2015

Extensive Genome Rearrangements of Caulobacter K31 and Genomic Diversity of type B3 Bacteriophages of Caulobacter Crescentus

Kurt Taylor Ash
University of South Carolina

Extensive genome rearrangements of *Caulobacter* K31 and Genomic diversity of type B3 bacteriophages of *Caulobacter Crescentus*

by

Kurt Taylor Ash

Bachelor of Science
University of Oklahoma, 2005

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Biological Sciences

College of Arts and Sciences

University of South Carolina

2015

Accepted by:

Bert Ely, Major Professor

Charles Lovell, Committee Member

Lydia Matesic, Committee Member

Bob Friedman, Committee Member

Elizabeth Wright, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

ABSTRACT

The genus Caulobacter is found in a variety of habitats and is known for its ability to thrive in low-nutrient conditions. K31 is a novel Caulobacter isolate that has the ability to tolerate copper and chlorophenols, and can grow at 48°C with a doubling time of 40 h. K31 contains a 5.5 Mb chromosome that codes for more than 5500 proteins and two large plasmids (234 and 178 kb) that code for 438 additional proteins. A comparison of the K31 genome and the Caulobacter crescentus NA1000 genome, the representative strain and most well studied isolate of the Caulobacter genus, revealed extensive rearrangements of gene order suggesting that the genomes had been randomly scrambled. However, a careful analysis revealed that the distance from the origin of replication was conserved for the majority of the genes and that many of the rearrangements involved inversions that included the origin of replication. On a finer scale, numerous small indels were observed. K31 proteins involved in essential functions shared 80 – 95% amino acid sequence identity with their C. crescentus homologues, while other homologue pairs tended to have lower levels of identity. In addition, the K31 chromosome contains more than 1600 genes with no homologue in NA1000.

The genomes of type B3 bacteriophage of Caulobacter crescentus are among the largest phage genomes thus far deposited into GenBank with sizes over 200 kb. The bacteriophage samples of our collection where first isolated by graduate students of Dr. Ely's lab in 1977 in an project aimed at discovering transducing bacteriophages of

Caulobacter *crescentus* (Johnson 1977). We began our genomic characterization of these bacteriophages in hopes of finding genomic rearrangements as observed in the host NA1000 and possibly a more clear understanding of the phenomenon. However, no major rearrangements were discovered and we changed our direction to be more of an evolutionary analysis of this group of bacteriophages. We introduce six new bacteriophage genomes which were obtained from phage collected from various water systems in the southeastern United States and from tropical locations across the globe. Evolutionary analysis of the genomes reveal a "core genome" which accounts for roughly 1/3 of the bacteriophage genomes and is predominately localized to the head, tail, and lysis gene regions. Despite being isolated from geographically distinct locations, the genomes of these bacteriophages were highly conserved in genome sequence and gene order. Here we present the results of our analysis which identify the insertions, deletions, translocations, and horizontal gene transfer events which are responsible for the genomic diversity of this group of bacteriophages.

PREFACE

"Never discourage anyone…who continually makes progress, no matter how slow."

-Plato.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Caulobacter crescentus is a gram negative member of the class alpha-proteobacter found in oligotrophic, freshwater environments (Poindexter 1964). They can thrive in nutrient-rich environments as well, since they are an important component of aerobic digestion systems at typical sewage treatment plants (Schmidt and Stanier 1965; MacRae and Smit, 1991). A novel characteristic of the genus *Caulobacter* is their asymmetrical cell division (Figure 1.). Each cell begins as a swarmer cell with an intact flagellum. The process of DNA replication does not start until the swarmer cell has lost its flagellum and formed a stalk. This stalk, which serves as extra surface area for nutrient uptake, can adhere to a surface using the holdfast material at its tip. The stalked cell is a mature cell that produces a daughter swarmer cell during each round of cell division (Stove and Stanier 1962). A 16S rRNA gene sequence analysis of bacteria previously defined as *Caulobacter* has revealed a division amongst the freshwater and marine bacterial species (Stahl 1992; Abraham 1999). The marine species were reassigned to the genus Maricaulis and the freshwater bacteria are divided into the *Caulobacter* and Brevundimonas genera which shared rRNA sequence similarity ranging from 93% to 95% (Abraham 1999). Further 16s rRNA analysis of the *Caulobacter* genus reveals the presence of two sub clades, one contains C. *crescentus* and C. segnis while the other contains C. *henricii* and *Caulobacter* sp. K31 (Abraham 1999; Maruyama 2009). *Caulobacter* K31 was discovered in an aquifer that was contaminated with pentachlorophenol. Pentachlorophenol is a wood preservative used by sawmills to

protect the logs from fungus growth. Analysis of the K31 genome revealed that it is closely related to the well-studied *Caulobacter crescentus* NA1000 genome. However, we discovered that gene order was poorly conserved between the two genomes. Therefore, we embarked on a project to compare the two bacteria and their genomes in detail as described in Chapter 1.

As the bacterial genome comparison project developed, we hypothesized that the kinds of genome rearrangements that we observed in the two bacteria may be common in *Caulobacter* bacteriophages as well. Bacteriophages rely upon the host machinery for replication and proliferation, therefore, the bacteriophages could be exposed to the same causative agents of genome rearrangement which has affected the host genome. We began to analyze the bacteriophage genomes for rearrangements. Bacteriophage genomes represent the largest reservoir of unexplored genes (Hatfull 2008). Estimates put the number of bacteriophage in the biosphere to be 10^{31}, which would make bacteriophages the most abundant life form on the planet (Hendrix 2002). The entire population of bacteriophages is believed to be resupplied every few days, which would require, at least, 10^{23} infections/second globally (Sutter 2007). Because of their small size, bacteriophages were the first organisms to have their genomes sequenced (Sanger 1977). The Escherichia coli phage Lambda λ was the first double stranded bacteriophage to be sequenced Sanger (1982) and it is also a member of the virus family *Siphoviridae*. The *Siphoviridae* possess long, non-contractile tails (type B), and they include the most common type of phage that infect *Caulobacter crescentus* (Johnson 1977). *Caulobacter* phages were first isolated in 1965 (Schmidt 1965) and phage φCbK has served as a prototype (Agabian-Keshishian 1970). It has the B3 morphotype with an elongated head

and a long, non-contractile tail. To infect its host, φCbK uses a unique head filament to bind to the flagellum and uses the rotation of the flagellum to move towards the cell body (Guerrero-Ferreira 2011). Subsequently, the long non-contractile tail binds to the pilus channel, which it uses for injection of the phage DNA (Lagenaur 1977; Guerrero-Ferreira 2011). In 1977, Johnson, Wood and Ely (1977) published a study describing a large collection of bacteriophages. Approximately three-fourths of these bacteriophage isolates had the B3 morphotype and appeared to be closely related to φCbK. Since most of these phages were still available, we began an evolutionary comparison of the genomes of these closely related bacteriophages as described in Chapter 2. In contrast to the bacterial genomes, the phage genomes are not subject to rearrangements, but the comparison of the phage genomes has provided insight into how these genomes have evolved.

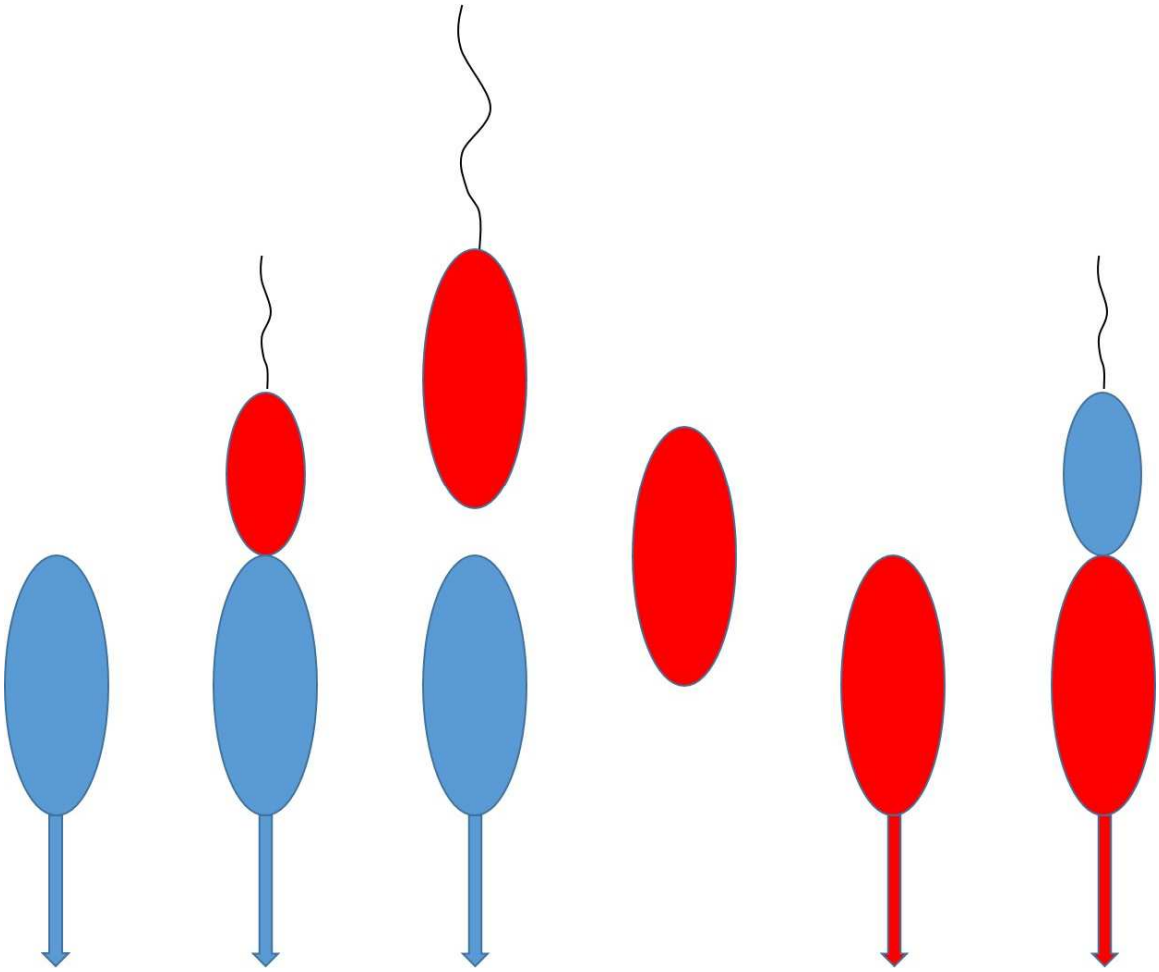

FIGURE 0.1. The cell cycle of *Caulobacter crescentus* CB15. The Caulobacter cell cycle begins in mature stalked cells (bottom row) which produce asymmetrically a swarming daughter cell. As the swarming cell matures, it loses its flagella and transitions into a mature stalked cell where cell cycle begins.

CHAPTER 1

A comparison of the *Caulobacter* NA1000 and K31 genomes reveals extensive genome rearrangements and differences in metabolic potential[1]

[1]Ash, K., T. Brown, T. Watford, L. Scott, C. Stephens, B. Ely (2014). Open Biology **4**(10).

INTRODUCTION

Stalked bacteria from the genus *Caulobacter* are ubiquitous inhabitants of aquatic ecosystems, particularly in oligotrophic habitats in which biologically available concentrations of organic matter and other nutrients are very low. To gain further insight into the physiological diversity of the genus and the evolution of *Caulobacter* genomes, we compared the genomic DNA sequences of two *Caulobacter* species isolated from different habitats: CB15 (also known as NA1000), isolated from a surface freshwater site (Poindexter 1964); and K31, a groundwater isolate of particular interest for its ability to tolerate and degrade chlorophenols (Mannisto 1999). The genomic DNA sequence of *Caulobacter crescentus* CB15 (Neirman 2001, Marks 2010), along with a corrected annotation (Ely 2014), is available in the GenBank database (http://www.ncbi.nlm.nih.gov). *Caulobacter*s are Gram-negative bacteria that are of interest to microbiologists for several reasons. Perhaps foremost, their unique dimorphic life cycle includes motile and non-motile (stalked) cell types, produced during an obligatory cell cycle progression (Stove 1962). The motile cell is an immature cell, which must lose its flagellum and differentiate into a stalked cell before it can replicate its chromosome and divide. This dimorphic *Caulobacter* life cycle may have evolved to allow exploitation of distinct strategies for nutrient acquisition. In addition to their adherent stalked stage, the chemotactic 'swarmer' cells can actively seek nutritionally optimal microhabitats. Swarmer cells are analogous to the G1 phase of the eukaryotic cell cycle in which DNA replication is blocked. Also, stalked cells do not initiate the next round of chromosome replication until after cell division has been completed. Therefore, *Caulobacter*s do not initiate overlapping rounds of chromosome replication.

The ease with which swarmer cells of *C. crescentus* can be isolated in high yield allows investigators to prepare cultures that are synchronous with respect to the cell cycle. In the presence of adequate nutrients, progression through the *Caulobacter* cell cycle is controlled by an internal clock that eventually directs the swarmer to eject its flagellum and progress to the sessile, stalked stage of the life cycle (Stove 1962). This maturation is accompanied by the initiation of DNA replication (analogous to S phase). Subsequent developmental events (flagellum and pili synthesis, cell division) depend on progression and completion of chromosome replication. Aided by the availability of the genome sequence and a suite of genetic techniques (Nierman 2001, Marks 2010, Ely 1991), excellent progress has been made in outlining the gen- etic regulatory network and signal transduction pathway controlling the *C. crescentus* cell cycle (Wang 1993, Holtzendorff 2004, Biondi 2006).

Less effort has gone into applying genomics to under- standing the environmental biology of *Caulobacter*s, which form tightly adherent, stable biofilms on submerged surfaces (Jordan 1975, Entcheva-Dimitrov 2004). Prokaryotes represent a vast reservoir of biomass interacting with the earth's water supplies and bear significant responsibility for the cycling of carbon, nitrogen and phosphorus in the biosphere. Members of the alpha subdivision of Proteobacteria, including the genus *Caulobacter*, are ubiquitous components of the microbiota of virtually every habitat examined. This group is especially prominent in oligotrophic habitats, which include vast tracts of open-ocean and subsurface aquifers. Heterotrophic microbes with high affinities for organic compounds complete the mineralization of carbon in such habitats (Poindexter 1981). However, our understanding of the molecular and genetic adaptations of microbes for oligotrophy is

very limited. The genomic DNA sequence of *C. crescentus* suggests a variety of possible adaptations, including an extraordinarily large set of membrane transport systems (Neirman 2001). An analysis of ribosomal RNA sequences and other data has shown that stalked bacteria previously assigned to the genus *Caulobacter* actually fall into two distinct branches comprising freshwater and marine genera (Stahl 1992, Abraham 1999). The freshwater genera include *Caulobacter* and *Brevundimonas*. A comparison of the *Caulobacter* 16S rDNA sequences indicates two well supported branches: one containing C. *segnis* and various *C. crescentus* isolates, and the other containing *Caulobacter* isolate FWC20 and C. *henricii* (Abraham 1999, Maruyama 2009). Although the genomes of 37 *Caulobacter* isolates are listed in the IMG database (img.jgi.doe.gov), only four are listed as complete, and two of these (CB15 and NA1000) are different versions of the same isolate. In addition, the 'finished' genome of C. segnis contains numerous sequencing errors and the annotation needs to be corrected to include more than 100 genes that are not present in the publically available version of the genome (Patel 2015). The other finished genome, described in this study, belongs to K31, a groundwater isolate that is most closely related to the C. *henricii* lineage, based on a partial 16S rRNA sequence (Mannisto 1999). K31 was poorly characterized, but its abilities to adapt to a low-oxygen groundwater habitat and to tolerate and degrade chlorophenols make it relevant to groundwater bioremediation (Mannisto 1999). As the K31 genome is from a different branch of the *Caulobacter* genus relative to the well-studied C. *crescentus* CB15/NA1000 genome, we chose to compare the two genomes to provide a representation of the genetic diversity of the genus. We also provide a preliminary characterization of the growth and metabolism of the K31 strain that

illustrates some of the adaptation this strain has made to life in a cooler subterranean habitat.

MATERIALS AND METHODS

Media and Growth Conditions

The *Caulobacter* K31 strain was obtained from Minna Mannisto (Arctic Microbiology Research Consortium, Ravoniemi Research Station, Finland), the investigator responsible for its original isolation. When K31 was first isolated, it was grown in PYGV medium, which contains equal amounts of peptone, yeast extract and glucose (0.025% w/v) plus a vitamin solution and 1.5% agar (Mannisto 1999). To determine the optimal growth conditions for K31, we varied the components of the growth medium. We found that riboflavin stimulates growth on minimal media plates (M2G) [18]. Previous experiments with C. *crescentus* strain CB15 showed that it grew better in PYE (0.2% Bacto pep- tone, 0.1% yeast extract, 0.75 mM MgSO4, 0.5 mM CaCl2) [18] in the presence of both glucose and glutamate than with either addition alone (Ely laboratory 1977, unpublished data). In similar experiments with K31, growth with PYE plus glucose caused the pH of the culture fluid to go down and growth with PYE plus glutamate caused the pH to go up. When glucose and glutamate were added together, the pH stayed between 7 and 8. Thus, the growth conditions for optimal yield of K31 are PYE supplemented with 10 mM glucose, 30 mM monosodium L-glutamate and 10 mM riboflavin (PYEGGR). The growth rate is essentially the same in PYE with no additions; however, the final yield was increased about threefold when the additives were present. In contrast to NA1000, which grows at maximum growth rate at 33 –35°C

[18], K31 failed to grow at 35°C. At 30°C, K31 had a doubling time of 160 min, significantly slower than the 110 min doubling time of NA1000 at 30°C. Our experiments also revealed the ability for K31 to grow at 4°C with a doubling time of about 40 h. The ability to grow at low temperatures is consistent with the fact that K31 was isolated from cold groundwater (Mannisto 1999). To determine growth in the presence of copper, a 60 mg/ml of copper sulfate stock solution was diluted to provide a range of concentrations, and 100 ml of each dilution was added to 10 ml of a K31 or NA1000 culture in PYEGGR. After growth at 30°C for 48 h, the optical density of each culture was measured using a Klett-Summerson colorimeter.

Genome sequence determination and annotation

Genomic DNA from *Caulobacter* strain K31 was isolated from a saturated culture using the Qiagen Genomic-tip system, following the manufacturer's protocol. Subsequent library construction and nucleotide sequencing were carried out at the Joint Genome Institute (JGI) of the US Department of Energy. Genomic DNA was sheared to generate the appropriate DNA size fragments (3, 8 and 40 kb) and was verified by gel electrophoresis. The sheared DNA was repaired to produce blunt-end fragments and was subsequently ligated into a series of bacterial plasmids: pUC18 for 3 kb inserts, pMCL200 for 8 kb inserts and pCC1Fos for 40 kb inserts. Plasmids were transformed by electroporation into Escherichia coli and plated on selective media, then picked and archived. To generate sequencing templates, plasmid DNA was isolated from cells and subjected to rolling-circle amplification (RCA). RCA products were subsequently used for Sanger sequencing and capillary gel electrophoresis. Assembly of the genome entailed alignment of sequences from all three libraries using PHRAP. After quality

assessment of the draft assembly, the sequence was automatically annotated by JGI using their standard annotation system. The GenBank [/EMBL/DDBJ] accession numbers for the genome sequence of *Caulobacter* strain K31 are CP000927, CP000928 and CP00029.

Genome Comparisons

For genome comparisons, the K31 genome was compared to the C. *crescentus* NA1000 genome, which is thought to be most closely related to the original isolate of the CB15 laboratory strain (Marks 2010). The two genomes were aligned and compared using the program PROGRESSIVEMAUVE (Darling 2002). Large inversions and translocation were estimated by a manual count of the aligned genomes. Annotations of the two genomes were viewed and analyzed in ARTEMIS (Rutherford 2000). A BLAST comparison to identify homologous genes in the NA1000 and K31 genomes was performed using the protein-coding sequences of all the CDS regions of the chromosomes for each bacterium. The comparisons were conducted using the BLASTSTATION 2 Windows-based software (http://www.blaststation.com). BLAST matches with an e-value that was less than 10^{-5} were considered significant.

Results and Discussion

Genome Overview

The *Caulobacter* K31 genome consists of a 5,477,872 base pair (bp) chromosome and two plasmids. The larger plasmid contains 233,649 bp, and the smaller contains 177,878 bp. The K31 chromosome has a 68.1% GC content and contains 5061 genes (table 1.1.). The accuracy of the genome assembly was confirmed by pulse field gel electrophoresis of SpeI-digested K31 genomic DNA (figure 1.1.). There was a one to

one correspondence between the bands observed on the gel and those predicted from the nucleotide sequence of the main chromosome. However, one extra 233 kbp band was present in the gel that corresponded to the predicted band from cleavage of pCAUL01 DNA at the single SpeI site predicted for pCAUL01. The pCaul02 plasmid is also predicted to have a single SpeI site that would generate a 178 kbp fragment that would co-migrate with a restriction fragment from the main chromosome. Therefore, to verify the presence of the pCAUL02 plasmid, we digested total K31 DNA with SnaBI, which also cut pCAUL02 DNA once. In this case, the resulting 178 kbp fragment was clearly visible on the PFGE gel because digestion of the main chromosome does not produce any fragments that are close in size. Thus, we were able to verify the presence of both plasmids using the PFGE analyses.

The K31 chromosome contains more than 1900 genes that do not appear to have a homologue in NA1000. Conversely, the NA1000 chromosome contains more than 200 genes that do not appear to have a homologue in K31. As with the NA1000 chromosome, the K31 chromosome has two identical ribosomal RNA operons. However, the K31 rRNA operons are adjacent to each other, whereas the NA1000 rRNA operons are separated by more than 90,000 bp. In contrast to the 97% sequence identity of the K31 and NA1000 rRNAs, the ribosomal internal transcribed spacer (ITS) regions of the two species are only 80% identical. Thus, considerable divergence has occurred between the K31 and NA1000 ITS regions even though the two copies of the ITS region are identical within each species.

When the COG (Cluster of orthologous groups) categories of the K31 genes were compared with those of NA1000, the K31 gene counts were higher in most categories

due to its larger genome, but genome percentages were similar for most categories (Table 1.2.). Exceptions were translation, motility and nucleotide transport, and metabolism categories where additional genes are not likely to serve a useful purpose. Examples where K31 has more genes include 14 additional sigma factors, more than 150 additional regulatory proteins, 19 additional response regulators, 32 additional TonB-like receptors, 22 additional major facilitator transporters, nine additional alcohol dehydrogenases, four additional ferredoxins and four additional cytochrome c class 1 genes. All of these genes code for proteins that are more similar to proteins produced in a variety of other bacteria than to any protein produced by NA1000. It remains to be determined whether NA1000 has lost these genes or K31 has gained these genes from other sources by horizontal gene transfer. It is more likely that some combination of the two processes has occurred. An important clue is that K31 has about the same number of transposase genes relative to the size of the genome as NA1000 does, but it only shares about one-third of them with NA1000. This comparison suggests that most of these transposases have been introduced from other species of bacteria and that horizontal gene transfer has played a major role in generating the differences in gene composition observed between these two species.

One example of horizontal gene transfer may be the two K31 plasmids that code for an additional 438 proteins. The larger plasmid (pCAUL01) contains genes for conjugal transfer, and includes a collection of transporter genes and genes for the regulation of transcription. In contrast to the smaller plasmid and the main chromosome, the large plasmid contains only two adjacent transposase genes. It also contains a gene, Caul_5182, that codes for an integration host factor (IHF) protein that is 88 – 90% identical to the IHF proteins produced by two homologous chromosomal genes

(CAUL_1806 and CAUL_2219). By contrast, NA1000 has a single homologous gene (CCNA_2416) that codes for the IHF protein. No other genes on pCAUL01 are homologous to any NA1000 gene. The smaller of the two plasmids (pCAUL02) contains genes for conjugal transfer (CAUL_5365 – 5374) that are present in the same order as those in pCAUL01 (CAUL_5213– 5223), with amino acid identities ranging from 47 to 60%. Thus, the two sets of conjugal transfer genes appear to be distantly related. In addition, genes Caul5296– 5310 comprise a region of pCAUL02 that codes for proteins responsible for the degradation of linear alkylbenzenesulfonate and is homologous to the corresponding genes in the alpha-proteobacterium *Parvibaculum lavamentivorans* (60–80% amino acid identity) with only one gene of the cluster falling below the percentage identity range shared by the rest of the genes. Thus, this set of degradative genes appears to have been acquired from a species of *Parvibaculum* or from some closely related genus. The pCAUL02 plasmid also appears to have experienced a number of transposition events as it contains 17 transposase genes, two of which have 95% amino acid identity to orfA and B of NA1000 IS511. However, they are located in two different regions of the plasmid and only 25% of the orfA gene is present. Only two of the other transposase genes code for proteins that are homologous to other NA1000 transposases. Two additional genes, Caul_5396 coding for a TonB-like receptor and Caul_5432 coding for an acyl carrier protein, are the only other pCAUL02 genes that code for proteins with more that 70% amino acid identity to a NA1000 protein. pCAUL02 also contains two anti-restriction genes that flank three phage integrase genes. These phage integrase genes are 78 – 91% identical to the corresponding genes from plasmid pACRY403 from *Acidiphilium cryptum* JF-5.

Phage genes are present in the main chromosome of K31 as well. A 12 kb region of phage genes is located adjacent to a gene that has 41% identity to dnaJ in both the K31 chromosome (beginning at nucleotide position 4220530) and the NA1000 chromosome (beginning at nucleotide position 3013149). As there is 66% nucleotide identity between the two regions and 76% nucleotide identity between the two dnaJ-like genes, these results suggest that the phage genes were present in the common ancestor of the two *Caulobacter* species. One difference between the two regions is a 400 base insertion in NA1000 (CCNA_2873) that disrupts a phage gene and codes for a bleomycin resistance protein that has 86% amino acid identity with a homologue found in Cystobacter fuscus. When the remaining homologous phage genes were compared, the amino acid identity ranged from 57% for a membrane protein of unknown function to 88% for the major capsid protein. For comparison, an 11 kb flagellar gene cluster that is required for motility has a 71% nucleotide identity when the two genomes are compared and the amino acid identity of the predicted proteins ranges from 59 to 93%. Thus, there appears to be a comparable level of nucleotide diversity and gene conservation between the phage and flagellar gene clusters. Sequence conservation in the flagellar gene cluster would be maintained by selection for motility, but it is not clear why the phage region should be conserved. Many of the phage genes code for structural proteins, but no phage particles have been observed in cultures of either strain.

Additional phage genes are found at four other locations in the K31 chromosome. One set of phage genes located in a 14 kb region beginning at nucleotide position 1648526 (CAUL_1559 – 1573) corresponds to a series of homologous genes found in the same order in Bradyrhizobium and Chloroflexus. A second set of phage genes located in

a 15 kb region beginning at nucleotide position 1935379 and proceeding on the reverse strand (CAUL_1824 – 1812) corresponds to a series of homologous genes found in the same order in *Pelobacter*. A third region contains phage genes scattered throughout a 40 kb region with lower GC content (CAUL_3468 – 3507) and the fourth region contains three phage tail collar protein genes (CAUL_1097 – 1099) followed by nine transposase genes (CAUL_1101 – 1109). All four of these K31 phage regions have no homologues on the NA1000 chromosome suggesting that they were acquired by some type of horizontal gene transfer.

Codon Usage

As the K31 chromosome contains a 68.1% genomic G+C content, the codons in the protein-coding regions should have a high G+C content, especially in the third codon position (GC3). In fact, 24 of the 25 most used codons contain either a G or a C in the third position. The overall GC3 percentage for K31 codons is 88.4%, slightly higher than the 86.9% observed in the NA1000 genome. The K31 codon usage table is similar to that of NA1000 as well, although K31 only has 49 tRNA genes compared with 51 in NA1000. Relative to NA1000, K31 has one copy instead of two copies of an asparagine tRNA, four copies instead of five copies of methionine tRNA, and three copies instead of two copies of an aspartate tRNA with a GTC anticodon. In addition, NA1000 has a leucine tRNA with a TAA anticodon that is not present in K31. The remaining tRNAs are present in both strains and have identical anticodons.

The GC content of pCAUL01 is 67.3% and the pattern of codon usage is similar to that of the K31 chromosome. By contrast, pCAUL02 has increased use of 29 of the 30

amino acid coding codons that end in A or U, consistent with its 64.3% GC content. The exceptional U-ending codon is UAU, where no increase is observed because UAU and UAC are used with almost equal frequency to code for tyro- sine in both the plasmid and the chromosomal genes. This nearly equal use of the two tyrosine codons occurs in most *Alphaproteobacteria* with GC-rich genomes as well, and is an exception to the observation that G- and C-ending codons are used preferentially in these high GC chromosomes. In 19 of the 29 codons with increased use, the frequency of use is more than double that found in the main chromosome. Thus, a difference of less than 4% in average GC content corresponds to a major difference in the use of A- or U-ending codons, with third position GC content dropping from 88.2% in the main chromosome to 79% in pCaul02.

Genome Rearrangements

When the genomes of closely related species are compared, they are usually collinear except where inversions have occurred. For example, Beare et al. (Beare 2009) identified 40 break- points in pairwise comparisons of four Coxiella strains and observed that 75% were within 100 bp of an insertion sequence. Similarly, Darling et al. (Darling 2008) identified 79 inversions in a comparison of eight strains of *Yersinia pestis*, and all of the breakpoints were adjacent to an rRNA operon or within 1500 bp of an insertion sequence. In a third study (Maruyama 2009), a comparison of two Streptococcus mutant strains revealed a single inversion. However, when 95 additional clinical isolates were examined, numerous other inversions were observed. By contrast, when the K31 chromosome was aligned to that of C. *crescentus* NA1000, more than 60 inversions and 45 large translocations were readily observed around the ori site (figure 1.2.). This level

of genomic rearrangements is more than an order of magnitude greater than the examples described above and leads to a well-scrambled genome. Although most inversions flank the origin of replication as observed in the genome comparisons described above, many others do not and instead lead to more local rearrangements.

Several constraints on genome reorganization have been proposed, including a conserved distance from the origin of replication for genes involved in transcription and translation due to copy number differences in fast growing bacteria (Couturier 2006). To determine whether this phenomenon was limited to fast-growing bacteria, we determined the relative positions of the first gene in each conserved sequence block in the K31 versus NA1000 comparison depicted in figure 1.2. Despite the extensive rearrangements present in the genome comparison, we found that most genes are in approximately the same place in both genomes or they are in the same place relative to the origin of replication but are on opposite sides of the origin (figure 1.3.). Thus, the position of a gene relative to the origin of replication appears to be conserved in a genus where multiple rounds of replication do not occur, and gene copy number differences do not exceed a factor of two. Furthermore, although it is obvious from figure 2 that the breakpoints of most chromosomal rearrangements do flank the origin of replication, the remainder must be limited to small changes in gene position or to compensating rearrangements to restore the distance between blocks of genes and the origin. Also, as we compared the position of genes in more than 100 conserved sequence blocks, genome position relative to the origin of replication must be important for a relatively large number of genes. One explanation for this phenomenon could be that DNA replication results in hemi-methylated GANTC sites in regions that are responsible for cell-cycle-

dependent gene expression. For example, the ctrA gene is located a large distance from the origin of replication and has increased expression when hemi-methylated (Reisenauer 2002). As CtrA is one of the master cell cycle regulators, this change in CtrA levels is necessary for normal cell growth and progression through the S-phase of the cell cycle. Thus, there may be a cascade of effects as chromosome replication progresses and additional gene promoters become hemi-methylated. These results also suggest that rearrangements that impact the distance of critical genes from the origin of replication are not tolerated so that they are lethal events or that they are immediately followed by a second rearrangement that restores the distance from the origin for those genes.

In addition to the rearrangements of major blocks of genes, we also observed eight instances where two or three translation related genes were in different locations in the two genomes with different flanking genes. Thus numerous small translocations seem to have occurred. One inversion that we examined in detail involved an inversion with breakpoints close to the origin of replication. The inverted segment was asymmetric, with less than 5 kb on one side of the origin and more than 54 kb on the other side. One of the breakpoints was adjacent to the dnaA gene so that it is 50 kb closer to the origin in NA1000 than in K31. The other breakpoint was between the genes coding for exodeoxyribonuclease III and the DNA polymerase III epsilon subunit in NA1000 such that the epsilon subunit is 50 kb farther from the origin of replication in K31. The inverted region differs in size by nearly 2 kb in the two species because five separate genes are present in K31 but not in NA1000, and one gene is missing from K31 that is present in NA1000. Thus the inverted region contains at least five small indels that occurred in one of the two genomes.

Another example of the complexity of *Caulobacter* genome rearrangements comes from the three genes for glutamyl-tRNA(Gln) amidotransferase, which comprise a single gatCAB operon in NA1000. In K31, the gatA and gatB genes are expressed from different promoters and are separated by two genes that are not found in the NA1000 genome (figure 1.4.). When other related genomes were compared, gatA and gatB were separated by one of the two K31 genes (designated X in figure 1.4.) in C. segnis TK0059, *C. crescentus* OR37, *Caulobacter* sp. AP07 genomes. However, they are separated by a third gene (designated Z in figure 1.4.) in the chromosomes of two species from closely related genera*, Brevundimonas* subvibrioides and *Phenylobacterium zucineum*. This third gene is found elsewhere in the NA1000 and K31 genomes with a relatively low level of amino acid identity. As K31 and AP07 are on the same branch of the tree, these data suggest that the ancestral *Caulobacter* genome contained gene X. If this hypothesis is true then gene Y must have been inserted into the K31 genome. The single *gatCAB* operon observed in the CB15 genome is found in the CB4 genome as well (D. Scott and B. Ely 2013, unpublished data). As CB4 is closely related to K31, it is likely that gene X has been lost independently in both NA1000 and CB4.

In contrast to the changes described above, it was striking that a conserved 14.5 kb block of 28 genes including 24 ribosomal protein genes remained intact at approximately the same place in the two genomes. This highly conserved block of genes was transcribed in a single direction and had 87% nucleotide identity plus small insertions or deletions that corresponded to approximately 1% of the total region. Part of the reason for this high level of conservation may be that the first 20 ribosomal genes

appear to be transcribed as a single operon. Thus, most disruptions of the operon would be lethal due to the failure to express some of these ribosomal protein genes.

Mobile genetic elements are often involved in genome rearrangements (Darmon 2014). Therefore, we examined the position of the transposase and integrase genes in the NA1000 (n = 50) and K31 (n = 90) genomes. In the NA1000 and K31 genomes, roughly one-third of the transposases and integrases interrupt regions of homology, another third are located in short non-homologous regions, and the remaining third are associated with a combined insertion and translocation. In most cases, additional genes that are not present in the other genome are adjacent to the transposases. Thus, in addition to mediating gene insertions, transposases may be associated with about one-third of the large translocations observed in the NA1000 and K31 genome comparison. These results indicate that although transposases are involved in some of the genome rearrangements, the majority of the observed rearrangements in the *Caulobacter* genomes must be due to some other mechanism.

Correlation between Genotype and Phenotype

With respect to the cell cycle, K31 grown in PYEGGR produces cells that are morphologically similar to those of C. *crescentus*. Both strains attach to surfaces and form biofilms. Comparison of the genes that code for cell cycle/developmental regulators and known components of developmentally regulated structures (flagellum, holdfast, pili and cytoskeletal proteins involved in division) reveals generally high (75–95%) amino acid sequence identity with C. *crescentus*, suggesting that the processes contributing to morphological development have been evolutionarily conserved. For

example, the 'master regulator' CtrA that controls many key cell cycle events is one of the most highly conserved proteins, with 97% amino acid identity to the corresponding NA1000 protein. The ctrA gene nucleotide sequence is also highly conserved (92%), suggesting that there is selection for codon usage as well as for conservation of amino acid sequence. In fact, both the ctrA amino acid and nucleotide sequences are highly conserved among the *Alphaproteobacteria*, with greater than 77% identity in both measures for the top 100 matches to the K31 gene in a BLAST search. By contrast, one interesting difference between the genomes is that three of the flagellin genes, *fljMNO*, are contiguous in NA1000, but they are located at three widely separated positions in the K31 genome. In the NA1000 annotation, these three genes are included in three separate mobile elements (Marks 2010). If these genes are truly contained in mobile elements, it would provide an explanation for their disparate locations in the K31 genome. The genome comparison between NA1000 and K31 also revealed the presence of additional copper resistance genes in the genome of K31. K31 has a total of six copper resistance genes in two clusters. One cluster contains copper resistance protein genes CAUL_2631 *copA* and Caul_2630 *copB*, and a similar cluster is found in NA1000. The second contains *copABCD* (CAUL_2346 – 47 and CAUL_2350 – 51) with no corresponding region in NA1000. Thus, the K31 genome contains an additional four-gene *copABCD* system that could confer a copper resistance phenotype. To determine whether K31 was actually more resistant to copper inhibition, we performed a series of growth experiments in the presence of a range of copper concentrations. After testing various dilutions of a 60 mg/mL stock solution of CuSO4, we discovered that 0.5 mg/mL was the highest level of copper that allowed K31 growth, and 0.1 mg/mL was the highest level that allowed

NA1000 growth. Thus, K31 is resistant to a five-fold higher concentration of copper than the level that NA1000 tolerates. As indicated above, experimental observations suggested that K31 grew better in the presence of riboflavin. Riboflavin, also known as vitamin B2, is synthesized in many plants and microorganisms, and is used to make FAD and FMN, two key cofactors in many enzymatic reactions. NA1000 contains the five enzymes necessary for the biosynthesis of riboflavin and FAD from GTP. The five enzymes and their various sub- units are encoded in five different genes: *ribAB, ribD, ribH, ribE and ribF*. Four of these genes, *ribD, ribE, ribAB* and *ribH*, are encoded in a single operon. This operon seems to be con- served in other closely related *Alphaproteobacteria* such as *Hyphomonas* and *Maricaulis maris*. However, the entire four- gene operon along with nine additional contiguous genes is absent from the K31 genome. The fifth gene, *ribF*, is located 200 kb from the riboflavin operon in C. *crescentus* and codes for a *RibF* protein that has riboflavin kinase and FAD synthetase activities. It is present in K31 as gene Caul_4052 and codes for a *RibF* protein that has 84% amino acid identity with the *C. crescentus* protein. Another gene annotated as a duplicate *ribH* gene (Caul_1421) is located 500 kb from the riboflavin operon in the *C. crescentus* genome and is also present in K31 as Caul_3045. In *C. crescentus*, the two *ribH* genes code for proteins that have little amino acid sequence similarity, but they are both identified as the beta subunit for riboflavin synthase. However, as K31 does not contain the gene for the alpha subunit, it is unlikely that the K31 *RibH* enzyme is involved in riboflavin synthesis.

Some freshwater *Caulobacter* isolates have been shown to have properties that are potentially useful for aquatic bioremediation applications, including the

degradation of aromatic compounds (Chaterjee 1987) and the reduction of arsenic (Macur 2001). The chlorophenol degradation capacity of *Caulobacter* strain K31 stands out among these. K31 was isolated in an enrichment culture for chlorophenol-tolerant bacteria in groundwater from Karkola, Finland (Mannisto 1999). The groundwater at Karkola percolates through nutrient-poor silt and clay, and is cold (7 –8°C), oxygen-deficient, iron-rich and mildly acidic (pH 6– 6.5). The aquifer had been contaminated with high levels of chlorophenols for at least two decades by a local sawmill that used polychlorophenol as a fungicide for lumber treatment. Chlorophenols have historically been used as fungicides in agricultural and industrial applications (e.g. wood preservatives applied at lumber mills), and also enter the environment as by-products of paper mills. K31 inhabited groundwater containing up to 190 mg/L total chlorophenols and was grown in the laboratory on complex media supplemented with 250 mg/L chlorophenols: 2,4,6-trichlorophenol (TCP), 2,3,4,6-tetrachlorophenol (TeCP) and pentachlorophenol (PCP). In limited testing, K31 was found to degrade TeCP, and to tolerate PCP at relatively high levels, though no evidence was found for degradation of PCP. K31 does not contain the *pcpB* gene (encoding PCP-4-monooxygenase) present in several chlorophenol-degrading *Sphingomonas* isolates from the same site. However, the K31 genome does contain several other genes that code for a variety of monoxygenases. In addition, it has genes for tert butyl ether degradation, alkane metabolism and a tannase/feruloyl esterase that are not found in the NA1000 genome.

In terms of physiology, the K31 genome appears to encode a much larger repertoire of enzymes involved in electron transfer reactions than C. *crescentus*, including more cytochrome c variants, respiratory nitrate reductase and many

dehydrogenases that are not present in the C. *crescentus* genome. As a ground-water resident growing slowly at low temperatures, K31 probably faces lower dissolved oxygen levels (if not outright anoxia) on a more routine basis than the surface-dwelling bacteria, and may find more respiratory versatility advantageous.

CONCLUSION

In summary, K31 is an interesting *Caulobacter* isolate that has the ability to tolerate copper and chlorophenols, and can grow at consistently low temperatures. The K31 chromosome appears to have been scrambled relative to that of C. *crescentus* NA1000. However, the positions of most genes relative to the distance from the origin of replication are unchanged, indicating that genome rearrangements are constrained. This genome scrambling also makes it difficult to identify individual chromosome rearrangement events. However, it is clear that additional mechanisms must be involved as transposases seem to be associated with only one-third of the observed events. In addition to the genome rearrangements, the K31 chromosome includes numerous insertions and deletions relative to the NA1000 chromosome, so that it contains 1200 more genes, plus an additional 400 genes are present on two very large plasmids. These extra genes provide K31 with increased metabolic versatility.

.

Table 1.1. A comparison of Caulobacter strains NA1000 and K31

Caulobacter isolate	NA1000	K31
source	pond water	groundwater sample
cell shape	crescent	crescent
genome features		
base pairs	4.02×10^6	5.48×10^6
plasmids	0	2
G/C content	67.2%	67.4%
protein-coding genes	3876	5443
tRNA genes	51	49
rRNA genes	6	6
5s	2	2
16s	2	2
23s	2	2
flagellar genes	43	44
phage genes	15	42
transposases	40	61
integrases	4	33
recombinases	2	7
pseudogenes	1	17

Table 1.2. K31 and NA1000 gene counts by COG category. Source: img.jgi.doe.gov.

COG category	K31 gene count	% of total $n = 4109$	NA1000 gene count	% of total $n = 3037$
amino acid transport and metabolism	260	6.33	222	7.31
carbohydrate transport and metabolism	219	5.33	157	5.17
cell cycle control, cell division and chromosome partitioning	33	0.80	23	0.76
cell motility	71	1.73	66	2.17
cell wall/membrane/envelope biogenesis	244	5.94	195	6.42
chromatin structure and dynamics	3	0.07	2	0.07
coenzyme transport and metabolism	132	3.21	114	3.75
defense mechanisms	71	1.73	45	1.48
energy production and conversion	231	5.62	165	5.43
function unknown	416	10.12	336	11.06
general function prediction only	483	11.75	344	11.33
inorganic ion transport and metabolism	220	5.35	153	5.04
intracellular trafficking, secretion and vesicular transport	137	3.33	83	2.73
lipid transport and metabolism	269	6.55	171	5.63
nucleotide transport and metabolism	75	1.83	71	2.34
post-translational modification, protein turnover and chaperones	155	3.77	130	4.28
replication, recombination and repair	222	5.40	127	4.18
secondary metabolites biosynthesis, transport and catabolism	163	3.97	100	3.29
signal transduction mechanisms	190	4.62	155	5.10
transcription	346	8.42	212	6.98
translation, ribosomal structure and biogenesis	169	4.11	166	5.47
not in COGs	1830	33.28	1178	29.95

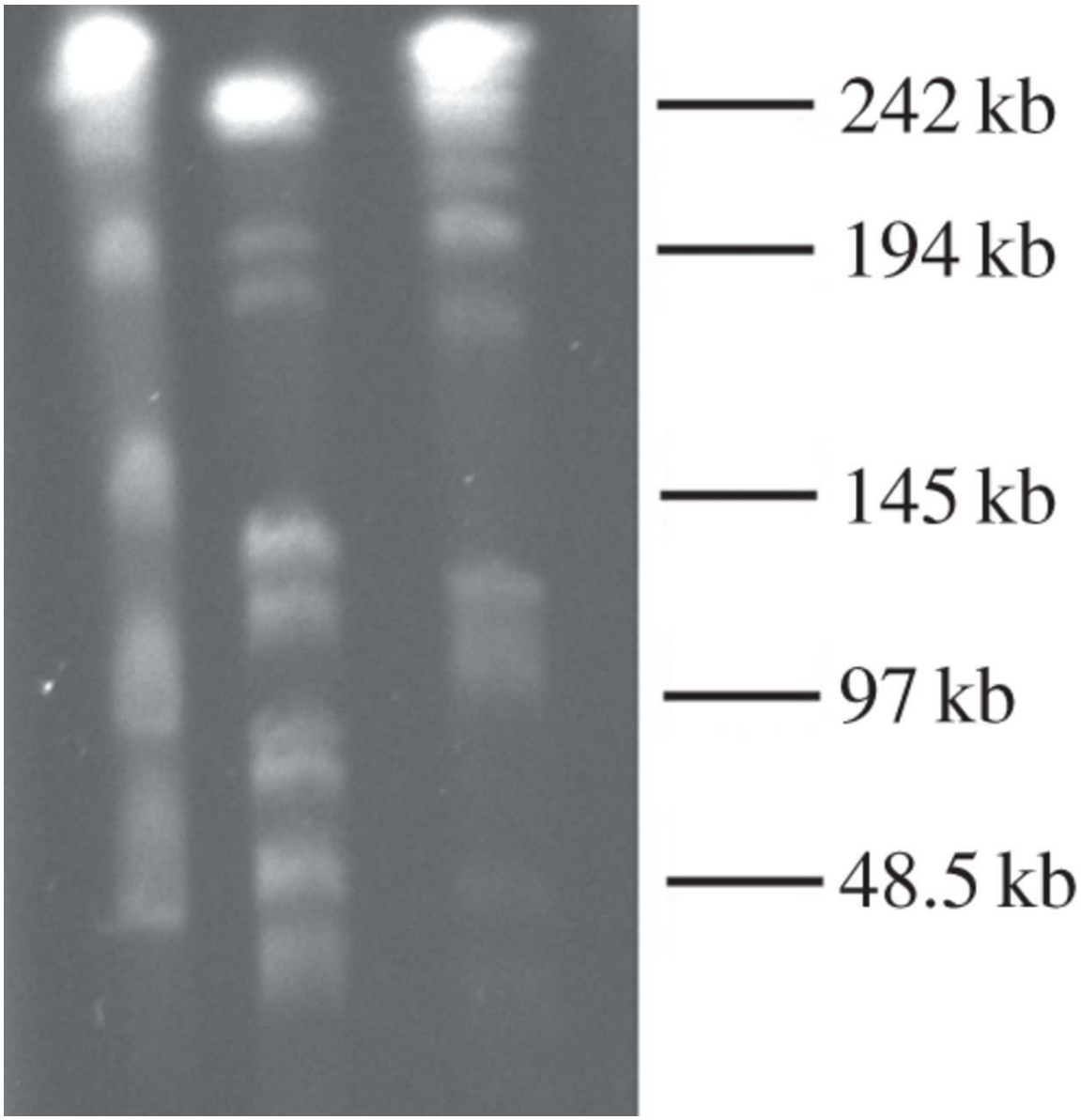

Figure 1.1. Pulse field gel electrophoresis of AseI and SpeI-digested DNA. Lane 1, lambda size ladder with sizes indicated to the right of the gel. Lane 2, AseI digest of K31 DNA. Lane 3, SpeI digest of K31 DNA.

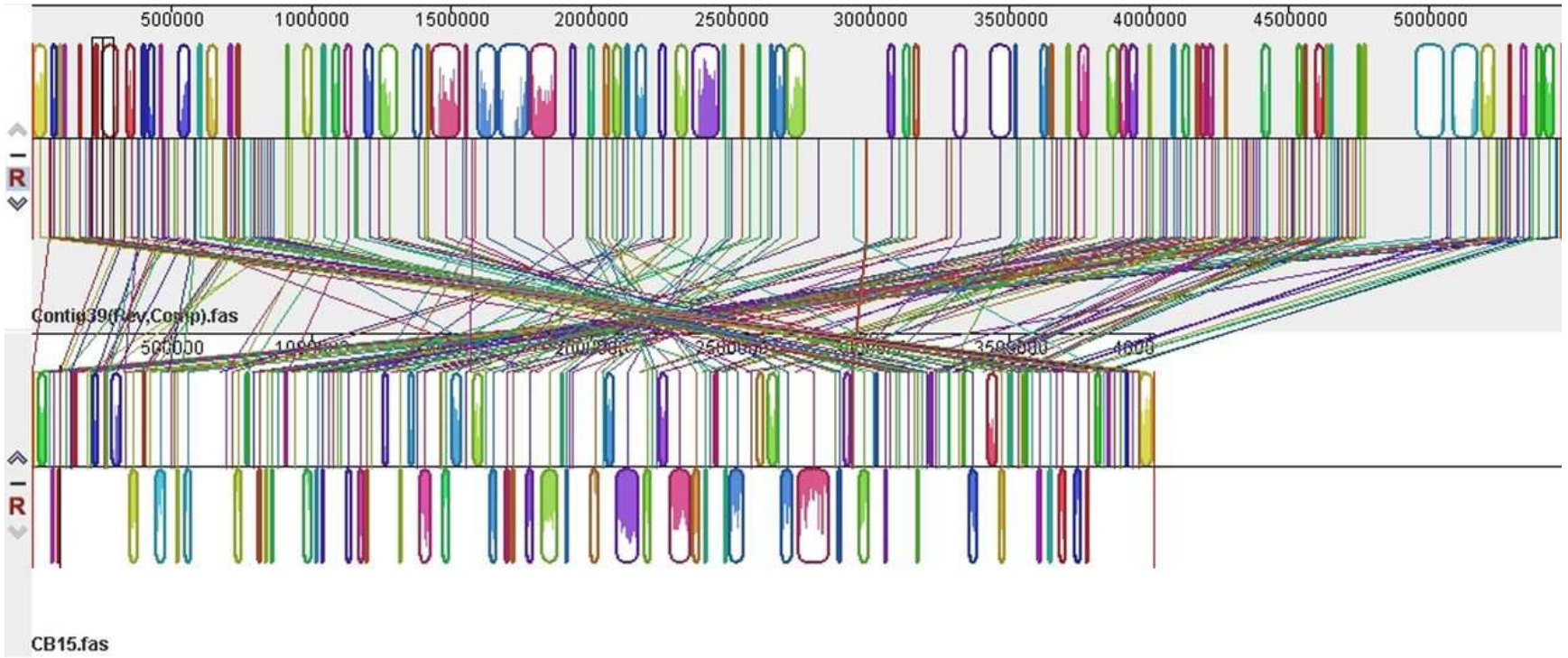

Figure 1.2. An alignment of the K31 chromosome with the C. crescentus NA1000 chromosome showing more than 60 inversions and 45 large translocations. Regions of contiguous homology have the same color and are connected by a line of that color.

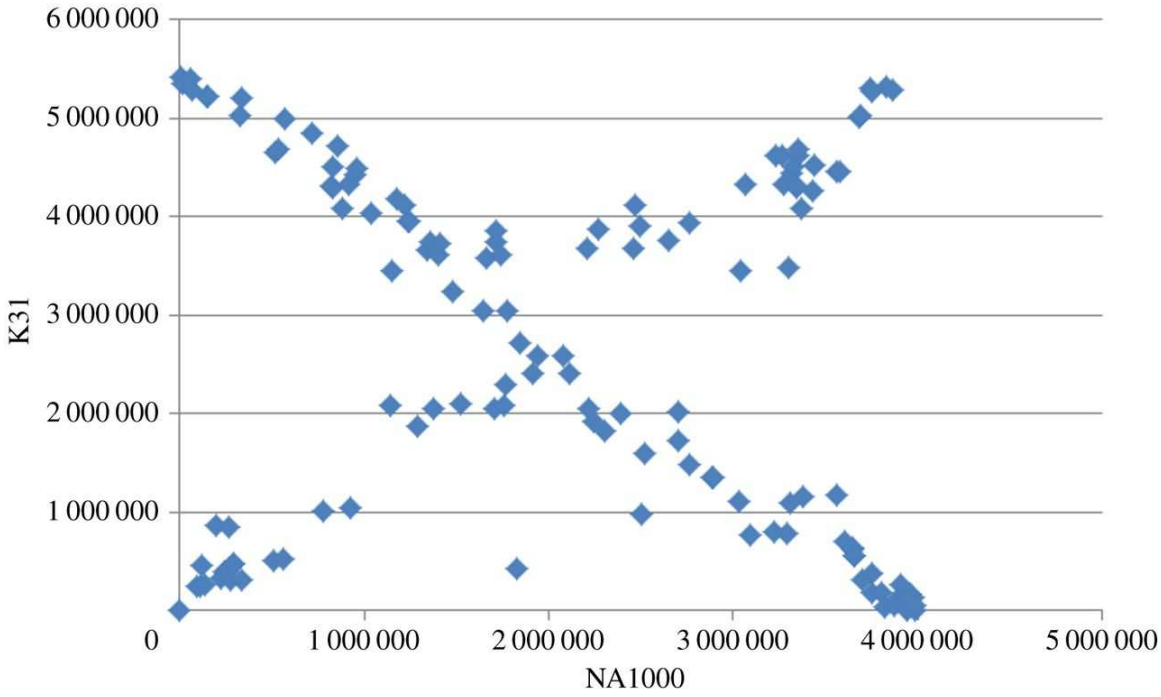

Figure 1.3. The relative positions of homologous genes in the K31 and the NA1000 chromosomes. The location in each chromosome of the first gene in each homologous block of genes identified in figure 2 was plotted. The origin of replication is located close to position 0 in both genomes.

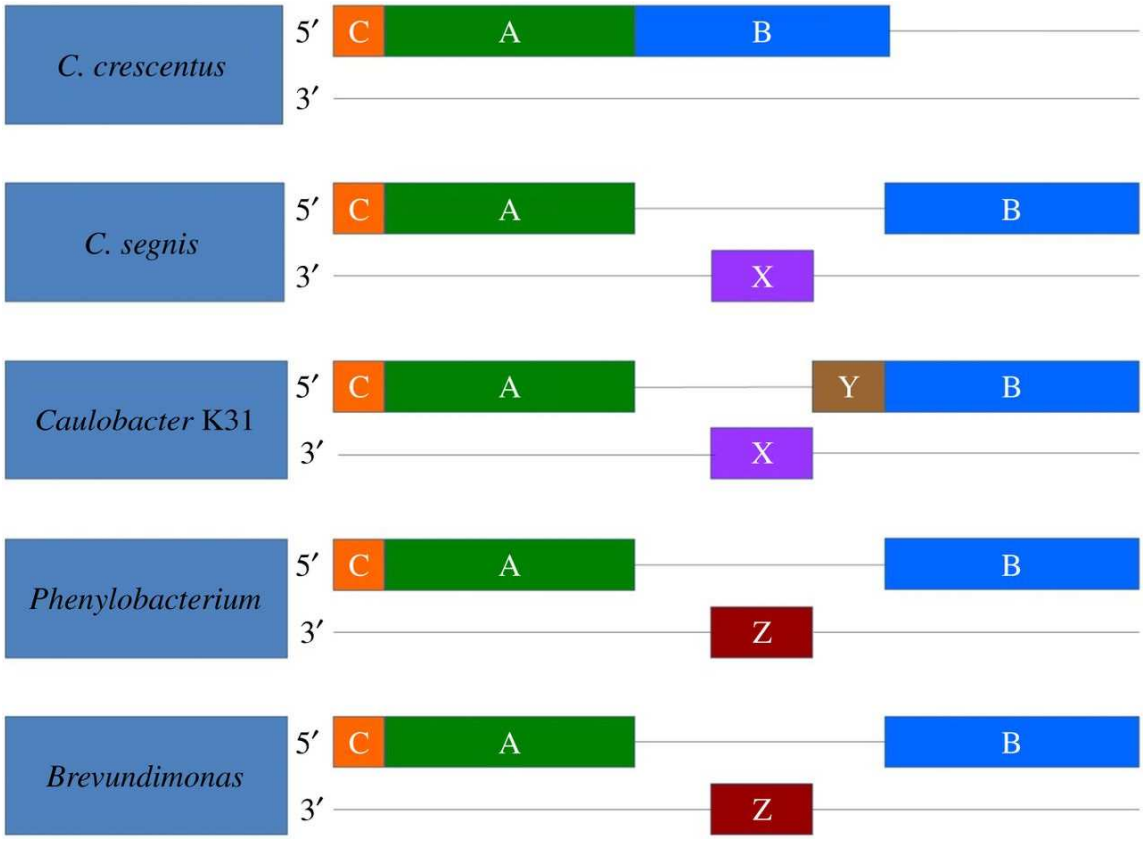

Figure 1.4. The gene arrangement in the gatCAB operon of NA1000, K31 and related bacteria. X, Y and Z represent genes that interrupt the operon that are not present in the NA1000 chromosome.

CHAPTER 2

Genomic diversity of type B3 bacteriophages of *Caulobacter crescentus*

Ash, K. and B. Ely. 2015. To Be Submitted to Current Microbiology

INTRODUCTION

Although they are still underrepresented in the GenBank database, bacteriophages are the most abundant organisms on the planet. Currently there are only ~1300 bacteriophage genome sequences in deposited in GenBank. The majority of the sequences deposited so far are of the *Siphoviridae*. This family is noted for having a non-enveloped head and a non-contractile tail. The majority of the known bacteriophages that infect *Caulobacter crescentus* are type B3 *Siphoviridae* which have an elongated head and a flexible tail (Johnson et al. 1977). The prototype phage φCbK has a large genome of about 200 kb and a 65% GC content (Agabian-Keshishian 1970; Panis 2012; Gill 2012). Recently, Gill et al. (2012) published an analysis of five additional φCbK-like *Caulobacter* bacteriophage genome sequences showing that these phage genomes vary in size and contain long terminal repeats. They also showed that these closely related genomes were organized into three primary modules, the Structural module, the Lysis module, and the DNA replication module. Phage genome comparisons can be used to examine how groups of phage genomes evolve. For example, the theory of modular evolution (Botstein 1980) proposes that bacteriophage genomes are a collection of interchangeable genetic elements (modules). Each module is responsible for a specific function and has the ability to evolve independently of the other modules in the genome. Thus, a collection of related bacteriophages would contain favorable combinations of the available modules. Since the φCbK-like phages contain three conserved modules, we decided to sequence six additional phage genomes to determine if modular evolution could be observed among these phages which were collected from surface water samples obtained from ponds and slow-moving streams across the south eastern United States, as

well as, samples from tropical fish tanks which represent diverse geographical locations since aquarium fish are captured and raised all over the world and then shipped in their own water to commercial dealers (Johnson 1977). Although our comparison of 12 φCbK-like phage genomes showed no support for the modular evolution model, we did find evidence of large numbers of insertions, deletions, and gene translocations. Thus, these phage genomes appear to evolve primarily via small changes rather than by generating recombinant genomes.

MATERIALS AND METHODS

Samples for Phage Isolation

The bacteriophage samples of our collection are ϕCr2, ϕCr5, ϕCr10, ϕCr29, ϕCr32, and ϕCr34.

Phage Isolation and Culture

Caulobacter crescentus CB15 was used as the bacterial host strain for the isolation and proliferation of the bacteriophages in this study (Johnson et al. 1977). Modified PYE growth medium (Johnson and Ely 1977) was used for all liquid cultures and soft-agar overlays. All incubations were at 30 °C with aeration. Agar plates were placed in plastic bags and refrigerated immediately after hardening, to prevent drying of the agar.

DNA Sequencing and Genome Assembly

DNA extraction was performed using a Qiagen QIAamp DNA Mini Kit. DNA sequencing was performed with both the Roche 454 and Illumina MiSeq sequencing

platforms. The resulting reads were aligned into contigs using the DNASTAR Seqman software program. Mauve Whole Genome Alignment software was used to align the contigs against the reference sequence (ϕCbK) to discern the orientation of the contigs and to assist in assembly of the phage genomes. The terminal repeats of these bacteriophage genomes were determined using the Tablet sequence viewer software to identify regions with twice as many reads as described by Gill et al. (2002). The newly sequenced phage genomes were annotated using the RAST automated annotation system and edited in the Artemis Genome browser. For gene comparisons, we used the BlastStation software and performed a BlastP search of the amino acid sequences of all coding regions for each genome. To be considered a match, the genes had to share an e-value less than e^{-05}.

RESULTS AND DISCUSSION

The CbK-like bacteriophages ϕCr2, ϕCr5, ϕCr10, ϕCr29, ϕCr32, and ϕCr34 were first described by Johnson et al. (1977), and are highly similar to the bacteriophage genomes introduced by Gill et al. (2012) in genetic make-up and gene location (Figure 2.1.). However, each genome has unique insertions and deletions. Genome size including the repeats ranged from 205 to 279 Kb (Table 2.1.). The BLASTn two sequence alignment (Tamura 2013) was used to determine the pairwise similarity of these genomes along with the six genomes described by Gill et al. (2012). Most phage genomes were 97-99% identical to each other over 94% to 100% of their genomes (Table 2.2.). In contrast, the Rogue genome was 80-84% identical to most of the other phage across 81-85% of its genome and the Colossus genome was 30% larger than the other genomes and had only 66-69% nucleotide identity across 31-33% of its genome. We also used a

genome-to-genome distance calculator (http://ggdc.dsmz.de/) to compare the 12 genomes and obtained similar results (Table 2.3.). This calculator is based upon the Genome Blast Distance Phylogeny approach (GBDP) which begins with a blast+ alignment between a query and subject sequence to establish the segments of sequence which are considered HSPs (High-scoring pairs; intergenomic matches). The distances between these pairs were calculated, and then converted to percent-wise similarities analogous to DNA-DNA Hybridization (Auch 2010). The data in Table 2.3. is the sum of all identities found in HSPs divided by total genome length. In this comparison Colossus had only 13% identity across its entire genome. Together these data indicate that the larger Colossus genome not only consists of 70% unique genetic material, but also the genes it shares with the other phage have only 67% identity at the nucleotide level. Since Colossus was clearly different from the other phage, we excluded it from many of the subsequent analyses but it provided important information about the core genome of these phages.

Taking advantage of the evolutionary distance between CbK and Colossus, we determined that only 108 genes were present in all 12 phage genomes suggesting that less than one-third of the genes in these genomes are necessary for a successful infection of the host bacterium. The location of these core genes is well conserved within the bacteriophage genomes of this group (Figure 2.2.). Even in a comparison between CbK and Colossus (Figure 2.3.), the location of the core genes is found to be in highly similar segments and arrangements. To begin an analysis of the remainder of the phage genomes, the non-core genes were classified into four categories (Table 2.4.). Genes that were shared in all genomes except for Colossus were classified as CbK-like. Since Rogue differed significantly from the other CbK-like phage (Tables 2.2. and 2.3., Figure

2.4.), we established a second category for genes which were present in all genomes except for Colossus and Rogue, designated CbK-like (-Rogue). Other genes which were present in at least 2 genomes, and did not fall into the categories mentioned above, were classified as INDELS. The final category designated Unique included genes which were present in a single genome. The locations of the genes in these five categories are summarized in a DNA plot image of the bacteriophage CbK genome with each gene category color-coded (Figure 2.5.). The genomic layout for each of the other CbK-like genomes is similar.

A majority of the core genes are contained within the three genomic modules defined by Gill et al. (2012), the structural genes, the replication genes, and the genes involved in lysis. In fact, only 42 of the 108 core genes lie outside of these modules. The location of these modules is well conserved across the all the genomes and the phylogenetic trees of the individual modules are identical to phylogenetic tree of the wholes genomes (Figure 2.4.). Thus, we see no evidence of alternate combinations of modules as proposed by Botstein (Botstein 1980). The DNA ligase gene is not included in our list of core phage genes since this CbK-like gene is not present in the Colossus genome. However Colossus does have a different DNA ligase gene (gp191). A Blast comparison of gp191 produced the best matches with genes in six bacterial genomes and two bacteriophage genomes. The best phage gene match was to the DNA ligase gene of Cr30, the T4-like bacteriophage used for transduction in *C. crescentus* genetic experiments (Ely and Johnson 1977; Ely 1991). The other matching phage gene was the DNA ligase gene from phiM12 which is the closest known relative of Cr30 (Ely et al. 2015). Gp191 is located in the same location of the Colossus genome as the DNA ligase

gene (CbK gp151) of the CbK-like group but complex gene rearrangements have occurred (Figure 2.6.). The Colossus DNA ligase gene gp191 is located between Colossus gp187 (corresponding to CbK gp149) and Colossus gp192 (CbK gp152). The homologues of CbK gp150 and gp151 genes are not present in the Colossus genome, but the CbK gp134 homologue, Colossus gp190, has been translocated adjacent to the new DNA ligase gene. In addition two other Colossus gene insertions at this location, gp188 and gp189, do not match any known phage genes. The Colossus region corresponding to the location of CbK gp134 is missing three genes in addition to the gp134 homologue that are present in the CBK genome and contains an insertion that codes for 11 proteins including gp169, a T4-like protein that has 65% amino acid identity to the corresponding phiM12 T4 30.3-like protein. Thus, it appears that the Colossus Cbk-like DNA ligase gene was replaced by a DNA ligase gene from a T4-like phage that co-infected a *Caulobacter* host along with a Colossus ancestor. At the same time, or in separate events, a translocation brought the Colossus gp190 gene from its position 12 kb away to its current location, and two additional genes were inserted as well. Since the distant locus also has a deletion and an insertion that includes a phiM12-like gene, there could have been a simultaneous complex rearrangement involving at least two phage genomes. This rearrangement could have created the Colossus current gene arrangement or the current arrangement could be the result of additional insertions or deletions that occurred after an initial complex rearrangement involving these two regions of the genome.

This scenario with the DNA ligase gene could be repeated in other gene sets which cannot be detected at this time due to the fact only about 20% of the genes in the phage genomes have a match to a gene with a predicted function and only 43% of the

core genes have a predicted function. In addition, the replacement of the DNA ligase gene does illustrate that the core phage genome is not equivalent to an essential phage genome. An essential phage genome would have to be determined experimentally, but it is likely that there would be substantial overlap between genes in the core genome and those in the essential genome.

Comparisons of the CbK-like genomes also provide evidence of gene fusion events. One example is Colossus gene gp212 which is 924 base pairs in length and codes for a 308 amino acid protein. The first 77 amino acids coded by this gene correspond to the first 77 amino acids coded by CbK gene, gp174 and amino acids 99 to 269 correspond to amino acids 27 to 195 coded by CbK gene gp169. Thus, the Colossus gp212 protein may combine the functions of both genes since it includes 77% CbK gp174 protein and the entire metal-dependent phosphohydrolase domain of the CbK gp169 protein plus some flanking amino acids. The CbK genes between gp174 and gp169 are not present in the Colossus genome suggesting that the fusion occurred as the result of a deletion event.

We also examined the region between CbK gp192 and gp193 where an insertion was observed in the genomes described by Gill et al. (2012). We found that one extra gene was present in this region in the Cr32 and Cr34 genomes, but nine genes were present in most of the other phage genomes, with 10 extra genes in Cr29. The tenth gene in Cr29, gp197, is a duplication of gene Cr29 gp144 which is classified as a core gene. Cr29 gp144 is 100% identical to its CbK counterpart (gp141), while Cr29 gp197 is only 30% identical to gp141 with 94% coverage. This duplication set is a classic example of an evolutionary event where one gene of the duplication is well conserved and maintains

the original function and the other quickly accumulates mutations (Force 1998). The one gene insert found in the Cr32 and Cr34 genomes suggests that a deletion event may have occurred. Evidence to support this idea was obtained when we examined the region between CbK gp192 and gp193 and found an open reading frame with an amino acid sequence that is identical to the first 50 amino acids of the 67 amino acids found in the corresponding gene in most of the other CbK-like phage genomes. The presence of a truncated gene indicates that, at least in this case, a deletion event is likely to have occurred. The matching gene in Rogue (Rogue_gp196) is nearly twice as long as the gene in the other phages suggesting the presence of another gene fusion. However, the distal portion of Rogue_gp196 does not have a significant match to any other gene in the GenBank database.

The inconsistencies seen in the numbers of genes in each category (Table 2.4.) is due to the gene duplications found within this collection of genomes. Rogue has two gene duplications, Colossus has three, and Cr29 contains seven gene duplication pairs. These duplications correspond to a total of 11 different CbK genes. Of the Cr29 duplications, two are core gene duplications that correspond to CbK_gp76 and CbK_gp141 (discussed above). Genes Cr29_gp237 – gp240 have been duplicated and the duplications are located directly downstream in genes Cr29_gp241 – gp245. A comparison of the CbK and Cr29 genomes indicates that duplication begins with the last 28 amino acids of Cr29_gp237, then continues with gp238, gp239, and gp240. With the exception of Cr29_gp243, which is an insert only carried by CbK, Cr32, Cr34, and Karma at relatively the same location, the duplications show very high conservation with the downstream duplication set 100% identical to the CbK homologs. We hypothesize

that the duplication was created by an HGT event from a CbK-like genome which begins with the second half of CbK_gp224 and continues through CbK-gp228. This CbK-like sequence was inserted into Cr29 immediately after gene gp240. The Rogue core gene duplicates correspond to CbK_gp67 and CbK_gp68 (the major capsid protein) and the core gene duplicates for Colossus correspond to CbK_gp68, CbK_gp73, and CbK_gp99 (a 128 kd tail protein). The duplications that correspond to core genes are moderately similar with percent identity ranging from 30% to 82%. The non-core gene duplications of Cr29 share high similarity ranging from 78% to 100% identity. This would suggest that the non-core gene duplications of Cr29 are relatively new on an evolutionary time scale. The core gene duplications of Rogue and Colossus include Rogue genes gp023 & gp070, gp051 & gp069, Colossus genes gp023 & gp070, gp036 & gp081, gp087 & gp088, and gp122 & gp123. The core gene duplications for Rogue and Colossus are all part of the structural module and match to major capsid protein and tail protein genes, but only one of Cr29 core gene duplicates (Cr29_gp81 & gp82) is part of the structural module. It is interesting that the largest genomes of this collection contain duplications of core structural genes. However, each genome contains different gene duplications making it unlikely that these duplications are associated with structural changes leading to increase in size of these bacteriophage genomes. It is unclear how these duplications affect the bacteriophages, for most of them fall are currently of unknown function.

The high degree of homogeneity amongst the CbK-like bacteriophage genomes was unexpected considering their diverse geographical origins and the level of genomic rearrangements observed within the genus *Caulobacter* (Ash 2014). An explanation could be linked to the lack of bacteriophage immunity genes within the genome of the

host *Caulobacter* crescentus CB15. CRISPR-cas adaptive immunity in *Streptococcus thermophilus* has been shown to be a driving force in the evolution of the bacteriophages which are specific for this host bacterium (Paez-Espino 2015). The lack of detectable phage immunity genes does not mean the host lacks bacteriophage immunity. Rather *Caulobacter* has an innate immunity to these bacteriophages since they only infect swarmer cells (Johnson 1977). The mechanism of infection for φCbK begins with attachment to the flagellum of the host bacterium via a head filament. The bacterial phage tail attaches to the pilus portal and it is hypothesized that retraction of the pilus filament facilitates genome insertion into the host cell (Guerrero-Ferreira 2011). Due to the asymmetrical cell division of *Caulobacter*, a bacteriophage particle would only be capable of infecting at most 50% of the cells in a population of *Caulobacter*. Further, the generation time of *Caulobacter crescentus* has been determined to be 1.5 – 2 hours with the swarmer cell stage lasting less than an hour (Poindexter 1964). Therefore, during the lifespan of a typical *Caulobacter crescentus* cell, it is only susceptible to infection by these bacteriophages for a brief window of time. Once the swarmer cell has matured into a stalked cell, it would never be aware of the presence of these bacteriophages. From the perspective of the bacteriophage, evolution and adaption would not be necessary for survival against the innate immunity of *Caulobacter* because the mature stalked cells are like stem cells that continuously produce daughter cells susceptible to ϕCbK-like phage infection, yet the stalked cells maintain their resistance to ϕCbk-like phage predation. Therefore, we propose that there is no need for an evolutionary arms race between *Caulobacter crescentus* and these type B3 bacteriophages. Instead, selection acts on the

bacteriophage genomes to maintain a system of genes and genome organization that works efficiently.

Table 2.1. Summary of the genomic characteristics of 11 *Caulobacter crescentus* bacteriophage genomes. Bacteriophage marked with an (*) are incomplete draft sequences.

	CbK	Cr2	Cr5	Cr10	Cr29*	Cr32*	Cr34*	Karma	Magneto	Swift	Rogue	Colossus
Genome Size	215710 bp	220299 bp	218729 bp	219348 bp	216027 bp	205467 bp	205907 bp	221828 bp	218929 bp	219216 bp	223720 bp	279967 bp
tRNA Genes	26	26	27	26	26	26	26	26	27	27	23	28
# of Genes	338	342	338	344	336	314	317	353	347	343	350	448

Table 2.2. Percent identity in pairwise comparisons 11 bacteriophage genomes (percent coverage).

	CbK	Cr2	Cr5	Cr10	Cr29	Cr32	Cr34	Karma	Magneto	Swift	Rogue
CbK											
Cr2	97% (98%)										
Cr5	98% (98%)	96% (97%)									
Cr10	97% (98%)	99% (98%)	96% (98%)								
Cr29	97% (99%)	99% (99%)	96% (98%)	99% (100%)							
Cr32	99% (100%)	97% (97%)	98% (97%)	97% (97%)	97% (97%)						
Cr34	99% (100%)	97% (97%)	98% (97%)	97% (97%)	97% (97%)	99% (100%)					
Karma	97% (98%)	97% (98%)	97% (98%)	97% (98%)	97% (98%)	97% (98%)	97% (98%)				
Magneto	97% (98%)	97% (98%)	96% (98%)	97% (97%)	97% (97%)	97% (98%)	97% (98%)	99% (97%)			
Swift	97% (94%)	97% (94%)	97% (94%)	97% (95%)	97% (96%)	97% (98%)	97% (96%)	98% (94%)	98% (95%)		
Rogue	83% (81%)	83% (81%)	80% (81%)	83% (81%)	83% (82%)	83% (82%)	83% (82%)	84% (81%)	83% (82%)	84% (85%)	
Colossus	67% (31%)	68% (30%)	66% (30%)	68% (30%)	68% (30%)	67% (31%)	67% (31%)	67% (30%)	67% (31%)	67% (31%)	69% (31%)

Table 2.3. Genome-to-Genome distances expressed as %DDH estimates with confidence intervals in parentheses.

	CbK	Cr2	Cr5	Cr10	Cr29	Cr32	Cr34	Karma	Magneto	Swift	Rogue
CbK											
Cr2	96.8% (1.2)										
Cr5	96.2% (1.4)	95.9% (1.5)									
Cr10	96.6% (1.3)	99.4 %(0.3)	96.1% (1.4)								
Cr29	97.2% (1.1)	100% (0.3)	96.3% (1.3)	100% (0.1)							
Cr32	99.3% (0.4)	96.9% (1.2)	96.3% (1.4)	96.8% (1.2)	97.2% (1.1)						
Cr34	99.4% (0.4)	96.7% (1.2)	96.1% (1.4)	96.6% (1.3)	97.1% (1.1)	100% (0.0)					
Karma	95.8% (1.5)	96% (1.4)	95.4% (1.6)	95% (1.6)	95.8% (1.5)	96% (1.4)	95.8% (1.5)				
Magneto	95.8% (1.5)	96.8% (1.2)	94.9% (1.7)	96.2% (1.4)	96.4% (1.3)	95.9% (1.5)	95.7% (1.5)	98.2% (0.8)			
Swift	89.3% (2.6)	89.9% (2.5)	89.6% (2.5)	91.6% (2.3)	92.4% (2.1)	91.2% (2.3)	90.9% (2.3)	93.6% (1.9)	93.4% (1.9)		
Rogue	37.1% (3.0)	38.4% (3.0)	37.4% (3.0)	39.6% (3.0)	38.8% (3.0)	37.5% (3.0)	37.4% (3.0)	38.1% (3.0)	38.2% (3.0)	42.2% (3.0)	
Colossus	13.2% (2.5)	13.2% (2.5)	13.2% (2.5)	13.1% (2.5)	13.1% (2.5)	13.2% (2.5)	13.2% (2.5)	13.2% (2.5)	13.3% (2.5)	13.2% (2.5)	13.2% (2.5)

Table 2.4.: Classifications of the bacteriophage genes based upon the prevalence of these genes amongst the entire group of 12 bacteriophages.

Gene Classifications	CbK	Cr2	Cr5	Cr10	Cr29	Cr32	Cr34	Karma	Magneto	Swift	Rogue	Colossus
Core Genes	110	111	111	111	111	108	110	110	110	110	110	112
CbK-like	147	147	147	147	140	140	140	147	147	147	146	N/A
CbK-like (- Rogue)	35	35	35	35	35	35	35	35	35	35	N/A	N/A
Indels	46	50	44	52	51	32	31	60	51	50	47	21
Unique	0	0	2	0	0	0	2	1	4	1	47	315

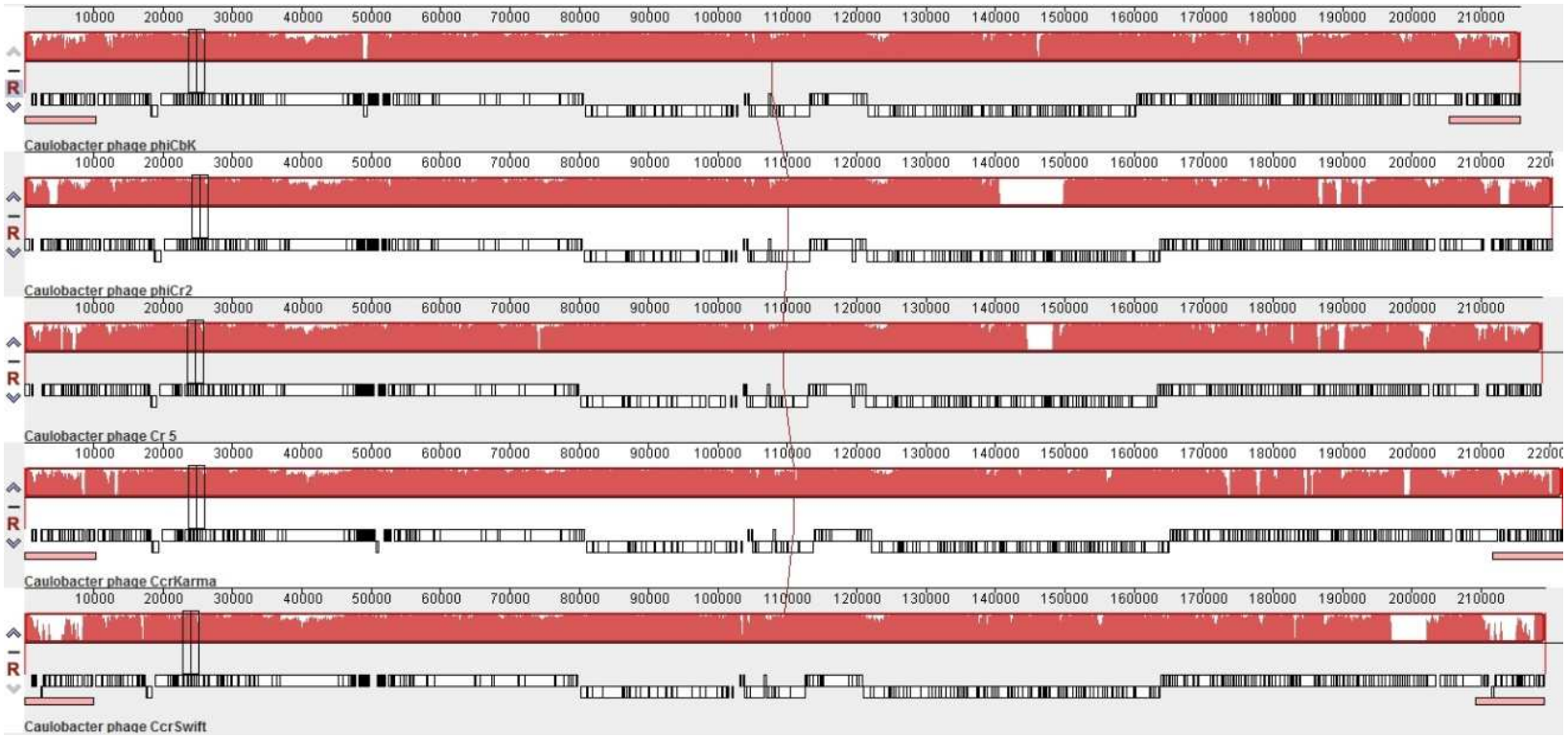

Figure 2.1. Mauve whole genome comparison of the CbK, Cr2, Cr5, CcrKarma, and CCrswift bacteriophage genomes. Red space corresponds to similar sequences. Regions that are entirely white correspond to inserted gene regions.

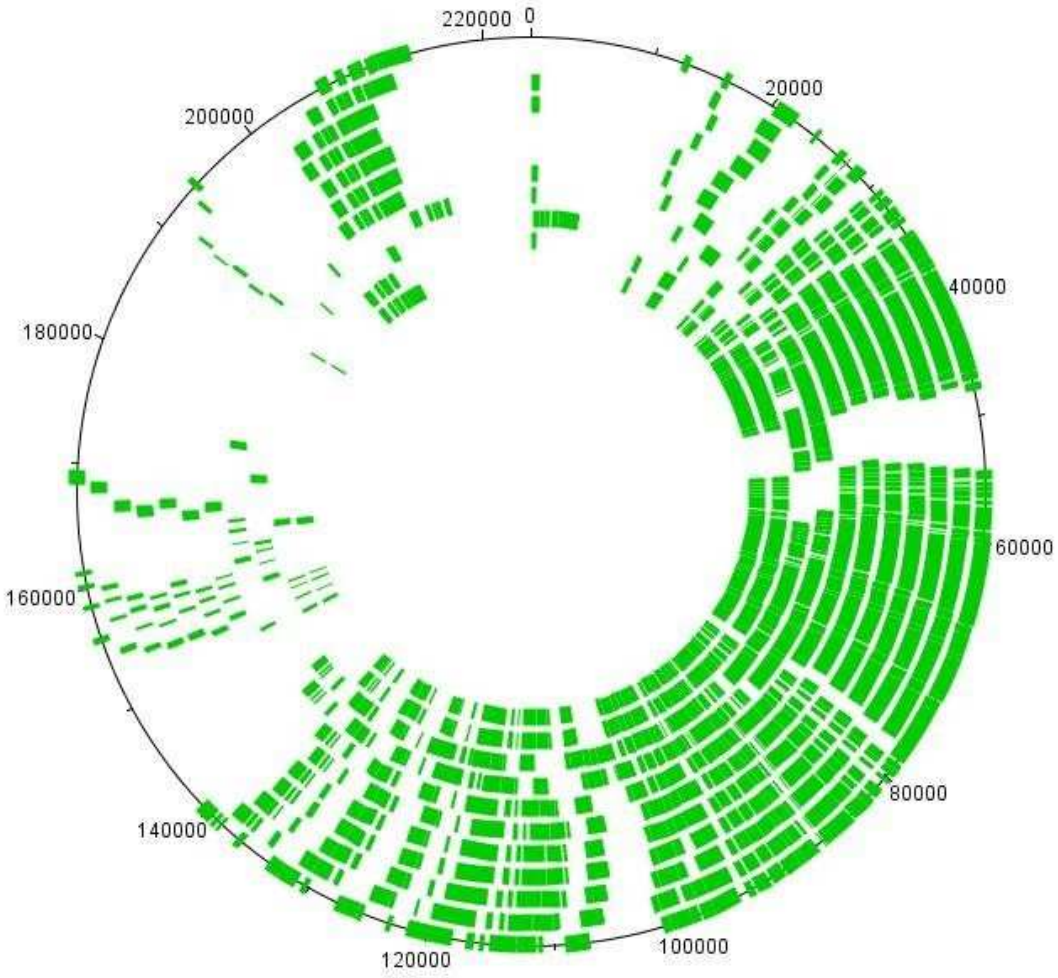

Figure 2.2. The core gene locations within each of the CbK-like genomes. The genomes are arranged in size from largest to smallest (Table 1), with Rogue on the outside track and Cr32 on the inner most track. The Colossus genome is omitted from this comparison.

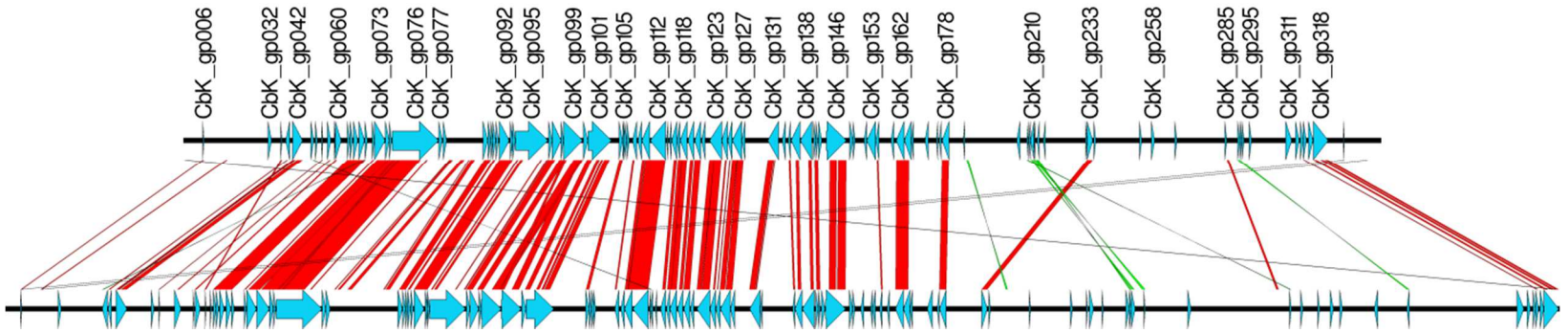

Figure 2.3. A comparison of CbK and Colossus showing that the gene order and location of most of the core genes are conserved within the respective genomes. Red lines indicate genes are in the same orientation, green lines indicate an inversion.

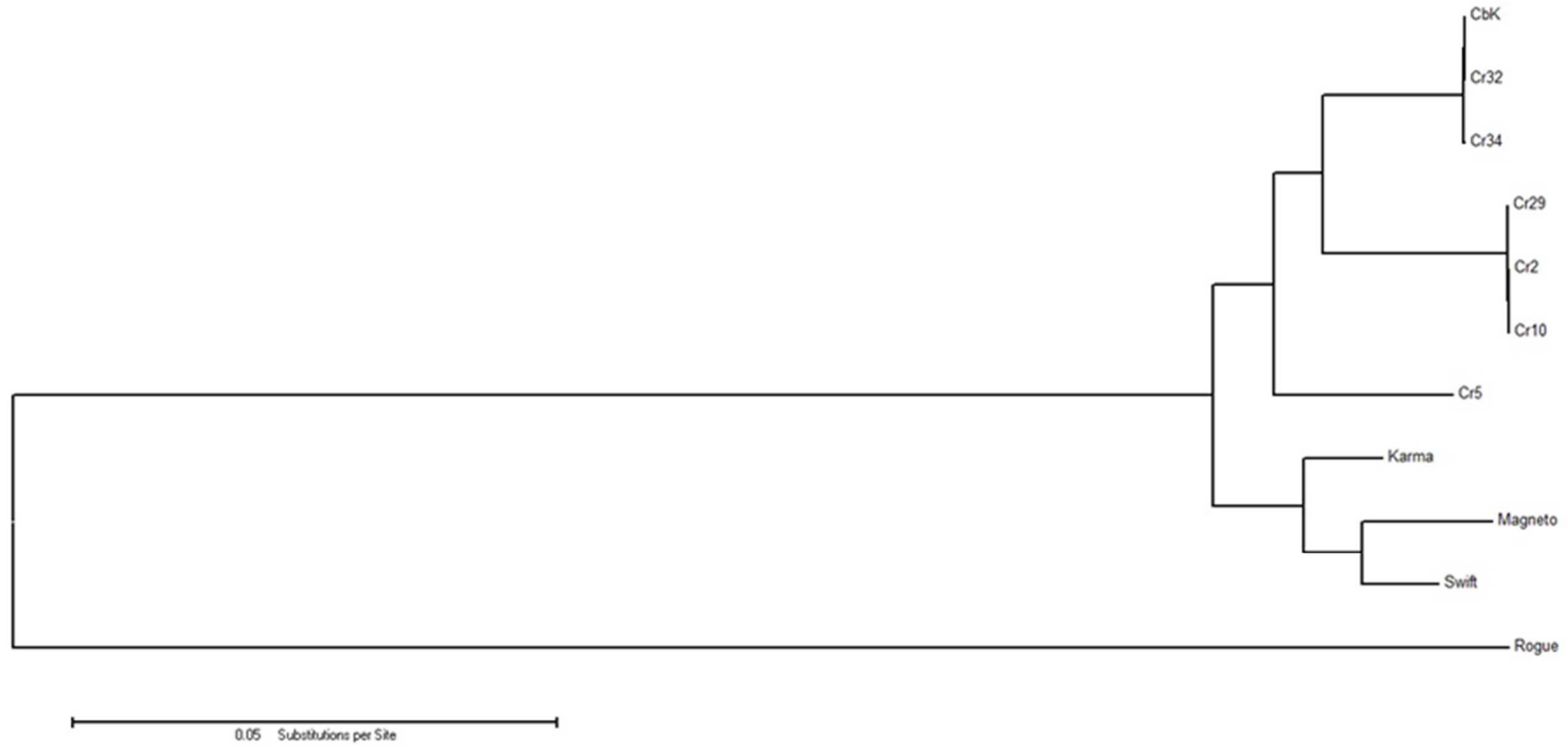

Figure 2.4. Molecular Phylogenetic analysis by Maximum Likelihood method. The evolutionary relationships of these genomes was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura 1993). The tree with the highest log likelihood (-498908.1958) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 11 genome nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 179597 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura 2013).

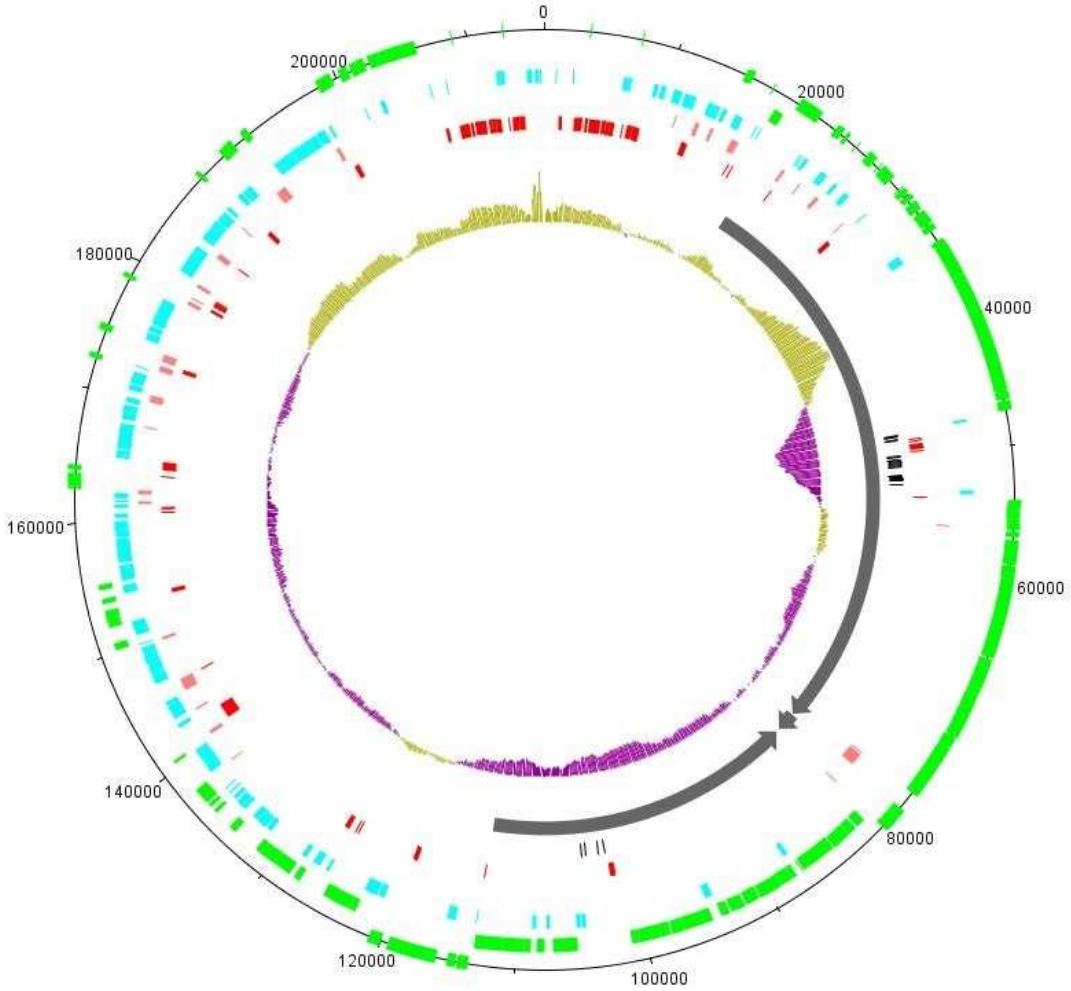

Figure 2.5. DNA plot image of the CbK genome gene categories. The green blocks represent the core genes, the light blue are conserved throughout the 11 CbK-like genomes. The light red blocks are conserved in all CbK-like genomes except that of Rogue. Dark red blocks represent the indels in the CbK genome. Black blocks are the tRNA locations. The grey arrows represent the location of the Genomic modules, the structural module, lysis module, and DNA replication module respectively, clockwise from the top of the circle. The center plot is the GC content of the CbK genome with gold bars marking regions of above average %GC and purple bars marking the regions with below average %GC.

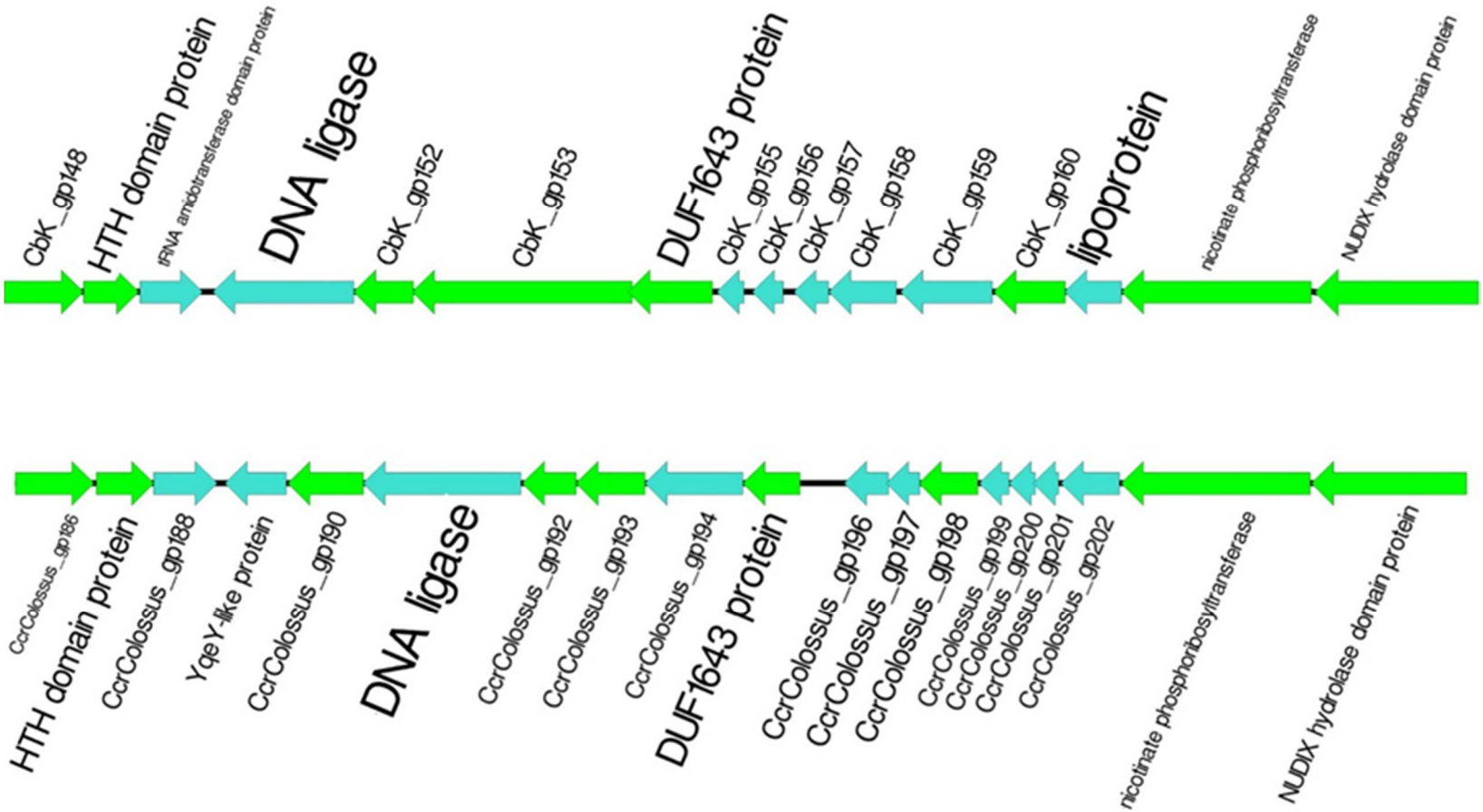

Figure 2.6. The DNA ligase regions of CbK genome and Colossus genome. The green arrows represent the genes which have been classified as "core genes"

CONCLUSION

This research originally began as a project to teach genetics to high school students of the Eau Claire HS Science Club. While teaching the students to use the NCBI BLAST program to identify genes, we discovered that the Caulobacter K31 genome had been submitted to the GenBank database. We then used the whole genome alignment software program Mauve to compare this new *Caulobacter* genome to the previously sequenced *Caulobacter crescentus* CB15/NA1000 genome and observed that there was a high degree of genomic rearrangement. Similar comparisons of other closely related bacterial genome sequences revealed that this high degree of genome rearrangements was unique and the comparison of the two *Caulobacter* genomes became the focus of the first chapter of this dissertation. Attempts were made to characterize and discern the cause of these genomic rearrangements but the sheer number of rearrangements made analyses challenging. We found no significant sequence similarity at rearrangement breakpoints and did not observe any rearrangements among NA1000 lines that had been sub-cultured separately for more than 1000 generations. Many of the rearrangements occurred in the vicinity of tRNA genes. tRNA genes have been shown to be associated with indels and non-homologous recombination in *Caulobacter segnis* (Patel 2015). The exact cause of the genomic rearrangements observed in Caulobacter is still unknown; however, there are a few conclusions, which can be drawn from our research. We do not think that the genome rearrangements occur under laboratory conditions due to results of an experiment

using ancestor Caulobacter cell lines which after numerous rounds of re-culturing did not produce observable changes in the DNA banding pattern. Also the rearrangements do not appear to be the result of recombination due to sequence similarity within the genome. The observed genome rearrangements most likely are the result of DNA repair caused by damage from an external DNA damage causing agent (e.g. sunlight). However, many more experiments would need to be conducted to fully characterize this phenomenon. We hypothesized that genomic rearrangements might be observable in bacteriophage genomes, as well. However, no genomic rearrangements were observed in the bacteriophage genomes that we sequenced. Instead, we found that the genomes sequences and gene order of these bacteriophages were highly conserved. This conservation allowed us to identify the 108 "core genes" that are present in all 12 of the bacteriophage genomes. More than 140 additional genes were found in all of the phage except Colossus, and another 35 were present in all of the phage except for Colossus and Rogue. We also identified numerous indels, unique genes, duplications, and translocations.

The research done for this dissertation was the beginning of an investigation into the extensive genome rearrangements discovered in genus of *Caulobacter*. Our work has demonstrated that the rearrangements are likely a response to an external agent and not a spontaneous event that occurs during growth in laboratory conditions. The bacteriophage genomes we sequenced allowed us to begin to define the core genome of the B3 type bacteriophages of *Caulobacter* and provide a starting point for future research to accurately define the essential genes of this bacteriophage group. However, a defined

core genome also leads to additional questions. Why are these phage genomes so large? Do the additional 200 genes serve a useful purpose(s)?

REFERENCES

Abraham, W. R., C. Strompl, et al. (1999). "Phylogeny and polyphasic taxonomy of Caulobacter species. Proposal of Maricaulis gen. nov. with Maricaulis maris (Poindexter) comb. nov. as the type species, and emended description of the genera Brevundimonas and Caulobacter." Int J Syst Bacteriol **49 Pt 3**: 1053-73.

Agabian-Keshishian, N. and L. Shapiro (1970). "Stalked bacteria: properties of deoxriybonucleic acid bacteriophage phiCbK." J Virol **5**(6): 795-800.

Ash, K., T. Brown, et al. (2014). "A comparison of the Caulobacter NA1000 and K31 genomes reveals extensive genome rearrangements and differences in metabolic potential." Open Biol **4**(10).

Auch, A. F., M. von Jan, et al. (2010). "Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison." Stand Genomic Sci **2**(1): 117-34.

Beare, P. A., N. Unsworth, et al. (2009). "Comparative Genomics Reveal Extensive Transposon-Mediated Genomic Plasticity and Diversity among Potential Effector Proteins within the Genus Coxiella." Infection and Immunity **77**(2): 642-656.

Bender, R. A., C. M. Refson, et al. (1989). "Role of the flagellum in cell-cycle-dependent expression of bacteriophage receptor activity in Caulobacter crescentus." J Bacteriol **171**(2): 1035-40.

Biondi, E. G., S. J. Reisinger, et al. (2006). "Regulation of the bacterial cell cycle by an integrated genetic circuit." Nature **444**(7121): 899-904.

Botstein, D. (1980). "A theory of modular evolution for bacteriophages." Ann N Y Acad Sci **354**: 484-90.

Chatterjee, D. K. and A. W. Bourquin (1987). "Metabolism of aromatic compounds by Caulobacter crescentus." J Bacteriol **169**(5): 1993-6.

Couturier, E. and E. P. Rocha (2006). "Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes." Mol Microbiol **59**(5): 1506-18.

Darling, A. E., B. Mau, et al. (2010). "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement." PLoS One **5**(6): e11147.

Darling, A. E., I. Miklos, et al. (2008). "Dynamics of genome rearrangement in bacterial populations." PLoS Genet **4**(7): e1000128.

Darmon, E. and D. R. Leach (2014). "Bacterial genome instability." Microbiol Mol Biol Rev **78**(1): 1-39.

Ely, B. (1991). "Genetics of Caulobacter crescentus." Methods Enzymol **204**: 372-84.

Ely, B., W. Gibbs, et al. (2015). "The Caulobacter crescentus Transducing Phage Cr30 is a Unique Member of the T4-Like Family of Myophages." Curr Microbiol **70**(6): 854-8.

Ely, B. and R. C. Johnson (1977). "Generalized Transduction in CAULOBACTER CRESCENTUS." Genetics **87**(3): 391-9.

Ely, B. and L. E. Scott (2014). "Correction of the Caulobacter crescentus NA1000 genome annotation." PLoS One **9**(3): e91668.

Entcheva-Dimitrov, P. and A. M. Spormann (2004). "Dynamics and control of biofilms of the oligotrophic bacterium Caulobacter crescentus." J Bacteriol **186**(24): 8254-66.

Force, A., M. Lynch, et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-45.

Gill, J. J., J. D. Berry, et al. (2012). "The Caulobacter crescentus phage phiCbK: genomics of a canonical phage." BMC Genomics **13**: 542.

Guerrero-Ferreira, R. C., P. H. Viollier, et al. (2011). "Alternative mechanism for bacteriophage adsorption to the motile bacterium Caulobacter crescentus." Proc Natl Acad Sci U S A **108**(24): 9963-8.

Hatfull, G. F. (2008). "Bacteriophage genomics." Curr Opin Microbiol **11**(5): 447-53.

Hendrix, R. W. (2002). "Bacteriophages: evolution of the majority." Theor Popul Biol **61**(4): 471-80.

Holtzendorff, J., D. Hung, et al. (2004). "Oscillating global regulators control the genetic circuit driving a bacterial cell cycle." Science **304**(5673): 983-7.

Johnson, R. C. and B. Ely (1977). "Isolation of spontaneously derived mutants of Caulobacter crescentus." Genetics **86**(1): 25-32.

Johnson, R. C., N. B. Wood, et al. (1977). "Isolation and Characterization of Bacteriophages for Caulobacter crescentus." Journal of General Virology **37**(2): 323-335.

Jordan, T. L. and J. T. Staley (1975). "Electron microscopic study of succession in the periphyton community of lake Washington." Microb Ecol **2**(4): 241-51.

Lagenaur, C., S. Farmer, et al. (1977). "Adsorption properties of stage-specific Caulobacter phage phiCbK." Virology **77**(1): 401-7.

MacRae, J. D. and J. Smit (1991). "Characterization of caulobacters isolated from wastewater treatment systems." Appl Environ Microbiol **57**(3): 751-8.

Macur, R. E., J. T. Wheeler, et al. (2001). "Microbial populations associated with the reduction and enhanced mobilization of arsenic in mine tailings." Environ Sci Technol **35**(18): 3676-82.

Mannisto, M. K., M. A. Tiirola, et al. (1999). "Diversity of chlorophenol-degrading bacteria isolated from contaminated boreal groundwater." Arch Microbiol **171**(3): 189-97.

Marks, M. E., C. M. Castro-Rojas, et al. (2010). "The genetic basis of laboratory adaptation in Caulobacter crescentus." J Bacteriol **192**(14): 3678-88.

Maruyama, F., M. Kobata, et al. (2009). "Comparative genomic analyses of Streptococcus mutans provide insights into chromosomal shuffling and species-specific content." BMC Genomics **10**: 358-358.

Nierman, W. C., T. V. Feldblyum, et al. (2001). "Complete genome sequence of Caulobacter crescentus." Proc Natl Acad Sci U S A **98**(7): 4136-41.

Paez-Espino, D., I. Sharon, et al. (2015). "CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus." MBio **6**(2).

Panis, G., C. Lambert, et al. (2012). "Complete genome sequence of Caulobacter crescentus bacteriophage phiCbK." J Virol **86**(18): 10234-5.

Patel, S., B. Fletcher, et al. (2014). "Genome sequence and phenotypic characterization of Caulobacter segnis." Curr Microbiol **70**(3): 355-63.

Poindexter, J. S. (1964). "Biological Properties and Classification of the Caulobacter Group." Bacteriol Rev **28**: 231-95.

Poindexter, J. S. (1981). "The caulobacters: ubiquitous unusual bacteria." Microbiol Rev **45**(1): 123-79.

Reisenauer, A. and L. Shapiro (2002). "DNA methylation affects the cell cycle transcription of the CtrA global regulator in Caulobacter." EMBO J **21**(18): 4969-77.

Rutherford, K., J. Parkhill, et al. (2000). "Artemis: sequence visualization and annotation." Bioinformatics **16**(10): 944-5.

Sanger, F., G. M. Air, et al. (1977). "Nucleotide sequence of bacteriophage phi X174 DNA." Nature **265**(5596): 687-95.

Sanger, F., A. R. Coulson, et al. (1982). "Nucleotide sequence of bacteriophage lambda DNA." J Mol Biol **162**(4): 729-73.

Schmidt, J. M. and R. Y. Stanier (1965). "Isolation and Characterization of Bacteriophages Active against Stalked Bacteria." J Gen Microbiol **39**: 95-107.

Stahl, D. A., R. Key, et al. (1992). "The phylogeny of marine and freshwater caulobacters reflects their habitat." Journal of Bacteriology **174**(7): 2193-2198.

Stove, J. L. and R. Y. Stanier (1962). "Cellular Differentiation in Stalked Bacteria." Nature **196**(4860): 1189-1192.

Suttle, C. A. (2007). "Marine viruses--major players in the global ecosystem." Nat Rev Microbiol **5**(10): 801-12.

Tamura, K. and M. Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." Mol Biol Evol **10**(3): 512-26.

Tamura, K., G. Stecher, et al. (2013). "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0." Mol Biol Evol **30**(12): 2725-9.

Tarleton, J. C. and B. Ely (1991). "Isolation and characterization of ilvA, ilvBN, and ilvD mutants of Caulobacter crescentus." J Bacteriol **173**(3): 1259-67.

Wang, S. P., P. L. Sharma, et al. (1993). "A histidine protein kinase is involved in polar organelle development in Caulobacter crescentus." Proceedings of the National Academy of Sciences **90**(2): 630-634.

APPENDIX A: COPYRIGHT RELEASE

Chapter 1 is a manuscript which has previously been published in *Open Biology* research journal. *Open Biology*'s policy states that the original authors maintain all copyright privileges.

What is your policy on copyright?

Authors publishing in *Open Biology* can disseminate their articles under a Creative Commons Attribution Licence, allowing them to post the final published version on repositories as soon as the article is published.

http://rsob.royalsocietypublishing.org/faq#question6

How will the journal operate?

The editorial team will be responsible for developing the editorial direction of the journal and will be the final authority on what is published. The journal will be run with the full support of the Society's Editorial Office and the Editorial Board. *Open Biology* will be published online on a continuous publication model where articles are immediately citable. Online archiving and information on article downloads will be available. Articles will be published under the terms of the Creative Commons Attribution Licence, leaving copyright with the authors, but allowing anyone to download, reuse, reprint, modify, distribute, and/or copy articles provided the original authors and source are cited.

Our vision is to provide a high quality publishing venue for biology at the molecular and cellular level. We urge you to submit your very best research to *Open Biology*. In return we will provide the peer-review, high quality editorial feedback, rapid publication and international dissemination of your research associated with all Royal Society journals.

http://rsob.royalsocietypublishing.org/content/1/1/110001