and DNA Replication. *Cell* **2015**, *162*, 328–337.
10. Reisenauer, A.; Shapiro, L. DNA methylation affects the cell cycle transcription of the CtrA global regulator in Caulobacter. *EMBO J.* **2002**, *21*, 4969–4977.
11. Panis, G.L.; Murray, S.R.; Viollier, P.H. Versatility of global transcriptional regulators in alpha-Proteobacteria: from essential cell cycle control to ancillary functions. *Microbiol. Rev.* **2015**, *39*, 120–133.
12. Brilli, M.; Fondi, M.; Fani, R.; Mengoni, A.; Ferri, L.; Bazzicalupo, M.; Biondi, E.G. The diversity and evolution of cell cycle regulation in alpha-proteobacteria: A comparative genomic analysis. *BMC Syst. Biol.* **2010**, *4*.
13. Koppenhöfer, S.; Wang, H.; Scharfe, M.; Kaever, V.; Wagner-Döbler, I.; Tomasch, J. Integrated Transcriptional Regulatory Network of Quorum Sensing, Replication Control, and SOS Response in *Dinoroseobacter shibae*. *Front. Microbiol.* **2019**, *10*.
14. Degnen, S.T.; Newton, A. Chromosome replication during development in *Caulobacter crescentus*. *J. Mol. Biol.* **1972**, *64*, 671–680.
15. Frandi, A.; Collier, J. Multilayered control of chromosome replication in *Caulobacter crescentus*. *Biochem. Soc. Trans.* **2019**, *47*, 187–196.
16. Laub, M.T.; McAdams, H.H.; Feldblyum, T.; Fraser, C.M.; Shapiro, L. Global analysis of the genetic network controlling a bacterial cell cycle. *Science*. **2000**, *290*, 2144–2148.

17. Laub, M. T. Genomic analysis of the genetic network controlling cell cycle progression in *Caulobacter crescentus*. **2002**.
18. Biondi, E.G.; Reisinger, S.J.; Skerker, J.M.; Arif, M.; Perchuk, B.S.; Ryan, K.R.; Laub, M.T. Regulation of the bacterial cell cycle by an integrated genetic circuit. *Nature* **2006**, *444*, 899–904.
19. Quon, K.C.; Yang, B.; Domian, I.J.; Shapiro, L.; Marczynski, G.T. Negative control of bacterial DNA replication by a cell cycle regulatory protein that binds at the chromosome origin. **1997**, 13600–13605.
20. Jenal, U. The role of proteolysis in the *Caulobacter crescentus* cell cycle and development. *Res. Microbiol.* **2009**, *160*, 687–695.
21. Yildirim, A.; Feig, M. High-resolution 3D models of *Caulobacter crescentus* chromosome reveal genome structural variability and organization. *Nucleic Acids Res.* **2018**, *46*, 3937–3952.
22. Zweiger, G.; Marczynski, G.; Shapiro, L. A *Caulobacter* DNA methyltransferase that functions only in the predivisional cell. *J. Mol. Biol.* **1994**, *235*, 472–485.
23. Laub, M.T.; McAdams, H.H.; Feldblyum, T.; Fraser, C.M.; Shapiro, L. Global analysis of the genetic network controlling a bacterial cell cycle. *Science* **2000**, *290*, 2144–2148.
24. Domian, I.J.; Reisenauer, A.; Shapiro, L. Feedback control of a master bacterial cell-cycle regulator. *Proc. Natl. Acad. Sci.* U. S. A. **1999**, *96*, 6648–6653.
25. Mohapatra, S.S.; Fioravanti, A.; Vandame, P.; Spriet, C.; Pini, F.; Bompard, C.; Blossey, R.; Valette, O.; Biondi, E.G. Methylation-dependent transcriptional regulation of crescentin gene (*creS*) by GcrA in *Caulobacter crescentus. Mol. Microbiol.* **2020**, *114*, 127–139.
26. Holtzendorff, J., Hung, D., Brende, P., Reisenauer, A., Viollier, P. H., McAdams, H. H., et al. Oscillating global regulators control the genetic circuit driving a bacterial cell cycle. *Science* **2004**, *304*, 983–987.
27. Fioravanti, A.; Fumeaux, C.; Mohapatra, S.S.; Bompard, C.; Brilli, M.; Frandi, A.; Castric, V.; Villeret, V.; Viollier, P.H.; Biondi, E.G. DNA Binding of the Cell Cycle Transcriptional regulator GcrA depends on N6-Adenosine methylation in *Caulobacter crescentus* and other

Alphaproteobacteria. **2013**, *9*.

28. Haakonsen, D.L.; Yuan, A.H.; Laub, M.T. The bacterial cell cycle regulator *gcrA* is a σ70 cofactor that drives gene expression from a subset of methylated promoters. *Genes Dev*. **2015**, *29*, 2272–2286.
29. Gonzalez, D.; Collier, J. DNA methylation by CcrM activates the transcription of two genes required for the division of *Caulobacter crescentus*. *Mol. Microbiol*. **2013**, *88*, 203–218.
30. Gonzalez, D.; Kozdon, J.; Mcadams, H.H.; Shapiro, L.; Collier, J. The functions of DNA methylation by CcrM in *Caulobacter crescentus*: a global approach. *Nucleic Acids Res*. **2014**, *42*, 3720–3735.
31. Angelastro, P.S.; Sliusarenko, O.; Jacobs-Wagner, C. Polar localization of the CckA histidine kinase and cell cycle periodicity of the essential master regulator CtrA in *Caulobacter crescentus*. *J. Bacteriol.* **2010**, *192*, 539–552.
32. Tsokos, C.G.; Perchuk, B.S.; Laub, M.T. A Dynamic Complex of Signaling Proteins Uses Polar Localization to Regulate Cell-Fate Asymmetry in *Caulobacter crescentus*. *Dev. Cell* **2011**, *20*, 329–341.
33. Mann, T.H.; Shapiro, L. Integration of cell cycle signals by multi-PAS domain kinases. *Proc. Natl. Acad. Sci.* U. S. A. **2018**, *115*, E7166–E7173.
34. Lasker, K.; von Diezmann, L.; Zhou, X.; Ahrens, D.G.; Mann, T.H.; Moerner, W.E.; Shapiro, L. Selective sequestration of signaling proteins in a membraneless organelle reinforces the spatial regulation of asymmetry in *Caulobacter crescentus*. *Nat. Microbiol*. **2020**, *5*, 418–429.
35. Teeseling, M.C.F. Van; Thanbichler, M. Generating asymmetry in a changing environment: cell cycle regulation in dimorphic alphaproteobacteria. *Biol. Chem.* **2020**.
36. Greene, S.E.; Brilli, M.; Biondi, E.G.; Komeili, A. Analysis of the CtrA Pathway in *Magnetospirillum* Reveals an Ancestral Role in Motility in Alphaproteobacteria. **2012**, *194*, 2973–2986.
37. Poncin, K.; Gillet, S.´ E.; De Bolle, X. Learning from the master: targets and functions of the

CtrA response regulator in *Brucella abortus* and other alphaproteobacteria. *FEMS Microbiol. Rev.* **2018**, *019*, 500–513.

38. Quon, K.C.; Marczynski, G.T.; Shapiro, L. Cell cycle control by an essential bacterial two-component signal transduction protein. *Cell* **1996**, *84*, 83–93.
39. Christen, B.; Abeliuk, E.; Collier, J.M.; Kalogeraki, V.S.; Passarelli, B.; Coller, J.A.; Fero, M.J.; McAdams, H.H.; Shapiro, L. The essential genome of a bacterium. *Mol. Syst. Biol*. **2011**, 7, 1–7.
40. Guzzo, M.; Castro, L.K.; Reisch, C.R.; Guo, M.S.; Laub, M.T. Molecular Biology and Physiology Knockdown in *Caulobacter crescentus*. **2020**, *11*, 1–16.
41. Leicht, O.; van Teeseling, M.C.F.; Panis, G.; Reif, C.; Wendt, H.; Viollier, P.H.; Thanbichler, M. Integrative and quantitative view of the CtrA regulatory network in a stalked budding bacterium. *PLoS genetics*. **2020**, *16*, 4, e1008724.
42. Barnett, M.J.; Hung, D.Y.; Reisenauer, A.; Shapiro, L.; Long, S.R. A homolog of the CtrA cell cycle regulator is present and essential in *Sinorhizobium meliloti*. *J. Bacteriol*. **2001**, *183*, 3204–3210.
43. Bellefontaine, A.F.; Pierreux, C.E.; Mertens, P.; Vandenhaute, J.; Letesson, J.J.; De Bolle, X. Plasticity of a transcriptional regulation network among alphaproteobacteria is supported by the identification of CtrA targets in *Brucella abortus*. *Mol. Microbiol*. **2002**, *43*, 945–960.
44. Francez-Charlot, A.; Kaczmarczyk, A.; Vorholt, J.A. The branched CcsA/CckA-ChpT-CtrA phosphorelay of *Sphingomonas melonis* controls motility and biofilm formation. *Mol. Microbiol.* **2015**, *97*, 47–63.
45. Bird, T.H.; MacKrell, A. A CtrA homolog affects swarming motility and encystment in *Rhodospirillum centenum. Arch. Microbiol*. **2011**, *193*, 451–459.
46. Miller, T.R.; Belas, R. Motility is involved in *Silicibacter* sp. TM1040 interaction with dinoflagellates. *Environ. Microbiol*. **2006**, *8*, 1648–1659.
47. Mercer, R.G.; Callister, S.J.; Lipton, M.S.; Pasa-tolic, L.; Strnad, H.; Paces, V.; Beatty, J.T.; Lang, A.S.; Acteriol, J.B. Loss of the response regulator CtrA causes pleiotropic effects on gene

expression but does not affect growth phase regulation in *Rhodobacter capsulatus. J. Bacteriol*. **2010**, *192*, 2701–2710.

48. Wang, H.; Ziesche, L.; Frank, O.; Michael, V.; Martin, M.; Petersen, J.; Schulz, S.; Wagner-Döbler, I.; Tomasch, J. The CtrA phosphorelay integrates differentiation and communication in the marine alphaproteobacterium *Dinoroseobacter shibae*. *BMC Genomics* **2014**, *15*.
49. Zan, J.; Heindl, J.E.; Liu, Y.; Fuqua, C.; Hill, R.T. The CckA-ChpT-CtrA phosphorelay system is regulated by quorum sensing and controls flagellar motility in the marine sponge symbiont *Ruegeria* sp. KLH11. *PLoS One* **2013**, *8*.
50. Hernández-Valle, J.; Sanchez-Flores, A.; Poggio, S.; Dreyfus, G.; Camarena, L. The CtrA regulon of *Rhodobacter sphaeroides* favors adaptation to a particular lifestyle. *J. Bacteriol.* **2020**, *202*, 1–16.
51. Gao, F.; Zhang, C.-T. Ori-Finder: A web-based system for finding *oriCs* in unannotated bacterial genomes. **2008**.
52. Lechner, M.; Findeiß, S.; Steiner, L.; Marz, M.; Stadler, P.F.; Prohaska, S.J. Proteinortho: Detection of (Co-)orthologs in large-scale analysis; *BMC Bioinformatics* **2011,** *12*, 1, 1-9**.**
53. Muñoz-Gómez, S.A.; Hess, S.; Burger, G.; Franz Lang, B.; Susko, E.; Slamovits, C.H.; Roger, A.J. An updated phylogeny of the alphaproteobacteria reveals that the parasitic Rickettsiales and holosporales have independent origins. *Elife* **2019**, *8*, 1–23.
54. Livny, J.; Yamaichi, Y.; Waldor, M.K. Distribution of centromere-like *parS* sites in bacteria: Insights from comparative genomics. *J. Bacteriol*. **2007**, *189*, 8693–8703.
55. Fogg, P.C.M. Identification and characterization of a direct activator of a gene transfer agent. *Nat. Commun.* **2019**, *10*.
56. Mercer, R.G.; Lang, A.S. 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.* **2014**, *14*.
57. Skovgaard, O.; Bak, M.; Løbner-Olesen, A.; Tommerup, N. Genome-wide detection of

chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome Res.* **2011**, *21*, 1388–1393.

58. Jung, A.; Raßbach, A.; Pulpetta, R.L.; van Teeseling, M.C.F.; Heinrich, K.; Sobetzko, P.; Serrania, J.; Becker, A.; Thanbichler, M. Two-step chromosome segregation in the stalked budding bacterium *Hyphomonas neptunium. Nat. Commun*. **2019**, *10*, 1–16.
59. Kopejtka, K.; Lin, Y.; Jakubovičová, M.; Koblížek, M.; Tomasch, J. Clustered core- and pan-genome content on Rhodobacteraceae chromosomes. *Genome Biol. Evol.* **2019**, *11*, 2208-2217.
60. Farrera-Calderon, R.G.; Pallegar, P.; Westbye, A.B.; Wiesmann, C.; Lang, A.S.; Beatty, J.T. The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer agent (RcGTA) production is bi-directional and regulated by cyclic-di-GMP. *J. Bacteriol.* **2020**, *5*, 1–17.
61. Leung, M.M.; Brimacombe, C.A.; Beatty, J.T.; Beatty, C.J.T. Transcriptional regulation of the *Rhodobacter capsulatus* response regulator CtrA. *Microbiology* **2013**, *159*, 96–106.
62. Ding, H.; Grüll, M.P.; Mulligan, M.E.; Lang, A.S.; Thomas Beatty, J. Induction of *Rhodobacter capsulatus* gene transfer agent gene expression is a bistable stochastic process repressed by an extracellular calcium-binding *rtx* protein homologue. *J. Bacteriol.* **2019**, *201*, 1–18.
63. Tomasch, J.; Wang, H.; Hall, A.T.K.; Patzelt, D.; Preuße, M.; Brinkmann, H.; Bhuju, S.; Jarek, M.; Geffers, R.; Lang, A.S. Packaging of *Dinoroseobacter shibae* DNA into Gene Transfer Agent particles is not random. *Genome Biol. Evol.* **2018**, *10*, 359–369.
64. Hynes, A.P.; Mercer, R.G.; Watton, D.E.; Buckley, C.B.; Lang, A.S. 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.* **2012**, *85*, 314–325.
65. Irvine, P.; Emanuel, K.; He, J.; Horowitz, L.W.; Vecchi, G.; Keith, D. Halving warming with idealized solar geoengineering moderates key climate hazards. *Nat. Clim. Chang*. **2019**, *9*, 295–299.

4 CHAPTER 4: Shared properties of gene transfer agent and core genes revealed by comparative genomics of Alphaproteobacteria

4.1 Abstract

Gene transfer agents (GTAs) are phage-like particles that transfer pieces of cellular genomic DNA to other cells. Homologs of the *Rhodobacter capsulatus* GTA (RcGTA) structural genes are widely distributed in the Alphaproteobacteria and particularly well conserved in the order Rhodobacterales. Possible reasons for their widespread conservation are still being discussed. It has been suggested that these alphaproteobacterial elements originate from a prophage that was present in an ancestral bacterium and subsequently evolved into a GTA that is now widely maintained in extant descendant lineages. Here, we analyzed genomic properties that might relate to the conservation of these alphaproteobacterial GTAs. This revealed that the chromosomal locations of the GTA gene clusters are biased. They primarily occur on the leading strand of DNA replication, at large distances from long repetitive elements and thus are in regions of lower plasticity, and in areas of extreme GC skew, which also accumulate core genes. These extreme GC skew regions arise from the preferential use of codons with an excess of G over C, a distinct phenomenon from the elevated GC content that has previously been found to be associated with GTA genes. The observed properties, along with their high level of conservation, show that GTA genes share multiple features with core genes in the examined lineages of the Alphaproteobacteria.

4.2 Introduction

Gene transfer agents (GTAs) are phage-like particles that transfer small pieces of genomic DNA between cells that have been identified in multiple Gram-negative bacteria and one archaeon [1]. Currently there are five distinct GTA types known, each appearing to have an independent evolutionary origin and varying breadths of taxonomic distribution [1]. Homologs of the *Rhodobacter capsulatus* GTA (RcGTA) genes are found in the genomes of members of multiple orders of the class Alphaproteobacteria [2,3], and functionality of these RcGTA-like elements has been confirmed in divergent members of the alphaproteobacterial order Rhodobacterales [1,4,5,6]. It has been suggested that these GTA elements are

descendants of a prophage that integrated into the genome of an ancestral alphaproteobacterium, subsequently lost multiple phage-specific features, such as DNA replication and packaging specificity, and acquired mutations that resulted in a reduced head size [1]. This proto-GTA was then maintained through to the evolution of the extant lineages where the GTA genes are under cellular control.

Most of the RcGTA structural genes are located in a gene cluster of approximately 14 kb [7] that is conserved in the genomes of almost all examined members of the order Rhodobacterales [3]. This set of genes is also conserved to varying degrees in about half of the members of the alphaproteobacterial orders Rhizobiales, Sphingomonadales and Caulobacterales [3]. Possible reasons for the widespread conservation of these GTA genes are still being discussed. On the one hand, GTAs might contribute to transfer of beneficial genes among cells [1,7,8,9]. This hypothesis is supported by findings on the unrelated GTA produced by *Bartonella* spp., where particle release and DNA uptake are restricted to the subpopulations with the highest fitness [10] and thus the GTAs are more likely to transfer genes that offer a benefit to the recipient cell. However, a modelling approach did not find support for any fitness advantage to GTA-producing over non-GTA-producing populations as the resulting gene transfer did not compensate for the loss caused by GTA release [11], which requires cell lysis of the producing subpopulation of cells [8,11]. Perhaps GTAs are simply defective remnants of previously functional prophages [8,12], but this is difficult to reconcile with the findings that the RcGTA-related genes are under purifying selection [13] and that the production of RcGTA is co-regulated with the ability of cells to receive DNA from the particles [14]. Alternatively, an immunological function of GTAs has been proposed where GTAs transfer prophage DNA that can be incorporated into a recipient cell's CRISPR-Cas array and thereby "vaccinate" the cells before an actual infection takes place [11].

The RcGTA family gene clusters show an increased GC content ((G+C)/(A+T+G+C)) relative to the rest of the genome, which results from a bias in the encoded proteins to contain amino acids that have a lower carbon content [3]. This could be important for the production of GTAs during nutrient limitation, as observed for RcGTA [12]. A previous analysis of different factors associated with GTA gene expression [15] drew our attention to the localization of the GTA gene clusters in regions of especially

high GC skew, which is the normalized ratio of guanine to cytosine ((G-C)/(G+C)) and different from absolute GC content, in two considered species, *R. capsulatus* and *Dinoroseobacter shibae*. Circular bacterial chromosomes can be divided into two halves, the right and left replichores, based on the orientation relative to the origin (*ori*) and terminus (*ter*) of replication. The GC skew typically has positive values on the right replichore and negative values on the left replichore as guanine and cytosine dominate on the leading and lagging strand, respectively [16,17,18,19,20,21]. This asymmetric distribution is thought to be largely driven by deamination of cytosine to thymine, which might be affected by DNA replication since the distribution pattern matches replication directionality [21,22]. This chromosomal composition bias is increased by some factors, such as an elevated growth rate [20], and decreased by others, such as recombination [23], and an overall more pronounced GC skew correlates with lower numbers of repeats [24]. Repetitive sequences and mobile genetic elements such as prophages can facilitate chromosomal rearrangements that reduce genomic stability, although these increase genomic plasticity and can provide an organism with greater adaptability [25–27].

Motivated by these previous observations related to DNA composition patterns, we performed a comprehensive genome sequence and structure analysis focused on patterns of GTA gene cluster conservation in four orders of the Alphaproteobacteria. This revealed trends in their localization, GC skew, codon usage, and potential DNA methylation and led to the overall conclusion that GTA genes share multiple properties with core genes in these bacteria.

4.3 Methods

Analyses were carried out with R studio version 4.0.3 and relevant packages (Table 4.1).

4.3.1 Genome dataset and chromosome reorientations

Closed genomic sequences from bacteria within four alphaproteobacterial orders, the Rhodobacterales (n=147), Sphingomonadales (n=114), Caulobacterales (n=30) and Rhizobiales (n=462), were obtained from the NCBI GenBank assembly database (e.g., https://www.ncbi.nlm.nih.gov/assembly/?term=Rhodobacterales) on 12 March 2019. Accession numbers of sequences used are provided in Table S4.1. We note that there have been subsequent taxonomic

revisions among these bacteria but do not believe these detract from the utility or meaning of our analyses as based on this previous, long-standing taxonomic organization. The origin of replication (*ori*) was identified on each chromosome using Ori-Finder and default settings [28]. The ptt files were generated from gbff files (https://github.com/sgivan/gb2ptt#gb2ptt), downloaded from NCBI on 23 April 2019. Only chromosomes where one *ori* could be unambiguously identified were subsequently included in the investigation. We next determined the locations of the gene encoding the GTA major capsid protein (MCP) and found that all were located on the presumed major chromosomes (largest replicons) that were used for subsequent analyses.

Depending on the analysis, the positions of *ori* or the GTA MCP gene were used to reorient the DNA fasta and gff files using custom R functions that are available within the newly developed package "reorientateCircGenomes" (https://github.com/SonjaElena/reorientateCircGenomes.git). This package simplifies reorientation of sequences within fasta and gff files that originate from NCBI or Prokka based on base pair location or ProteinID. It can also be used to visualize circular chromosomes with strand information, GC skew, and locations of selected genes.

4.3.2 Homology analysis

To identify homologous proteins, and thus gene families, in the genomes from the different orders, all proteins were blasted against each other and a matrix was generated for each order using Proteinortho version 5.16b [29]. The criteria to be considered a homolog were an e-value ≤1e-05, identity ≥30%, and coverage ≥75%. Based on the identification of specific genes of interest (e.g., the GTA major capsid protein gene) in reference organisms' genomes, this database was then used to identify homologs in the other genomes. The reference organisms for the Rhodobacterales, Sphingomonadales, Caulobacterales, and Rhizobiales are *Dinoroseobacter shibae* DFL 12 (= DSM 16493), *Sphingopyxis alaskensis* RB2256, *Brevundimonas subcrescentus* ATCC 15264, and *Brucella suis* 1330, respectively. A protein was designated a core protein if it was present in ≥90% of all genomes within an order.

4.3.3 DNA composition analysis

The GC skew was calculated as (G-C)/(G+C) for a sliding window of 10000 bp. For cumulative

visualizations, the sliding window size was set to 0.1% of the chromosome lengths and then the mean GC skew of all organisms being considered was calculated.

GC skew peaks were identified independently of the reversal of the right and left replichores by using sliding quantiles to identify the local GC skew minima and maxima. The positions on the chromosome with GC skew values that belonged to the upper or lower 3% of the GC skew values in a sliding window of 150000 bp were identified. Genes located at these locations were then identified for further analyses. The relative GC skew was calculated as (GCsample–GCcontrol)/GCcontrol [3]. The GC content of genes was calculated as G+C over the length of each protein coding region.

To examine the GC skew of prophages in relation to their respective host genomes, the insertion positions on the chromosome were determined using PHASTER [30]. Hits that overlapped with GTA locations, identified by ProteinOrtho, were attributed to be part of GTAs. To compare the GC contents and skews of GTAs and phages, the genomic sequences of phages that infect bacteria in the four considered orders were downloaded from the NCBI virus database (on 4 November 2020). The GC contents and skews were determined over sliding windows of 1000 bp for the phages and per gene for the GTAs.

4.3.4 Identification of repeats, methylation motif sites and large-scale inversions

Repetitive elements were identified with RepSeek [31]. Only repeats with a length >800 bp and identity >90% were included to focus the analyses on long and highly similar repeats that are more prone to recombination [32]. We excluded overlapping repeats because these are probably not a major reason for large-scale chromosomal rearrangements. CcrM methylation potential was examined by searching for the GANTC motif in the DNA sequences. To rule out that the pattern we observed was caused by base composition instead of a possible methylation site we also examined variations of this pattern that have equal base compositions (CGANT, CTGAN) (Figure S1). Large-scale chromosome rearrangements were identified by visual inspection using MAUVE with the option progressiveMauve [33].

4.3.5 Codon usage

The DNA sequences for each open reading frame (ORF) were extracted from the genome fasta

nucleotide acid files, based on the position information in the annotation (gff) files using the Biostrings and Genomic Ranges packages in R (Table 4.1). After sorting the ORFs of each organism according to whether they were found in GC skew peaks or not, the occurrence of each codon in each group was counted. The means per codon for both groups and for each organism were then calculated. The relative codon usage was determined as (peak–not-peak)/not-peak. For visualization, codons were grouped based on the encoded amino acids.

4.3.6 Phylogenetics

To identify closely related strains in which the location of the GTA gene cluster was switched between right and left replichores, a phylogenetic tree was generated for each order using RNA polymerase β protein (RpoB) amino acid sequences. Those RpoB sequences were identified with ProteinOrtho using sequences AAV96733.1, AAN30162.1, ANF54622.1 and ABF53199 as references for the Rhodobacterales, Rhizobiales, Caulobacterales, and Sphingomonadales, respectively. The alignments and trees were generated with MEGAX [34]. The default settings were used for pairwise and multiple alignments. Partial deletion with a delay divergent cutoff of 30% was used for gaps and missing data. The trees were constructed with the maximum-likelihood method. The branching patterns were evaluated using 100 bootstrap replicates, and the LG model was applied with gamma distribution at invariant sites. The site coverage cutoff was 95%.

Table 4.1. List of R (version 4.0.3) packages used in this study.

Package name	Version	Reference
Tidyverse	1.3.0	
Biostrings	2.54.0	
GenomicRanges	1.38.0	
Ggbio	1.34.0	[60]
XML	3.99-0.3	
RCurl	1.98-1.1	

Ringo	1.50.0	[61]
BSgenome	1.54.0	
ggExtra	0.9	
DescTools	0.99.34	
coRdon	1.4.0	[62]
rlist	0.4.6.1	
genoPlotR	0.8.9	[63]
reorientateCircGenomes	0.0.1	This study

4.4 Results and Discussion

4.4.1 Dataset generation

Closed genomic sequences were downloaded from the NCBI database for four orders of the class Alphaproteobacteria. These orders were chosen based on their high numbers of available complete genomic sequences and their high level of GTA gene cluster conservation [3]. The Rhizobiales had the most genomes (462), followed by the Rhodobacterales (147), Sphingomonadales (114) and Caulobacterales (30) (Table 4.2). To standardize analyses of gene localization, repeats and methylation motifs, all chromosomes were reoriented to the origin of replication (*ori*). Genomes were excluded when a single *ori* could not be unambiguously identified. For GTA gene cluster-related analyses, further dataset reduction was made based on presence of the major capsid protein (MCP) gene. These selection criteria resulted in a reduced set of genomes available for analysis (Table 4.2). Repeating the analyses with only one representative strain per species for the Rhodobacterales and Rhizobiales showed that there was no bias in the results caused by overrepresentation of strains for certain species (data not shown).

Table 4.2. Number of genomes available for analysis based on selection criteria.

Order	Closed genomes	One unambiguous origin of replication (*ori*)	GTA major capsid protein (MCP) gene[a]	One representative strain per species
Rhizobiales	462	133	76	71
Rhodobacterales	147	70	59	45
Sphingomonadales	114	17	6	17
Caulobacterales	30	8	4	8

[a] Based on more than one representative strain per species; final number of genomes included in analysis.

4.4.2 GTA gene clusters are located on the leading strand, close to the terminus of replication, and far from repeat regions

As expected, based on previous analyses, presence of complete GTA gene clusters was the most well conserved in genomes of Rhodobacterales members, followed by the Rhizobiales, Sphingomonadales and Caulobacterales (Figure 4.1). The localization was conserved near the chromosomal replication terminus (*ter*) in the Rhodobacterales (Figure 4.1) and to a lesser extent in the Rhizobiales. The clusters were more scattered in the Sphingomonadales and tended to be halfway between *ori* and *ter* in the Caulobacterales (Figure 4.1).

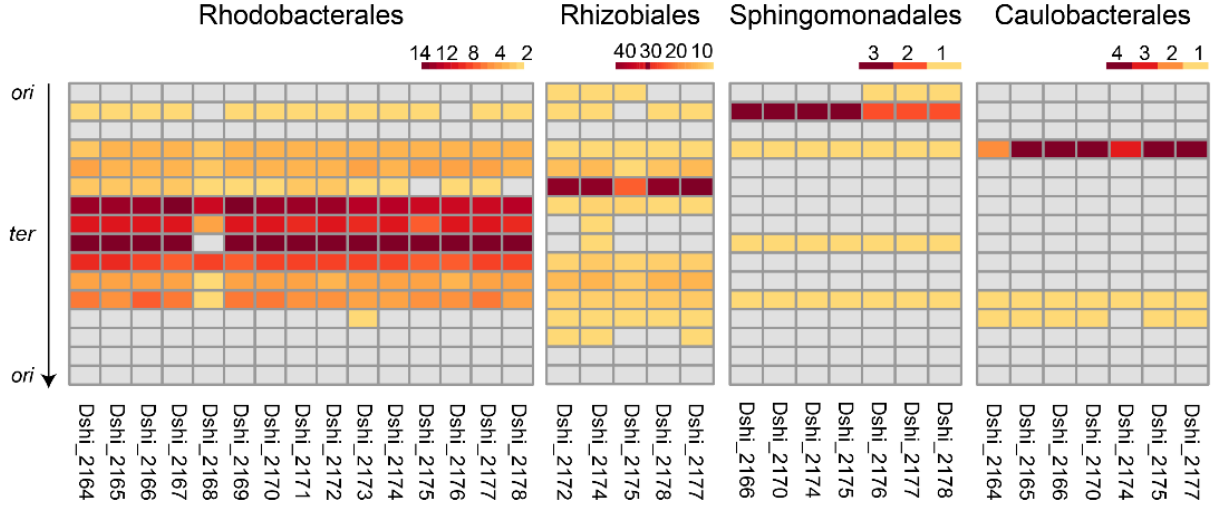

Figure 4.1: Localization of GTA gene cluster genes on alphaproteobacterial chromosomes. The chromosomes were normalized for scale, *ori*-oriented (top and bottom), and divided in 16 parts. Each gene is represented in one column and labeled according to the *D. shibae* locus tags (bottom); for reference, Dshi_2174 encodes the major capsid protein. The heatmap indicates the number of a gene's homologs located in the region of the chromosome, according to the scales above, and grey indicates no homolog was found in a region. The differences in the numbers of individual genes detected in the orders is due to differences in the degrees of conservation of the gene cluster among the orders.

Long repeats (>800 bp) enable homologous recombination [34] and can be responsible for extensive chromosomal rearrangements. Indeed, recombination events between repetitive elements were identified as the likely explanations for the GTA gene cluster's location on different replichores in closely related species in the Rhodobacterales and Rhizobiales. For example, the replichore switch observed between *Phaeobacter inhibens* 2.10 and *P. galleciensis* was associated with regions containing many transposable elements (Figure S2), which are often the cause of homologous recombination [35]. Similarly, the recombination event associated with replichore differences between *Brucella suis* and *B. abortus* was due to a region with paralogous genes (Figure S2).

The MCP gene was found on the leading strand of DNA replication in all but five genomes of the Rhodobacterales and Rhizobiales, irrespective of the replichore the cluster was located on (Figure S3). DNA replication and transcription occur simultaneously in bacteria and head-on collisions between the replisome and RNA polymerase lead to disruptions of both processes that require conflict resolution mechanisms [36]. This has a strong effect on genome organization and evolution because conflicts occur more often on the lagging strand (head-on), and genes oriented this way have higher mutation rates [37]. These disruptions are less common for genes on the leading strand (co-directional) [38,39,40] and slower evolving core genes tend to have this orientation [39,41]. Therefore, the GTA genes are like core genes with respect to gene orientation on the chromosome and although recombination events switch GTA gene clusters between replichores, the orientation of the clusters on the leading strand of DNA replication is maintained.

4.4.3 Cumulative genomic GC skew reveals a unique pattern associated with Rhodobacterales GTA clusters

A typical GC skew pattern for a circular genome is positive on the right replichore and negative on the left replichore. It has been 81ypothesized that inversions from the co-directional to head-on orientation can be traced by a sign change of the GC skew (e.g. right replichore, leading strand, positive GC skew to right replichore, lagging strand, negative GC skew) (Table 4.3) [39]. We applied this categorization to all genes whereby not following the typical GC skew indicates an inversion from the other strand. By determining the proportion of genes present in the different genome orientations and with different GC skews we found that co-directional localization was predominant in all four orders, with 8-10% more genes located on the leading strand than on the lagging strand (Figure S4A). The overall proportion of genes that follow the typical GC skew was similar in the four orders and ranged from 62.6% to 66.8% (Figure S4B). This trend was most pronounced on the leading strand in the Rhodobacterales (Figure 4.2A) and on the lagging strand in the Rhizobiales and Sphingomonadales, while equal numbers of genes followed the typical GC skew on the leading and lagging strands in the Caulobacterales.

Table 4.3: Definitions and characteristics of terms related to GC skew and inversions.

Replichore	Strand	Orientations of transcription and translation	GC skew	Typical GC skew[a]	Potential inversion[b]
Right	Leading +	Co-directional	Positive	Yes	No
Right	Lagging -	Head-on	Negative	No	Yes
Right	Leading -	Co-directional	Negative	No	Yes
Right	Lagging +	Head-on	Positive	Yes	No

Left	Leading +	Co-directional	Positive	No	Yes
Left	Lagging -	Head-on	Negative	Yes	No
Left	Leading -	Co-directional	Negative	Yes	No
Left	Lagging +	Head-on	Positive	No	Yes

[a] Positive on the right replichore and negative on the left replichore.

[b] Those genes with GC skews that do not follow the typical GC skew pattern.

Approximately equal numbers of genes that follow the typical GC skew or are inverted were found on the lagging strand in the Rhodobacterales (Figure 4.2A). Thus, the Rhodobacterales genomes have the strongest conservation of the typical GC skew pattern on the leading strand but the lowest conservation on the lagging strand. The same distribution pattern, although with a slightly stronger preference for the leading strand, was found for core genes (Figure S4C). The differences between the gene proportions (Figure S4C) was significant in all orders (Kruskal-Wallis rank sum test p-values: Rhizobiales, <2.2x10-16; Rhodobacterales, <2.2x10-16; Caulobacterales, 3.8x10-8; Sphingomondales, 1.3x10-11) and also when only core genes were considered (Kruskal-Wallis rank sum test p-values: Rhizobiales, <2.2x10-16; Rhodobacterales, <2.2x10-16; Caulobacterales, 2.2x10-6; Sphingomondales, 1.8x10-8). Thus, overall, the majority and similar proportions of genes follow the typical GC skew in all four orders. However, these are mainly located on the leading strand in the Rhodobacterales and on the lagging strand or equally distributed in the other three orders, indicating a distinct gene orientation trend among the Rhodobacterales.

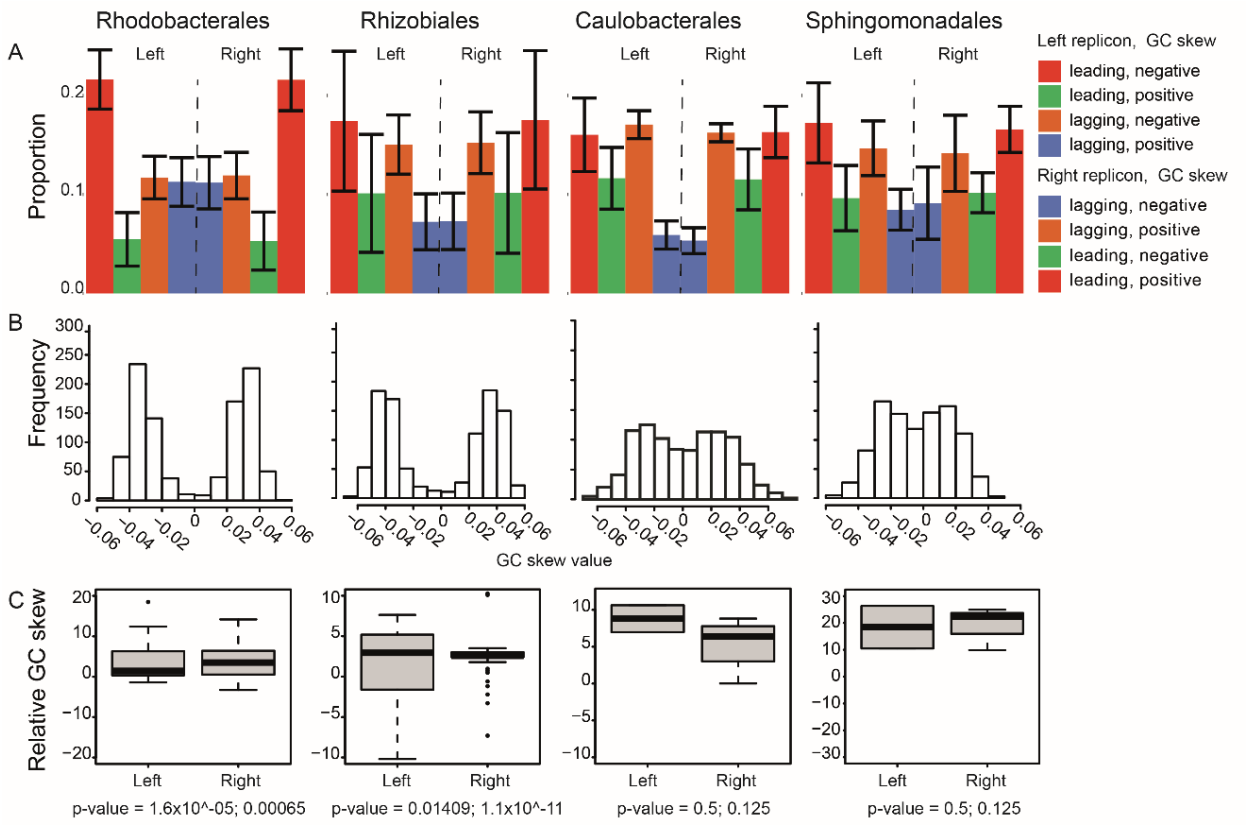

Figure 4.2: GC skew analyses. A. Genes located on the leading or lagging strand following the typical GC skew (positive or negative on the right or left replichore, respectively). The data are plotted as the percentage with the standard deviation. B. Frequency of GC skew mean at all 1,000 locations of the cumulative genome. C. GC skew of the GTA gene clusters relative to their respective "host" genome ((GC skew GTA-GC skew host)/GC skew host). The genome was divided into a left and a right replichore, as the values of the GTAs on the right or left replichore would otherwise equalize. The significance of the differences in distributions between GTA and "host" values was tested using the pairwise Wilcoxon test for the left and right replichore separately.

Most organisms' genomes follow the typical GC skew pattern, and it is considered an archetypal genomic property. A high GC skew might reflect the original ancestral genome of the bacteria considered here. Indeed, it was recently proposed that a deviation from this pattern could be used to detect misassembled genomic sequences [40]. In the Rhodobacterales the expected skew is reduced on the lagging strand due to increased inversions, with genes changed from co-directional to head-on orientation. It is unclear why this pattern is found for this group, but it has been suggested that switching genes from co-directional to the head-on orientation might have benefits by increasing evolvability [39] although this

is still under debate [42,43]. There are three hypotheses on the evolutionary history of genome architecture [39] whereby there is either a reduction of head-on genes, a reduction of co-directional genes, or the original ratio is retained. While all four orders studied here have similar proportions of head-on and co-directional genes (Figure 4.2A), the inversions on the leading and lagging strands change in different directions in the Rhodobacterales compared to the other three orders, making it difficult to draw a general conclusion from our analysis.

Next, we used cumulative representations of the GC skew (determined over a sliding window) to evaluate the magnitudes of and patterns in the GC skews. The maximal deviations from zero were similar in all four orders, but the patterns varied among them (Figure 4.2B). The GC skew values deviated $>\pm 0.03$ in the Rhodobacterales and Rhizobiales, resulting in clear bimodal distributions (Figure 4.2B). In contrast, the distributions were closer to normal in the Sphingomonadales and Caulobacterales although some bimodality could still be seen. The GTA gene clusters were predominantly located within GC skew peaks in all four orders (Figure S5). Indeed, comparing the absolute GC skews of the regions where GTA gene clusters are located to the remainders of the genomes showed that they have greater GC skews than average (Figure 4.2C). Interestingly, the strongest deviations could be observed in the Sphingomonadales and Caulobacterales, which have more genes with GC skews around zero (Figure 4.2B). This indicates that GTA clusters tend to have high GC skews irrespective of the rest of the genome's overall properties.

4.4.4 Correlation between GC skew and codon usage in the Rhodobacterales

A difference in GC skew value could affect the codon usage, so we examined which codons are enriched in genes with a high GC skew. Therefore, we checked for potential differences in codon usage between genes in GC skew peak and non-peak regions. In the Rhodobacterales we found that codons that were overrepresented in GC skew peaks also had a significantly higher GC content (Figure 4.3). This contradicts findings of a negative correlation between GC content and composition bias [23]. Although the Rhizobiales and Rhodobacterales displayed very similar bimodal GC skew distributions, no differences between the codon usage in peak and non-peak regions were observed in the Rhizobiales. Similarly, no significant correlations were found for the other two orders. Next, we compared codon

usage to codon GC skew. Significant positive correlations were found for the Rhodobacterales and Caulobacterales (Spearman's correlation coefficient p-values = 0.0002 and 0.012, respectively) (Figure 4.3). Hence, in the Caulobacterales GC content and in the Rhodobacterales the GC content and GC skew positively correlate with codon usage, which suggests special codons are used in GC skew peaks and that there is a distinct evolutionary process occurring in the Rhodobacterales.

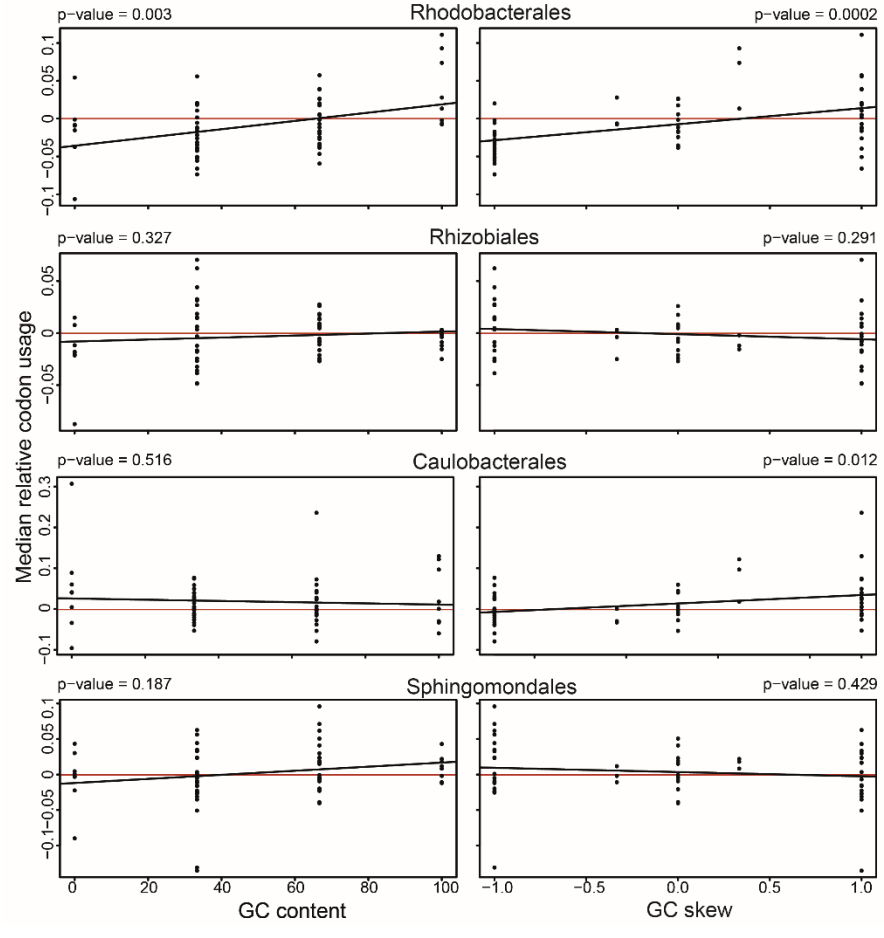

Figure 4.3: Relationships between codon usage and GC content or GC skew. The median relative codon usages for each order were calculated from the mean codon usage values per genome (peak–non-peak/non-peak) and correlated with the GC content (left) or GC skew (right). The Spearman correlation was used to test significance and p-values are given above the plots.

To determine what influence GC content and GC skew might have on codon usage, we selected codons with identical GC content that encode the same amino acid and compared them based on their GC skew and usage (Figure 4.4). We found that codons with higher GC skew were more predominant in GC

skew peaks for the Rhodobacterales and Caulobacterales. Thus, in GC skew peaks of those orders, codons with a higher proportion of guanine are preferred instead of cytosine. This was also true for GTA genes compared to non-peak genes in all four orders (Figure 4.4). It was previously shown that GTA genes preferentially encode amino acids with a lower carbon content, which results in increased GC content [15,3]. However, our results show that these genes also preferentially use codons with increased GC skew, which is responsible for the location of GTA gene clusters in GC skew peak regions.

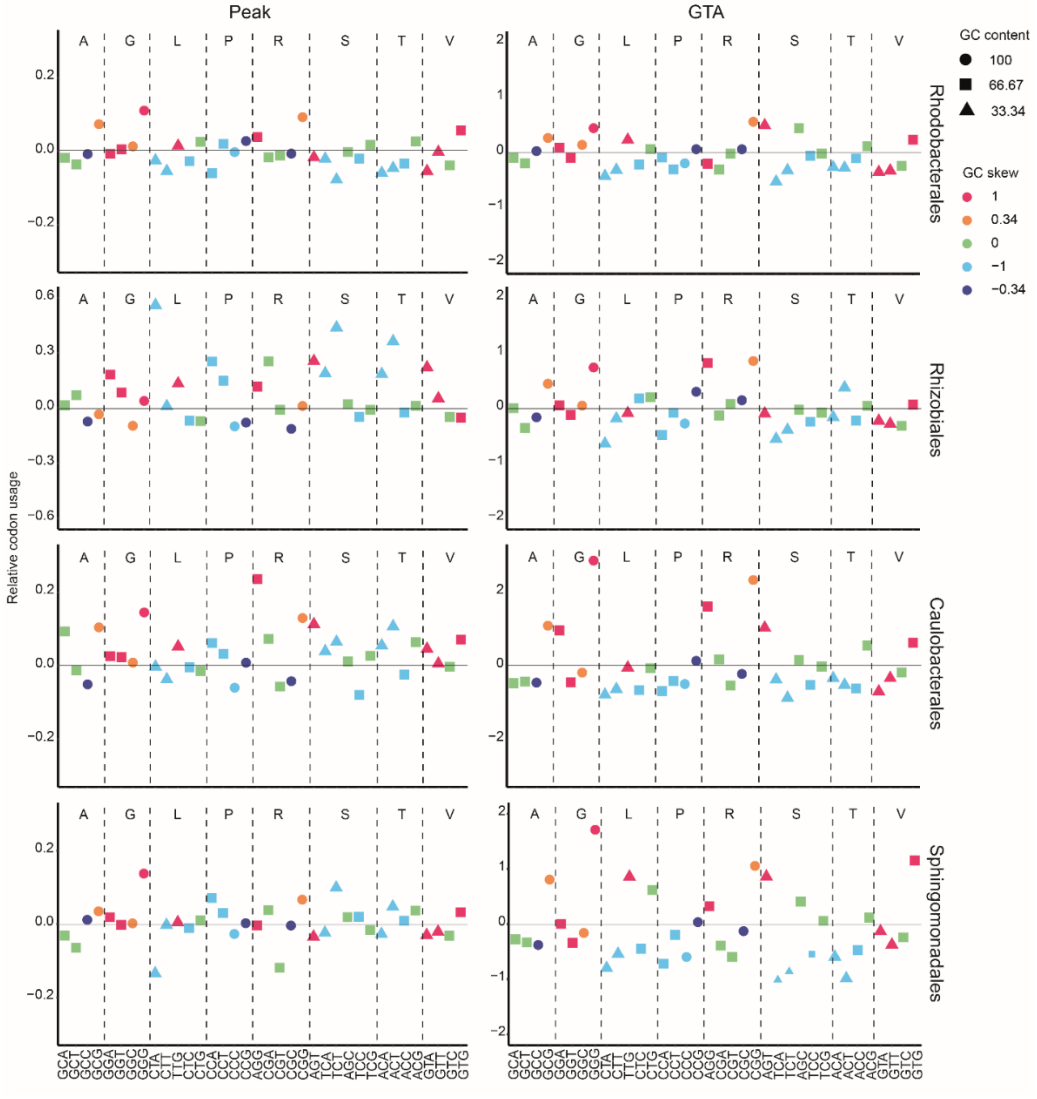

Figure 4.4: Relative codon usage in GC skew peak genes and GTA genes. Only codons that have a GC content greater than zero and codons specifying a particular amino acid that have the same GC content were considered. The relative codon usage values for peak (or GTA) versus non-peak genes were calculated as (peak (or GTA))/non-peak. The GC content and skew for each codon in each order are indicated based on their shape and color, respectively, according to the legend at the top right.

The codons are indicated at the bottom of the plots with dashed vertical lines separating the different amino acids, which are indicated above the plots using their one-letter codes.

Nucleotide strand bias is generally attributed to cytosine deamination events that produce thymine [21,44]. Therefore, cytosine levels should decrease as they are converted into thymine if deamination is responsible for the GC skew. To see if there is a lower occurrence of cytosine in the GC skew peaks, we examined codons for their G and C content and their relative usage. We found an increased codon usage correlated with an increased G content in the Rhodobacterales and Caulobacterales, while the C content tended to decrease (Figure S6). In the GTA genes, both the use of guanine and cytosine correlated positively with the increase in codon usage (Figure S7). There was no significant change in guanine or cytosine levels in the peak versus non-peak regions of the Sphingomonadales or Rhizobiales. This could mean that deamination processes are not responsible for the GC skew or at least that they play a smaller role than the preferred selection for guanine-containing codons in the Rhodobacterales and Caulobacterales.

4.4.5 Core and GTA genes are commonly found in GC skew peaks

We found that core genes, defined as those present in ≥90% of all genomes of an order, were accumulated in regions with GC skew peaks in all four orders (Figure 4.5A). The ratios of core genes in peak to non-peak regions (calculated as percentage peak/percentage non-peak) were the most extreme in the Rhodobacterales (32%) and Rhizobiales (14%) but were also >1% in the Sphingomonadales (2.4%) and Caulobacterales (2.4%).

To investigate which genes are in GC skew peaks and how conserved their presence in peaks is, we compared the numbers of members of gene families found inside and outside of peaks (without differentiation according to the direction of the peaks) (Figure 4.5B). Besides the GTA gene cluster genes that are strongly enriched in GC skew peaks in all four orders, there were multiple examples of genes associated with central physiological processes, such as protein processing (i.e., chaperones; *dnaK* and *clpB*), translation (ribosomes), cell division, nicotinamide adenine dinucleotide metabolism, flagellar

motility, and recombination, that were frequently found in GC skew peaks (Table S2). Overall, although core genes do not necessarily have to be in GC skew peaks, they are overrepresented in these peaks.

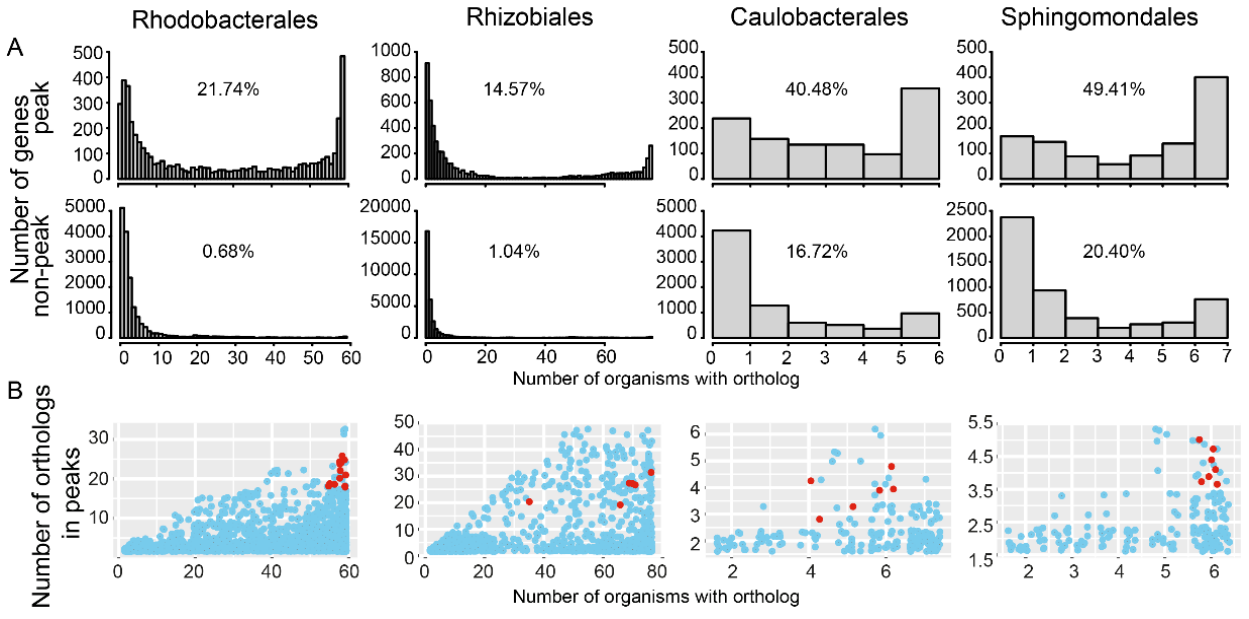

Figure 4.5: Gene conservation in GC skew peak and non-peak locations. A. Number of orthologs in and outside of GC skew peaks. The percentage of genes found in ≥90% of genomes in the orders Rhodobacterales and Rhizobiales or six and five genomes of the Caulobacterales and Sphingomonadales, respectively (representing the numbers of genomes closest to 90%), is shown inside the plot. B. Number of genomes in which an ortholog was found compared to how often the genes were in GC skew peaks. GTA gene cluster genes are indicated in red.

The localization of GTA gene clusters as well as many core genes in genomic areas with GC skew peaks was especially evident for the Rhodobacterales and Rhizobiales, which also have the most pronounced GC skews of the considered orders. This is particularly interesting for two reasons. First, these core genes have been conserved in a wide range of genomes over a long period of time and thus originate from a common ancestor [45,42]. Second, the GTA gene cluster is believed to have evolved from a prophage that integrated into the genome of a shared ancestor of multiple alphaproteobacterial orders [2], which means that it has been present in these genomes as long as many of the core genes [13] and also shares the property of being preferentially located in GC skew peaks. Thus, it is possible that the original ancestor of these bacteria had a more extreme GC skew, which is supported by the following

points. GTAs are evolutionarily related to phages, which have been shown to integrate into the host genome in ways that maximize their success of replication, such as preferential integration on the leading strand and closer to *ter*, and with counter-selection for motifs that could result in disruption of macrodomain structures of the host genome (e.g., motifs associated with the *ter* macrodomain are not found in prophages) [43]. The GTA gene clusters showed similar trends (co-directional orientation and localization closer to *ter*), and possibly these properties have been maintained since evolving from the original prophage. Regarding GC skew, however, prophages and the GTA gene cluster differ. We found no preferential localization of phages in GC skew peaks (Figure S8), which also fits with their place as accessory mobile genetic elements as opposed to core genes. Moreover, our results have some commonality with findings for eukaryotes, in which highly conserved genes are also contained in genomic regions with strong GC skew [46,47,48], and it was also shown that those highly conserved genes encode proteins with longer half-lives [48].

4.4.6 Core and GTA genes are located far from repetitive elements

Although GTA gene clusters are occasionally translocated between replichores (Figures S2 and S3), the cluster itself and its orientation relative to the direction of DNA replication are well preserved. Therefore, we investigated the stability of these localizations. Regions with repeats, especially long repeats (>800 bp), represent hotspots for chromosome rearrangement and thus increase the local genome plasticity. We found that the distance of GTA gene clusters to long repeats is higher than for most other genes (Table S3). To investigate whether this characteristic is also shared by core genes we sorted all genes into two groups according to their distances from the nearest repeat, with “far” being considered as >1/1000 of the genome size (approximately >4 kb for most organisms) and “near” being considered as <1/1000 of the genome size (approximately <4 kb for most organisms). Core and GTA genes were predominantly found further away from repeats in all four orders (Figure 4.6) and therefore are localized in regions with lower plasticity. Interestingly, a previous analysis of genomes from different phyla showed that genomes with a stronger overall GC skew tend to have fewer repeats and it was concluded that stable chromosomes have higher GC skews and less repeats [27]. As discussed above, it was

previously documented that GTA genes have a higher GC content [3], and high GC content regions have also been shown to have lower rearrangement frequencies [15]. Studies on the gammaproteobacterium *Vibrio parahaemolyticus* and the epsilonproteobacterium *Helicobacter pylori* also showed that GC content and plasticity regions are negatively correlated [49,50]. This could also contribute to the conserved GTA gene cluster localization.

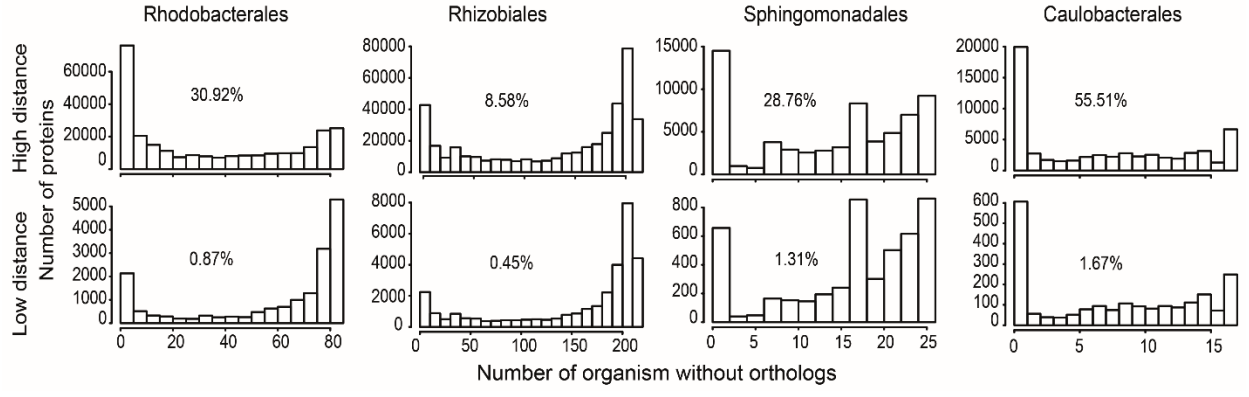

Figure 4.6: Relationship of gene conservation to distance from repeats. Genes were classified as far from, >1/1000 of the genome size (approximately >4 kb for most organisms), or near to, <1/1000 of the genome size (approximately <4 kb for most organisms), repeats. The percentages of the number of genes found in ≥90% of the genomes are shown for the Rhodobacterales and Rhizobiales. The percentages of genes found in more than six or eight genomes (representing the numbers of genomes closest to 90%) are shown for the Caulobacterales and Sphingomonadales, respectively.

4.4.7 Relationship between DNA methylation and GTA gene cluster localization

One of the key RcGTA regulators is the response regulator protein CtrA. This regulator is almost universally conserved throughout the Alphaproteobacteria [7,51], with shared and unique roles in different lineages and acting as a key cell cycle regulator in some. In Caulobacter crescentus, where CtrA has been best characterized, the *ctrA* promoter and multiple promoters of the CtrA regulon are targeted by the methyltransferase CcrM and the transcriptional regulator GcrA. CcrM methylates the adenine residue of the motif 5′-GANTC-3′, which the protein GcrA then binds to and recruits the RNA polymerase to initiate transcription from the associated promoter [52,53]. While it is well documented that methylation has a strong influence on the CtrA phosphorelay and its regulon, and that CtrA controls the GTA gene

cluster in Rhodobacterales members, a connection between all three components (CtrA, methylation by CcrM and the GTA gene cluster) has not yet been investigated to our knowledge. Therefore, we analyzed potential CcrM methylation patterns by determining the occurrence of the GANTC recognition motif over the length of the chromosomes (normalized to 1000 bp). A strong increase in methylation motifs in the region around *ori* was observed in all four orders (Figure 4.7). Strong and slight additional increases in GANTC motif numbers towards *ter* also exist in the Caulobacterales and Rhodobacterales, respectively.

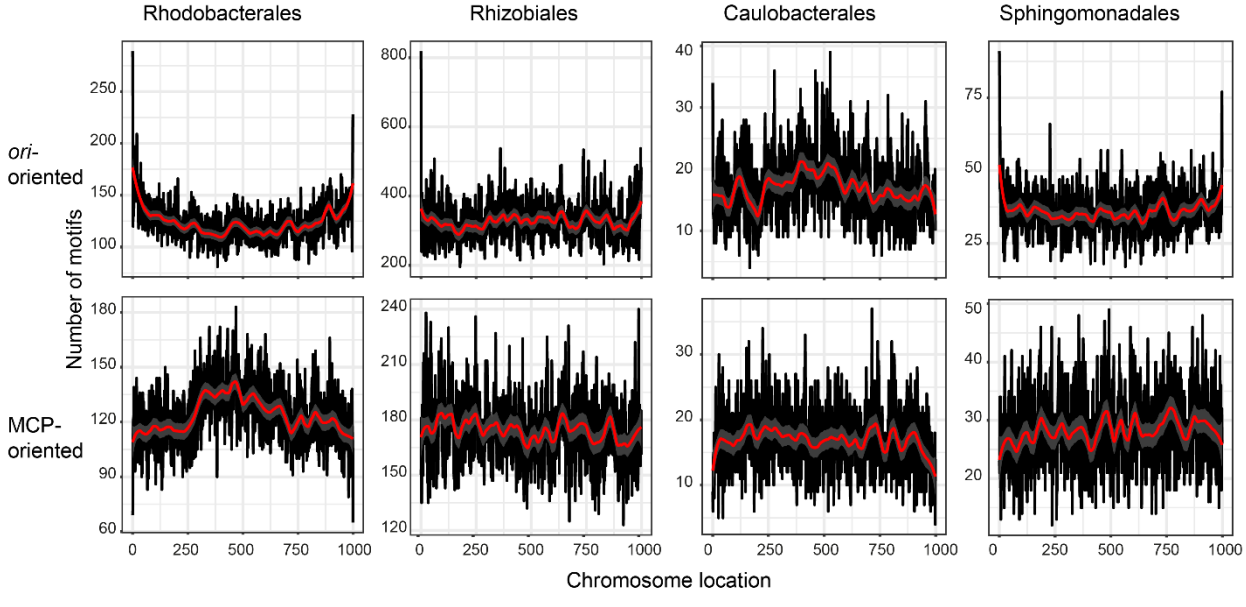

Figure 4.7: CcrM GANTC methylation motif occurrence across chromosomes. The cumulative occurrence of the methylation motif on chromosomes is plotted with chromosomes normalized to a length of 1000. The genomes were either oriented with position 0 and 1000 representing the origin of replication (*ori*, top) or the GTA major capsid protein gene (MCP, bottom). A LOESS curve was fitted to visualize local trends (red). The normalized genomes were divided into 100 parts in which the occurrences of the GANTC motif were quantified, with the cumulative number of motifs plotted.

The reorientation of each chromosome such that the position of the MCP was the first gene showed greater overall fluctuations in the GANTC motif pattern occurrences (Figure 4.7), probably due to the greater variability in the localization of the GTA genes compared to *ori* (Figure 4.1). A slight drop in GANTC numbers at the MCP gene can also be seen in the Caulobacterales and Sphingomonadales genomes. In the Rhodobacterales, reorientation based on the MCP gene showed that this region had the

lowest methylation potential across genomes, even though there is a slight increase in motifs near *ter* in this group. The strongest conservation of GTA gene cluster localization is also found in this order, where it is biased towards *ter*. This is in accordance with our previous study where the Rhodobacterales *ctrA* gene was also found conserved near to *ter* and showed the lowest number of GANTC motifs in *ctrA* promoter regions [54]. Hemi-methylation of this motif during replication has been found to be a signal for transcriptional activation [55]. The presence of less GANTC sequences and its location near *ter* might indicate that *ctrA* transcriptional control is uncoupled from replication in this group. However, future studies are needed to show the potential significance of the low number of GANTC motifs in the GTA gene cluster region.

4.5 Conclusion

In this study, we performed a comprehensive genome structure analysis for four orders of the class Alphaproteobacteria to examine patterns of the RcGTA-type gene cluster localization, genomic GC skew, and DNA methylation. We found that the GTA gene cluster shares properties with core genes, such as localization in low plasticity regions, gene orientation on the leading strand of DNA replication, and localization in regions of especially strong GC skew. These high GC skew regions at least partly arise due to a selection for codons with higher GC skew and are not necessarily associated with codon GC content enrichment. The GTAs studied here are proposed to have evolved from a phage that integrated into the genome of an ancestral alphaproteobacterial host. Generally, phages try to mimic their host's genome structure [56], but we did not find any notable elevation of GC skew among phages and prophages. Therefore, it seems that part of the evolutionary process of becoming a GTA included gaining this elevated GC skew, and this might be connected to other properties these genes have in common with core genes, such as their location in regions distant from repetitive elements.

4.6 Supplementary materials

This chapter includes two supplementary files that are available in digital format using this link:

The Word file contains the supplementary figures S1-S6. The Excel file S4.1 states the presence of Ori and MCP in the analyzed genomes. Excel file S4.2 shows the number of homologues genes and

presence in GC skew peaks per gene. This information is presented in four tabs, sorted by order. Excel file S4.3 shows the distance of the major capsid protein to the closest repeat in four tabs, sorted by order.

4.7 References

1. Lang, A.S.; Westbye, A.B.; Beatty, J.T. The Distribution, Evolution, and Roles of Gene Transfer Agents in Prokaryotic Genetic Exchange. *Annu. Rev. Virol.* **2017**, *4*, 87–104.
2. Lang, A.S.; Beatty, J.T. Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* **2007**, *15*, 54–62.
3. Shakya, M.; Soucy, S.M.; Zhaxybayeva, O. Insights into Origin and Evolution of α-proteobacterial Gene Transfer Agents. *Virus Evol.* **2017**, *3*, 1–13.
4. Tomasch, J.; Wang, H.; Hall, A.T.K.; Patzelt, D.; Preuße, M.; Brinkmann, H.; Bhuju, S.; Jarek, M.; Geffers, R.; Lang, A.S. Packaging of *Dinoroseobacter shibae* DNA into Gene Transfer Agent particles is not random. *Genome Biol. Evol.* **2018**, *10*, 359–369.
5. Biers, E.J.; Wang, K.; Pennington, C.; Belas, R.; Chen, F.; Moran, M.A. Occurrence and expression of gene transfer agent genes in marine bacterioplankton. *Appl. Environ. Microbiol.* **2008**, *74*, 2933–2939.
6. Nagao, N.; Yamamoto, J.; Komatsu, H.; Suzuki, H.; Hirose, Y.; Umekage, S.; Ohyama, T.; Kikuchi, Y. The gene transfer agent-like particle of the marine phototrophic bacterium *Rhodovulum sulfidophilum*. *Biochem. Biophys. Reports* **2015**, *4*, 369–374.
7. Lang, A.S.; Beatty, J.T. Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. *Proc Natl Acad Sci U S A* **2000**, *97*, 859–864.
8. Redfield, R.J. Do Bacteria Have Sex? *Microbes Evol.* **2014**, *2*, 139–144.
9. Hynes, A.P.; Shakya, M.; Mercer, R.G.; Grüll, M.P.; Bown, L.; Davidson, F.; Steffen, E.; Matchem, H.; Peach, M.E.; Berger, T.; et al. Functional and Evolutionary Characterization of a Gene Transfer Agent's Multilocus "Genome." *Mol. Biol. Evol.* **2016**, *33*, 2530–2543.
10. Québatte, M.; Christen, M.; Harms, A.; Körner, J.; Christen, B.; Dehio, C. Gene Transfer Agent Promotes Evolvability within the Fittest Subpopulation of a Bacterial Pathogen. *Cell Syst.* **2017**, *4*, 611-621.e6.
11. Redfield, R.J.; Soucy, S.M. Evolution of Bacterial gene transfer agents. *Front. Microbiol.* **2018**, *9*, 1–41.

12. Westbye, A.B.; Beatty, J.T.; Lang, A.S.; Rice, P. Guaranteeing a captive audience: coordinated regulation of gene transfer agent (GTA) production and recipient capability by cellular regulators. *Curr. Opin. Microbiol.* **2017**, *38*, 122–129.

13. Zhao, Y.; Wang, K.; Ackermann, H.-W.; Halden, R.U.; Jiao, N.; Chen, F. Searching for a "hidden" prophage in a marine bacterium. *Appl. Environ. Microbiol.* **2010**, *76*, 589–595.

14. Lang, A.S.; Zhaxybayeva, O.; Beatty, J.T. Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* **2012**, *10*, 472–82.

15. Kogay, R.; Wolf, Y.I.; Koonin, E. V.; Zhaxybayeva, O. Erratum: Selection for reducing energy cost of protein production drives the gc content and amino acid composition bias in gene transfer agents. *Mbio.* **2020,** *11*, 4, e01206-20.

16. Koppenhöfer, S.; Wang, H.; Scharfe, M.; Kaever, V.; Wagner-Döbler, I.; Tomasch, J. Integrated transcriptional regulatory network of quorum sensing, replication control, and SOS response in *Dinoroseobacter shibae*. *Front. Microbiol.* **2019**, *10*.

17. Lobry, J.R. Asymmetric Substitution Patterns in the Two DNA Strands of Bacteria. *Mol. Biol. Evol.* **1996**, *13*, 660–665.

18. Freeman, J.M.; Plasterer, T.N.; Smith, T.F.; Mohr, S.C. Patterns of genome organization in Bacteria. *Science,* **1998**, *279*, 5358, 1827-1827.

19. Rocha, E.P.C. Order and disorder in bacterial genomes. *Curr. Opin. Microbiol.* **2004**, *7*, 519–527.

20. Rocha, E.P.C. The replication-related organization of bacterial genomes. *Microbiology* **2004**, *150*, 1609–1627.

21. Bhagwat, A.S.; Hao, W.; Townes, J.P.; Lee, H.; Tang, H.; Foster, P.L. Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in *Escherichia coli. Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 2176–2181.

22. Kono, N.; Tomita, M.; Arakawa, K. Accelerated laboratory evolution reveals the influence of replication on the GC skew in *Escherichia coli. Genome Biol. Evol.* **2018**, *10*, 3110–3117.

23. Zhao, H.L.; Xia, Z.K.; Zhang, F.Z.; Ye, Y.N.; Guo, F.B. Multiple factors drive replicating strand composition bias in bacterial genomes. *Int. J. Mol. Sci.* **2015**, *16*, 23111–23126.

24. Achaz, G.; Coissac, E.; Netter, P.; Rocha, E.P.C. Associations between inverted repeats and the structural evolution of bacterial genomes; *Nucleic acids research*, **2002,** *30*, 9, 2031-42.

25. Vandecraen, J.; Chandler, M.; Aertsen, A.; Van Houdt, R. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit. Rev. Microbiol.* **2017**, *43*, 709–730.

26. Sela, I.; Wolf, Y.I.; Koonin, E. V. Selection and genome plasticity as the key factors in the evolution of Bacteria. *Phys. Rev. X* **2019**, *9*, 31018.

27. Rocha, E.P.C.; Blanchard, A. Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution; **2002**, *30*, 9, 2031-42.

28. Gao, F.; Zhang, C.-T. Ori-Finder: A web-based system for finding *oriCs* in unannotated bacterial genomes. **2008**.

29. Lechner, M.; Findeiß, S.; Steiner, L.; Marz, M.; Stadler, P.F.; Prohaska, S.J. Proteinortho: Detection of (Co) orthologs in large-scale analysis. *BMC Bioinformatics* **2011**, *12*.

30. Arndt, D.; Grant, J.R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D.S. PHASTER: A better, faster version of the PHAST phage search tool. **2016**, *44*, 16–21.

31. Achaz, G.; Dé Ric Boyer, F.; Rocha, E.P.C.; Viari, A.; Coissac, E. Repseek, a tool to retrieve approximate repeats from large DNA sequences. **2007**, *23*, 119–121.

32. Kung, S.H.; Retchless, A.C.; Kwan, J.Y.; Almeida, R.P.P. Effects of DNA size on transformation and recombination efficiencies in *Xylella fastidiosa*. *Appl. Environ. Microbiol.* **2013**, *79*, 1712–1717.

33. Darling, A.C.E.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. **2004**, 1394–1403.

34. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549.

35. Hawkey, J.; Hamidian, M.; Wick, R.R.; Edwards, D.J.; Billman-Jacobe, H.; Hall, R.M.; Holt, K.E. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. **2011**.

36. Mangiameli, S.M.; Merrikh, C.N.; Wiggins, P.A.; Merrikh, H. Transcription leads to pervasive replisome instability in bacteria. *Elife* **2017**, *6*, 1–27.

37. Lin, Y.-L.; Pasero, P. Interference Between DNA Replication and Transcription as a Cause of Genomic Instability. *Curr. Genomics* **2012**, *13*, 65–73.

38. Rocha, E.P.C.; Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **2003**, *31*, 6570–6577.

39. Merrikh, C.N.; Merrikh, H. Gene inversion potentiates bacterial evolvability and virulence. *Nat. Commun.* **2018**, *9*.

40. Lu, J.; Salzberg, S.L. SkewIT: The Skew Index Test for large-scale GC Skew analysis of bacterial genomes. *PLoS Comput. Biol.* **2020**, *16*, 1–16.

41. Liu, H.; Zhang, J. January 11, 2020 No support for the adaptive hypothesis of lagging-strand encoding in bacterial genomes from Merrikh & Merrikh. **2020**, 1–8.

42. Dewey, C.N.; Pachter, L. Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Hum. Mol. Genet.* **2006**, *15 Spec No*, 51–56.

43. Bobay, L.M.; Rocha, E.P.C.; Touchon, M. The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.* **2013**, *30*, 737–751.

44. Duncan, B.K.; Miller, J.H. Mutagenic deamination of cytosine residues in DNA. *Nature* **1980**, *287*, 560–561.

45. Daubin, V.; Gouy, M.; Perrière, G. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res.* **2002**, *12*, 1080–1090.

46. Hartono, S.R.; Korf, I.F.; Chédin, F. GC skew is a conserved property of unmethylated CpG island promoters across vertebrates. *Nucleic Acids Res.* **2015**, *43*, 9729–9741.

47. Ginno, P.A.; Lim, Y.W.; Lott, P.L.; Korf, I.; Chédin, F. GC skew at the 59 and 39 ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.* **2013**, *23*, 1590–1600.

48. Dai, Y.; Holland, P.W.H. The Interaction of natural selection and GC skew my drive the fast evolution of a sand rat homeobox gene. *Mol. Biol. Evol* **2019**, *36*, 1473–1480.

49. Fischer, W.; Windhager, L.; Rohrer, S.; Zeiller, M.; Karnholz, A.; Hoffmann, R.; Zimmer, R.; Haas, R. Strain-specific genes of *Helicobacter pylori:* Genome evolution driven by a novel type IV secretion system and genomic island transfer. *Nucleic Acids Res.* **2010**, *38*, 6089–6101.

50. Han, H.; Wong, H.C.; Kan, B.; Guo, Z.; Zeng, X.; Yin, S.; Liu, X.; Yang, R.; Zhou, D. Genome plasticity of *Vibrio parahaemolyticus*: Microevolution of the "pandemic group." *BMC Genomics* **2008**, *9*, 1–12.

51. Brimacombe, C.A.; Ding, H.; Johnson, J.A.; Thomas Beatty, J. 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.* **2015**, *197*, 2653–2663.

52. Fioravanti, A.; Fumeaux, C.; Mohapatra, S.S.; Bompard, C.; Brilli, M.; Frandi, A.; Castric, V.; Villeret, V.; Viollier, P.H.; Biondi, E.G. DNA Binding of the Cell Cycle Transcriptional Regulator GcrA Depends on N6-Adenosine Methylation in *Caulobacter crescentus* and Other Alphaproteobacteria. **2013**, *9*.

53. Haakonsen, D.L.; Yuan, A.H.; Laub, M.T. The bacterial cell cycle regulator *gcrA* is a σ70 cofactor that drives gene expression from a subset of methylated promoters. *Genes Dev.* **2015**, *29*, 2272–2286.

54. Tomasch, J.; Koppenhöfer, S.; Lang, A.S. Connection between chromosomal location and function of CtrA phosphorelay genes in Alphaproteobacteria. *Front. Microbiol.* **2021**, *12*, 1–8.

55. Mohapatra, S.S.; Fioravanti, A.; Vandame, P.; Spriet, C.; Pini, F.; Bompard, C.; Blossey, R.; Valette, O.; Biondi, E.G. Methylation-dependent transcriptional regulation of crescentin gene (*creS*) by GcrA in *Caulobacter crescentus*. *Mol. Microbiol.* **2020**, *114*, 127–139.

56. Samson, J.E.; Magadán, A.H.; Sabri, M.; Moineau, S. Revenge of the phages: Defeating bacterial defences. *Nat. Rev. Microbiol.* **2013**, *11*, 675–687.

57. Yin, T.; Cook, D.; Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **2012**, *13*, R77.

58. Toedling, J.; Sklyar, O.; Huber, W. Ringo - An R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* 2007, 8, 1–4.

59. Elek, A.; Kuzman, M.; Vlahovicek, K. coRdon: Codon Usage Analysis and Prediction of Gene Expressivity. R package version 1.4.0. Available online: https://github.com/BioinfoHR/coRdon.

5 CHAPTER 5: Patterns of abundance, chromosomal localization, and domain organization among c-di-GMP-metabolizing genes revealed by comparative genomics of five alphaproteobacterial orders

5.1 Abstract

Background

Bis-(3′-5′)-cyclic dimeric guanosine monophosphate (c-di-GMP) is a bacterial second messenger that affects diverse processes in different bacteria, including the cell cycle, motility, and biofilm formation. Its cellular levels are controlled by the opposing activities of two types of enzymes, with synthesis by diguanylate cyclases containing a GGDEF domain and degradation by phosphodiesterases containing either an HD-GYP or an EAL domain. These enzymes are ubiquitous in bacteria with up to 50 encoded in some genomes, the specific functions of which are mostly unknown.

Results

We used comparative analyses to identify genomic patterns among genes encoding proteins with GGDEF, EAL, and HD-GYP domains in five orders of the class Alphaproteobacteria. GGDEF-containing sequences and GGDEF-EAL hybrids were the most abundant and had the highest diversity of co-occurring auxiliary domains while EAL and HD-GYP containing sequences were less abundant and less diverse with respect to auxiliary domains. There were striking patterns in the chromosomal localizations of the genes found in two of the orders. The Rhodobacterales' EAL-encoding genes and Rhizobiales' GGDEF-EAL-encoding genes showed opposing patterns of distribution compared to the GGDEF-encoding genes. In the Rhodobacterales, the GGDEF-encoding genes showed a tri-modal distribution with peaks mid-way between the origin (*ori*) and terminus (*ter*) of replication and at *ter* while the EAL-encoding genes peaked near *ori*. The patterns were more complex in the Rhizobiales, but the GGDEF-encoding genes were biased for localization near ter.

Conclusions

The observed patterns in the chromosomal localizations of these genes suggest a coupling of synthesis and hydrolysis of c-di-GMP with the cell cycle. Moreover, the higher proportions and

diversities of auxiliary domains associated with GGDEF domains and GGDEF-EAL hybrids compared to EAL or HD-GYP domains could indicate that more stimuli affect synthesis compared to hydrolysis of c-di-GMP.

5.2 Introduction

Bis-(3′-5′)-cyclic dimeric guanosine monophosphate (c-di-GMP) is a near-ubiquitous second Bis-(3′-5′)-cyclic dimeric guanosine monophosphate (c-di-GMP) is a second messenger that was first described for its role in regulating cellulose biosynthesis in *Gluconacetobacter xylinus* [1, 2], but which is now recognized as near-ubiquitous and affecting a large variety of processes in bacteria [3, 4]. Cellular concentrations of c-di-GMP are regulated in response to internal and external stimuli, and the resulting changes can be part of bacterial adaptation to changes in their environment [5]. The cellular levels of c-di-GMP are controlled by two groups of enzymes with opposing activities, where it is synthesized by diguanylate cyclases (DGCs) and degraded by c-di-GMP-specific phosphodiesterases (PDEs). DGCs have conserved GGDEF domains and synthesize c-di-GMP from two molecules of guanosine triphosphate (GTP) [6]. There are two distinct types of PDEs, with either EAL or HD-GYP domains, that degrade c-di-GMP. Both types are able to break c-di-GMP into the linear 5′-phosphoguanylyl (3′-5′) guanosine (pGpG) dinucleotide [7, 8] which is then further broken down to two molecules of guanosine monophosphate (GMP) by an oligo-ribonuclease [9, 10]. PDEs of the HD-GYP type can also break c-di-GMP into two GMPs in one step [11, 12]. In addition, there are hybrid proteins that have both GGDEF and EAL domains and thus represent a “biochemical conundrum”. It has been suggested that these proteins can switch between synthesis and hydrolysis of c-di-GMP [13], with a protein’s activity controlled by, for example, phosphorylation status [14] or dimerization [15]. However, it is also possible that one of the domains is not functional. Based on the proteins characterized in detail, the most common scenarios are that only the EAL domain or both domains are functional [16].

C-di-GMP levels can be controlled via transcriptional and translational regulation of gene expression, or through post-translational modification of the synthesis and degradation enzymes as a quicker response. Auxiliary domains can be present on the enzymes and include sensory, signaling and

protein binding domains, and these can allow for rapid adaptation [17]. Cellular changes in c-di-GMP concentration can result from a variety of input and output signals that are detected by the enzymes or their regulators and that affect the production or degradation of c-di-GMP [18]. An analysis of genomic sequences from different bacterial phyla found that members of the phylum Proteobacteria encode the highest numbers of c-di-GMP-modulating enzymes [18].

In the Alphaproteobacteria, c-di-GMP has been examined for its role in many different processes, such as the symbiosis of Sinorhizobium meliloti with plant roots [19] and related to its effects on the regulatory network associated with the transcriptional regulator CtrA [20], which is highly conserved in this class [21]. The CtrA phosphorelay consists of the histidine kinase CckA, the phosphotransferase ChpT and the transcriptional regulator CtrA [22]. It has been suggested that its ancestral role in alphaproteobacteria was related to the control of motility and recombination [23, 24], but there has also been work establishing a link between this phosphorelay and c-di-GMP with respect to regulation of the cell cycle and cell differentiation in Caulobacter crescentus [20, 25] and gene transfer agent (GTA) production in Rhodobacter capsulatus and Dinoroseobacter shibae [26, 27]. C-di-GMP affects the CtrA phosphorelay directly through effects on the enzymatic activity of CckA, which changes the phosphorylation level of CtrA and thus its activity [28, 29]. The concentration of c-di-GMP also appears to be affected by CtrA because loss of CtrA results in changes in the transcript levels of genes encoding c-di-GMP-metabolizing enzymes [30].

The chromosomal positioning of genes can affect their functions in different ways and have effects on multiple cellular processes. For example, gene location can influence the spatial distribution of proteins within cells due to transcription-coupled translation [31]. Positioning can also have effects with respect to the cell cycle because genes that are close to the origin of replication (*ori*) are replicated earlier and are therefore temporarily present in higher copies than genes that are closer to the terminus of replication (*ter*) [32]. An example where this has important implications was found in Bacillus subtilis, where it was shown that the temporal copy number imbalances due to the opposite localization of genes encoding members of a regulatory network influenced its output [33, 34]. Additionally, gene location can

influence expression due to the state of DNA methylation through the cell cycle. The partially replicated portions of the chromosome are hemi-methylated during replication starting at *ori*, and methylation status can affect regulatory protein binding and transcription [35]. For example, and directly related to regulatory systems already discussed above, one of the promoters where transcription of *ctrA* initiates is only activated in the hemi-methylated state in *C. crescentus* [36]. It seems likely there are additional and broader implications of gene location related to CtrA because a previous analysis also showed that numerous genes that are connected to CtrA have conserved chromosome positions in members of the Alphaproteobacteria [37].

C-di-GMP-modulating enzymes are broadly distributed in phylogenetically and metabolically diverse bacteria. They are also very diverse with respect to their roles and regulation, with a wide range of stimuli affecting c-di-GMP levels, and only a small proportion of the total diversity of these enzymes has been characterized in detail [38]. Therefore, we were interested in identifying any underlying genomic properties that these enzymes might share. We performed a comparative analysis of sequences containing GGDEF, EAL, and HD-GYP domains from five orders of the Alphaproteobacteria, the Rhodospirillales, Sphingomonadales, Rhodobacterales, Rhizobiales and Caulobacterales. We identified the auxiliary domains present with these c-di-GMP-metabolizing domains and attempted to identify patterns regarding enzyme occurrences, distributions, and chromosomal localizations.

5.3 Methods

5.3.1 Dataset

Protein sequences with identified EAL (PF00563), GGDEF (PF00990) or HD (PF01966) domains from genomes of bacteria within five orders of the Alphaproteobacteria (Rhodospirillales, Sphingomonadales, Rhodobacterales, Rhizobiales and Caulobacterales) were downloaded from the EMBL website on 6 August 2020 (GGDEF: http://pfam.xfam.org/family/PF00990#tabview=tab7; EAL: http://pfam.xfam.org/family/PF00563#tabview=tab7; HD: http://pfam.xfam.org/family/PF01966#tabview=tab7) [39]. Proteins with the HHExxxxxGYP motif from within the HD sequences were then selected and considered PDEs while the remaining HD sequences

were considered auxiliary domains if they co-occurred with a c-di-GMP-metabolizing domain. Proteins with both EAL and GGDEF domains were placed in their own group (GGDEF_EAL).

All analyses were done in R version 4.0.3 with the appropriate packages as needed (Table S1).

5.3.2 Organism, domain, and genomic annotations

Sequence identifiers were extracted from the EMBL fasta files and used to access the respective organism information from UniProt (e.g., https://www.uniprot.org/uniprot/V4RSF5.txt) or EBI (e.g., https://www.ebi.ac.uk/ena/browser/api/summary/PNQ76602) [40]. The identifiers were also used to withdraw the domain information from Pfam (e.g., http://pfam.xfam.org/protein/A0A0N0K049#tabview=tab0). Domain annotations could not be withdrawn for all sequences due to inconsistent html path formatting, which reduced the dataset (Table 1). The identifiers were also used to obtain the NCBI protein identifiers from UniProt (e.g., https://www.uniprot.org/uniprot/A0A0N0K3V8.txt). Due to inconsistencies some sequences have different version numbers (e.g., https://www.uniprot.org/uniprot/A0A0N0K3V8.txt?version=1) and only version 1 was subsequently considered in such cases. All sequence identifiers and html paths can be found in Table S2. The NCBI identifiers were used to obtain genomic information from the gff and fasta files, downloaded from NCBI in GenBank format.

Table 5.1: Genomes, genera, and species/strains available for analyses.

Order	Closed genomes	One unambiguously identified *ori*	C-di-GMP-metabolizing domains	By genera[a]	By species or strains
Rhodospirillales	132	75	EAL	42	62
			GGDEF	47	71

			GGDEF_EAL	48	74
			HD-GYP	23	35
Sphingomonadales	27	17	EAL	22	145
			GGDEF	25	172
			GGDEF_EAL	26	174
			HD-GYP	8	18
Rhodobacterales	187	69	EAL	123	308
			GGDEF	132	333
			GGDEF_EAL	121	278
			HD-GYP	11	16
Rhizobiales	424	133	EAL	77	227
			GGDEF	90	257
			GGDEF_EAL	93	266
			HD-GYP	38	105
Caulobacterales	44	8	EAL	5	20
			GGDEF	5	28
			GGDEF_EAL	5	28
			HD-GYP	1	1

[a] Genera numbers include any Sphingomonadales sp., Rhodobacteraceae sp., Rhizobiales sp., and Caulobacteraceae sp. designations as one each.

5.3.3 Identification of chromosomal origins of replication

The origin of replication (*ori*) was identified for each chromosome using Ori-Finder and default settings [41]. The ptt files were generated (https://github.com/sgivan/gb2ptt#gb2ptt) from gbff files, downloaded from NCBI on 23 April 2019. Only chromosomes with one unambiguously identified *ori* were subsequently included in the investigation, which reduced the dataset (Table 1). The terminus of

replication (*ter*) was assumed to be opposite *ori* on the circular chromosomes [42].

5.3.4 Phylogenetic analysis

RpoB sequences (PF05000, RNA polymerase Rpb1, domain 4) were downloaded for the members of each order and their NCBI identifiers were determined. Alignments were done using MAFFT with L-INS-i option [43] in Geneious version 11.0.5 [44]. Phylogenetic trees were reconstructed using IQ-TREE version 2.1.4 [45], with the best substitution matrix identified using ModelFinder. The robustness of the analysis was tested using a bootstrap test (1000 replicates) [46] and a hill-climbing nearest-neighbor interchange search [45, 47]. Trees were modified and annotated in iTOL version 5 [48].

5.4 Results

5.4.1 Occurrence of c-di-GMP-modulating domains

We quantified the genes encoding the domains associated with c-di-GMP synthesis and degradation in members of the five alphaproteobacterial orders. This included those that contained one of the GGDEF, EAL, or HD-GYP domains or both GGDEF and EAL domains. The GGDEF and GGDEF_EAL sequences accounted for the highest proportions in all five orders at 35-48%, followed by proteins containing an EAL domain that ranged between 8.9% and 23.4% of all sequences (Figure 1). The HD-GYP domain-containing sequences made up the smallest share, accounting for only 0.3-5% of all sequences, and co-occurrence of GGDEF or EAL with an HD-GYP domain was not observed (Figure 1). Each c-di-GMP-metabolizing domain was found almost exclusively once per sequence, but there were a few exceptions (Table S4).

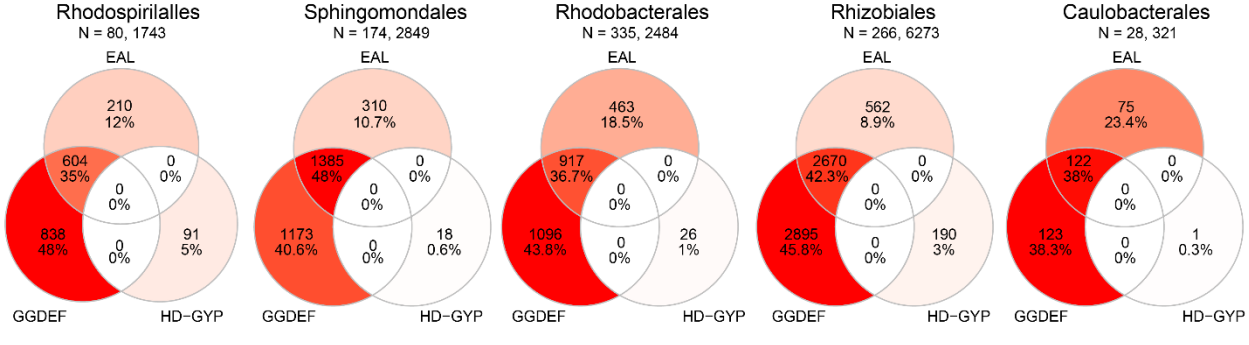

Figure 5.1: Numbers of sequences with GGDEF, EAL or HD-GYP sequences in the five orders. The number of genomes and the total number of sequences for each order are above the diagrams. The Venn diagrams show the numbers of sequences with both GGDEF and EAL domains in the corresponding overlapping circles. The coloration is a gradient from the highest (red) to lowest (white) values within each order.

Next, the numbers of c-di-GMP-metabolizing sequences in different genera were compared by calculating the mean number of sequences per genus (Figure 2, Table S3). The c-di-GMP-metabolizing sequences per genus decreased from the Rhizobiales, Rhodospirillales, Caulobacterales, Sphingomonadales to the Rhodobacterales, but ranges of 1-72 (Rhodomicrobium and Neorhizobium), 1.7-51 (Ferruginivarius and Thalassospira), 3.6-14.2 (Phenylobacterium and Caulobacter), 3-25.4 (Croceicoccus and Novosphingobium), and 1-49 (Salicibibacter and Roseibium) were observed in the respective individual orders.

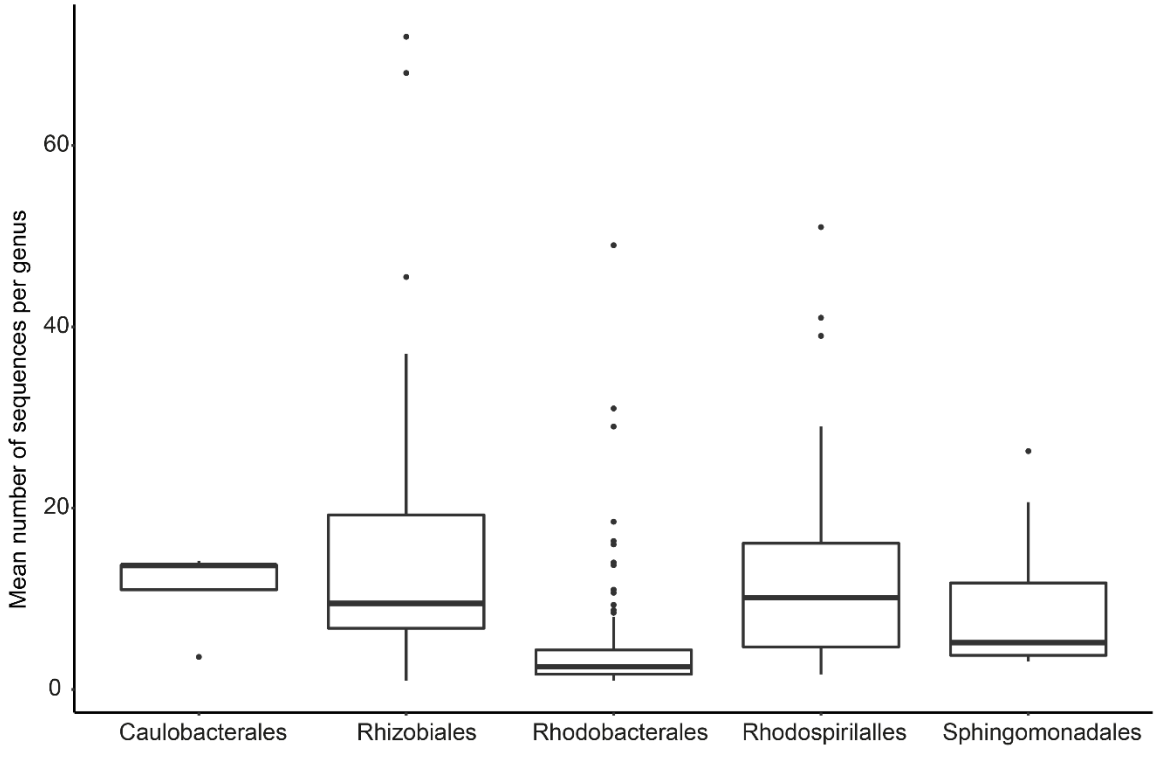

Figure 5.2: Mean number of c-di-GMP-metabolizing sequences per genome per genus in the different orders. The number of c-di-GMP-metabolizing genes in a genus was divided by the number of strains considered in the respective genus. The mean values from all genera of each order were used to make the box plot.

For the subsequent investigation of the numerical relationships among the various domains, all orders were analyzed (Figure S1), but due to the larger number of available sequences and therefore more unambiguous results, we focused in particular on the Rhodobacterales and Rhizobiales. Examination of the per genus ratios of genes encoding synthesizing enzymes to those encoding hydrolyzing enzymes, i.e., GGDEF:EAL, revealed that this ratio was always 2 or higher (Figure 3A). However, the ratio was more consistently close to 2 across the Rhodobacterales (0.5-6) as compared to the Rhizobiales (1-16), where there were more frequently higher numbers of GGDEF sequences and more variation among members of this order. When the numbers of GGDEF and EAL domain sequences per genome were examined (Figure 3B), we found that the medians were 2 and 11 for GGDEF sequences and 1 and 2 for EAL sequences in the Rhodobacterales and Rhizobiales, respectively. This again shows that genes encoding the synthesizing enzymes occur more frequently than those encoding hydrolyzing enzymes in both orders. The GGDEF:GGDEF_EAL ratios peaked at 1 in the studied orders except the Rhodobacterales where more variability was observed and a higher proportion of members showed higher ratios (Figure 3A, Figure S2). Interestingly, the relationships of the GGDEF:EAL and GGDEF:GGDEF_EAL ratios showed opposite patterns in the Rhodobacterales and Rhizobiales. While the GGDEF:EAL ratios were less variable and most consistently at 2 in the Rhodobacterales, there was much greater variability in the Rhizobiales. Conversely, there was more variability in the GGDEF:GGDEF_EAL ratios in the Rhodobacterales but a distinct peak at 1 in the Rhizobiales. The relationship of GGDEF:HD-GYP domains was found to be fairly consistent at 2.5:1 in the Rhodobacterales but highly variable in the Rhizobiales (Figure 3A).

The large variability in numbers of c-di-GMP-metabolizing proteins among organisms stimulated us to investigate their evolutionary relationships. Therefore, the number of c-di-GMP enzymes present in different species was evaluated in a phylogenetic context (Figure S4). Some closely related groups were found in which the numbers of c-di-GMP genes were similar. In the Rhizobiales there was a large cluster in which the c-di-GMP-metabolizing gene numbers were elevated, and which consisted of several genera, including *Devosia*, *Fulvimarina* and Rhizobium. Smaller additional clusters with increased c-di-GMP

numbers that were less closely related were also observed. In the Rhodobacterales, the closely related genera *Stapia* and *Labrenzia* stood out with their high c-di-GMP-metabolizing gene numbers. A connection between phylogeny and c-di-GMP-metabolizing gene number could also be observed in the Rhodospirillales. Here there were three clusters of organisms that had increased gene numbers and one notable group was made up of three genera including *Magnetospirillum*, *Magnetovibrio* and *Telmatosprillum*. A clear connection between phylogenetic relationships and numbers of c-di-GMP-metabolizing genes was not observed in the Sphingomonadales, and it is difficult to make any statement for the Caulobacterales because of the lower genome and gene numbers.

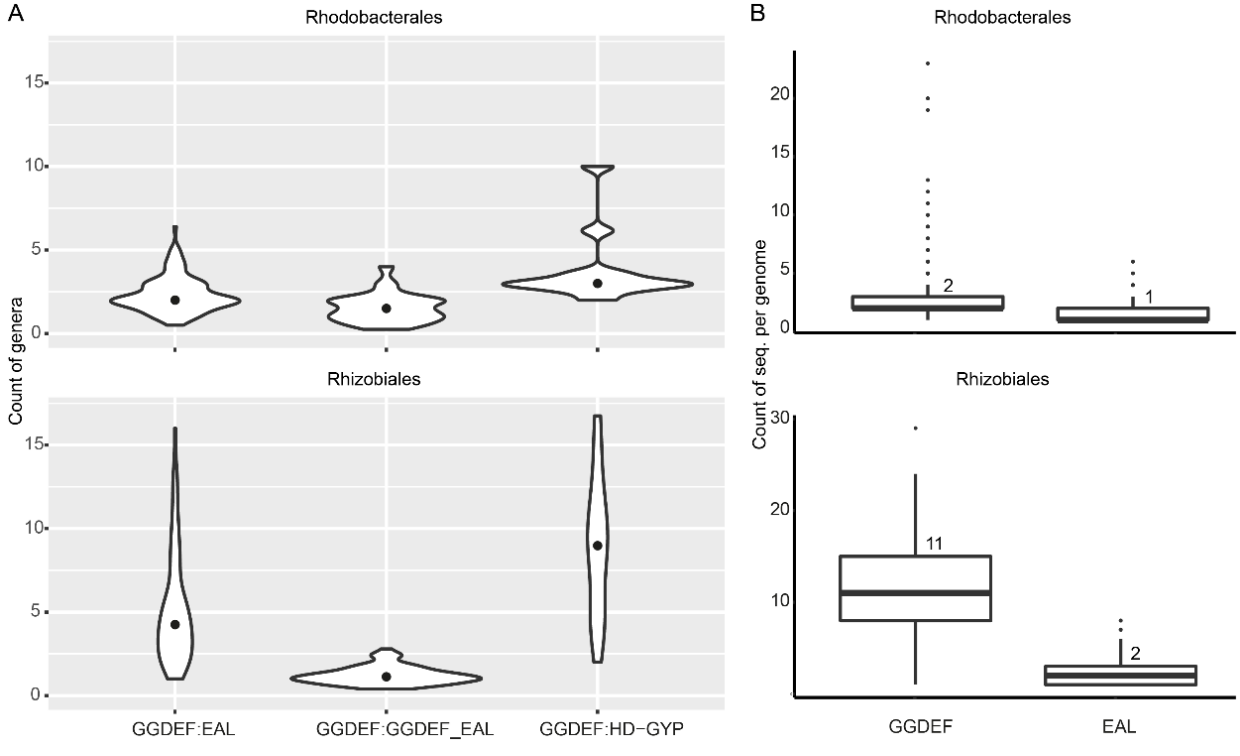

Figure 5.3: Numerical relationships among c-di-GMP-metabolizing sequences. A. Ratios for GGDEF:EAL, GGDEF:GGDEF_EAL, and GGDEF:HD_GYP sequences for the orders Rhodobacterales and Rhizobiales. The ratios were calculated per genome and the mean per genus was plotted. The median is indicated by the black dot. B. Counts of sequences with only a GGDEF domain or only an EAL domain per genome. The median value (50% quantile) is given on top of each box.

5.4.2 Relationship between gene numbers, genome size, and location of c-di-GMP-associated genes on secondary chromosomes

There was a statistically significant positive correlation between chromosome size and the number of c-di-GMP-metabolizing genes in all five orders (Figure S5). We only included the largest replicon in this analysis, although c-di-GMP-metabolizing genes were also found on secondary chromosomes and extrachromosomal replicons. In five genomes from different genera of the Rhodospirillales, six genomes from three genera in the Sphingomonadales, five genomes from five different genera of the Rhodobacterales, 23 genomes from 13 genera of the Rhizobiales, and one genome of the Caulobacterales c-di-GMP-metabolizing genes were found outside of the largest replicon (Table S5). In Nitrospirillum amazonense CBAmc (Rhodospirillales), Rhizobium sp. NXC24 (Rhizobiales) and Asticcacaulis excentricus CB 48 (Caulobacterales) more c-di-GMP genes were found on the second-largest replicon and in Paracoccus denitrificans PD1222 (Rhodobacterales) equal numbers of c-di-GMP-metabolizing genes were found on the largest and second-largest replicons.

Secondary chromosomes (defined as replicons >800 kb that are not the largest replicons in the genome) contain genes that evolve faster [49] and are more common in the Rhizobiales (Figure 4). We investigated if c-di-GMP-metabolizing genes were found outside of the largest chromosome more often when secondary chromosomes were present. We found that only a small fraction of the genomes examined in this study had secondary chromosomes in four of the orders (14.9% or 21 genomes of the Rhodospirillales, 8.8% or 10 genomes of the Sphingomonadales, 10% or 15 genomes of the Rhodobacterales, and 6.7% or 2 genomes of the Caulobacterales) whereas this was higher for the Rhizobiales (44% or 204 genomes). There were c-di-GMP-metabolizing genes on the secondary chromosomes in all orders and these accounted for 21.3%, 21.1%, 30%, 31.7% and 73.3% of all c-di-GMP-metabolizing genes in the Rhodospirillales, Sphingomonadales, Rhodobacterales, Rhizobiales, and Caulobacterales, respectively. We note that the high percentage of c-di-GMP-metabolizing genes identified on secondary chromosomes in the Caulobacterales is based on only two genomes. Overall, the results indicate that the presence of secondary chromosomes did not result in a greater proportion of c-di-

GMP-metabolizing genes located there.

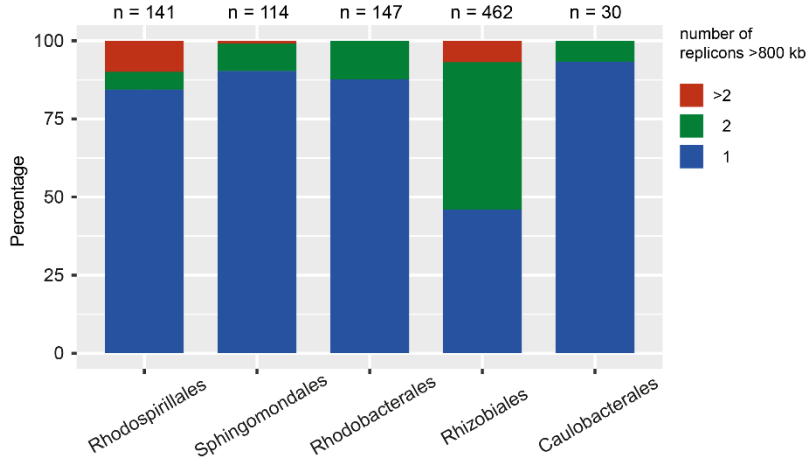

Figure 5.4: Proportions of genomes with one, two or more than two replicons >800 kb in the five orders. The total numbers of genomes in each order are above the plot.

5.4.3 Chromosomal organization patterns of c-di-GMP-associated genes

As discussed above, location on the chromosome can affect gene expression. We therefore wanted to examine the localization of c-di-GMP-metabolizing genes on chromosomes relative to the origin (*ori*) and terminus (*ter*) of replication. No obvious trend was observed in the Rhodospirillales, while GGDEF and GGDEF_EAL sequences seemed less prevalent near *ter* in the Sphingomonadales (Figure S6). The number of genes included in the analysis for the Sphingomonadales EAL group and all groups for the Caulobacterales were so low that patterns might not be obvious even if present. However, interesting patterns were evident in the Rhizobiales and Rhodobacterales (Figure 5). In the Rhodobacterales the EAL and GGDEF_EAL sequences were predominately found near *ori* whereas GGDEF sequences were predominately not close to *ori* and showed a tri-modal distribution with peaks mid-way between *ori* and *ter* and around *ter*. In the Rhizobiales, clear patterns were observed for the GGDEF and GGDEF_EAL sequences, which both showed multiple peaks but with opposing patterns. The distribution of the GGDEF sequences showed three peaks, with the largest near *ter* and two smaller peaks near *ori*. The GGDEF_EAL sequences peaked where the GGDEF sequences were lowest, mid-way between *ori* and *ter*. Although there were far fewer sequences, the Rhizobiales HD-GYP group showed a

similar trend as the GGDEF_EAL sequences, while there was no obvious pattern for the EAL sequences.

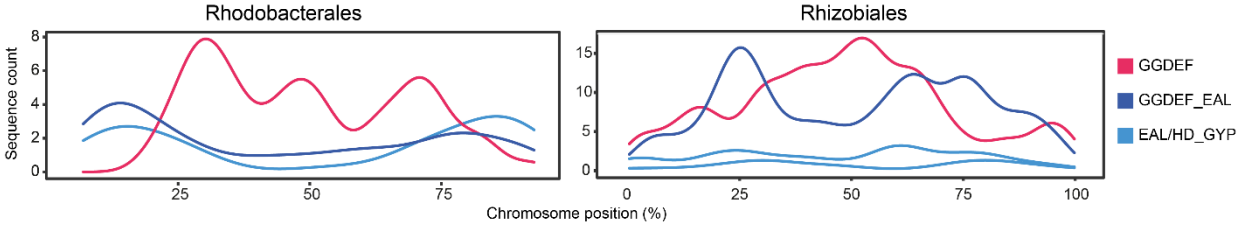

Figure 5.5: Chromosomal locations of c-di-GMP-metabolizing genes. Cumulative distributions of c-di-GMP-metabolizing genes on the chromosomes of Rhodobacterales and Rhizobiales, with lengths normalized to 100% where *ori* is at 0% and *ter* is at 50%. The color-coded lines represent the estimate of the kernel density. Only closed genomes with one unambiguously determined *ori* were used in this analysis.

Comparison of the similarities of distributions among the groups of genes indicated that the Rhodobacterales EAL and GGDEF_EAL genes were similarly distributed (two-sample Kolmogorov-Smirnov test; p-value = 0.16) while the EAL and GGDEF as well as the GGDEF and GGDEF_EAL pairs were distributed differently (two-sample Kolmogorov-Smirnov test; p-values = 0.009 and 0.0008, respectively). The Rhizobiales GGDEF and GGDEF_EAL genes were also distributed differently (two-sample Kolmogorov-Smirnov test; p-value = 0.04).

5.4.4 Additional domains on c-di-GMP-associated proteins

It has previously been documented that proteins with c-di-GMP-metabolizing domains frequently contain additional domains [20], hereafter referred to as auxiliary domains, which presumably function in many cases to regulate the c-di-GMP-related enzymatic activities. Only the Rhodobacterales and Rhizobiales are discussed in detail here because of the larger numbers of sequences available for these orders, but similar trends were also observed in the other three (Table S6, Figure S7). Auxiliary domains were associated with all four c-di-GMP sequence groups and there were 101 different auxiliary domains found across all five orders and sequence groups. We note that the auxiliary domains analyzed here are those that are identified and specified in databases but recognize that some of the sequences will have uncharacterized domains that are not captured there. We plotted the length of EAL-containing sequences

and this showed that all those with identified auxiliary domains were >375 amino acids long (Figure S8). The proportions of those without identified auxiliary domains that were <375 amino acids long were 49% in the Rhizobiales and 82% in the Rhodobacterales, indicating that some of these proteins likely contain auxiliary domains but these remain to be recognized and annotated in the sequence databases. The same analysis with GGDEF sequences revealed that all sequences containing identified auxiliary domains were >275 amino acids long (Figure S8). The proportions of those without identified auxiliary domains that were <275 amino acids long were 13% in the Rhizobiales and 30% in the Rhodobacterales and, therefore, most of these sequences likely also contain currently unannotated auxiliary domains.

In both the Rhodobacterales and Rhizobiales the GGDEF group had the highest variability among auxiliary domains, followed by GGDEF_EAL, EAL and HD-GYP sequences (Figure 6A). However, this could be driven by the higher number of sequences containing GGDEF domains compared to other domains (Figure 1). The GGDEF and GGDEF_EAL groups had the greatest overlap of auxiliary domains whereas there were only a few unique domains present with the EAL and HD-GYP domain-containing sequences. Overall, there were uniform distributions of sequences that contain none, one, or more than one auxiliary domain (Figure 6B, Figure S7). The HD-GYP group had the highest proportion of sequences with auxiliary domains, followed by the GGDEF_EAL, GGDEF and EAL groups (Figure 6B). The GGDEF_EAL group had the biggest proportion of sequences that had more than one auxiliary domain on individual proteins.

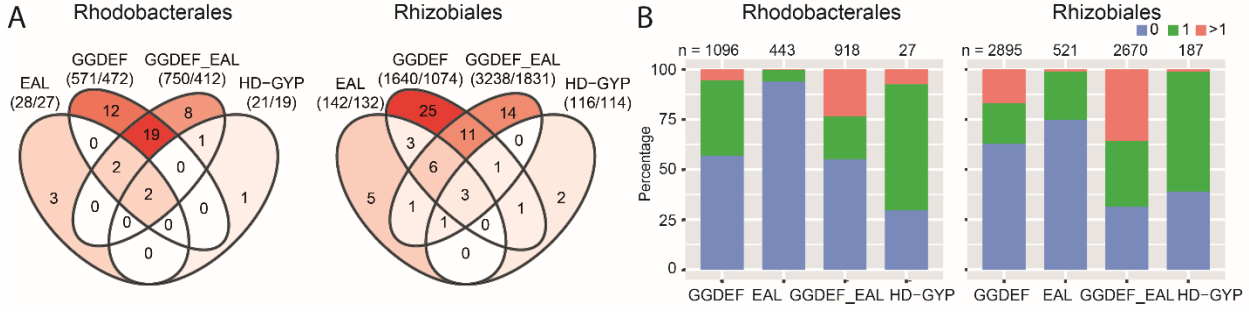

Figure 5.6: Occurrence of auxiliary domains on c-di-GMP-metabolizing proteins of the different enzyme groups. A. Numbers of different auxiliary domains that can be found for each group and shared among groups. The first number below the group identification (EAL, GGDEF, GGDEF_EAL, HD-GYP) indicates the number of auxiliary domains in the respective group and the second number indicates the number of sequences these domains are found in. The c-di-GMP-metabolizing domains

themselves are not included in this analysis. The color code of the Venn diagram represents the domain counts from the highest (red) to zero (white). B. Percentage of sequences with none, one, or more than one auxiliary domain. The number of sequences included in this analysis is given above the group identification. Repeated occurrence of a domain in a sequence was counted as one.

Some auxiliary domains were more commonly found in certain groups and some of these co-occurrences were conserved across the five orders (Tables S6 and S7). A previous study reported that cGMP-specific phosphodiesterases, adenylyl cyclases and FhlA (GAF) and Per-Arnt-Sim (PAS) were the most common auxiliary domains associated with GGDEF domains in various bacterial species [17]. The GAF domain is a sensory domain involved in light sensing and it and the PAS domain have been found in phytochromes [17, 50]. In our GGDEF sequences, the response regulator receiver (REC) domain and PAS domain variants dominated. Cognate histidine kinases modulate REC domain-containing proteins through their phosphorylation status via their kinase and phosphatase activities, which are themselves regulated by various signals. The phosphorylation status of the REC domain then controls the activity of the associated output domain (e.g., GGDEF). In the EAL group REC domains, CSS-motif (Pfam PF12792) domains and GAF_2 domains were most common. CSS-motif domains are known for roles in redox sensing [17]. The Caulobacterales EAL sequences were an exception, because these were most often associated with histidine kinase and phosphotransferase domains that act upstream of REC domains in histidyl-aspartyl phosphorelay systems. In the GGDEF_EAL sequences, the PAS subfamilies PAS_3, PAS_4, PAS_7 and PAS_9, as well as the MHYT domain were most common. The MHYT domain consists of six transmembrane segments and it has been suggested to function in O2, NO and CO sensing [51]. In the HD-GYP sequences HD_5 and two domains of unknown function, DUF3369 and DUF3391, were the most prevalent.

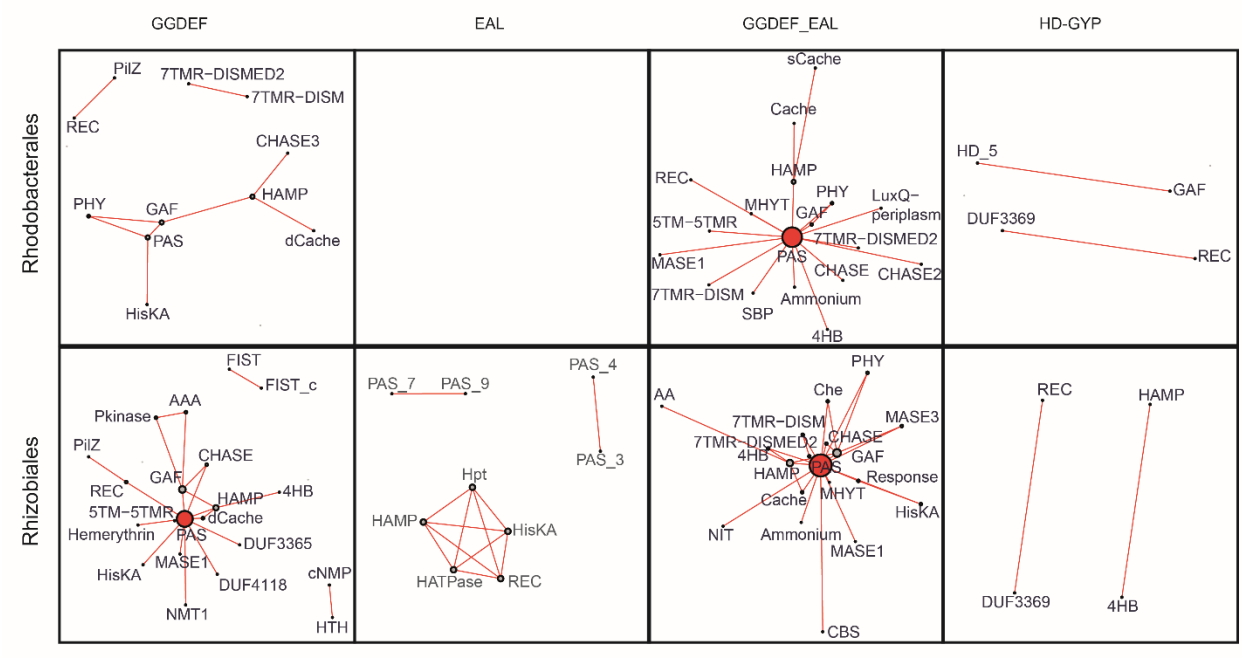

Figure 5.7: Weighted graphs representing the co-occurrences of auxiliary domains with c-di-GMP-metabolizing sequences. Auxiliary domains occurring together are connected by lines with the size and red color of the node indicating higher frequency of co-occurrence with other domains. Lengths of edges represent the number of times the connected domains co-occur, and the sizes of the points indicate the number of times these domains occur. All full domain names are provided in Table S8.

Despite detailed knowledge on the structure and function of DGCs and PDEs, it has remained challenging to assign physiological roles to individual proteins. Analysis of the co-occurrence of additional domains might aid in assigning those roles. Therefore, we next investigated which additional domains occurred together and constructed co-occurrence networks (Figure 6, Table S8). We focused on the Rhodobacterales and Rhizobiales because more sequences with more than one auxiliary domain were available for these orders. Most of the Rhodobacterales GGDEF sequences that had more than one auxiliary domain had co-occurrences of two specific auxiliary domains (Figure 7). Exceptions were phytochrome (PHY), PAS, GAF, histidine kinase, adenylate cyclase, methyl-accepting protein and phosphatase (HAMP) domains, which co-occurred with two or three other domains. PAS domains were dominant in co-occurrences with many other domains in the GGDEF and GGDEF_EAL groups of both orders as well as the Rhizobiales' EAL group (Figure 7). Linkage of one domain with a variety of others

creates complex patterns, such as found for the GGDEF sequences of both orders where calcium channels and chemotaxis receptors (dCache_1), GAF_2, HAMP and cyclase/histidine kinase-associated sensory extracellular (CHASE3) domains formed a network. The Cache and CHASE domains are extra-cytoplasmic sensory domains [52, 53] while the HAMP domain is usually found in integral membrane proteins that transmit conformational changes from periplasmic ligand-binding domains to cytoplasmic domains as part of histidyl-aspartyl phosphorelay signaling [54]. In the GGDEF_EAL sequences of both orders and the GGDEF sequences of the Rhizobiales, the PAS domains were notable because they are the domains connected with the most other auxiliary domains. Interestingly, the EAL sequences of the Rhizobiales had one cluster composed of the same domains that are most prevalent in the EAL sequences of the Caulobacterales (Table S7). These are the HisKA domain (activated via dimerization and able to transfer a phosphoryl group often as part of histidyl-aspartyl phosphorelay systems [55]), the Hpt domain that mediates phosphotransfer in histidyl-aspartyl phosphorelay systems [56], the HAMP domain, and HATPase that is found in multiple ATPases such as histidine kinases [57]. This shows that the EAL sequences, when linked to auxiliary domains, are often part of signaling cascades, especially in the Rhizobiales and Caulobacterales. The HD-GYP sequences showed two connections per order, one of which seemed to be conserved in the Rhodobacterales and Rhizobiales and consisted of the DUF3369 and REC domains.

5.5 Discussion

5.5.1 Association with diverse secondary domains suggests a wide variety of signals affect DGC activity

Our analysis of the occurrence of the EAL, GGDEF, GGDEF_EAL and HD-GYP sequences in orders of the Alphaproteobacteria showed that the GGDEF and GGDEF_EAL domains made up the biggest proportions in all orders, followed by the EAL domains, while the HD-GYP domains accounted for the smallest share. Compared to results from a study on c-di-GMP-metabolizing gene distributions among prokaryotes, which found the overall proportions to be 50.4% GGDEF, 16.1% EAL and 33.5% GGDEF_EAL [11], the alphaproteobacterial orders have slightly lower GGDEF and higher

GGDEF_EAL proportions. Moreover, the GGDEF and GGDEF_EAL sequences are associated with more different types of auxiliary domains and have a proportionally higher occurrence of auxiliary domains, respectively. This suggests that the GGDEF_EAL proteins more frequently respond to signals/stimuli, but the GGDEF-only proteins integrate a broader variety of signals. Thus, since GGDEF domain sequences are more abundant and seem to have more and more diverse auxiliary domains than PDE domain sequences, it could be that the synthesis of c-di-GMP is mainly controlled in response to extracellular and intracellular signals while its degradation is more unspecific. Since the GGDEF_EAL sequences of the Rhizobiales, like the EAL sequences of the Rhodobacterales, show a lower diversity of auxiliary domains, they too could be responsible for unspecific degradation while increases in c-di-GMP are more regulated. However, we note that this analysis is limited by its reliance on detecting recognized auxiliary domains while it is likely that some of these proteins contain currently unrecognized auxiliary domains.

5.5.2 Importance of EAL-type PDE domains in Proteobacteria

Proteins with only EAL domains outnumbered those with HD-GYP domains at least two-fold in all orders. This agrees with a previous analysis of these domains in several phyla where the Proteobacteria, with the exception of the Deltaproteobacteria, and Oligoflexia were the only investigated phyla in which EAL domains outnumbered HD-GYP domains [58]. The driving forces behind the trends for relative abundances of these two different types of PDEs are not clear and likely require a larger phylogenetic analysis to untangle. More information on the specific roles of individual proteins is also required. The possible activities of proteins with both GGDEF and EAL domains, which are even more abundant than PDEs without GGDEF domains, further complicates the situation.

5.5.3 Shared genomic features of the Rhizobiales GGDEF_EAL and Rhodobacterales EAL sequences

Interestingly, multiple commonalities exist between the GGDEF_EAL sequences of the Rhizobiales and the EAL sequences of the Rhodobacterales. Both gene groups are biased for localization away from *ter*, and their relative abundances compared to GGDEF sequences are reversed in the two

orders. The GGDEF:EAL ratio is very consistent in the Rhodobacterales but there is no such consistency in the Rhizobiales. Conversely, while the GGDEF:GGDEF_EAL ratios were more varied in the Rhodobacterales, they were much more consistent in the Rhizobiales. This could indicate that the roles of the EAL sequences in the Rhodobacterales are swapped with GGDEF_EAL sequences in the Rhizobiales. However, the hybrid nature of GGDEF_EAL sequences makes this difficult to conclude. The two activities can be switched, e.g., by dimerization, which is required for GGDEF but not for EAL activity [15], or through regulation by auxiliary domains [14, 59, 60]. However, a study of the conservation of amino acid patterns showed that the catalytic activity in hybrid sequences is most often preserved in both domains or only in the EAL domain [16]. Future studies must show whether the Rhizobiales hybrid sequences have mainly retained EAL activity and thereby compensate for the lack of EAL sequences near *ori*, assuming they are involved in the same functions as the Rhodobacterales EAL sequences that are also positioned near *ori*. This could potentially be initially evaluated through a large-scale bioinformatic analysis of the enzymatic domains in the Rhizobiales GGDEF_EAL hybrids to look for conservation of known critical residues required for DGC and PDE activity.

5.5.4 Conserved chromosomal positioning

In the two orders with the most available data, the Rhodobacterales and Rhizobiales, there is a clear conservation of the Rhodobacterales EAL- and GGDEF_EAL- and the Rhizobiales GGDEF_EAL-encoding genes away from *ter* while the GGDEF-encoding genes are predominant on the *ter*-proximate half of the chromosome in both orders. Overall, the concentrations of GGDEF genes peak when the EAL and GGDEF_EAL genes in the Rhodobacterales and the GGDEF_EAL genes in the Rhizobiales drop. This could indicate that there is more c-di-GMP degradation in the cell near the *ori* and more synthesis near the *ter* in the Rhodobacterales, which could also apply to the Rhizobiales should it turn out that the hybrid sequences primarily act as PDEs (discussed above).

There are multiple potential effects caused by the chromosomal locations of specific genes. The observed chromosomal localization patterns revealed in this study might affect cellular c-di-GMP concentrations during the cell cycle. Genes that are close to *ori* are replicated earlier than genes that are

close to *ter*, which leads to a temporary copy number imbalance between genes at these two locations [32]. In *B. subtilis*, the opposite location of two genes encoding components of a phosphorelay with respect to *ori* and *ter* leads to temporal copy number imbalances, and this allows spore formation to only take place at the end of the cell cycle when the balance between the regulators is restored [33, 34]. In Vibrio cholerae, moving genes from *ori* to *ter* and thus reducing their copy number during the cell cycle has an impact on growth and infectivity [61, 62]. Such copy number imbalances can be pronounced in organisms that initiate multiple rounds of DNA replication within individual cells, such as *Escherichia coli* [63], although there is no evidence this occurs in members of the alphaproteobacteria. Regardless, it is possible that the biased localizations of genes encoding c-di-GMP-metabolizing enzymes we observed could have some effects on cellular c-di-GMP concentrations through temporary copy number imbalances, but future experimental work is required to evaluate this.

Another effect of localization could be manifested through DNA methylation, where the chromosomal DNA changes from fully methylated to hemi-methylated during replication. This change in methylation can affect gene transcription. For example, the p1 promoter of the *ctrA* gene in *C. crescentus* is only active in the hemi-methylated state. Thus, *ctrA*, which is localized near *ori*, is transcribed more during DNA replication because it is hemi-methylated right at the beginning of the cycle. However, any broad role of methylation in regulating transcription of genes encoding c-di-GMP-metabolizing enzymes is currently unknown and future work is required to investigate this possibility.

5.6 Conclusions

C-di-GMP-metabolizing enzymes are very diverse, and the specific roles and functions of only a few of these proteins are known. In this study new patterns and common properties for these proteins were identified in members of the alphaproteobacteria. We systematically examined gene occurrence, localization on the genome, and the presence of auxiliary domains. In the Rhodobacterales and Rhizobiales, the EAL and GGDEF_EAL sequences, respectively, are primarily located away from *ter* while GGDEF sequences are biased towards ter. Additionally, the EAL and GGDEF_EAL domain-containing sequences show lower diversity and occurrence of auxiliary

domains compared to the GGDEF sequences. There are several known examples in which chromosome localization of genes is important, and this can manifest in different ways such as through changes in copy number and methylation status during the cell cycle. The patterns we found support the suggestion that the chromosomal localization of c-di-GMP-metabolizing genes is important in these bacteria. Our findings also support the notion that the synthesis of c-di-GMP is more regulated and responsive to a variety of specific signals whereas its degradation might be less regulated and dependent on different stimuli.

5.7 Supplementary materials

This chapter includes eight supplementary files that are available in digital format using this link:

The Word file contains the supplementary figures 1-7. Excel file S5.1 represents the sequences identifiers withdrawn from the different databases, Pfam, EBI and NCBI. Excel file S5.2 shows the number of sequences per genus. Excel file S5.3 shows the identified domains in each sequence. Excel file S5.4 represents the count of c-di-GMP-associated genes on chromosomes and plasmids. Excel file S5.5 shows the occurrence of secondary domains with cyclic di-GMP-modulating domain sequences. Excel file S5.6 displays the frequency of association of all secondary domains that co-occur with c-di-GMP-associated domains. Excel file S5.7 shows all domains found associated with one of the examined c-di-GMP-associated domains.

5.8 References

1. Ross P, Weinhouse H, Aloni Y, Michaeli D, Weinberger-Ohana P, Mayer R, et al. Regulation of cellulose synthesis in *Acetobacter xylinum* by cyclic diguanylic acid. *Nature*. **1987**, *325*, 279–81.

2. Ross P, Aloni Y, Weinhouse H, Michaeli D, Weinberger-Ohana P, Mayer R, et al. Control of cellulose synthesis *Acetobacter xylinum*. A unique guanyl oligonucleotide is the immediate activator of the cellulose synthase. *Carbohydr Res*. **1986**, *149*, 01–17.

3. Valentini M, Filloux A. Multiple Roles of c-di-GMP Signaling in Bacterial Pathogenesis. *Annu Rev Microbiol.* **2019**, *73*, 387–406.

4. Römling U, Galperin MY, Gomelsky M. Cyclic di-GMP: the First 25 Years of a Universal Bacterial Second Messenger. *Microbiol Mol Biol Rev*. **2013**, *77*, 1–52.

5. Krasteva PV, Sondermann H. Versatile modes of cellular regulation via cyclic dinucleotides. *Nat Chem Biol*. **2017**, *13*, 350–9.

6. Schirmer T. C-di-GMP Synthesis: Structural Aspects of Evolution, Catalysis and Regulation. *J Mol Biol*. **2016**, *428*, 3683–701.

7. Stelitano V, Giardina G, Paiardini A, Castiglione N, Cutruzzolà F, Rinaldo S. C-di-GMP Hydrolysis by *Pseudomonas aeruginosa* HD-GYP Phosphodiesterases: Analysis of the Reaction Mechanism and Novel Roles for pGpG. *PLoS One*. **2013**, *8*, e74920.

8. Christen M, Christen B, Folcher M, Schauerte A, Jenal U. Identification and Characterization of a Cyclic di-GMP-specific Phosphodiesterase and Its Allosteric Control by GTP*. *J Biol Chem*. **2005**, *280*, 30829–37.

9. Cohen D, Mechold U, Nevenzal H, Yarmiyhu Y, Randall TE, Bay DC, et al. Oligoribonuclease is a central feature of cyclic diguanylate signaling in *Pseudomonas aeruginosa*. *Proc Natl Acad Sci*. **2015**, *112*, 11359–64.

10. Orr MW, Donaldson GP, Severin GB, Wang J, Sintim HO, Waters CM, et al. Oligoribonuclease is the primary degradative enzyme for pGpG in *Pseudomonas aeruginosa* that is required for cyclic-di-GMP turnover. *Proc Natl Acad Sci*. **2015**, *12*, e5048–57.

11. Bellini D, Caly DL, McCarthy Y, Bumann M, An S-Q, Dow JM, et al. Crystal structure of an HD-GYP domain cyclic-di-GMP phosphodiesterase reveals an enzyme with a novel trinuclear catalytic iron centre. *Mol Microbiol*. **2014**, *91*, 26–38.

12. Ryan RP, Fouhy Y, Lucey JF, Crossman LC, Spiro S, He Y-W, et al. Cell–cell signaling in *Xanthomonas campestris* involves an HD-GYP domain protein that functions in cyclic di-GMP turnover. *Proc Natl Acad Sci.* **2006**, *103*, 6712–7.

13. P. RR, Yvonne F, F. LJ, Maxwell DJ. Cyclic Di-GMP Signaling in Bacteria: Recent Advances and New Puzzles. *J Bacteriol*. **2006**, *188*, 8327–34.

14. Levet-Paulo M, Lazzaroni J-C, Gilbert C, Atlan D, Doublet P, Vianney A. The Atypical Two-component Sensor Kinase Lpl0330 from *Legionella pneumophila* Controls the Bifunctional Diguanylate Cyclase-Phosphodiesterase Lpl0329 to Modulate Bis-(3'-5')-cyclic Dimeric GMP Synthesis. *J Biol Chem*. **2011**, *286*, 31136–44.

15. Jenal U, Malone J. Mechanisms of Cyclic-di-GMP Signaling in Bacteria. *Annu Rev Genet*. **2006**, *40*, 385–407.

16. Seshasayee ASN, Fraser GM, Luscombe NM. Comparative genomics of cyclic-di-GMP signalling in bacteria: post-translational regulation and catalytic activity. *Nucleic Acids Res.* **2010**, *38*, 5970–81.

17. Randall TE, Eckartt K, Kakumanu S, Price-Whelan A, Dietrich LEP, Harrison JJ. Sensory Perception in Bacterial Cyclic Diguanylate Signal Transduction. *J Bacteriol.* **2022**, *204*, e00433-21.

18. Römling U, Gomelsky M, Galperin MY. C-di-GMP: the dawning of a novel bacterial signalling system. *Mol Microbiol*. **2005**, *57*, 629–39.

19. Krol E, Schäper S, Becker A. Cyclic di-GMP signaling controlling the free-living lifestyle of alpha-proteobacterial rhizobia. *Biol Chem*. **2020**, *401*, 1335–48.

20. Jenal U, Reinders A, Lori C. Cyclic di-GMP: second messenger extraordinaire. *Nat Rev Microbiol.* **2017**, *15*, 271–84.

21. Brilli M, Fondi M, Fani R, Mengoni A, Ferri L, Bazzicalupo M, et al. The diversity and evolution of cell cycle regulation in alpha-proteobacteria: a comparative genomic analysis. *BMC Syst Biol*. **2010,** *4*, 52.

22. Biondi EG, Reisinger SJ, Skerker JM, Arif M, Perchuk BS, Ryan KR, et al. Regulation of the bacterial cell cycle by an integrated genetic circuit. *Nature*. **2006**, *444*, 899–904.

23. Greene SE, Brilli M, Biondi EG, Komeili A. Analysis of the CtrA pathway in *Magnetospirillum* reveals an ancestral role in motility in alphaproteobacteria. *J Bacteriol*. **2012**, *194*, 2973–86.

24. Poncin K, Gillet S, De Bolle X. Learning from the master: targets and functions of the CtrA response regulator in *Brucella abortus* and other alpha-proteobacteria. *FEMS Microbiol Rev*. **2018**, *42*, 500–13.

25. Lori C, Ozaki S, Steiner S, Bohm R, Abel S, Dubey BN, et al. Cyclic di-GMP acts as a cell cycle oscillator to drive chromosome replication. *Nature*. **2015**, *523*, 236–9.

26. Pallegar P, Peña-Castillo L, Langille E, Gomelsky M, Lang AS. Cyclic di-GMP-mediated regulation of gene transfer and motility in *Rhodobacter capsulatus*. *J Bacteriol*. **2020**, *202*, e00554-19.

27. Koppenhöfer S, Lang AS. Interactions among redox regulators and the CtrA phosphorelay in *Dinoroseobacter shibae* and *Rhodobacter capsulatus*. *Microorganisms*. **2020**, *8*, 562.

28. Mann TH, Seth Childers W, Blair JA, Eckart MR, Shapiro L. A cell cycle kinase with tandem sensory PAS domains integrates cell fate cues. *Nat Commun*. **2016**, *7*.

29. Farrera-Calderon RG, Pallegar P, Westbye AB, Wiesmann C, Lang AS, Beatty JT. The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer agent production is bidirectional and regulated by cyclic di-GMP. *J Bacteriol*. **2021**, *203*, e00525-20.

30. Mercer RG, Callister SJ, Lipton MS, Pasa-Tolic L, Strnad H, Paces V, et al. Loss of the response regulator CtrA causes pleiotropic effects on gene expression but does not affect growth phase regulation in *Rhodobacter capsulatus*. *J Bacteriol*. **2010**, *192*, 2701–10.

31. Gowrishankar J, Harinarayanan R. Why is transcription coupled to translation in bacteria? *Mol Microbiol*. **2004**, *54*, 598–603.

32. Slager J, Veening J-W. Hard-Wired Control of Bacterial Processes by Chromosomal Gene Location. *Trends Microbiol*. **2016**, *24*, 788–800.

33. Narula J, Kuchina A, Lee DD, Fujita M, Süel GM, Igoshin OA. Chromosomal Arrangement of Phosphorelay Genes Couples Sporulation and DNA Replication. *Cell*. **2015**, *162*, 328–37.

34. Lazazzera BA, Hughes D. Location affects sporulation. Nature. 2015;525:42–3.

35. Marinus MG, Løbner-Olesen A. DNA Methylation. *EcoSal Plus.* **2014**, *6*.

36. Reisenauer A, Shapiro L. DNA methylation affects the cell cycle transcription of the CtrA global regulator in Caulobacter. *EMBO J.* **2002**, *21*, 4969–77.

37. Tomasch J, Koppenhöfer S, Lang AS. Connection between chromosomal location and function of CtrA phosphorelay genes in Alphaproteobacteria. *Front Microbiol*. **2021,** 1–8.

38. Sondermann H, Shikuma NJ, Yildiz FH. You've come a long way: c-di-GMP signaling. *Curr Opin Microbiol*. **2012,** *15*, 140–6.

39. PFAM Database BT - Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 1392.

40. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49 D480–9.

41. Luo H, Quan C-L, Peng C, Gao F. Recent development of Ori-Finder system and DoriC database for microbial replication origins. *Brief Bioinform*. 2019;20:1114–24.

42. Rocha EPC. The replication-related organization of bacterial genomes. *Microbiology*. 2004;150:1609–27.

43. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.

44. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647–9.

45. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*. 2020;37:1530–4.

46. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evol*. 1985;39:783–91.

47. Sankoff D, Abel Y, Hein J. A tree · a window · a hill; generalization of nearest-neighbor interchange in phylogenetic optimization. *J Classif*. 1994;11:209–32.

48. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49:W293–6.49. Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ. Why Genes Evolve Faster on Secondary Chromosomes in Bacteria. PLOS Comput Biol. 2010;6:e1000732.

50. Nagatani A. Phytochrome: structural basis for its functions. *Curr Opin Plant Biol*. 2010;13:565–70.

51. Galperin MY, Gaidenko TA, Mulkidjanian AY, Nakano M, Price CW. MHYT, a new integral membrane sensor domain. *FEMS Microbiol Lett.* 2001;205:17–23.

52. Anantharaman V, Aravind L. Cache – a signaling domain common to animal Ca2+-channel subunits and a class of prokaryotic chemotaxis receptors. *Trends Biochem Sci*. 2000;25:535–7.

53. Zhulin IB, Nikolskaya AN, Galperin MY. Common Extracellular Sensory Domains in Transmembrane Receptors for Diverse Signal Transduction Pathways in Bacteria and Archaea. J Bacteriol. 2003;185:285–94.

54. Dunin-Horkawicz S, Lupas AN. Comprehensive Analysis of HAMP Domains: Implications for Transmembrane Signal Transduction. J Mol Biol. 2010;397:1156–74.

55. Mascher T, Helmann JD, Unden G. Stimulus Perception in Bacterial Signal-Transducing Histidine Kinases. Microbiol Mol Biol Rev. 2006;70:910–38.

56. Kato M, Mizuno T, Shimizu T, Hakoshima T. Insights into Multistep Phosphorelay from the Crystal Structure of the C-Terminal HPt Domain of ArcB. *Cell*. 1997;88:717–23. doi:10.1016/S0092-8674(00)81914-5.

57. Dutta R, Inouye M. GHKL, an emergent ATPase/kinase superfamily. *Trends Biochem Sci*. 2000;25:24–8.

58. Galperin MY, Chou S-H. Sequence Conservation, Domain Architectures, and Phylogenetic Distribution of the HD-GYP Type c-di-GMP Phosphodiesterases. *J Bacteriol*. 2022;204:e00561-21.

59. Pallegar P, Canuti M, Langille E, Peña-Castillo L, Lang AS. A two-component system acquired by horizontal gene transfer modulates gene transfer and motility via cyclic dimeric GMP. *J Mol Biol*. 2020;432:4840–55.

60. Patterson DC, Ruiz MP, Yoon H, Walker JA, Armache J-P, Yennawar NH, et al. Differential ligand-selective control of opposing enzymatic activities within a bifunctional c-di-GMP enzyme. *Proc Natl Acad Sci.* 2021;118:e2100657118.

61. Soler-Bistué A, Aguilar-Pierlé S, Garcia-Garcerá M, Val M-E, Sismeiro O, Varet H, et al. Macromolecular crowding links ribosomal protein gene dosage to growth rate in *Vibrio cholerae*. *BMC Biol*. 2020;18:43.

62. Soler-Bistué A, Mondotte JA, Bland MJ, Val M-E, Saleh M-C, Mazel D. Genomic Location of the Major Ribosomal Protein Gene Locus Determines *Vibrio cholerae* Global Growth and Infectivity. *PLOS Genet*. 2015;11:e1005156.

63. Helmstetter CE, Cooper S. DNA synthesis during the division cycle of rapidly growing *Escherichia coli* Br. *J Mol Biol.* 1968;31:507–18.

6 CHAPTER 6: Summary and future directions

GTAs are found in multiple bacterial phyla, and one type of GTA appears to be particularly widely conserved in the large and diverse subphylum of the Alphaproteobacteria [1]. As such, they are an important element of some of the bacteria dominating the oceans, like those of the Roseobacter group, the metabolism of which has important consequences, e.g., on the global climate [2]. GTAs can transport host genes between cells and thereby make an important contribution to the evolution and environmental adaptation of species. Therefore, it is of interest to determine under which conditions GTAs are expressed and what factors contribute to their conservation in bacterial genomes. To follow up on these questions I applied computational, comparative analyses to publicly available genomic and transcriptomic datasets.

In the second chapter of this thesis, I used datasets from different studies, which include mutant strains such as knockouts of genes encoding oxygen sensors and related physiological changes to investigate the influence of environmental signals on the expression of known GTA regulators and the GTA gene cluster itself in the two model organisms D. shibae and R. capsulatus. The analysis of these transcriptomic data revealed that the Dnr/Fnr family of regulators, which are able to sense concentrations of oxygen and nitrogen in the environment [3], repress the CtrA phosphorelay genes, which in turn repress three of the four Dnr/Fnr regulators (fnrL, dnrD, and especially dnrF). In fact, direct binding of FnrL to some promoters of the CtrA regulon was found in R. capsulatus. These results suggest that there are direct and indirect interactions between these redox regulators and the CtrA phosphorelay that regulates GTA gene cluster expression. This is consistent with findings by Pallegar et al. [4], who found that oxygen influences the activity of a c-di-GMP-synthesising enzyme that is involved in the regulation of GTAs in R. capsulatus. There are numerous other organisms in which oxygen and nitrogen availability affects physiological traits [3,5]. This study provided the first links between Dnr/Fnr regulators and the CtrA phosphorelay and GTA gene expression and showed their influence on various traits, such as motility, that might be conserved in Alphaproteobacteria. However, due to the inclusion of datasets, that were originally obtained for different purposes the results were not necessarily consistent in terms of activation and inhibition or intensity of regulation. This makes it difficult to predict the actual reactions

that take place in the cells and clearly highlights the importance of combining bioinformatic studies with further laboratory investigations to determine from where these discrepancies originate. Despite these difficulties, bioinformatic analyses allowed me to generate hypotheses and to perform preliminary assessments to better define research questions and identify promising research lines, such as, in this case, how oxygen concentrations affect GTA gene expression.

Another factor that influences gene expression and, ultimately, cell traits, is the localization of genes on the chromosome. Therefore, in the third chapter I investigated the localization of genes related to the CtrA phosphorelay, some of which involved in GTA gene regulation, in various alphaproteobacterial orders. I found different degrees of localization conservation. In the Caulobacterales, the essential CtrA phosphorelay connects replication and cellular differentiation. Here, the ctrA gene is located close to ori and its expression is activated after the replication fork has passed this locus. Thereby, CtrA activity takes place during the on-going cell cycle. By contrast, in the Rhodobacterales, ctrA is not essential, suggesting that CtrA is not embedded in control of the cell cycle. It was found to be preferentially located near ter in this group, which could indicate that the gene is expressed and can become an active protein once replication is finished. Thereby it would be able to activate GTA gene expression when DNA replication is over so that GTA production cannot start before DNA replication is completed. This study indicates that there are selection pressures on many genes such as CtrA to have a preferential localization on the genome.

The needed in vivo follow up, laboratory experiment where ctrA is moved to the proximity of ori in one of these Rhodobacterales members could provide further insights on the importance of the location of CtrA on the chromosome. If CtrA would then be expressed earlier during replication, the produced GTA particles would reflect this by containing a higher concentration of ori-proximate DNA fragments.

In the fourth chapter I examined multiple genomic properties related to GTA gene cluster location. As of now, no consensus has been reached to explain the strong sequence conservation of the alphaproteobacterial GTAs. Using a computational investigation of genomes of five alphaproteobacterial orders, I found that these clusters are located mainly on the leading strand of DNA replication and at high

distances from long repetitive elements and therefore are positioned in genomic regions characterized by lower plasticity, and therefore should evolve at lower evolutionary rates. Moreover, the locations of the GTA gene clusters distinguish themselves by an extreme GC skew – a property that they share with core genes. Especially in the Rhodobacterales, these GC skew regions arise from the preferential use of codons with high GC skew values, i.e., if one amino acid is encoded by two codons with the same GC content, the one with a higher GC skew value is preferably used. While not much is known about the impact of GC skew in prokaryotes, GC-skewed regions in eukaryotes accumulate genes that encode proteins with increased half-lives [6]. Thus, these results could help to determine the importance of GC skew in bacteria in the future, although the hypotheses from this computational analysis should be tested by subsequent long-term evolutionary experiments. For example, two codons encoding the same amino acid and with the same GC content but with different GC skews could be swapped and the effect on the evolutionary rate, mutational rate and protein stability monitored. However, it should be noted that not all properties could be detected in all orders, e.g., the usage of codons with high GC skew was particularly prominent in the Rhodobacterales and the location of the GTA gene clusters varied among the orders. This analysis was also limited by the low number of genomes available for some groups. However, multiple genomic characteristics about GTA gene clusters were revealed that should be considered as relevant factors in their conservation.

In the last research chapter, I investigated in more detail the localization of genes encoding c-di-GMP synthesizing and hydrolyzing enzymes for five alphaproteobacterial orders. These enzymes are involved in the regulation of many important cellular processes such as motility, the cell cycle and GTA production. I investigated their occurrence and chromosomal positions, as well as the presence of secondary domains. The genes encoding GGDEF domains, which are involved in c-di-GMP synthesis, were the most abundant, followed by genes encoding GGDEF and EAL hybrids (GGDEF_EAL). These hybride proteins also had the highest diversity of secondary domains. The phosphodiesterase domain-containing proteins (EAL and HD-GYP) were the least abundant and diverse. On the chromosome, proteins with GGDEF sequences dominated at the ter-proximate half, while those containing the other

domains (GGDEF_EAL, EAL, HD-GYP) were preferentially found at the ori-proximate half in the orders Rhodobacterales and Rhizobiales. This could lead to a degradation of c-di-GMP at the beginning of DNA replication by PDEs followed by a build-up towards the end of the cell cycle by DGCs. An alternative explanation for the distinct chromosomal locations of these genes could be a build-up of a c-di-GMP gradient along the ori-ter axis. More secondary domains that can be used to sense environmental signals or facilitate protein-protein interactions are found in the DGC containing sequences. This could indicate that stimuli are mainly integrated at the synthesis phase of c-di-GMP while the hydrolysis is more unspecific. However, this analysis was limited by the fact that in many cases the entries in the various available databases for genomes, sequences and domain structures do not share the same identifiers and annotations or are inconsistently formatted. In my analysis I used a database in which DNA sequences are sorted based on domain occurrence. However, the identifiers of these sequences were not compatible with the database that contains the associated genome information. Therefore, a third database had to be interposed in order to combine both. This procedure reduced the available dataset since not all sequences or organisms were present in all databases.

6.1 Conclusions

Overall, the results of my work offer new insights into the genetic and gene transcription mechanisms in Alphaproteobacteria associated with GTA gene clusters. This has contributed to a better understanding of the importance of these GTAs and their evolutionary histories in these bacteria. My dissertation identified several approaches that have proven useful to study these important aspects of alphaproteobacterial genetics and biology. However, additional investigations must be carried out to clarify further questions. For example, I found that signals from oxygen and nitrogen concentrations and cell densities control a common gene regulon, but how exactly the networks are connected to each other must be clarified in future investigations. In addition, I was able to show that the localization of genes encoding many important regulators of the GTA gene cluster are strongly conserved on the chromosome and the implications of this distribution need to be investigated *in vivo*. The localization of the GTA gene cluster shows characteristics that are shared with core genes. However, why core and GTA genes are

accumulated in regions with pronounced GC skew still needs to be clarified.

6.2 References

1. Lang, A.S.; Beatty, J.T. Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* **2007**, *15*, 54–62.
2. Gonzalez, J.M.; Simó, R.; Massana, R.; Covert, J.S.; Casamayor, E.O.; Pedrós-Alió, C.; Moran, M.A. Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl. Environ. Microbiol.* **2000**, *66*, 4237–4246.
3. Van Alst, N.E.; Picardo, K.F.; Iglewski, B.H.; Haidaris, C.G. Nitrate sensing and metabolism modulate motility, biofilm formation, and virulence in *Pseudomonas aeruginosa*. *Infect. Immun*. **2007**, *75*, 3780–3790.
4. Pallegar, P.; Peña-Castillo, L.; Langille, E.; Gomelsky, M.; Lang, A.S. C-di-GMP-mediated regulation of gene transfer and motility in *Rhodobacter capsulatus*. *J. Bacteriol.* **2020**, *202*, 1–17.
5. Henares, B.M.; Higgins, K.E.; Boon, E.M. Discovery of a nitric oxide responsive quorum sensing circuit in *Vibrio harveyi*. *ACS Chem. Biol*. **2012**, *7*, 28.
6. Dai, Y.; Holland, P.W.H. The interaction of natural selection and GC skew may drive the fast evolution of a sand rat homeobox gene. *Mol. Biol. Evol.* **2019**, *36*, 1473–1480.
7. Tomasch, J.; Wang, H.; Hall, A.T.K.; Patzelt, D.; Preuße, M.; Brinkmann, H.; Bhuju, S.; Jarek, M.; Geffers, R.; Lang, A.S. Packaging of *Dinoroseobacter shibae* DNA into Gene Transfer Agent particles is not random. *Genome Biol. Evol.* **2018**, *10*, 359–369.
8. Koppenhöfer, S.; Wang, H.; Scharfe, M.; Kaever, V.; Wagner-Döbler, I.; Tomasch, J. Integrated transcriptional regulatory network of quorum Sensing, Replication Control, and SOS Response in *Dinoroseobacter shibae*. *Front. Microbiol*. **2019**, *10*.
9. Hynes, A.P.; Mercer, R.G.; Watton, D.E.; Buckley, C.B.; Lang, A.S. 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.* **2012**, *85*, 314–325.

10. Farrera-Calderon, R.G.; Pallegar, P.; Westbye, A.B.; Wiesmann, C.; Lang, A.S.; Beatty, J.T. The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter capsulatus* gene transfer agent (RcGTA) production is bi-directional and regulated by cyclic-di-GMP. *J. Bacteriol.* **2020**, *203*, e00525-20.
11. Narula, J.; Kuchina, A.; Lee, D.D.; Fujita, M.; Süel, G.M.; Igoshin, O.A. Chromosomal arrangement of phosphorelay genes couples sporulation and DNA replication. *Cell* **2015**, *162*, 328–337.
12. Rodriguez Ayala, F.; Bartolini, M.; Grau, R. The stress-responsive alternative sigma factor SigB of *Bacillus subtilis* and its relatives: An old friend with new functions. *Front. Microbiol*. **2020**, *11*, 1–20.