8-9-2014

Utilizing Next Generation Sequencing to Generate Bacterial Genomic Sequences for Evolutionary Analysis

Derrick C. Scott
University of South Carolina - Columbia

Utilizing Next Generation Sequencing to Generate Bacterial Genomic Sequences for Evolutionary Analysis

by

Derrick C. Scott

Bachelor of Science
Virginia State University, 2004

Master of Science
Virginia Tech, 2009

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

2014

Accepted by:

Bert Ely, Major Professor

Johannes Stratmann, Committee Member

Lucia Pirisi-Creek, Committee Member

Richard Showman, Committee Member

Robert Friedman, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

DEDICATION

To my wife, Dr. LaTia Scott: Thank you for your wisdom and support. You push me further even when I don't want you to! I am a better man because of it. I can't wait to start this next phase of our lives with you beside me. To Zoey and my unborn son: Daddy finally finished! I will use this platform to make sure that you guys are in a position to go even further than I have.

Love, Daddy

ACKNOWLEDGEMENTS

All thanks to God for this journey and placing me in Columbia to meet some great people. I wouldn't change a thing. Daddy and Momma, I know you have heard me say before that I am about to graduate but this time it is for real. To Reggie, Michelle, Coretta and all of my family; I love you. Shout out to the Fellas Crew! JD, I wish you were here for this man.

I thanked almost everyone when I wrote my thesis for my Masters at Virginia Tech with one glaring typo. Dr. Larry Brown was the department chair at Virginia State (now a Dean!) where I did my Bachelors. He was a huge supporter who first made me believe that I could go to grad school and be successful. I typed the wrong name in my thesis! Sorry for that! Thanks to Dr. Stratmann for the guidance in my first years here and putting me in a position to succeed with Dr. Ely. Dr. Ely has been instrumental to my growth as a Scientist and also as a man. Thanks for all the drops of wisdom you've imparted, for the early morning fishing trips, and giving my wife purpose while here in Columbia. We love you. I give special thanks to Mrs. Charlene for keeping my paperwork straight and especially to Mrs. Pat. I would not have graduated without you! Thanks to Dr. Marton for his invaluable insight during my comprehensive exam and Dr. Friedman for his expertise. Also I would like to give a special thank you to Dr. Showman and Dr. Pirisi-Creek for the last minute assist!

ABSTRACT

Many important questions in the field of prokaryotic biology cannot be answered due to the low availability of sequenced and finished genomes. Recent improvements in technology and decreases in price have made the ambition of *de novo* bacterial genomic sequencing a reality for a wide range of researchers. However, with the advancement of sequencing technology comes the need for an evaluation to determine the most reliable bioinformatics methods in generating a complete and accurate assembly. Biases inherent in the sequencing technology and GC-rich genomes complicate genome assemblies. Here, we sequenced bacterial strains from the GC-rich *Caulobacter* genus and the closely related *Brevundimonas* genus. We found that the Pacific Biosciences RS II sequencing systems was the best sequencer to use in conjunction with the HGAP2 assembler. Using our newly acquired sequences, we found that the genus *Caulobacter* exhibits extensive genome rearrangements giving the appearance of "Genome Scrambling". We found that these extensive rearrangements had no correlation to genome relatedness within the genus and that they did not disrupt the conservation of NA1000 essential genes between the species. We also found that using the 16S rRNA region to group these bacteria were as accurate as using entire conserved operons spanning thousands of base pairs.

PREFACE

Chapters two and three are based on two studies submitted/in preparation for publications that discuss a comparison of genome sequencers and assemblers and also the evolution of *Caulobacters*. As such, they are presented in whole in this dissertation including each reference section specific to each study. A key feature of both projects was a genomics based approach that utilized the novel DNA sequences elucidated in the first paper to shed light on the genome comparison studies done in the second paper. The Introduction contains its References immediately following the conclusion of the chapter.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION TO CAULOBACTER

SCOTT, D.C. AND ELY, B. TO BE SUBMITTED TO *CURRENT MICROBIOLOGY*.

Bacteria are a large domain of prokaryotic microorganisms. Typically a few micrometers in length, bacteria have a wide range of shapes, ranging from spheres to rods and spirals (Fredrickson et al. 2004). Bacteria are present in most habitats on Earth, growing in soil, acidic hot springs, radioactive waste, water, and deep in the Earth's crust. They also grow in organic matter and the live bodies of plants and animals. There are typically 40 million bacterial cells in a gram of soil and a million bacterial cells in a milliliter of fresh water. In all, there are approximately five nonillion (5×10^{30}) bacteria on Earth (Whitman, Coleman, & Wiebe 1998) forming a biomass that exceeds that of all plants and animals (Hogan, 2010). Bacteria are vital in recycling nutrients, with many steps in nutrient cycles depending on these organisms, such as the fixation of nitrogen from the atmosphere and putrefaction. Most bacteria have not been characterized, and only about half of the phyla of bacteria have species that can be grown in the laboratory (Rappé & Giovannoni 2003).

Caulobacteriales is an important order of bacteria (Henrici and Johnson, 1935). These bacteria were isolated by submerging slides into freshwater Lake Alexander in Minnesota. After drying the slides, these bacteria remained attached to the glass surface and were subsequently stained and described thoroughly. Henrici and Johnson described the *Caulobacters* as gram-negative unicellular stalked bacteria which produced stalks at one end of the cell. They multiply exclusively by binary fission and are rod-shaped, fusiform, or vibrioid. The bacteria were divided into four families based on morphological differences. They were discovered to be aerobic and though most strains were isolated from fresh water sources, they were found also in seawater, soil, and insects (Table 1.1).

Table 1.1: Pointdexter 1964

TABLE 1. *Sources of Caulobacter isolates**

Strain no.	Source	Sampling date	Isolated by
CB1, 2	Tap water	July, 1959	Author
CB4, 5	Pond water, SW	June, 1959	Author
CB6, 7	Lake water	June, 1959	Author
CB8, 9	Tap water	June, 1959	Author
CB10, 11, 13	Pond water, SC	August, 1959	Author
AC12	Pond water, SC	August, 1959	Author
CB15–18	Pond water, SE	January, 1960	Author
CB21, 23–29, 31, 35, 36	Pond water	February, 1962	Bacteriology class
CB37	Millipede hind-gut	March, 1962	Author
AC47, 48	Pond water, NC	October, 1962	Author
CB51, 57, 63	Pond water, NC	October, 1962	Author
CB65, 70, 79, 81, 82, 86	Pond water, SE-a	October, 1962	Author
CB66, 83	Pond water, SE-b	October, 1962	Author
CB88, 89	Soil	November, 1962	M. Macpherson
CB91	Stream water	February, 1963	Author
CB93	Millipede hind-gut	April, 1963	Author
CB-G	Well water, Kentucky	ca. 1953	Bowers et al.
CB-H	Tap water, Delft	ca. 1951	A. L. Houwink
CB-R	Contaminated *Chlorella* culture, Moscow	ca. 1960	G. A. Zavarzin
KA1–4	Pond water-tap water mixtures	1953	B. J. Bachmann
KA5, 6	Pond water-tap water mixtures	1954	Bacteriology class
CM11, 13	Filtered seawater	February, 1963	Author

* Sources of isolates obtained by the classes, Dr. Bachmann, and the author are in California. Abbreviations refer to various ponds.

Highly motile, (Jones, 1905; Omeliansky, 1914) members of the *Caulobacteriaceae* family are stalked, with the long axis of the elongated cells coinciding with the axis of the stalk. Stalks are slender and unbranched and are often attached to the substrate by a button-like holdfast. The most well studied genus in this family, including early electron microscopy (Houwink, 1952) is *Caulobacter*. *Caulobacter crescentus* is a gram-negative bacterium that thrives in low nutrient environments. It exists in either of two forms, a stalked cell form and a swarmer cell form, and it invariably differentiates into the two cell types and divides asymmetrically at each cell cycle (Figure 1.1). *C. crescentus* strain CB15 has a well-developed system of genetics (Ely, 1991; Shapiro, 1976) and was the first *Caulobacter* to have its genome sequenced (Nierman et al., 2001). Being simple and having its genome amenable to alterations, it was developed into a single-celled model system to study cellular differentiation, asymmetric division, and their coordination with cell cycle progression (Brun & Janakiraman, 2000; Laub et al., 2000). Studies initiated in the Ely and Shapiro laboratories of this obligatory differentiation during the cell cycle

eventually made *Caulobacter* the dominant prokaryotic model system for studying the molecular mechanisms of cell-cycle control and cellular differentiation (Figure 1.2). In fact, many of the regulatory circuits identified in *C. crescentus* are present throughout the family of *Alphaproteobacteria* (Christen et al., 2011). To better understand the context of these regulatory circuits, whole genome studies are needed. For example, comparisons of closely related genomes could help determine how horizontal gene transfer and other genome rearrangements impact the regulatory circuits that govern the progression through the cell cycle.

Figure 1.1: Electron micrograph of a C. crescentus predivisional cell **(**Brun, 2001).

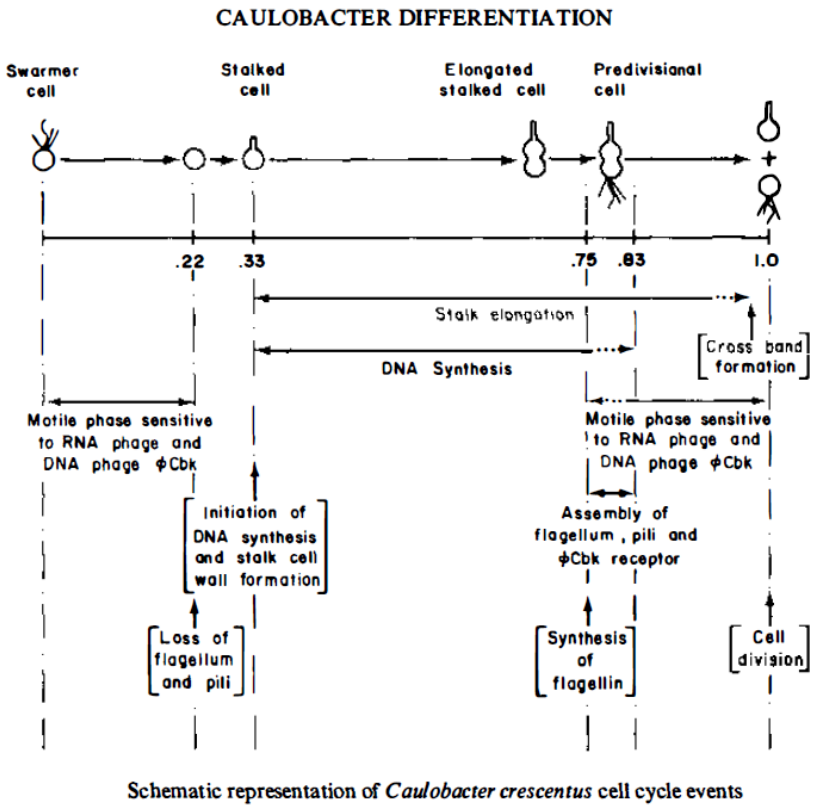

Figure 1.2: Basis of *Caulobacter* cell cycle (Shapiro, 1976)

Presently, there are three *Caulobacter* genomes available in public databases. They are *C. vibrioides* (formerly *C. crescentus*) CB15 or NA1000 (two laboratory versions of the same original isolate, Marks et al. 2010), *C. segnis* TK0059 (Brown et al. 2011; Patel et al. submitted for publication), and *Caulobacter* strain K31 (Ash et al. submitted for publication). *C. crescentus* is the best studied of the three species. Its genome consists of 4,016,942 base pairs in a single circular chromosome encoding 3,767 genes (see Table 1.2 from Nierman et al. 2001). The genome contains multiple clusters of genes encoding proteins essential for survival in a nutrient poor habitat in addition to many extra cytoplasmic function sigma factors, providing the organism with the ability to respond to a wide range of environmental fluctuations. *C. segnis* was isolated from soil by Takahashi and Komahara in 1973 and is closely related to *C. crescentus* (Urakami et

al., 1990). K31 was isolated from an enrichment culture for chlorophenol-tolerant bacteria in groundwater from Karkola, Finland and does not have a species designation (Gregoriev et al., 2011). However, our analysis of the K31 16S ribosomal RNA gene revealed that K31 was closely related to *C. henricii* strain CB4 (Figure 1.3).

Table 1.2: General Features of the *C. crescentus* genome. Nierman, 2001

Size, bp	4,016,942
G+C percent	67.2
Total no. ORFs	3,767
ORF size, bp	969
Percent coding	90.6
No. rRNA operons (16S-23S-5S)	2
No. tRNA	51
No. similar to known proteins	2,030 (53.9%)
No. similar to proteins of unknown function	721 (19.2%)
No. hypothetical proteins	1,012 (26.9%)
No. of ORFs in paralogous families	1,801 (47.8%)

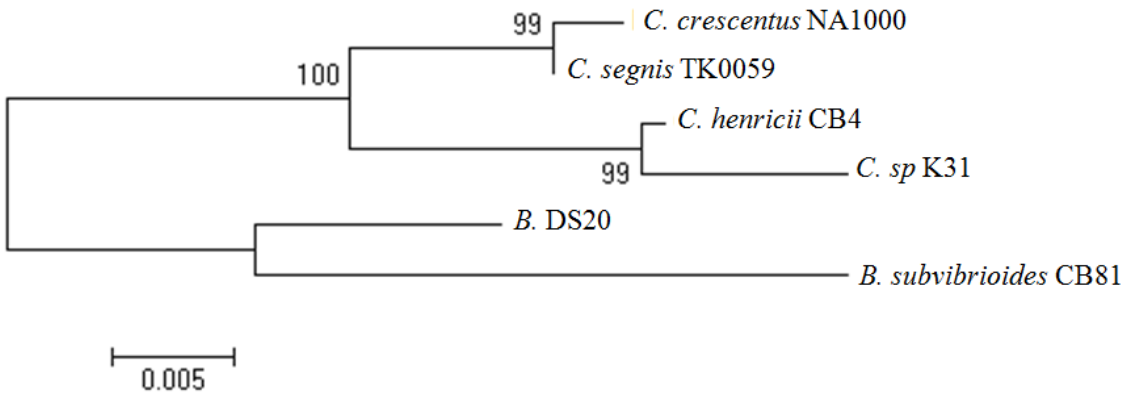

Figure 1.3: A phylogenetic tree based on 16S ribosomal RNA gene sequences of *Caulobacter* isolates

Since bacterial genomes change over time and they exchange genetic material with other species, our laboratory compared the three available *Caulobacter* genomes to determine the level of genome conservation. To our surprise, we found an extremely high

level of genome rearrangements (Figure 1.4). We designated this high level of genome rearrangement “genome scrambling”.

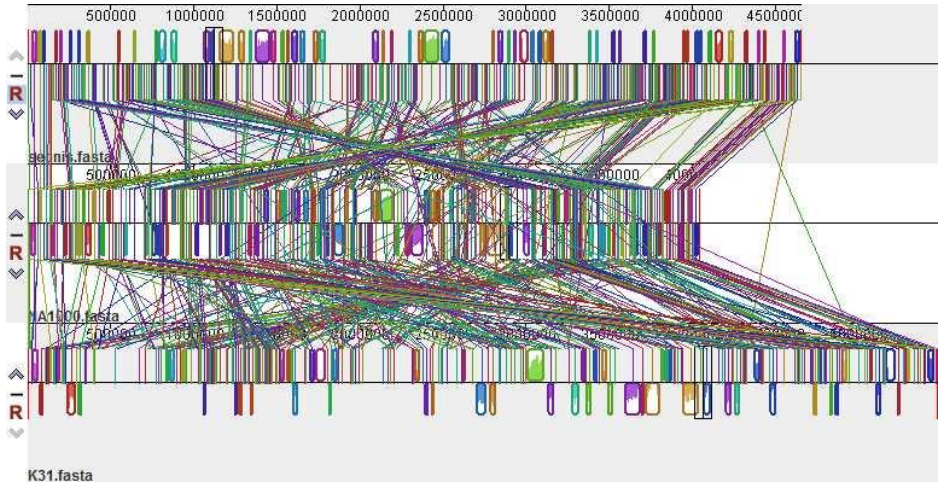

Figure 1.4: A comparison of the *C. segnis* TK0059, *C. crescentus* CB15, and *C. spp.* K31 genome. Ash and Ely, unpublished. Each line represents a rearrangement event.

In contrast, a similar comparison between two closely related enteric bacteria (*Escherichia coli and Salmonella typhi*) from the family *Enterobacteriaceae* revealed very few genome rearrangements (Figure 1.5).

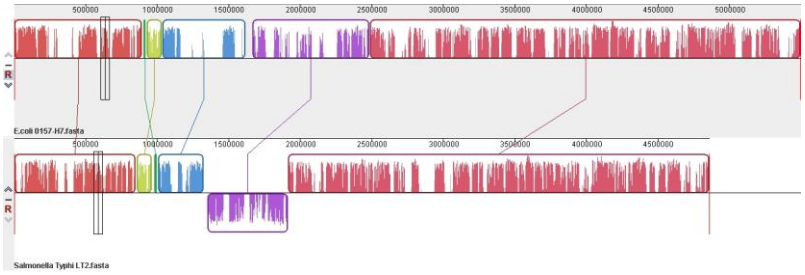

Figure 1.5: A comparison of the *Escherichia coli* and *Salmonella typhi* genomes. Ash and Ely, unpublished. Each line represents a rearrangement event.

To demonstrate that the observed genome scrambling was not due to improper assembly of the *Caulobacter* genomes, we performed a series of pulsed field gel electrophoresis (PFGE) experiments on the CB15 and K31 chromosomes after digestion with rare cutting restriction endonucleases and found a one to one correspondence

between the fragments predicted from the database sequences and the fragments observed in the PFGE gels. Thus, genome scrambling appears to be a real phenomenon in this genus and not an assembly artifact.

tRNA genes and transposable elements are often associated with genome rearrangements. In many cases, they are associated with horizontal gene transfer (HGT) events as well. However, fewer than 25% of *Caulobacter* rearrangements can be attributed to these elements (Ely, unpublished). Thus, it is likely that a careful analysis of *Caulobacter* genome rearrangements would lead to the discovery of new mechanisms that result in genome rearrangements. Furthermore, recent approaches employing codon usage patterns to identify HGT events have been found to greatly overestimate these events while missing a large fraction of the genuine HGT events (Friedman and Ely, 2012). Therefore, a phylogenetic approach may be the only way to identify HGT events with confidence and the availability of conspecific *Caulobacter* genome sequences would allow us to examine this phenomenon in more detail and provide an accurate estimate of the contribution of HGT events to *Caulobacter* genome evolution.

The massive number of rearrangements observed in the *Caulobacter* genome comparisons (Figure 1.4) makes it extremely difficult to pinpoint distinct rearrangement events. Since we observed fewer rearrangements between the two more closely related species (*C. segnis* and *C. crescentus)*, we expect the trend to continue with only a small number of rearrangements observed in a comparison of independent isolates of a single species so that we could identify individual rearrangement events. Once identified, we can examine the rearrangements to see what types of sequences are present at the rearrangement junctions. Therefore, we propose to compare different isolates of the same

species to investigate divergent genetic changes. Similar research has been done in pathogenic bacteria (Wilson, 2012), but genome scrambling has not been observed in these studies.

To meet the need for additional closely-related *Caulobacter* genome sequences, we propose that the *Caulobacter henricii* (CB4) genome would provide a good comparison for the existing *sp*. K31 genome sequence. CB4 is closely related to K31 (Figure 1.3) and we hypothesize that the comparison with CB4 will yield fewer genome rearrangement events so that both ends of individual rearrangement events can be identified. Both bacteria are crescent shaped (Figure 1.6), require riboflavin for growth, and produce a yellow pigment; however, CB4 requires vitamin B12 for growth. They also have different growth rates at optimal conditions. At 30°C, K31 had a doubling time of 160 minutes. The doubling time for CB4 at 30°C is 150 minutes.

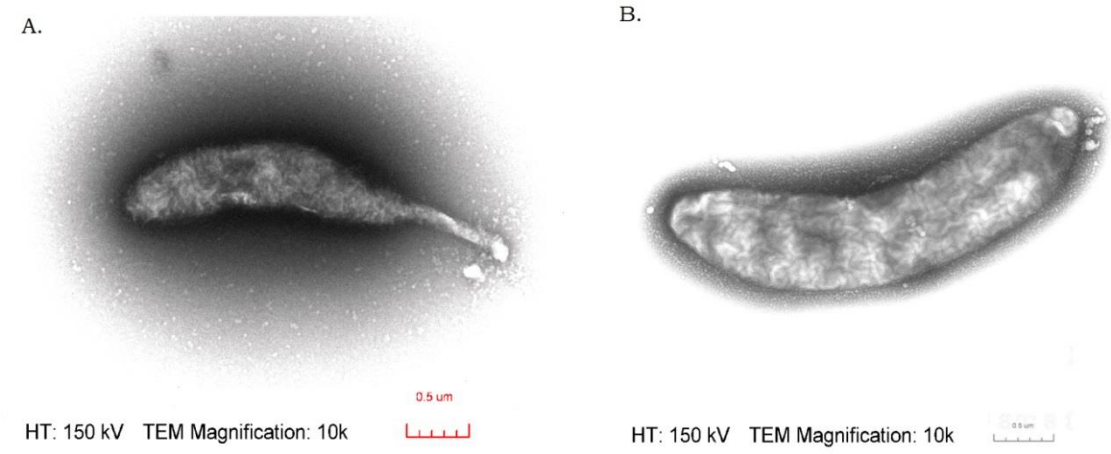

Figure 1.6: **A.** Transmission Electron Microscopy image of CB4 at 4000x Magnification. **B.** Transmission Electron Microscopy image of K31 at 20000x magnification.

To use CB4 in our genome comparison studies, we first had to obtain the DNA sequence of the bacteria, annotate it, and assemble it into one contig. This task previously needed extensive amounts of money, resources, and man power. However, the advent of

better and cheaper DNA sequencing meant that the aim of sequencing and assembling a bacterial genome would be one increasingly attempted by scientists with little to no bioinformatics expertise. This notion led us to compare the available options of sequencing a bacterial genome, providing the pros and cons of each choice, and providing straight forward and budget friendly recommendations that the average scientist could employ.

REFERENCES

Brown, P.J.B., Kysela, D.T., et al. (2011) "Genome sequences of eight morphologically diverse alphaproteobacteria." Journal of Bacteriology, 193(17):4567

Brun, Y. V. and Janakiraman, R. (2000) in *Prokaryotic Development*, eds. Brun, Y. V. & Shimkets, L. J. American Society of Microbiology: pp. 297–317.

Christen, B., E. Abeliuk, JM Collier, et al. (2011). "The essential genome of a bacterium." Molecular Systems Biology 7:528

Ely, B. 1991. Genetics of **Caulobacter crescentus**. Methods Enzymol. 204**:**372-384.

Ely, B., and R. C. Johnson. 1977. Generalized transduction in **Caulobacter crescentus**. Genetics 87**:**391-399.

Fleischmann, R.D., Adams, M.D. et al (1995) "Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd" Science 269(5223): 496-498+507-512

Fredrickson, J. K., J. M. Zachara, et al. (2004). "Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the hanford site, washington state." Applied Environmental Microbiology 70(7): 4230-41.

Friedman, R., and Ely, B. (2012) "Codon usage methods for horizontal gene transfer detection generate an abundance of false positive and false negative results." Current Microbiology Nov; 65(5):639-42.

Henrici, A. T. and D. E. Johnson (1935). "Studies of Freshwater Bacteria: II. Stalked Bacteria, a New Order of Schizomycetes." Journal of Bacteriology 30(1): 61-93.

Hogan, C. M. (2010). "Bacteria." Encyclopedia of Earth Sidney Draggan and C.J. Cleveland, National Council for Science and the Environment, Washington DC.

Houwink, A. L. (1952). "Contamination of electron microscope preparations." Experientia 8: 385.

Jones, M. (1905). "A peculiar microorganism showing rosette formation." Zentr. Bakteriol. Parasitenk. (Abt. II): 14:459-463.

Laub, M. T., McAdams, Het al. (2000) Science 290, 2144–2148.

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

Nierman, W.C., Feldblyum, T.V. (2001) "Complete genome sequence of Caulobacter crescentus" Proceedings of the National Academy of Sciences USA 98(7): 4136–4141

Omeliansky, V. L. (1914). "A new bacillus: Bacillus flagellatus." Omel. Zh. Mikrobiol. Epidemiol. Immunobiol.: 1:24.

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

Shapiro, L. (1976). "Differentiation in the Caulobacter cell cycle." Annual Review of Microbiol 30: 377-407.

Stahl, D. A., Key, R., Flesher, B. & Smit, J. (1992). J Bacteriol 174, 2193-2198.

Urakami, T., Oyanag, H., Arak, H., et al. (1990) "Recharacterization and Emended Description *Mycoplana* and Description of Two New of the Genus Species, *Mycoplana ramosa* and *Mycoplana segnis*." International Journal Of Systematic Bacteriology Pg: 434-442

Whitman, W. B., D. C. Coleman, et al. (1998). "Prokaryotes: the unseen majority." Proceedings of the National Academy of Science USA 95(12): 6578-83.

Wilson, D. J. (2012). "Insights from genomics into bacterial pathogen populations." Public Library of Science Pathogens 8(9): e1002874.

CHAPTER 2

COMPARISON OF GENOME SEQUENCING TECHNOLOGY AND ASSEMBLY METHODS FOR THE ANALYSIS OF A GC-RICH BACTERIAL GENOME

SCOTT, D.C. AND ELY, B. TO BE SUBMITTED TO *CURRENT MICROBIOLOGY*.

ABSTRACT

Motivation: Improvements in technology and decreases in price have made *de novo* bacterial genomic sequencing a reality for many researchers. With the increase of sequencing comes the need to evaluate the most reliable methods in generating a complete and accurate assembly. Certain biases complicate these methods when working with GC-rich genomes.

Results: We sequenced the GC-rich *Caulobacter henricii* on the Illumina MiSeq, Roche 454, and Pacific Biosciences RS II sequencing systems. We performed assemblies using eight readily available programs and found that builds using 2nd generation data produced accurate yet numerous contigs. SPAdes performed the best followed by PANDAseq. Celera Assembler produced a build using 3rd generation data error corrected with 2nd generation data. We duplicated this build using 3rd generation data with HGAP2.0. We authenticated these builds by enzymatic digestion and Pulsed Field Gel Electrophoresis (PFGE) and designated the HGAP2.0 build as the reference.

Availability and Implementation: Software used in this study can be found at

http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.1/
http://www.clcbio.com/products/clc-genomics-workbench/
https://github.com/PacificBiosciences/SMRT-Analysis/wiki/SMRT-Analysis-Release-Notes-v2.1
http://www.genome.umd.edu/masurca.html
http://454.com/contact-us/software-request.asp
https://github.com/neufeld/pandaseq/wiki/Installation
http://www.dnastar.com/
http://bioinf.spbau.ru/spades

Contact: Derrick C. Scott: Scottdc@mailbox.sc.edu

INTRODUCTION

Despite ground breaking advances in the field of prokaryotic biology, there are many unanswered questions left to be studied. Many of these questions cannot begin to be elucidated without the assembly of high quality genome sequences. Larger than most viruses but smaller than most eukaryotic genomes, bacterial genomes have been sequenced to understand pathogen-host interactions, to understand the environment specific evolution of species, and also to trace the source of bacterial related disease outbreaks. For example the Human Microbiome Project, which aims to establish a comprehensive baseline of the microbial diversity at 18 different human body sites, has identified thousands of new microbial strains and has radically increased the number of bacterial genomes that are currently being sequenced (Consortium, 2012). Also, the Wellcome Trust Sanger Institute has recently collaborated with Public Health England to complete the sequences of 3,000 bacterial genome strains from PHE's National Collection of Type Cultures (NCTC) reference.

Second- and third-generation sequencing technology can now generate high quality, fast, high throughput sequencing data. However, there are advantages and disadvantages associated with each individual technology. Things to consider when choosing a technology are read lengths, accuracy, price of sequencing, and the time needed to complete a sequencing run. Pacific Biosciences (www.pacificbiosciences.com) has developed instrumentation that creates unprecedented read lengths of up to 20,000 bp with a 99.9% accuracy rate. Illumina (www.illumina.com) has technology that can routinely generate 600 billion base pairs (GB) in a single run as well as other iterations

that can go from sample to data in as little as 8 hours. A major consideration for all current technologies is cost. In addition to the cost of actual sequence runs, added costs include the cost of sample library preparation. This cost can be greater than the cost of sequencing. As such, many researchers have begun to adopt a strategy of sequencing just a single library while relying on deep coverage of the genome to compensate for the lack of multiple libraries.

Typically, whole genome assembly projects have begun by using a combination of two or more short and long read libraries (Fleischmann, et al., 1995). Short libraries are paired-end reads of relatively short fragments less than 800bp in length. However, if repeated sequences are longer than the read lengths, they cannot be sequenced by short reads. This problem created the need for long fragments that could span the long repeats. In tandem, multiple libraries create much data for next-generation sequencing algorithms and are especially useful for large genomes (Schatz, et al. 2010). Since bacterial genomes tend to be orders of magnitude smaller than eukaryotic genomes and have few repeated sequences, the need for large fragment libraries is lessened. Also, long fragment libraries are difficult to create, leaving small fragment libraries paired with deep coverage technologies such as the Illumina HiSeq or MiSeq an attractive approach of *de novo* sequencing.

While 2nd generation sequencing has been tremendously successful at obtaining robust data used in assembly algorithms, it is not without bias. Bias in sequencing data is extremely detrimental and can lead to inaccurate results and incorrect assemblies. Accuracy at this step is crucial as it irreversibly affects downstream analyses such as functional annotation, comparative genomics (Rappuoli, 2001), single nucleotide

polymorphism (SNP) discovery, and any further application exploiting primary genomic sequence data (Metzker, 2009; Janssens, et al. 2010; de Magalhaes, et al. 2009). One reason for the bias is that 2nd generation sequencing technology depends on polymerase chain reaction (PCR) to amplify the fragments before further analysis. However, guanine and cytosine rich (GC-rich) regions are more difficult to amplify and subsequently sequence. PCR polymerases need single stranded, linear copies of DNA to copy, but GC-rich regions are extremely stable and non-linear due to stacking interactions and secondary structure. This secondary structure leads to bias where regions of high GC content can have extremely sparse coverage compared to that obtained for other regions. This incomplete sampling can lead to gaps that prevent the full genomic assembly of an organism or incorrect base calling which passes off erroneous raw read data as authentic. GC-rich regions also include biologically relevant target sequences, particularly for epigenetics. For example, CpG islands—large clusters of cytosine and guanine combinations near gene promoters—are located in GC-rich regions. CpG islands play important roles in aging, colorectal cancer, retinoblastoma, and other forms of tumorigenesis (Cohen, et al. 2008; Karpinski, et al. 2008; Issa and Toyota, 1999). Thus bias against GC-rich regions could hinder the novel discovery of this functionally important motif in future sequencing projects. It is possible to alleviate the GC-rich bias with deeper coverage of the genome and by using improved protocols (Quail, et al. 2009; Aird, et. Al, 2011), but the answer to truly eliminate this bias lies in 3rd generation sequencing which eliminates the need for fragment PCR amplification, but introduces different disadvantages to the sequencing discussion. Pacific Biosciences employs this type of sequencing with its Single Molecule Real Time sequencing (SMRT). Instead of

cycles of template amplification, the incorporation of dNTPs by the replicating DNA polymerase is observed in real time. Each nucleotide is attached to a fluorescent dye that is cleaved at the moment of incorporation. The base call is made according to the observed fluorescence of the released dye. This method allows for extremely long read lengths but suffers from less accurate base calling as compared to that of 2nd generation technologies.

The cost to sequence a genome has lowered dramatically in responses to technological advances (Jackman and Birol, 2010). Many reviews have been published that assess and compare different strategies for the assembly of genomes and novel metrics have been designed to maximize the quality of the assemblies (Quail, et al. 2012; Loman, et al. 2012; Schatz, et al. 2011, Narzisi, 2011). These studies demonstrate that there is no one size fits all approach to a quality genome assembly. Each researcher has different needs and queries. Also, as sequencing becomes routine, more researchers with little to no experience in bioinformatics and limited access to assembly experts will be attempting the process of assembly. These problems influenced us to compare the efficacy and accuracy of a panel of assembly programs that use input data derived from the GC-rich *Caulobacter henricii* genome sequenced with the Illumina MiSeq benchtop sequencer, Roche GS FLX 454 sequencer, and the Pacific Biosciences RS II DNA Sequencing System. Each software program attempts a *de novo* assembly using a single, short fragment, deep coverage library; a single molecule sequence long-read deep coverage library; or a combination of both.

We used standard FLX chemistry on the Roche 454 system to begin our sequencing project. This system has been used in over 3000 publications

(www.454.com/publications; Finotello, et al. 2011). Next we used version two chemistry with the Illumina MiSeq benchtop sequencer to obtain 2x250bp runs. This instrument is widely used in *de novo* sequencing projects as well as tracking the spread of contagious diseases (Harris, et al. 2012; Smith, et al. 2013; Chen, et al. 2013; Eyre, et al. 2012). Finally, we used the PacBio RS II single molecule sequencing system which can generate long reads and does not amplify the template thereby eliminating low to nonexistent coverage in GC-rich regions. This feature made it an attractive alternative in terms of sequencing a genome that is 66% GC. It has been used to sequence the genomes of mycorrhizal fungus as well as elucidating the methylome of human bacterial pathogens and *Caulobacter* (Powers et al., 2013; Caporaso et al., 2012; Kozdon et al., 2013; Tisserant et al., 2013).

To assemble the sequence data, we compared eight assembly programs ranging from free open access software to proprietary pay for use programs. Previous studies showed that long-read assemblers do not perform well on our data sets (Huang, 2003 & 2006) and short-read assemblers such as Velvet (Zerbino and Birney, 2008) and AbySS (Simpson, et al., 2009) produce builds with a large number of small contigs and a lower reference sequence reconstruction (Kumar, 2010 & Zhang, 2011). As such, these assemblers were not tested. AllPaths-LG was the best performing assembler on large genomes in the GAGE evaluation (Salzberg, et al., 2011), but it requires a minimum of a short and long-read jumping library so we were not able to evaluate its performance in this study either. The free software included the PAired-eND Assembler for DNA sequences (PANDAseq) which is designed to assemble Illumina overlapping pair-end reads; the Maryland Super Read Cabog Assembler (MaSuRCA), which combines the

benefits of deBruijn graph and Overlap-Layout-Consensus assembly approaches; the St. Petersburg genome assembler (SPAdes), intended for use with Illumina reads on both standard isolates and single-cell MDA bacteria assemblies; the Celera Assembler, first designed to assemble Sanger sequencing reads; Newbler, designed exclusively to handle 454 reads but now expanded to support multiple sequencing technologies. It also has a graphical user interface version; and Hierarchal Genome Assembly Process 2.0 (HGAP2), an assembler developed by Pacific Biosciences to automatically error correct and assemble PacBio long-read data. SPAdes support for other technologies is currently in progress. The pay for use programs we employed were SeqMan NGen, a software suite from DNASTAR that is easy to use and includes applications for traditional sequence analysis, all next-generation sequence assembly and analysis, gene expression studies, RNA-Seq, ChIP-Seq, and transcriptome analysis; and CLC Genomics Workbench 6, a comprehensive and user-friendly analysis package for analyzing, comparing, and visualizing next generation sequencing data. Both companies were generous enough to provide us with an extended fully functional trial period. Some software delivered more flexibility by taking advantage of the ability to customize each assembly using command line inputs while the other software offered exceptionally user friendly operation and the lack of requiring extensive knowledge of Linux based coding. The Celera Assembler was also used in the hybrid assembly approach where we utilized high accuracy short reads from Illumina data to error correct the lower accuracy long-reads from PacBio. Our bacterial data set was a novel GC-rich genome that lacked a reference. These characteristics make it an ideal candidate to test GC bias and the ability of our data sets to produce an accurate and contiguous reference.

METHODS

Genome sequencing

The alphaproteobacterium *Caulobacter henricii* (ATCC® 15253™) designated CB4 was ordered from the American Type Culture Collection Genomic. Genomic DNA for **CB4** was prepared with QIAGEN DNeasy Blood and Tissue Kit (www.qiagen.com/~). For PacBio **RSII** sequencing, the library prep template for the 10kb protocol was used but the DNA was sheared for 20kb fragments using a Covaris tube and a final 0.4x bead wash for a finished library. The collection protocols for the P4-C2 chemistry were

Protocol: MagBead Standard Seq v2

Movie Time: 120 min

Insert Size (bp): 20000

Stage Start: True

Control: DNA Control 3kb-10kb.

A 500-bp paired end library for Illumina MiSeq v2 chemistry and 7-kb paired end library for GS-FLX titanium were prepared, and sequencing was performed according to the manufacturers' instructions. The 454 and Illumina sequencing processes were performed using the services of ***EnGenCore*** LLC. The PacBio sequencing was done using the services of University of Washington PacBio Sequencing Services.

Assembler Inputs

In creating the PBcR build using **Celera Assembler 8.0**. We first created a frg file based on our Illumina MiSeq short read fastq data:

fastqToCA -libraryname illumina -insertsize 500 50 -technology illumina-long -mates CB4_S10_L001_R1_001.fastq,CB4_S10_L001_R2_001.fastq > cb4both500.frg

We then used our newly created frg file of short reads to error correct the fastq file of long reads that was obtained from our Pacbio RSII sequencing data:

pacBioToCA -length 500 -partitions 200 -l pacbio -t 16 -s pacbio.spec -fastq filtered_subreads.fastq cb4both500.frg > run.out 2>&1

If there is higher than 25X coverage of long read data, it is recommended to use only 25X of the longest post-correction sequences in the assembly step. To estimate the coverage and average corrected read size we ran:

cat pacbio.log |sort -nk6 |awk '{SUM+=$NF; TOTAL++; } END {print SUM/4000000" "SUM/TOTAL}'

Since we had over 25X, we downsampled to use only the longest 25X. Our genome size is approximately 4 Mbp so we used 105 Mbp in PBcR sequences:

gatekeeper -T -F -o asm.gkpstore pacbio.frg

Followed by:

gatekeeper -dumpfrg -longestlength 0 105000000 asm.gkpstore > pacbio.25x.frg

Lastly, we assembled the 25X of corrected data:

runCA -p asm2 -d asm2 -s asm.spec pacbio.25X.frg > asm2.out 2>&1
Pacbio.spec file specified the parameters:

utgErrorRate = 0.25
utgErrorLimit = 6.5
cnsErrorRate = 0.25

cgwErrorRate = 0.25
ovlErrorRate = 0.25
merSize=14
merylMemory = 128000
merylThreads = 16
ovlStoreMemory = 8192
# grid info
useGrid = 0
scriptOnGrid = 0
frgCorrOnGrid = 0
ovlCorrOnGrid = 0
ovlHashBits = 24
ovlThreads = 2
ovlHashBlockLength = 20000000
ovlRefBlockSize = 50000000
frgCorrThreads = 2
frgCorrBatchSize = 100000
ovlCorrBatchSize = 100000
ovlConcurrency = 6
cnsConcurrency = 16
frgCorrConcurrency = 8
ovlCorrConcurrency = 16
cnsConcurrency = 16
Asm.spec file specified the parameters:
overlapper = ovl
unitigger = bogart
merSize = 14
ovlErrorRate = 0.03
obtErrorRate = 0.03
obtErrorLimit = 4.5
utgErrorRate=0.015
utgGraphErrorRate=0.015
utgGraphErrorLimit=0
utgMergeErrorRate=0.03
utgMergeErrorLimit=0
merylMemory = 32000
merylThreads = 2
ovlStoreMemory = 1192
ovlHashBits=24
ovlThreads = 2
ovlHashBlockLength = 20000000
ovlRefBlockSize = 5000000
frgCorrThreads = 2
frgCorrBatchSize = 100000
ovlCorrBatchSize = 100000
ovlCorrConcurrency = 2

ovlCorrConcurreny=1
ovlConcurrency=1
ovlCorrConcurrency = 2
cnsConcurrency = 2

For **CLC Genomics Workbench 6**, we used the default settings of the de novo assembly tool from the toolbox. For **HGAP 2.0**, we used the default settings in SMRT Analysis 2.1 and an estimated genome size of 4MB. For the pre-assembly, we targeted 30X coverage and had the algorithm compute the minimum seed read length (5418 bp). **MaSuRCA** v2.1.0 ran as runSRCA.pl config.txt followed by bash assemble.sh. The contents of the config.txt file were:

PATHS
JELLYFISH_PATH=/share/apps/MaSuRCA/MaSuRCA-2.1.0/bin
SR_PATH=/share/apps/MaSuRCA/MaSuRCA-2.1.0/bin
CA_PATH=/share/apps/MaSuRCA/MaSuRCA-2.1.0/CA/Linux-amd64/bin
END
DATA
PE= pe 450 50 /home/~/CB4_S10_L001_R1_001.fastq
/home/~/CB4_S10_L001_R2_001.fastq
END
PARAMETERS
GRAPH_KMER_SIZE=auto
USE_LINKING_MATES=1
CA_PARAMETERS = ovlMerSize=30 cgwErrorRate=0.25 ovlMemory=4GB
KMER_COUNT_THRESHOLD = 1
NUM_THREADS= 24
JF_SIZE=200000000
DO_HOMOPOLYMER_TRIM=1
END

Newbler was executed with the following command line:

runAssembly /home~/CB4_S10_L001_R1_001.fna

/home~/CB4_S10_L001_R2_001.fna. **PANDAseq** was executed with the following

command line: pandaseq -f /home~/CB4_S10_L001_R1_001.fastq -r

/home/~/CB4_S10_L001_R2_001.fastq > CB4panda.fasta. **SeqMan NGen 11.2.1** was

run as a de novo assembly specifying Illumina >50nt paired end reads of 500bp. The

expected genome size and coverage were inputted and the options were run at default

settings. **SPAdes** was executed with the command line input

spades.py -1 /home/~/CB4_S10_L001_R1_001.fastq.-2

/home/~/CB4_S10_L001_R2_001.fastq

–k 21,33,55,77,99,127 –careful –o CB4spades

Assembly and evaluation

All assembly metrics were calculated with QUAST. The command was executed with the following input using HGAP2 as the reference:

quast.py --gene-finding --gage -R HGAP2.fasta Celera.fasta CLCGenomics.fasta DNAStar.fasta HGAP2.fasta MaSuRCA.fasta Newbler.fasta PANDAseq.fasta PBcR.fasta SPAdes.fasta

RESULTS

Data

The first data set was produced by the Roche GS FLX 454 system using standard FLX chemistry to sequence the genome of *Caulobacter henricii* strain CB4. The second data set consists of 2x250bp Illumina MiSeq paired-end reads that was obtained using Reagent Kit v2. The third data set was generated using the P4-C2 chemistry of the PacBio RS II system. The data sets were post processed using BLASTn in order to discard contaminating and plasmid sequences.

The Assemblers

Eight genome assemblers were used in this study:

- Celera Assembler 8.0 (Myers et al., 2000)
- CLC Genomics Workbench 6 (CLC Bio)
- HGAP 2.0 (Chin et al., 2013)
- MaSuRCA v2.1.0 (Zimin, et al., 2013)

- Newbler v2.6 (Margulies et al., 2005)
- PANDAseq (Bartram et al., 2011)
- DNAStar SeqMan NGen 11.2.1 (DNASTAR, Madison, WI, USA)
- SPAdes v2.5.1 (Bankevich, et al., 2012)

As some assemblers can only be used with specific data sets, we ran all assemblers for any data set that was compatible. Our 454 data set was assembled using Newbler. Our Pacbio data set was assembled using HGAP 2.0 and polished using Quiver. The MiSeq data set was assembled using Celera, CLC Genomics, SeqMan, MaSuRCA, and SPAdes. We used one hybrid approach as well. We used the Celera error correction module to error correct our long-read Pacbio data set with the high accuracy short-reads of our MiSeq data set then assembled the error-corrected data set using Celera. This assembly and the HGAP2 assembly both generated the same build which we designated as the reference genome. We used the reference along with the NGA50 contig size (if a contig is misassembled with respect to the reference, it is broken down into smaller pieces) to determine which software produced the best assembly.

Depth of Coverage. We used the 250bp reads from the *C. henricii* MiSeq data set which yielded approximately 100X coverage. We used the 600bp reads from the 454 data set which produced approximately 100X coverage. Our Pacbio data set yielded average read lengths of 4289 bp with a coverage of approximately 55X.

The Assemblies

We examined various metrics on the performance of each assembler as described in Magoc, et al., 2013. They are as follows:

•The number of contigs at least 500 bp.

•The total length which is the total number of bases in the assembly.

• N50 size, which is the size of the smallest contig such that 50% of the genome is contained in contigs of size N50 or larger. For example, if the genome size is 4 Mb, the contig N50 size would be computed by adding up the contig sizes from largest to smallest until the cumulative size was greater than 2 Mb. The size of the smallest contig in this set is the N50 size.

• Nx statistics are similar to N50, where the Nx size is the length of the smallest contig such that x% of the genome is contained in the contigs of size Nx or larger.

• Misassemblies, determined by comparison to the reference genome and defined as the sum of the number of relocations, translocations and inversions affecting at least 1000 bp. A relocation is defined as a misjoin in a contig/scaffold such that if the contig/scaffold is split into two pieces at the misjoin, then the left and right pieces map to distinct locations on the reference genome that are separated by at least 1000 bp, or that overlap by at least 1000 bp. A translocation is defined as a misjoin where the left and the right pieces map to different chromosomes or plasmids. An inversion is defined as a misjoin such that the left and the right pieces map to opposite strands on the same chromosome.

• Local errors, defined as misjoins where the left and right pieces map onto the reference genome to distinct locations that are less than 1000 bp apart, or that overlap by less than 1000 bp.

• The number of unaligned contigs, computed as the number of contigs that MUMmer (Delcher, et al., 1999; Delcher, et al., 2002; Kurtz, et al., 2004) was not able to align, even partially, to the reference genome.

• Corrected N50 size (NGA50), defined as the N50 size obtained after splitting contigs/scaffolds at each error. Note that local errors were not used for the purpose of calculating corrected N50 values.

• The Genome fraction which is the fraction of the reference genome covered by contigs/scaffolds.

• Duplication ratio, an approximation of the amount of overlaps among contigs/scaffolds that should have been merged. Failure to merge overlaps leads to overestimation of the genome size and creates two copies of sequences that exist in just one copy.

All metrics were calculated using the Quality ASsement Tool for genome assembly (Gurevich, et al., 2013)

Generation of the Reference Genome

One of our main goals in this research was to generate a finished genome with no gaps. We were able to accomplish this goal using two different methods. As described in Koren, 2012, we used an approach that utilized the short, high-accuracy sequences of MiSeq to correct the error inherent in the long, single-molecule sequence reads generated by the Pacbio RS II using different modules found in the Celera Assembler 8.0. The corrected "hybrid" PBcR (PacBio corrected Reads) were then assembled *de novo* into 2 contigs consisting of a 3,870,958 bp contig and a 100,699 bp plasmid (3,971,657 total bp). We also used the HGAP 2.0 assembler that self-corrected the PacBio long-reads to create a draft assembly. This draft was then polished with Quiver to generate a more highly accurate consensus sequence. It produced 2 contigs, the first being 3,868,732 bp and a 97,894 bp plasmid (3,966,626 total bp). The plasmid sequences in all builds were easily identified through comparison with the reference and BLASTn and were

subsequently removed in all downstream analyses. We used Mauve (Darling, et al., 2010) to compare the two assemblies and discovered that at 99.99999879% similarity, they were virtually identical to each other (Figure 2.1). We determined the extra base pairs from the PBcR were simply repeats of the ends of the genome reinforcing its circular nature.

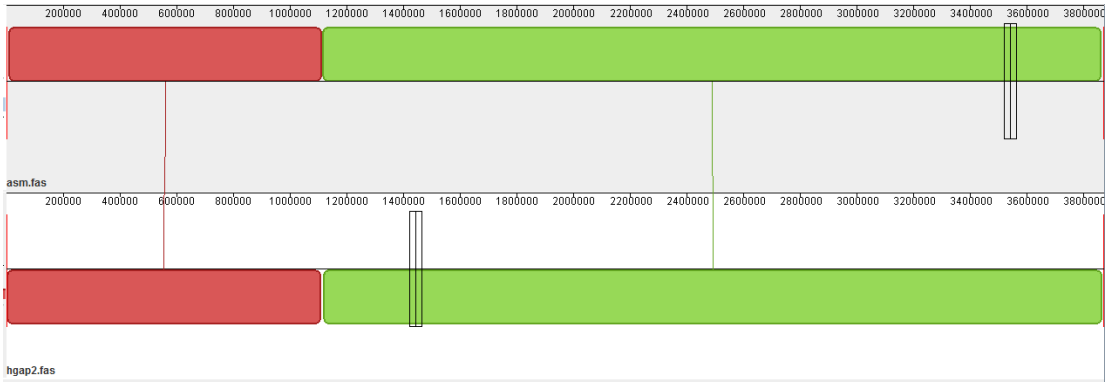

Figure 2.1: Mauve visualization of two assemblies. The top bar represents the PBcR assembly. The bottom bar represents the HGAP 2.0 assembly. The colors represent identical DNA sequence that is shared between the two builds.

We tested the accuracy of each build by downloading the consensus of each assembly into the Webcutter 2.0 (bio.biomedicine.gu.se/cutter2/) program (Figure 2.1) and generating a theoretical digest using the *Sna*BI enzyme which cut the genomic sequence 15 times.

Enzyme name	No. cuts	Positions of sites	Recognition sequence
PmeI	4	1506426 2579389 2730581 3161448	gttt/aaac
SnaBI	15	750053 1221895 1312727 1536306 1607656 1830807 2102422 2339210 2344120 2858345 2935897 2964776 3369596 3621757 3804953	tac/gta
SwaI	4	740929 2230691 2449696 3296931	attt/aaat

Figure 2.2: Positions of reported of HGAP2 cut sites after Webcutter 2.0 analysis

This digest would produce moderate to large fragments of the genomic DNA that could be easily identified via Pulsed Field Gel Electrophoreses (PFGE) (Figure 2.3). We also predicted the *Pme*I and *Swa*I digestion patterns which both cut the genome 4 times and would confirm the legitimacy of the assembly in conjunction with the *Sna*BI data (Figures 2.2 and 2.3).

	PmeI	**SnaBI**	**SwaI**
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	**3,868,732**	**3,868,732**	**3,868,732**

Figure 2.3: Predicted fragment sizes of HGAP2 build after enzymatic digestion

When the *C. henricii* DNA was digested with each of the restriction enzymes and the resulting fragments were resolved by PFGE, there was a one to one correspondence

between the bands observed on the gel and those predicted from the assembled nucleotide sequence (Figure 2.4). These data indicate that these genome assemblies matched the organization of the actual *C. henricii* chromosome. We decided to use the HGAP 2.0 assembly as our reference based on the fact that it used the default settings in SMRT Analysis 2.1 to generate this build and would be easier to duplicate as opposed to the PBcR assembly which used many steps to achieve the final output. Research also showed that PacBio consensus accuracy always exceeded that of the second-generation sequencing data and consistently matched or exceeded the quality of both short-read and hybrid assemblies (Koren et al., 2013). However, we found that at both "ends" of the HGAP2 assembly there were fragmented proteins due to misaligned reading frames

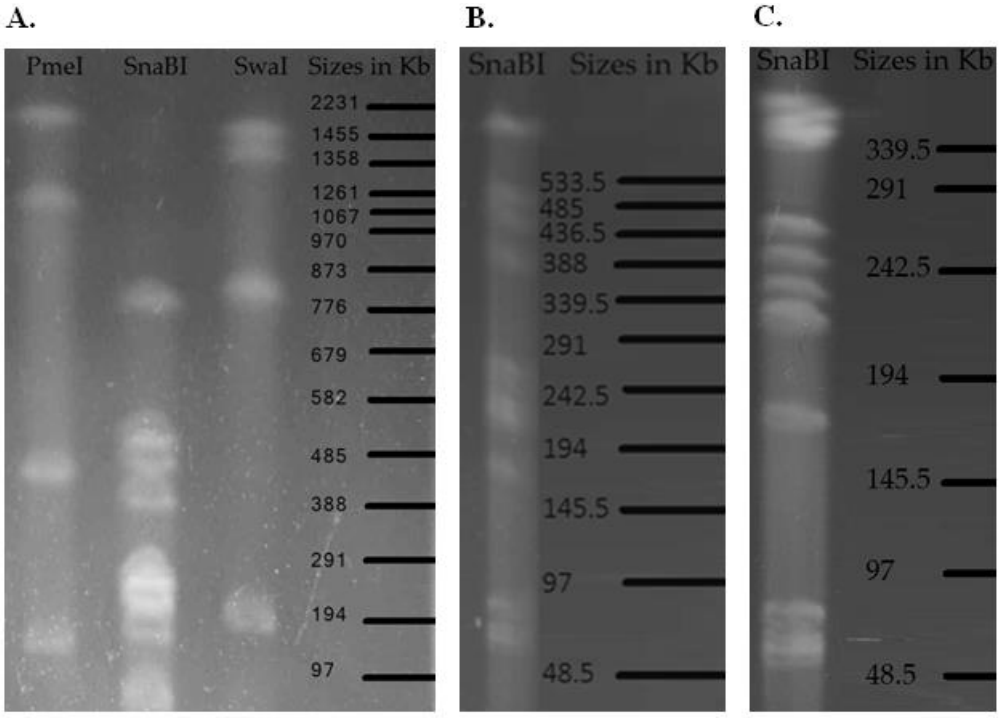

Figure 2.4: *Caulobacter henricii* CB4 genomic DNA digested with *Pme*I, *Sna*BI, and *Swa*I followed by PFGE under varying conditions. A. 6 V/cm for 16 hours. Switch times ramped from 20-120 seconds. B. 6 V/cm for 16 hours. Switch times ramped from 1-45 seconds. C. 6 V/cm for 16 hours. Switch times ramped from 10-20 seconds. All agarose

gels were 1% in SBA buffer and ran at a temperature of 14 degrees Celsius. The black lines indicate the positions and sizes of the fragments generated by a lambda DNA reference ladder.

Comparison of Assemblies

Using the HGAP 2.0 build as a reference, we computed the NGA50 (corrected N50) sizes of our assemblies. NGA50 values convey more information about a build because the program breaks the misassembled contigs at perceived misjoins to provide a superior gauge of assembly quality. If an assembler incorrectly merges two contigs, then this results in a larger N50 size. Since N50 is often used to determine how well an assembler performed, these incorrect builds appear to be better than they actually are. Using the 454 data set of the *C. henricii* CB4 genome, Newbler generated 69 contigs with a N50 and NGA50 value of 128 Kb and a genome fraction of 99.721% (Table 2.1). The combined length of all 69 contigs was 3,950,077 bp.

Using the MiSeq data set, SPAdes generated the assembly with the highest N50 and NGA50 scores of 849 Kb and 720 Kb respectively. PANDAseq was next with a N50 and NGA50 value of 349 Kb. It also generated the build with the fewest contigs while DNAStar produced the build with the largest total genome length at 3,954,246 bp. All assemblies displayed genome fraction percentages of over 99.4 with the MaSuRCA build reaching 99.981%.

Table 2.1: Comparison of assembler builds using the HGAP2 assembly as the reference. NGA50 values are boxed in red. The data were generated with QUAST.

Statistics without reference	Celera	CLCGenomics	DNAStar	HGAP2	MaSuRCA	Newbler	PANDAseq	PBcR	SPAdes
# contigs	210	119	78	1	60	69	27	1	28
Largest contig	96 335	228 321	512 281	3 868 732	501 495	414 950	683 332	3 870 958	1 717 074
Total length	3 885 508	3 872 940	3 954 246	3 868 732	3 931 679	3 950 077	3 954 266	3 870 958	3 875 493
N50	31 035	68 799	311 910	3 868 732	283 078	128 030	349 035	3 870 958	849 521
Misassemblies									
# misassemblies	0	2	1	0	3	1	1	2	1
Misassembled contigs length	0	148 706	428 086	0	632 172	211 829	523 614	3 870 958	1 717 074
Mismatches									
# mismatches per 100 kbp	0.65	3.81	0.93	0	25.08	0.41	0.23	0.36	0.91
# indels per 100 kbp	0.52	1.27	0.91	0	1.65	1.74	0.91	0.85	0.91
# N's per 100 kbp	0.03	0	0.13	0	0	0.99	0	0	0
Genome statistics									
Genome fraction (%)	99.393	99.704	99.836	100	99.981	99.721	99.729	100	99.834
Duplication ratio	1.01	1.002	1.003	1	1.017	1.006	1	1.002	1
NGA50	31 035	66 099	262 235	3 868 732	217 552	127 991	349 035	2 754 062	720 050
Predicted genes									
# predicted genes (unique)	3843	3743	3750	3643	3720	3745	3764	3639	3696
# predicted genes (>= 0 bp)	3843	3743	3783	3644	3744	3745	3764	3643	3697
# predicted genes (>= 300 bp)	3427	3370	3411	3313	3380	3389	3405	3314	3342
# predicted genes (>= 1500 bp)	536	552	558	557	561	574	575	558	562
# predicted genes (>= 3000 bp)	40	43	46	46	44	45	48	47	46

DISCUSSION

We compared the efficacy and accuracy of a panel of assemblers on a 66% GC bacterial genome consisting of input data derived from next generation sequencing technologies. In terms of 2nd generation data, the SPAdes assembler had the largest contig sizes in terms of N50 and NGA50 as compared to the other assemblers. The build generated from PANDAseq created the second best results. As expected, all the assemblies were improved when the data was mapped to the reference genome. The MaSuRCA assembly generated only one gap in reference coverage and a 99.981% genome fraction. SPAdes managed to leave two gaps in coverage while DNASTAR produced five gaps in coverage.

In terms of assembly errors, the Celera Assembler produced none although no assembler had more than three. However, the Celera Assembler performed the poorest in terms of the number of contigs and N50 scores. This is unsurprising as Celera utilizes the Overlap-Layout-Consensus method of contig generation which favors long-read input. The DNASTAR assembly had the largest number of uncalled bases per 100 kbp with a

score of 0.99 while CLC Genomics, MaSuRCA, and PANDAseq had the lowest with the score of zero.

The number of misjoined contigs did not greatly reduce the NGA50 values but we did find false detection of errors in some builds. The *C. henricii* genome is circular and in some instances, a contig started at the end or beginning of the reference and "wrapped around" to the other end of the reference. This resulted in that contig falsely being labeled as misjoined. Such was the case with the PBcR and SPAdes assembly (Figure 2.5).

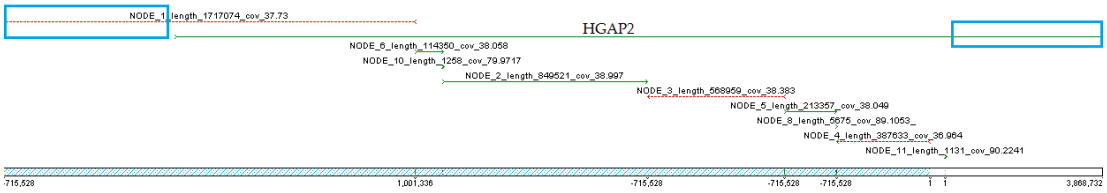

Figure 2.5: Mapping of SPAdes contigs to the HGAP2 reference genome. The blue boxes represent identical DNA sequences. Node 1 of SPAdes build was labeled as misjoined in QUAST analysis even though it is identical to reference sequence.

With an average genomic GC content of 65.72% and some genomic regions reaching 80% GC, the *C. crescentus* CB4 genome was a good choice to test the performance of assemblies where GC bias was a problem. Interestingly, each build covers at least 99% of the reference suggesting that the reason for multiple contigs and unfinished assemblies was not incomplete coverage from GC bias, but the ability of the algorithms to process and reconcile repetitive regions. In the SPAdes build, for example, it produced twenty-eight contigs after assembly but two contigs when aligned to the reference. We analyzed the ends of contigs that should have aligned but discovered that they contained sequences that were repeated at the ends of multiple contigs found in the assembly. Since there were multiple ways these contigs could be assembled, they could not be assembled further.

These issues were addressed in the builds of the 3rd generation PacBio RS II data. With a mean read length of 4,289 bp and maximum read lengths approaching 20,000 bp, these repetitive regions could easily be resolved. The weakness of this technology has traditionally been its low accuracy but this problem has been addressed with assemblers such as Celera that includes steps to error correct these reads with high accuracy short reads. Recently, PacBio developed the HGAP 2.0 assembler that self-corrects these errors and further polishes the consensus with Quiver to produce a result that is equal to the Celera correction method. Thus the short read data is no longer necessary since we achieved a complete and accurate (QV 39=99.987% accuracy) genome assembly using only PacBio data.

Overall, we conclude that the latest genome assemblers can produce very good yet incomplete *de novo* assemblies using single, deep coverage, short-read libraries of 2nd generation sequencers. However, these assemblers are limited by repetitive regions that can be difficult to resolve with the short-reads of these libraries. This result verifies the findings that repeated sequence in the genome induces complexity and poses the greatest challenge to all assembly algorithms (Phillippy, et al., 2008). Therefore, consistent with Koren, et al., 2013 and Shin, et al., 2013, we suggest that the simplest and most effective way to produce a *de novo* GC-rich bacterial genome assembly is with PacBio RS II long-reads using HGAP 2.0 assembler to self-correct the reads. However, we were able to complete a reference genome by using only one SMRT cell. This negates the need for multiple libraries and decreases the cost of a sequencing project.

ACKNOWLEDGEMENTS

This work was funded in part by a fellowship from The Southern Region Educational Board (SREB) to DS and NIH grant GM076277 to BE. We would like to thank Nicole Rapicavoli at Pacific Biosciences for her assistance with the HGAP2 assembly, Alexey Gurevich and Anton Korobeynikov at the Algorithmic Biology Lab, St. Petersburg, Russia for their support with the SPAdes and QUAST programs, and special thanks to Nathan Elger and Paul Sagona who are a part of the Research Cyberinfrastructure at The University of South Carolina.

REFERENCES

Aird, D., et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, *Genome Biol*, 12, R18.

Bankevich, A., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J Comput Biol*, 19, 455-477.

Bartram, A.K., et al. (2011) Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads, *Applied and Environmental Microbiology*, 77, 3846-3852.

bio.biomedicine.gu.se/cutter2/

Caporaso, J.G., et al. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms, *ISME J*, 6, 1621-1624.

Chen, E., et al. (2013) Human infection with avian influenza A(H7N9) virus re-emerges in China in winter 2013, *Euro Surveill*, 18.

Chin, C.S., et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat Methods*, 10, 563-569.

CLC Bio

Cohen, Y., et al. (2008) Hypermethylation of CpG island loci of multiple tumor suppressor genes in retinoblastoma, *Exp Eye Res*, 86, 201-206.

Consortium, T.H.M.P. (2012) A framework for human microbiome research, *Nature*, 486, 215-221.

Consortium, T.H.M.P. (2012) Structure, function and diversity of the healthy human microbiome, *Nature*, 486, 207-214.

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

de Magalhaes, J.P., et al. (2009) Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions, *Ageing Res Rev*, 9, 315-323.

Delcher, A.L., et al. (1999) Alignment of whole genomes, *Nucleic Acids Res*, 27, 2369-2376.

Delcher, A.L., et al. (2002) Fast algorithms for large-scale genome alignment and comparison, *Nucleic Acids Res*, 30, 2478-2483.

DNASTAR, Madison, WI, USA

Eyre, D.W et al. (2012) A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance, *BMJ Open*, 2.

Finotello, F., et al. (2011) Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data, *Brief Bioinform*, 13, 269-280.

Fleischmann, R.D., et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science*, 269, 496-512.

Gurevich, A., et al. (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, 29, 1072-1075.

Harris, S.R., et al. (2012) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study, *Lancet Infect Dis*, 13, 130-136.

Huang, X., et al. (2003) PCAP: a whole-genome assembly program, *Genome Res*, 13, 2164-2170.

Huang X, Y.S., et al. (2006) Application of a superword array in genome assembly, *Nucleic Acids Research*, 34, 201-205.

Jackman, S.D. and Birol, I. (2010) Assembling genomes using short-read sequencing technology, *Genome Biol*, 11, 202.

Janssens, W., et al. (2010) Genomic copy number determines functional expression of {beta}-defensin 2 in airway epithelial cells and associates with chronic obstructive pulmonary disease, *Am J Respir Crit Care Med*, 182, 163-169.

Karpinski, P., et al. (2008) The CpG island methylator phenotype correlates with long-range epigenetic silencing in colorectal cancer, *Mol Cancer Res*, 6, 585-591.

Koren, S., et al. (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing, *Genome Biology*, 14, R101.

Koren, S., et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads, *Nature Biotechnology*, 30, 693-700.

Kozdon, J.B., et al. (2013) Global methylation state at base-pair resolution of the Caulobacter genome throughout the cell cycle, *Proc Natl Acad Sci U S A*, 110, E4658-4667.

Kumar, S. and Blaxter, M.L. (2010) Comparing de novo assemblers for 454 transcriptome data, *BMC Genomics*, 11, 571.

Kurtz, S., et al. (2004) Versatile and open software for comparing large genomes, *Genome Biol*, 5, R12.

Loman, N.J., et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms, *Nat Biotechnol*, 30, 434-439.

Magoc, et al. (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms, *Bioinformatics*, 29, 1718-1725.

Margulies, M., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, 437, 376-380.

Metzker, M.L. (2009) Sequencing technologies - the next generation, *Nat Rev Genet*, 11, 31-46.

Myers, E.W., et al. (2000) A whole-genome assembly of Drosophila, Science, 287, 2196-2204.

Narzisi, G. and Mishra, B. (2011) Comparing de novo genome assembly: the long and short of it, *PLoS One*, 6, e19175.

Phillippy, A.M., et al. (2008) Genome assembly forensics: finding the elusive mis-assembly, *Genome Biol*, 9, R55.

Powers, J.G., et al. (2013) Efficient and accurate whole genome assembly and methylome profiling of E. coli, *BMC Genomics*, 14, 675.

Quail, M.A., et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, 13, 341.

Quail, M.A., et al. (2009) Improved protocols for the illumina genome analyzer sequencing system, *Curr Protoc Hum Genet*, Chapter 18, Unit 18 12.

Rappuoli, R. (2001) Reverse vaccinology, a genome-based approach to vaccine development, *Vaccine*, 19, 2688-2691.

Salzberg, S.L., et al. (2011) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res*, 22, 557-567.

Schatz, et al. (2010) Assembly of large genomes using second-generation sequencing, *Genome Res*, 20, 1165-1173.

Schatz, et al. (2011) Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies, *Brief Bioinform*, 14, 213-224.

Smith, D., et al. (2013) The Hospital Microbiome Project: Meeting Report for the 1st Hospital Microbiome Project Workshop on sampling design and building science measurements, Chicago, USA, June 7th-8th 2012, *Stand Genomic Sci*, 8, 112-117.

Simpson, J.T., et al. (2009) ABySS: a parallel assembler for short read sequence data, *Genome Res*, 19, 1117-1123.

Tisserant, E., et al. (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis, *Proc Natl Acad Sci U S A*, 110, 20117-20122.

Toyota, M. and Issa, J.P. (1999) CpG island methylator phenotypes in aging and cancer, *Semin Cancer Biol*, 9, 349-357.

www.454.com/publications

www.illumina.com

www.pacificbiosciences.com

www.qiagen.com/products/catalog/sample-technologies/dna-sample-technologies/genomic-dna/dneasy-blood-and-tissue-kit

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res*, 18, 821-829.

Zhang, W., et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies, *PLoS One*, 6, e17915.

Zimin, A.V et al. (2013) The MaSuRCA genome assembler, *Bioinformatics*, 29, 2669-2677.

CHAPTER 3

Genome Rearrangement in *Caulobacters* Do Not Affect the Essential Genome

SCOTT, D.C. AND ELY, B. TO BE SUBMITTED TO *CURRENT MICROBIOLOGY*.

ABSTRACT

Bacteria in the *Caulobacter* genus are oligotrophic, found in various diverse habitats, and flourish in conditions with minimal available nutrients. K31 is a novel *Caulobacter* which was isolated from a research station in Finland. When its genome was compared to that of *C. crescentus* NA1000, numerous genome rearrangements were observed. Similar experiments in other closely related bacteria revealed nominal rearrangements. A phylogenetic analysis of 16S rRNA indicates that K31 is more closely related to *Caulobacter henricii* CB4 ATC15253 than other known *Caulobacters*. We sequenced CB4 and compared the genomes of all available *Caulobacters* to study rearrangements, discern the conservation of the NA1000 essential genome, and address concerns of using 16S rRNA in species grouping *Caulobacter* species. We also sequenced *Brevundimonas* DS20, a novel relative of *Caulobacter* and used it as part of an outgroup for phylogenetic comparisons. We expected to find that there would be fewer rearrangements when comparing more closely related *Caulobacters*. However we found that relatedness had no impact on the observed "genome scrambling". We also discovered that the essential genomes of the *Caulobacters* are similar but not identical. Some genes were only found in NA1000, some were missing in a combination of one or more species, and some proteins were 100% identical across species. Also, phylogenetic comparisons of highly conserved regions reveal clades similar to 16S rRNA-based phylogenies, suggesting that 16S rRNA comparisons are a good method to group *Caulobacters*.

INTRODUCTION

Alphaproteobacteria comprise a large and metabolically diverse group that includes the genus *Caulobacter*. Caulobacters are found in essentially all habitats ranging from fresh and salt water, soil, root systems, and water treatment plants. They thrive in low nutrient conditions and generally share the same phenotypic properties. Caulobacters are Gram-negative bacteria that have piqued the interest of microbiologists for several decades. They exhibit a rare dimorphic phenotype consisting of a stalked non-motile cell and a motile swarmer cell produced at cell division. The motile cell is immature and must first shed its flagellum and differentiate into the stalked form before it replicates its chromosome and divides asymmetrically to regenerate itself and produce a flagellated daughter cell, thus continuing its life cycle. The ability to synchronize this cell cycle has allowed great advancement in comprehending the genetic regulatory network and signal transduction pathway controlling the *C. crescentus* cell cycle (Holtzendorff *et al.*, 2004; Wang *et al.*, 1993).

Compared to the wealth of information available to support cell cycle research, the amount of research dedicated to understanding the environmental and evolutionary biology of Caulobacters is minimal. However, as the genomic sequences of more Caulobacters are becoming available, a significant opportunity has arisen to add to this literature. Found abundantly in virtually all habitats, the study of these bacteria can potentially enhance our understanding of the molecular and genetic adaptations of microbes from varying environmental niches.

Ribosomal RNA analyses show that bacteria previously defined as *Caulobacter* are actually grouped into two separate branches consisting of freshwater and marine

species, *Caulobacter* and *Maricaulis*, respectively (Abraham *et al*., 1999; Stahl *et al*., 1992). Further 16S rRNA comparisons by Abraham *et al*. (1999) revealed that the freshwater branch is clearly defined into two species, *Caulobacter* and *Brevundimonas* (Figure 3.1). Thus *Brevundimonas* genomes are ideal for use as an outgroup for the analysis of *Caulobacter* genomes. The genus *Caulobacter* can be divided into branches as well based on their 16S rRNA gene sequences (Abraham *et al.,* 1999). One branch contains *C. crescentus* and *C. segnis* while the other contains *C. henricii* and *Caulobacter sp*. K31. This separation influenced us to compare these genomes to the essential genome that has been experimentally defined for the *C. crescentus* strain NA1000 (Christen *et al*., 2011). The genomic DNA sequences of both *C. crescentus* strain CB15 and its derivative NA1000 (Nierman *et al*., 2001; Marks *et al*., 2010) and *C. segnis* strain TK0059 have been published (Brown *et al*. 2011; Patel *et al*., submitted for publication). In addition, *Caulobacter* strain K31, a groundwater isolate of particular interest for its ability to tolerate and degrade chlorophenols (Mannisto *et al*., 1999), has also had its sequence elucidated (Ash *et al*., submitted for publication). To provide a fourth strain for this genome comparison, the genome nucleotide sequence of *C. henricii* strain CB4 was determined as part of this study. Although many other *Caulobacter* isolates are listed in the IMG genome database (img.jgi.doe.gov), no other *Caulobacter* genome sequences have been fully assembled. Similarly, *Brevundimonas subvibrioides* strain CB81 is the only *Brevundimonas* with an available genome sequence (Lucas *et al*., 2010). Therefore, we have determined the nucleotide sequence of the *Brevundimonas* DS20 genome to provide a second *Brevundimonas* genome. Even though the sequences of the 16S rRNA genes in the four *Caulobacter* species vary by no more than 3% from each other and less

than 7% from the corresponding *Brevundimonas* sequences, we found extensive genome rearrangements among these six genomes.

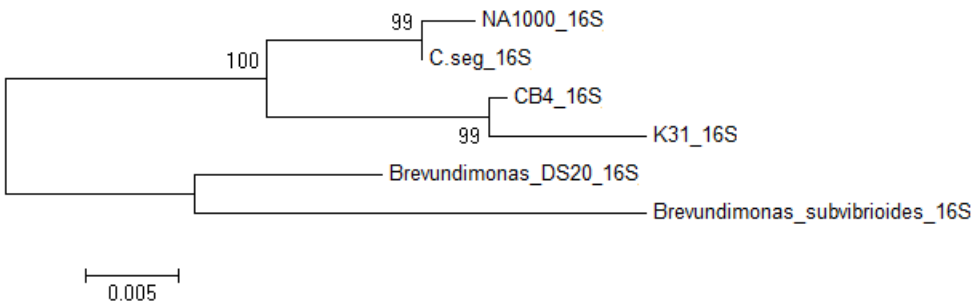

Figure 3.1: Phylogenetic tree indicating the relationship between NA1000, *C. segnis*, K31, and CB4 based on 16S rRNA gene sequence analysis. The tree was constructed using the maximum likelihood method (Felsenstein, 1981). *Brevundimonas subvibrioides* and *Brevundimonas* DS20 were used as outgroup taxa. All branches were recovered in both neighbor-joining and maximum-parsimony trees (Fitch, 1971; Saitou & Nei, 1987). Bootstrap values greater than 50% are given at branch points. Bar, 0.005 substitutions per nucleotide position.

METHODS

Media and Growth conditions. The *Caulobacter henricii* strain CB4 (ATCC 15253) was obtained from the American Type Culture Collection. It was grown at 30°C for 48 hours in PYE medium (Johnson and Ely 1977) that contains 2 g Bacto Peptone, 1 g Yeast Extract, 0.5M $MgSO_4$, and 0.5M $CaCl_2$ per L. We found that CB4 would grow on minimal media glucose plates (Johnson and Ely 1977), in the presence of vitamin B12. At 30°C, CB4 had a doubling time of 190 minutes. In addition it had a doubling time of 2 weeks at 10°C which is close to the average temperature of the groundwater where it was found. The bacterial cells appeared healthy and highly motile at the lower temperature at this slower growth rate.

In addition, we isolated a *Caulobacter*-like bacterium from a contaminated culture of *Caulobacter* FWC20. It was grown at 30°C under the same conditions as CB4. Based

on genome comparisons, we determined this bacterium to be a novel member of the genus *Brevundimonas*. We named this isolate *Brevundimonas* DS20. It had a doubling time of 12 days at 10℃ and 120 minutes at 30℃.

Genome sequence determination and annotation.

Genomic DNA from *Caulobacter henricii* CB4 and *Brevundimonas* DS20 was isolated from a saturated PYE culture using the Qiagen DNeasy Tissue Kit following the manufacturer's protocol. Primers were used to amplify the 16S rRNA region of the genome and the amplified DNA was sequenced at Selah Genomics using Sanger Sequencing on an ABI 3730. BLAST comparisons of the resulting sequences to those in the NCBI GenBank database were used to confirm species identification. The DS20 16S rRNA sequence matched closely to multiple *Brevundimonas* species but was not identical to that of any species present in the database. Genomic DNA library construction and nucleotide sequencing were carried out by the University of Washington Pacbio Sequencing Services using the Pacific Biosciences RSII sequencing system. The library prep template for the 10kb protocol was used but the DNA was sheared for 20kb fragments using a Covaris tube and a final 0.4x bead wash for a finished library. The collection protocols for the P4-C2 chemistry were

Protocol: MagBead Standard Seq v2

Movie Time: 120 min

Insert Size (bp): 20000

Stage Start: True

Control: DNA Control 3kb-10kb**.**

The reads were assembled using HGAP2.0 and each build was verified using Pulsed Field Gel Electrophoresis (PFGE) to separate DNA cut with the *Pme*I and *Swa*I restriction enzymes using the protocols described by Ely and Gerardot (1988). Annotation was performed using The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST) (Overbeek et al., 2014).

Genome comparisons.

Whole genome comparisons were performed using *MAUVE* Multiple Genome Alignment (Aaron *et al*, 2004). A BLAST comparison was performed using the BlastStation version 1.3 software (www.blaststation.com) to identify homologs of the 480 essential NA1000 genes (Christen *et al*. 2011) in the protein coding sequences of all the CDS regions of the chromosomes for genomes of strains CB4, TK0059, and K31. *Brevundimonas subvibrioides* CB81 and *Brevundimonas* DS20 were used as a closely related outgroup. BLAST matches with an e-value that was less than e^{-5} were considered significant. Phylogenetic and molecular evolutionary analyses were conducted using *MEGA* version 5.1 (Tamura, Dudley, Nei, and Kumar 2007). Genome comparisons were based on 16S rRNA gene sequence, 23S rRNA gene sequence, ITS region gene sequence, a highly conserved ribosomal protein operon, the highly conserved dcw cluster, and a highly conserved prophage region. Phylogenetic trees were constructed using the maximum likelihood method (Felsenstein, 1981). All branches were recovered in both neighbor-joining and maximum-parsimony trees (Fitch, 1971; Saitou & Nei, 1987).

RESULTS AND DISCUSSION

Genome overview

The *Caulobacter henricii* CB4 genome consists of a 3,868,732 bp chromosome and a 97,894 bp plasmid. It contains two identical rRNA operons that are separated by 787,533 bp. The plasmid has a GC content of 65% and contains one integrase gene but no transposases. Since the K31 genome contains two megaplasmids (Ash *et al*. submitted), we compared the predicted amino acid sequences of the CB4 plasmid genes to those of the K31 plasmids and found that only four of the plasmid genes are homologous to any of the genes in either of the two K31 plasmids. The CB4 chromosome contains 3751 genes and has a GC content of 66%. As such, the codons in the protein coding regions should have a high G+C content, especially in the third codon position (GC3). Indeed, 29 of the 30 most used codons contain either a G or a C in the third position. The overall GC3 percentage for CB4 is 83.2%. CB4 has been characterized and described previously (Pointdexter, 1964).

The *Brevundimonas* DS20 genome consists of 3,487,386 bp and does not include a plasmid. It has a GC content of 67% and contains 3479 genes with two identical rRNA operons that are separated by 223,973 bp. In DS20, 29 of the 30 most used codons contain either a G or a C in the third position. The overall GC3 percentage is 86.3%. DS20 forms yellow mucoidal colonies which are round, smooth, slightly raised, and glistening. The cells are rod shaped and lack the curved phenotype found in many Caulobacters (Figure 3.2).

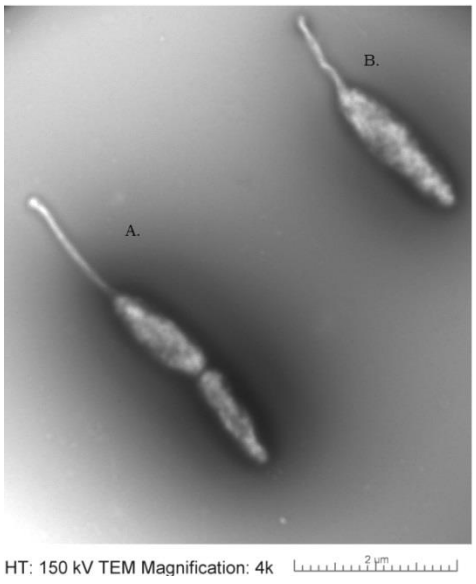

Figure 3.2: Transmission Electron Microscopy image of *Brevundimonas* DS20 at 4000x magnification. A predivisional cell (A) can be seen just prior to the completion of binary fission that would produce a stalked cell and a swarmer cell. A stalked cell (B) is also present.

Genome Rearrangements

The genomes of closely related species are usually very similar to one another except for where inversions have occurred. Genome comparisons in various bacteria have found that in the cases where a rearrangement has occurred, it is usually found near an insertion sequence (Beare *et al*. 2009) or next to an rRNA operon (Darling *et al*. 2008). However, Ash *et al*. (submitted) have demonstrated that a comparison of NA1000 and K31 reveals rearrangements an order of magnitude greater than previously described in other bacteria. When the K31 chromosome was aligned to that of NA1000, more than 60 inversions and 45 large translocations were readily observed (Ash *et al*., submitted). Since this level of genome scrambling makes it difficult to identify the endpoints of individual inversion events, we hypothesized that the level of observed genome scrambling would decrease in comparisons of more closely related genomes. When the *C.*

segnis genome was compared to the NA1000 genome, only 35 inversions and 11 translocations were identified (Figure 3.3).

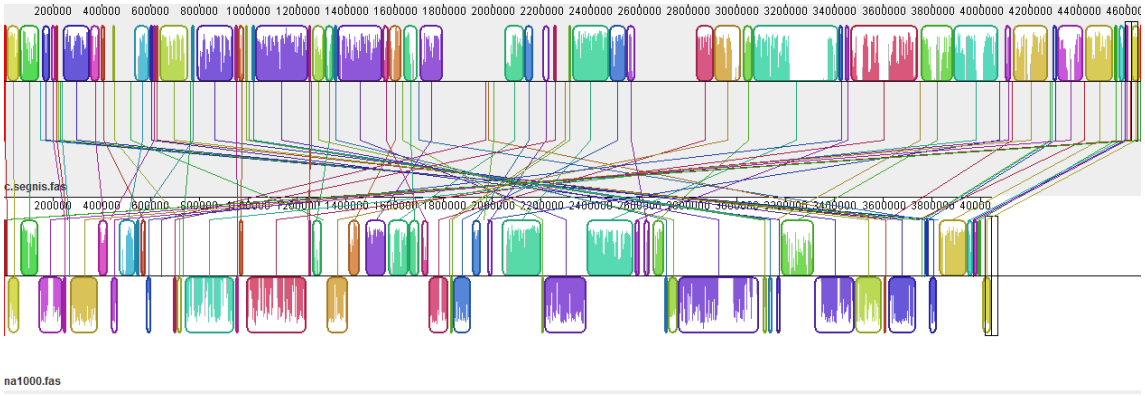

Figure 3.3: *MAUVE* comparison of *C. segnis* TK0059 (top) with NA1000 (bottom).

Since TK0059 is more closely related to NA1000 than to K31 (Abraham *et al.*, 1999), this reduced level of genome scrambling was consistent with our hypothesis. However, when we compared the *Caulobacter henricii* CB4 genome to the closely related to K31 genome, we observed more than 75 inversions and over 45 translocation events (Figure 3.4). Most of these translocations were small with only five being over 100,000 bp. The rearrangements were also mostly organized around the origin of replication as shown previously for the NA1000 and K31 comparison (Ash *et al.*, submitted). Intriguingly, when compared to the TK0059 genome, the CB4 genome had 3 large translocations and over 30 inversions (Figure 3.5). Thus the number of inversions and translocations appears to be unrelated to genetic distance. The two Brevundimonads also exhibit these high levels of genome rearrangement (Figure 3.6).

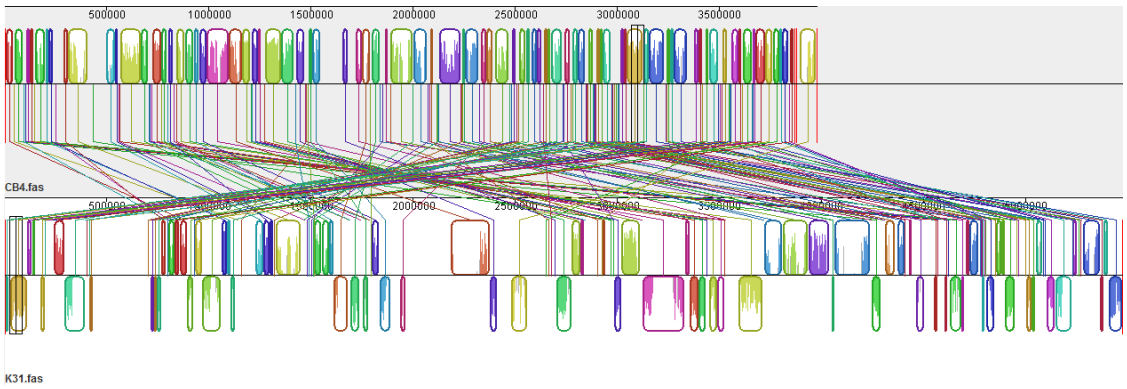

Figure 3.4: *MAUVE* alignment of CB4 (top) and K31 (bottom).

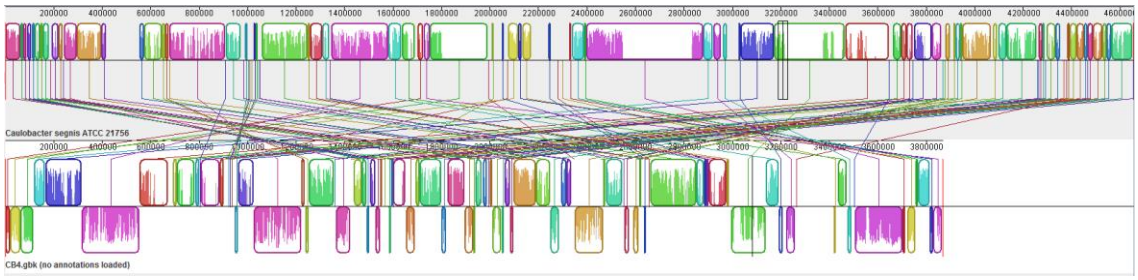

Figure 3.5: *MAUVE* alignment of *C. segnis* TK0059 (top) and CB4 (bottom).

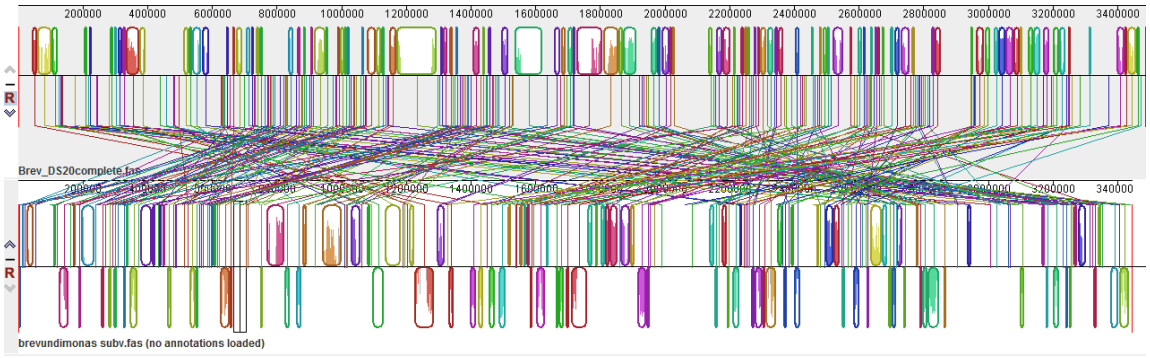

Figure 3.6: *MAUVE* alignment of *Brevundimonas* DS20 (top) and *B. subvibrioides* CB81 (bottom).

The Caulobacter Essential Genome

The identification of all essential DNA elements is necessary for a complete understanding of the regulatory networks that run a bacterial cell. Therefore, we decided to compare the 480 ORFs that comprise the essential genome of NA1000 (Christen *et al*,

2011) to the other 5 genomes in this study. We hypothesized that genes that were essential for growth of NA1000 would also be essential and highly conserved for the other bacteria as well. We used Blaststation software to BLAST these 480 ORFs against the CDS regions of the other genomes and determined that most of the 480 ORFs coded for proteins that had homologs in the other species. In fact, 17 of the essential genes coded for proteins with 100% amino acid identity in two *Caulobacter* species, two more had 100% identity in three species and one had 100% identity in all four species (Table 3.1). Most of these highly conserved genes code for ribosomal proteins where amino acid sequence conservation is expected due to the fact that these proteins bind to the rRNA and each other to form a very precise protein manufacturing machine.

There were nine NA1000 essential genes that were absent in the other genomes (Table 3.2). Four of these genes have unknown function, one codes for an antitoxin protein, and four others code for proteins involved in cell wall synthesis. In addition 23 *C. crescentus* essential genes are present in at least one other species, but are missing in at least one other species. It is unsurprising that an antitoxin protein would be essential. These genes translate proteins with the ability to neutralize a specific toxin. The absence of these genes paired with the presence of the corresponding toxin gene would prove fatal for the organism. Four of these essential genes are also present on an operon that was gained from a prophage. It is unlikely that genes gained from another organism through gene transfer would become essential without conferring some type of selective advantage to the host. What is more likely is that these genes were homologous to genes already present in NA1000 making the native NA1000 genes expendable. This is seen in the instance of the NA1000 essential gene IF-2. It is found in every genome that we

compared except for in *C. segnis* TK0059. However, TK0059 has an ortholog that functions the same as IF-2, making this gene dispensable in TK0059.

Table 3.1: Identification of the NA1000 essential genes that are %100 identical in TK0059, CB4, and K31

CCNA_number	start_of_ORF	end_of ORF	annotation	Identical Proteins
CCNA_01305	1431343	1431651	SSU ribosomal protein S10P	All Four Caulobacters
CCNA_00808	872310	872176	LSU ribosomal protein L34P	Cseg & CB4
CCNA_00722	779933	779643	chaperonin GroES	NA1000 & CB4
CCNA_00320	335209	334940	LSU ribosomal protein L27P	NA1000 & Cseg
CCNA_00699	758081	758380	LSU ribosomal protein L28P	NA1000 & Cseg
CCNA_01308	1433070	1433366	LSU ribosomal protein L23P	NA1000 & Cseg
CCNA_01310	1434213	1434491	SSU ribosomal protein S19P	NA1000 & Cseg
CCNA_01316	1436536	1436904	LSU ribosomal protein L14P	NA1000 & Cseg
CCNA_01323	1439421	1440026	SSU ribosomal protein S5P	NA1000 & Cseg
CCNA_01441	1553343	1553816	SSU ribosomal protein S9P	NA1000 & Cseg
CCNA_01741	1870539	1870159	SSU ribosomal protein S6P	NA1000 & Cseg
CCNA_01749	1877029	1877265	acyl carrier protein	NA1000 & Cseg
CCNA_01792	1919654	1919836	LSU ribosomal protein L32P	NA1000 & Cseg
CCNA_03130	3279647	3278952	cell cycle response regulator ctrA	NA1000 & Cseg
CCNA_03305	3483656	3483183	SSU ribosomal protein S7P	NA1000 & Cseg
CCNA_03310	3487006	3486449	transcription antitermination protein nusG	NA1000 & Cseg
CCNA_03430	3594108	3593983	LSU ribosomal protein L36P	NA1000 & Cseg
CCNA_01740	1870144	1869866	SSU ribosomal protein S18P	NA1000 & K31
CCNA_03306	3484041	3483670	SSU ribosomal protein S12P	NA1000, CB4, & Cseg
CCNA_02428	2570904	2570674	bacterial protein translation initiation factor 1 (II	NA1000, Cseg, K31

Table 3.2: Identification of the essential NA1000 genes present in TK0059, CB4, K31, DS20, and CB81

CCNA_number	start_of_ORF	end_of ORF	annotation	essential in NA1000	Found in C. segnis TK0059	Found in CB4	Found in K31	Found in Brev. DS20	Found in B. subvibrioides CB81
CCNA_00465	477921	479033	UDP-galactopyranose mutase	essential	NO	NO	NO	NO	NO
CCNA_00466	479191	480435	glycosyltransferase	essential	NO	NO	NO	NO	NO
CCNA_00467	480439	481710	oligosaccharide translocase/flippase	essential	NO	NO	NO	NO	NO
CCNA_00469	483454	482231	glycosyltransferase	essential	NO	NO	NO	NO	NO
CCNA_00761	820864	820655	hypothetical protein	essential	NO	NO	NO	NO	NO
CCNA_01304	1431129	1431329	hypothetical protein	essential	NO	NO	NO	NO	NO
CCNA_02841	2995269	2995508	hypothetical protein	essential	NO	NO	NO	NO	NO
CCNA_02844	2997483	2997265	antitoxin protein parD-3	essential	NO	NO	YES	NO	NO
CCNA_03307	3484065	3484331	hypothetical protein	essential	NO	NO	NO	NO	NO
CCNA_03630	3786790	3786224	socA antitoxin protein	essential	NO	NO	NO	NO	NO
CCNA_03474	3639765	3639538	SpoVT-AbrB family transcription factor, phd antitox	essential	NO	YES	YES	NO	NO
CCNA_00364	381273	380179	deoxyhypusine synthase	essential	YES	YES	YES	NO	YES
CCNA_01211	1338662	1337787	hypothetical protein	essential	YES	YES	YES	NO	NO
CCNA_01380	1494812	1495345	pole-organizing protein popZ	essential	YES	YES	YES	NO	NO
CCNA_02294	2441149	2442567	argininosuccinate lyase	essential	YES	YES	YES	NO	YES
CCNA_02644	2798562	2798119	putative cell division protein	essential	YES	YES	YES	NO	NO
CCNA_03213	3375439	3375747	putative polyhydroxyalkanoic acid system protein	essential	YES	YES	YES	NO	NO
CCNA_03277	3445041	3443992	glycosyltransferase	essential	YES	YES	YES	NO	NO
CCNA_03339	3521543	3520731	TolA protein	essential	YES	YES	YES	NO	NO
CCNA_03274	3442755	3442639	hypothetical protein	essential	NO	NO	NO	YES	NO
CCNA_00684	741111	740473	transcriptional activator chrR	essential	YES	NO	YES	YES	NO
CCNA_01864	1998726	1999349	transcriptional regulator, TetR family	essential	YES	NO	NO	YES	YES
CCNA_00041	45698	42585	bacterial protein translation initiation factor 2 IF-2	essential	NO	YES	YES	YES	YES

Phylogenetic relationships

Comparative sequence analysis of 16S ribosomal RNAs is currently the most widely used approach for the reconstruction of microbial phylogeny since the rRNA operon size, nucleotide sequence, and secondary structures of the three rRNAs (16S, 23S, 5S) are highly conserved within a bacterial species (Maidak *et al*., 1997). The 16S is the most conserved of these subunits and has been used widely as a sort of "evolutionary clock" (Woese, 1987). However, the use of a single marker gene to assess diversity is challenging, given the prevalence of horizontal gene transfer and the difficulty inherent in defining bacterial species (McDonald *et al*; 2005; Konstantinidis *et al*, 2006) as well as the limited ability of the 16S rRNA gene to resolve the relationships among closely related species. To evaluate the accuracy of the 16S rRNA tree (Figure 3.1), we constructed phylogenetic trees using the ITS and 23S gene regions of *C. crescentus*

NA1000, *Caulobacter segnis* TK0059, *Caulobacter* strain K31, *Caulobacter henricii* CB4, *Brevundimonas subvibrioides* CB81, and *Brevundimonas* DS20. We show all trees using the Maximum Likelihood model but all branch groupings and bootstrap values were identical when the Maximum Parsimony, Neighbor Joining, or Minimum Evolution models were used.

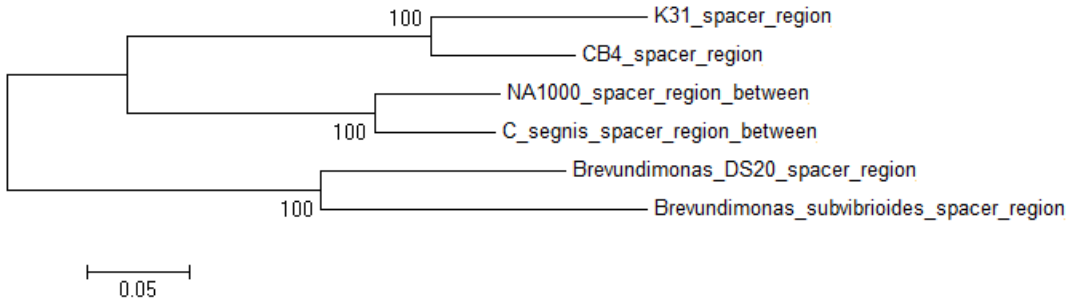

Figure 3.7: Maximum Likelihood tree based on comparison of ITS rRNA gene sequences. The numbers immediately to the left of branches indicate the number of times out of 100 the clade was recovered by bootstrap resampling. Bar, 0.05 substitutions per nucleotide position.

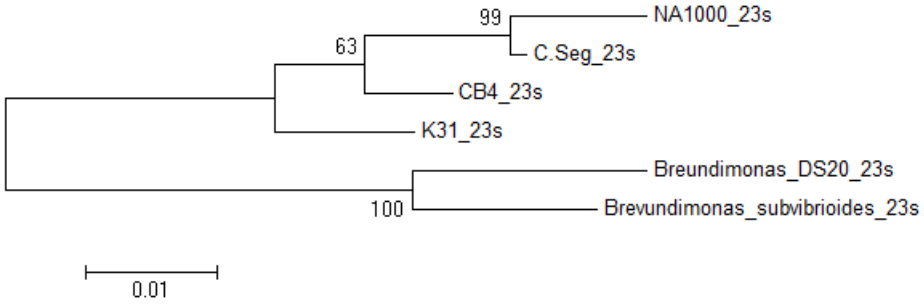

Figure 3.8: Maximum Likelihood tree based on comparison of 23S rRNA gene sequences. The numbers immediately to the left of branches indicate the number of times out of 100 the clade was recovered by bootstrap resampling. Bar, 0.01 substitutions per nucleotide position.

The agreement of all three trees corresponds to the findings by Abraham *et al* (1999) and suggests that any region of the rRNA may be used to establish relationships among these species of *Alphaproteobacteria*. To assess the consistency of these trees with other parts of the genome, we decided to do phylogenetic analyses of three large operons that span thousands of base pairs: the beginning operon of the divisional cell wall

(dcw) cluster (Ayala *et al*, 1994), a ribosomal protein operon containing 28 genes, and a prophage that is found in all six species. Although every tree produced in this study is essentially the same, the dcw operon and the conserved phage region were the only regions that did not show CB4 and K31 as monophyletic group. It will be interesting to study these regions in more *Caulobacters* once more genomes become available.

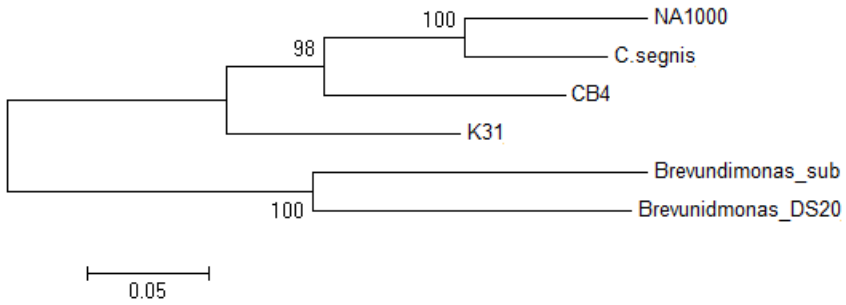

Figure 3.9: Maximum Likelihood tree based on a comparison of the dcw operon sequences. The operon sizes were approximately 10,000 bp. The numbers immediately to the left of branches indicate the number of times out of 100 the clade was recovered by bootstrap resampling. Bar, 0.05 substitutions per nucleotide position.

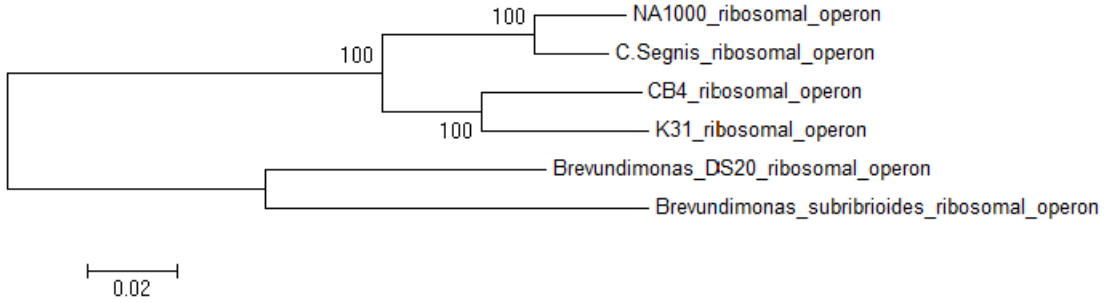

Figure 3.10: Maximum Likelihood tree based on a comparison of an 8900 bp ribosomal protein operon nucleotide sequence for each species. The numbers immediately to the left of branches indicate the number of times out of 100 the clade was recovered by bootstrap resampling. Bar, 0.02 substitutions per nucleotide position.

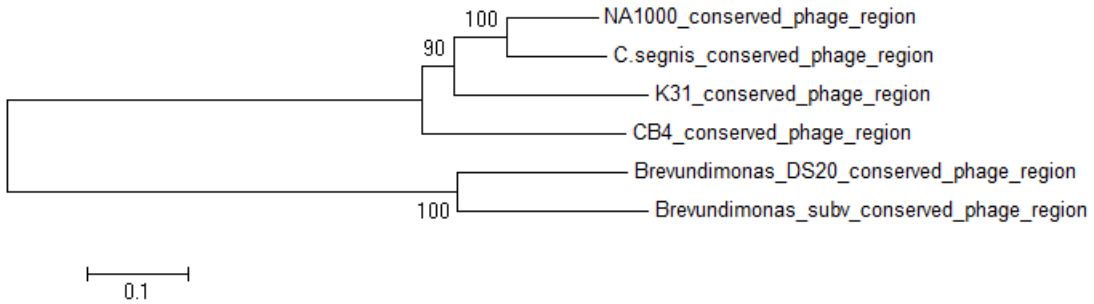

Figure 3.11: Maximum Likelihood tree based on a comparison of approximately 14,000 bp of conserved phage gene nucleotide sequences. The numbers immediately to the left of branches indicate the number of times out of 100 the clade was recovered by bootstrap resampling. Bar, 0.1 substitutions per nucleotide position.

A comparison of the 16S rRNA nucleotide sequences shows within genus differences ranging up to 3% and between genus differences in the range of 5-7% (Table 3.3). A similar range of within genus differences was observed when the 23S rRNA base pair sequences are compared (Table 3.4). However the between genus differences were slightly larger (7-8%).

Table 3.3: A comparison of 16S rRNA nucleotide sequences among the species included in this study (percent identity).

	NA1000	C. segnis	CB4	K31	B. sub	B. DS20
NA1000	100%	99%	98%	97%	93%	94%
C. segnis	99%	100%	98%	97%	94%	95%
CB4	98%	98%	100%	99%	93%	94%
K31	97%	97%	99%	100%	93%	93%
B. sub	93%	94%	93%	93%	100%	97%

B. DS20	94%	95%	94%	93%	97%	100%

Table 3.4: A comparison of 23S rRNA nucleotide sequences among the species included in this study (percent identity).

	NA1000	C. segnis	CB4	K31	B. sub	B. DS20
NA1000	100%	97%	97%	97%	93%	92%
C. segnis	97%	100%	98%	98%	93%	92%
CB4	97%	98%	100%	98%	93%	92%
K31	97%	98%	98%	100%	93%	93%
B. sub	93%	93%	93%	93%	100%	97%
B. DS20	92%	92%	92%	93%	97%	100%

In contrast, the nucleotide sequences of the region between the 16S and 23S rRNA genes (ITS) have within genus differences that range from 12-21% and between genus differences that are in a similar range (Table 3.5). This increase in diversity is unsurprising given that the ITS region of the rRNA is not used to make one of the three ribosomal subunits so it is not highly conserved and is vulnerable to indels and point mutations that have little or no impact on rRNA function.

Table 3.5: A comparison of ITS region nucleotide sequences among the species included in this study (percent identity).

	NA1000	C. segnis	CB4	K31	B. sub	B. DS20
NA1000	100%	88%	81%	80%	81%	86%

C. segnis	88%	100%	80%	80%	80%	85%
CB4	81%	80%	100%	83%	80%	79%
K31	80%	80%	83%	100%	81%	81%
B. sub	81%	80%	80%	81%	100%	80%
B. DS20	86%	85%	79%	81%	80%	100%

The *Caulobacter* dcw clusters differ by as much 19% in pairwise comparisons and 27% at the nucleotide level when compared with those of the two Brevundimonads (Table 3.6). This region is highly conserved among bacteria as it is involved in cell division and cell wall synthesis. However, these coding regions can still produce functional proteins even if codon changes are present so they are not as constrained as the rRNA region at the nucleotide sequence level.

Table 3.6: A comparison of dcw cluster nucleotide sequences among the species included in this study (percent identity).

	NA1000	C. segnis	CB4	K31	B. sub	B. DS20
NA1000	100%	88%	82%	83%	73%	75%
C. segnis	88%	100%	82%	81%	75%	73%
CB4	82%	82%	100%	84%	76%	76%
K31	83%	81%	84%	100%	73%	75%
B. sub	73%	75%	76%	73%	100%	79%
B. DS20	75%	73%	76%	75%	79%	100%

The *Caulobacter* ribosomal protein operons differ by as much 10% in pairwise comparisons and 21% when compared with those of the Brevundimonads (Table 3.7). Thus this region is more highly conserved than the dcw cluster probably because the amino acid sequences of ribosomal proteins are more constrained since they are involved in complex intermolecular interactions.

Table 3.7: A comparison of ribosomal protein operon nucleotide sequences among the species included in this study (percent identity).

	NA1000	C. segnis	CB4	K31	B. sub	B. DS20
NA1000	100%	96%	90%	90%	79%	80%
C. segnis	96%	100%	91%	90%	79%	80%
CB4	90%	91%	100%	93%	79%	80%
K31	90%	90%	93%	100%	79%	80%
B. sub	79%	79%	79%	80%	100%	86%
B. DS20	80%	80%	80%	80%	86%	100%

The *Caulobacter* and *Brevundimonas* conserved prophage region spans approximately 20 genes and the nucleotide sequence differs by as much 17% in pairwise comparisons among the Caulobacters. However, none of the *Caulobacter* prophage nucleotide sequences had significant identity to the corresponding *Brevundimonas* sequences (Table 3.8). Upon closer inspection, it was found that there was substantial shared identity among the genes in this region in all six genomes at the amino acid level. Part of the explanation for the disparity between the nucleotide and amino acid sequence identities may be that the *Caulobacter* phage regions display a codon usage bias for GGG

(Glycine), GCG (Alanine), and CGG (Arginine) in contrast to the *Brevundimonas* phage regions which display a bias towards CTC (Leucine), CGC (Arginine), and GCC (Alanine). This difference in codon usage would facilitate diversity at the nucleotide level while conserving the amino acid sequence. There may be some environmental or evolutionary pressure that is influencing this trend. Also, even though the *Brevundimonas subvibrioides* genome contained all genes of the conserved phage operon, we detected two translocations of genes to locations away from this region. Three genes were grouped together in a four gene operon along with a recombinase gene that is absent in four of the strains in this study but is found in K31. A fourth gene was found in a different four gene operon along with 3 other genes not found in any of the other bacteria in our study. Since these translocated genes were found 21,084 and 25,619 base pairs before the start of the prophage region, it is also possible that there was a single translocation event followed by an insertion between the first and second genes. In either case, we can conclude that the prophage region was present in the common ancestor of *Caulobacter* and *Brevundimonas* and has remained intact until recently despite the high level of genome rearrangements observed in these species.

Table 3.8: A comparison of conserved phage region nucleotide sequences among the species included in this study (percent identity).

	NA1000	C. segnis	CB4	K31	B. sub	B. DS20
NA1000	100%	83%	90%	87%	N/A	N/A
C. segnis	83%	100%	84%	87%	N/A	N/A
CB4	90%	84%	100%	75%	N/A	N/A
K31	87%	87%	75%	100%	N/A	N/A

B. sub	N/A	N/A	N/A	N/A	N/A	N/A
B. DS20	N/A	N/A	N/A	N/A	N/A	N/A

Conclusions

In summary, we found that despite the extensive scrambling of the *Caulobacter and Brevundimonas* genomes, the phylogenetic relationships of several large conserved gene clusters were identical to those of the 16S rRNA genes confirming that rRNA gene sequence comparisons are a valid mechanism for establishing species relationships in *Caulobacter*. In addition, most genes shown previously to be essential for *C. crescentus* (Christen et al. 2011) are highly conserved in other species of *Caulobacter* and *Brevundimonas*.

ACKNOWLEDGEMENTS

This work was funded in part by a fellowship from The Southern Region Educational Board (SREB) to DS and NIH grant GM076277 to BE.

REFERENCES

Aaron E. Darling, Bob Mau, and Nicole T. Perna. 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement. PLoS One. 5(6):e11147.

Ayala, J.A., Garrido, T., de Pedro, M.A., Vicente, M. Molecular biology of bacterial septation *J.M*

Abraham, W. R., Strompl, C., Meyer, H. & other authors (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, 1053-1073.

Beare, P. A., Unsworth, N., Andoh, M. & other authors (2009). Comparative genomics reveal extensive transposon-mediated genomic plasticity and diversity among potential effector proteins within the genus Coxiella. *Infect Immun* 77, 642-656.

B.L. Maidak, G.J. Olsen, N. Larsen, R. Overbeek, M.J. McCaughey, C.R. Woese "The RDP (Ribosomal Database Project)" *Nucleic Acids Res*, 25 (1997), pp. 109–111

Brown, P.J.B., Kysela, D.T., Buechlein, A., Hemmerich, C., and Brun, Y.V. "Genome sequences of eight morphologically diverse Alphaproteobacteria." *J. Bacteriol*. (2011) 193:4567-4568.

Christen, B., Abeliuk E., Colier, J.M., et al."The essential genome of a bacterium." *Mol Syst Biol.* 2011; 7: 528.

Darling, A. E., Miklos, I. & Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. PLoS *Genet* 4, e1000128.

Ely B, Gerardot CJ. Use of pulsed-field-gradient gel electrophoresis to construct a physical map of the Caulobacter crescentus genome. *Gene*. 1988 Sep 7;68(2):323–333.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17, 368–376.

Fitch, W. M. (1971). Towards defining the course of evolution: minimum change for a specific tree topology. *Syst Zool* 20, 406–416.

Holtzendorff J, Hung D, Brende P, Reisenauer A, Viollier PH, McAdams HH, Shapiro L (2004) Oscillating global regulators control the genetic circuit driving a bacterial cell cycle. *Science* 304: 983–987

IMG genome database (img.jgi.doe.gov)

Ghuysen, R Hakenbeck (Eds.), *Bacterial Cell Wall*, Elsevier, Amsterdam (1994), pp. 73–102

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

Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361: 1929–1940

Lucas,S., Copeland,A., Lapidus,A., et al. US DOE Joint Genome Institute Complete sequence of Brevundimonas subvibrioides ATCC 15264 Unpublished

Mannisto, M. K., Tiirola, M. A., Salkinoja-Salonen, M. S., Kulomaa, M. S. & Puhakka, J. A. (1999). Diversity of chlorophenol-degrading bacteria isolated from contaminated boreal groundwater. *Arch Microbiol* 171, 189-197.

Marks, M. E., Castro-Rojas, C. M., Teiling, C., Du, L., Kapatral, V., Walunas, T. L. & Crosson, S. (2010). The genetic basis of laboratory adaptation in Caulobacter crescentus. *J Bacteriol* 192, 3678-3688.

McDonald IR, Kampfer P, Topp E, Warner KL, Cox MJ, et al. (2005) Aminobacter ciceronei sp. nov. and Aminobacter lissarensis sp. nov., isolated from various terrestrial environments. *Int J Syst Evol Microbiol* 55: 1827–1832. doi: 10.1099/ijs.0.63716-0

Nierman, W. C., Feldblyum, T. V., Laub, M. T. & other authors (2001). Complete genome sequence of Caulobacter crescentus. *Proc Natl Acad Sci U S A* 98, 4136-4141.

Overbeek et al., "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)" *Nucleic Acids Res* 33(17)

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

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4, 406-425.

Stahl, D. A., Key, R., Flesher, B. & Smit, J. (1992). The phylogeny of marine and freshwater caulobacters reflects their habitat. *J Bacteriol* 174, 2193-2198.

Tamura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596-1599.

Wang SP, Sharma PL, Schoenlein PV, Ely B (1993) A histidine protein kinase is involved in polar organelle development in Caulobacter crescentus. *Proc Natl Acad Sci USA* 90: 630–634

www.blaststation.com

CHAPTER 4

CONCLUSION

It is an exciting new age in the field of prokaryotic genomics. Genome sequencing projects that once took millions of dollars and teams working together across many continents can now be done on a laboratory benchtop. However, these new advances have not come without setbacks. With the plethora of sequencing options available, the technology of assembling that data is trailing behind. The task of choosing the right path to assemble a finished genome is daunting for the average scientist who lacks a background doing bioinformatics type research. We took the task head on and although there was a steep learning curve, we believe that the blueprint we laid out is reproducible and feasible for anyone with a basic understanding of modern computing.

As such, we have determined that best way to sequence and assemble a bacterial genome, especially one with a high GC-content genome, is currently by using the Pacific Biosciences RSII system and the accompanying Hierarchical Genome Assembly Process 2 (HGAP2). Recently, there was an update to this assembler making HGAP3 the most current edition.

Obtaining new genomic sequences has allowed us to study problems that were impossible to answer. We decided to compare the genomes of *Caulobacters* and the closely related genus of *Brevundimonas.* We confirm the phenomenon of "Genome Scrambling" across both *Caulobacters* and *Brevundimonas* and also use the genomic data to compare the phylogenetic relationship of conserved regions throughout these genomes. In spite of the extensive scrambling, the phylogenetic relationships of these large conserved gene clusters were identical to those of the 16S rRNA genes confirming that rRNA gene sequence comparisons are a valid mechanism for establishing species

relationships in *Caulobacter*. We also found that most genes deemed essential in the NA1000 genome were conserved with the *Caulobacter* family.

With the advantage of a reliable method to obtain new bacterial genomes to expand our analysis, this is an exciting time to study genome arrangements. These findings help us to begin to understand the mechanisms of gene conservation and evolution and by which these genome rearrangements take place.

REFERENCES

Aaron E. Darling, Bob Mau, and Nicole T. Perna. 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement. PLoS One. 5(6):e11147.

Abraham, W. R., Strompl, C., Meyer, H. & other authors (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, 1053-1073.

Aird, D., et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, *Genome Biol*, 12, R18.

Ayala, J.A., Garrido, T., de Pedro, M.A., Vicente, M. Molecular biology of bacterial septation *J.M*

B.L. Maidak, G.J. Olsen, N. Larsen, R. Overbeek, M.J. McCaughey, C.R. Woese "The RDP (Ribosomal Database Project)" *Nucleic Acids Res*, 25 (1997), pp. 109–111

Bankevich, A., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J Comput Biol*, 19, 455-477.

Bartram, A.K., et al. (2011) Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads, *Applied and Environmental Microbiology*, 77, 3846-3852.

Beare, P. A., Unsworth, N., Andoh, M. & other authors (2009). Comparative genomics reveal extensive transposon-mediated genomic plasticity and diversity among potential effector proteins within the genus Coxiella. *Infect Immun* 77, 642-656.

bio.biomedicine.gu.se/cutter2/

Brown, P.J.B., Kysela, D.T., Buechlein, A., Hemmerich, C., and Brun, Y.V. "Genome sequences of eight morphologically diverse Alphaproteobacteria." *J. Bacteriol*. (2011) 193:4567-4568.

Brown, P.J.B., Kysela, D.T., et al. (2011) "Genome sequences of eight morphologically diverse alphaproteobacteria." Journal of Bacteriology, 193(17):4567

Brun, Y. V. and Janakiraman, R. (2000) in *Prokaryotic Development*, eds. Brun, Y. V. & Shimkets, L. J. American Society of Microbiology: pp. 297–317.

Caporaso, J.G., et al. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms, *ISME J*, 6, 1621-1624.

Chen, E., et al. (2013) Human infection with avian influenza A(H7N9) virus re-emerges in China in winter 2013, *Euro Surveill*, 18.

Chin, C.S., et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat Methods*, 10, 563-569.

Christen, B., Abeliuk E., Colier, J.M., et al."The essential genome of a bacterium." *Mol Syst Biol.* 2011; 7: 528.

Christen, B., E. Abeliuk, JM Collier, et al. (2011). "The essential genome of a bacterium." Molecular Systems Biology 7:528
CLC Bio

Cohen, Y., et al. (2008) Hypermethylation of CpG island loci of multiple tumor suppressor genes in retinoblastoma, *Exp Eye Res*, 86, 201-206.

Consortium, T.H.M.P. (2012) A framework for human microbiome research, *Nature*, 486, 215-221.

Consortium, T.H.M.P. (2012) Structure, function and diversity of the healthy human microbiome, *Nature*, 486, 207-214.

Darling, A. E., Miklos, I. & Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. PLoS *Genet* 4, e1000128.

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

de Magalhaes, J.P., et al. (2009) Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions, *Ageing Res Rev*, 9, 315-323.

Delcher, A.L., et al. (1999) Alignment of whole genomes, *Nucleic Acids Res*, 27, 2369-2376.

Delcher, A.L., et al. (2002) Fast algorithms for large-scale genome alignment and comparison, *Nucleic Acids Res*, 30, 2478-2483.

DNASTAR, Madison, WI, USA

Ely B, Gerardot CJ. Use of pulsed-field-gradient gel electrophoresis to construct a physical map of the Caulobacter crescentus genome. *Gene*. 1988 Sep 7;68(2):323–333.

Ely, B. 1991. Genetics of Caulobacter crescentus. Methods Enzymol. 204:372-384.
Ely, B., and R. C. Johnson. 1977. Generalized transduction in Caulobacter crescentus. Genetics 87:391-399.

Eyre, D.W et al. (2012) A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance, *BMJ Open*, 2.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17, 368–376.

Finotello, F., et al. (2011) Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data, *Brief Bioinform*, 13, 269-280.

Fitch, W. M. (1971). Towards defining the course of evolution: minimum change for a specific tree topology. *Syst Zool* 20, 406–416.

Fleischmann, R.D., Adams, M.D. et al (1995) "Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd" Science 269(5223): 496-498+507-512

Fleischmann, R.D., et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science*, 269, 496-512.

Fredrickson, J. K., J. M. Zachara, et al. (2004). "Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the hanford site, washington state." Applied Environmental Microbiology 70(7): 4230-41.

Friedman, R., and Ely, B. (2012) "Codon usage methods for horizontal gene transfer detection generate an abundance of false positive and false negative results." Current Microbiology Nov; 65(5):639-42.

Ghuysen, R Hakenbeck (Eds.), *Bacterial Cell Wall*, Elsevier, Amsterdam (1994), pp. 73–102

Gurevich, A., et al. (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, 29, 1072-1075.

Harris, S.R., et al. (2012) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study, *Lancet Infect Dis*, 13, 130-136.

Henrici, A. T. and D. E. Johnson (1935). "Studies of Freshwater Bacteria: II. Stalked Bacteria, a New Order of Schizomycetes." Journal of Bacteriology 30(1): 61-93.

Hogan, C. M. (2010). "Bacteria." Encyclopedia of Earth Sidney Draggan and C.J. Cleveland, National Council for Science and the Environment, Washington DC.

Holtzendorff J, Hung D, Brende P, Reisenauer A, Viollier PH, McAdams HH, Shapiro L (2004) Oscillating global regulators control the genetic circuit driving a bacterial cell cycle. *Science* 304: 983–987

Houwink, A. L. (1952). "Contamination of electron microscope preparations." Experientia 8: 385.

Huang X, Y.S., et al. (2006) Application of a superword array in genome assembly, *Nucleic Acids Research*, 34, 201-205.

Huang, X., et al. (2003) PCAP: a whole-genome assembly program, *Genome Res*, 13, 2164-2170.

IMG genome database (img.jgi.doe.gov)

Jackman, S.D. and Birol, I. (2010) Assembling genomes using short-read sequencing technology, *Genome Biol*, 11, 202.

Janssens, W., et al. (2010) Genomic copy number determines functional expression of {beta}-defensin 2 in airway epithelial cells and associates with chronic obstructive pulmonary disease, *Am J Respir Crit Care Med*, 182, 163-169.

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

Jones, M. (1905). "A peculiar microorganism showing rosette formation." Zentr. Bakteriol. Parasitenk. (Abt. II): 14:459-463.

Karpinski, P., et al. (2008) The CpG island methylator phenotype correlates with long-range epigenetic silencing in colorectal cancer, *Mol Cancer Res*, 6, 585-591.

Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361: 1929–1940

Koren, S., et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads, *Nature Biotechnology*, 30, 693-700.

Koren, S., et al. (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing, *Genome Biology*, 14, R101.

Kozdon, J.B., et al. (2013) Global methylation state at base-pair resolution of the Caulobacter genome throughout the cell cycle, *Proc Natl Acad Sci U S A*, 110, E4658-4667.

Kumar, S. and Blaxter, M.L. (2010) Comparing de novo assemblers for 454 transcriptome data, *BMC Genomics*, 11, 571.

Kurtz, S., et al. (2004) Versatile and open software for comparing large genomes, *Genome Biol*, 5, R12.

Laub, M. T., McAdams, Het al. (2000) Science 290, 2144–2148.
Loman, N.J., et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms, *Nat Biotechnol*, 30, 434-439.

Lucas,S., Copeland,A., Lapidus,A., et al. US DOE Joint Genome Institute Complete sequence of Brevundimonas subvibrioides ATCC 15264 Unpublished

Magoc, et al. (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms, *Bioinformatics*, 29, 1718-1725.

Mannisto, M. K., Tiirola, M. A., Salkinoja-Salonen, M. S., Kulomaa, M. S. & Puhakka, J. A. (1999). Diversity of chlorophenol-degrading bacteria isolated from contaminated boreal groundwater. *Arch Microbiol* 171, 189-197.

Margulies, M., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, 437, 376-380.

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

Marks, M. E., Castro-Rojas, C. M., Teiling, C., Du, L., Kapatral, V., Walunas, T. L. & Crosson, S. (2010). The genetic basis of laboratory adaptation in Caulobacter crescentus. *J Bacteriol* 192, 3678-3688.

McDonald IR, Kampfer P, Topp E, Warner KL, Cox MJ, et al. (2005) Aminobacter ciceronei sp. nov. and Aminobacter lissarensis sp. nov., isolated from various terrestrial environments. *Int J Syst Evol Microbiol* 55: 1827–1832. doi: 10.1099/ijs.0.63716-0

Metzker, M.L. (2009) Sequencing technologies - the next generation, *Nat Rev Genet*, 11, 31-46.

Myers, E.W., et al. (2000) A whole-genome assembly of Drosophila, Science, 287, 2196-2204.

Narzisi, G. and Mishra, B. (2011) Comparing de novo genome assembly: the long and short of it, *PLoS One*, 6, e19175.

Nierman, W. C., Feldblyum, T. V., Laub, M. T. & other authors (2001). Complete genome sequence of Caulobacter crescentus. *Proc Natl Acad Sci U S A* 98, 4136-4141.

Nierman, W.C., Feldblyum, T.V. (2001) “Complete genome sequence of Caulobacter crescentus” Proceedings of the National Academy of Sciences USA 98(7): 4136–4141

Omeliansky, V. L. (1914). “A new bacillus: Bacillus flagellatus." Omel. Zh. Mikrobiol. Epidemiol. Immunobiol.: 1:24.

Overbeek et al., “The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)” *Nucleic Acids Res* 33(17)
Phillippy, A.M., et al. (2008) Genome assembly forensics: finding the elusive mis-assembly, *Genome Biol*, 9, R55.

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

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

Powers, J.G., et al. (2013) Efficient and accurate whole genome assembly and methylome profiling of E. coli, *BMC Genomics*, 14, 675.

Quail, M.A., et al. (2009) Improved protocols for the illumina genome analyzer sequencing system, *Curr Protoc Hum Genet*, Chapter 18, Unit 18 12.

Quail, M.A., et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, 13, 341.

Rappuoli, R. (2001) Reverse vaccinology, a genome-based approach to vaccine development, *Vaccine*, 19, 2688-2691.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4, 406-425.

Salzberg, S.L., et al. (2011) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res*, 22, 557-567.

Schatz, et al. (2010) Assembly of large genomes using second-generation sequencing, *Genome Res*, 20, 1165-1173.

Schatz, et al. (2011) Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies, *Brief Bioinform*, 14, 213-224.

Shapiro, L. (1976). "Differentiation in the Caulobacter cell cycle." Annual Review of Microbiol 30: 377-407.

Simpson, J.T., et al. (2009) ABySS: a parallel assembler for short read sequence data, *Genome Res*, 19, 1117-1123.

Smith, D., et al. (2013) The Hospital Microbiome Project: Meeting Report for the 1st Hospital Microbiome Project Workshop on sampling design and building science measurements, Chicago, USA, June 7th-8th 2012, *Stand Genomic Sci*, 8, 112-117.

Stahl, D. A., Key, R., Flesher, B. & Smit, J. (1992). J Bacteriol 174, 2193-2198.

Stahl, D. A., Key, R., Flesher, B. & Smit, J. (1992). The phylogeny of marine and freshwater caulobacters reflects their habitat. *J Bacteriol* 174, 2193-2198.

Tamura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596-1599.

Tisserant, E., et al. (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis, *Proc Natl Acad Sci U S A*, 110, 20117-20122.

Toyota, M. and Issa, J.P. (1999) CpG island methylator phenotypes in aging and cancer, *Semin Cancer Biol*, 9, 349-357.

Urakami, T., Oyanag, H., Arak, H., et al. (1990) "Recharacterization and Emended Description *Mycoplana* and Description of Two New of the Genus Species, *Mycoplana ramosa* and *Mycoplana segnis*." International Journal Of Systematic Bacteriology Pg: 434-442

Wang SP, Sharma PL, Schoenlein PV, Ely B (1993) A histidine protein kinase is involved in polar organelle development in Caulobacter crescentus. *Proc Natl Acad Sci USA* 90: 630–634

Whitman, W. B., D. C. Coleman, et al. (1998). "Prokaryotes: the unseen majority." Proceedings of the National Academy of Science USA 95(12): 6578-83.

Wilson, D. J. (2012). "Insights from genomics into bacterial pathogen populations." Public Library of Science Pathogens 8(9): e1002874.

www.454.com/publications

www.blaststation.com

www.illumina.com

www.pacificbiosciences.com

www.qiagen.com/products/catalog/sample-technologies/dna-sample-technologies/genomic-dna/dneasy-blood-and-tissue-kit

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res*, 18, 821-829.

Zhang, W., et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies, *PLoS One*, 6, e17915.

Zimin, A.V et al. (2013) The MaSuRCA genome assembler, *Bioinformatics*, 29, 2669-2677.