Spring 5-10-2014

Genome Sequence and Phenotypic Characterization of Caulobacter Segnis

Sagar Patel
University of South Carolina-Columbia

Follow this and additional works at: https://scholarcommons.sc.edu/senior_theses

Part of the [Biology Commons](#), [Genetics Commons](#), and the [Genomics Commons](#)

Recommended Citation

Patel, Sagar, "Genome Sequence and Phenotypic Characterization of Caulobacter Segnis" (2014). *Senior Theses*. 8.

https://scholarcommons.sc.edu/senior_theses/8

This Thesis is brought to you by the Honors College at Scholar Commons. It has been accepted for inclusion in Senior Theses by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

GENOME SEQUENCE AND PHENOTYPIC CHARACTERIZATION OF *CAULOBACTER*
SEGNIS

By

Sagar Patel

Submitted in Partial Fulfillment
of the Requirements for
Graduation with Honors from the
South Carolina Honors College

May 2014

Approved:

Dr. Bert Ely
Director of Thesis

Dr. Latia Scott
Second Reader

Steve Lynn, Dean
For South Carolina Honors College

TABLE OF CONTENTS

Thesis Summary.....	3
Abstract.....	6
Introduction.....	7
Materials and Methods.....	9
Correction of the <i>Caulobacter segnis</i> Genbank Sequence	9
Feature Creation.....	10
Growth of <i>C. segnis</i>	10
TEM Staining Protocol	11
DNA Isolation.....	12
Pulsed Field Gel Electrophoresis (PFGE).....	12
Preparation of genomic DNA in agarose plugs	12
Restriction digest of genomic DNA in agarose plugs.....	13
PFGE parameters	14
Origin Comparison.....	15
Results and Discussion	15
Confirmation of the <i>C. segnis</i> contig assembly	23
Phenotypic Characterization	25
Conclusion	27
References.....	36

THESIS SUMMARY

The reason I chose to pursue my thesis in a bacterial genomics research lab was two-fold. First bacteria fascinate me because of their ability to grow in the most extreme of conditions. They evolve rapidly and speciation occurs often. Secondly genomics is the future of biology and knowing the evolutionary patterns of bacteria through their genomes gives us a window of insight into the mechanisms and factors which cause these organisms to adapt to new environments and in the long run, survive.

The purpose of this inquiry was to correct the genome sequence of *Caulobacter segnis* through reversion to the original sequencing data, which allowed us to reannotate the corrected sequence (add features which indicate highly probable genes) and drastically reduce the number of pseudogenes (partial remnants of functional genes). Originally there were 126 pseudogenes present in the GenBank sequence annotation. This large number led us to question whether they were truly pseudogenes or whether they were due to sequencing or data interpretation errors. One pseudogene was resequenced by designing DNA primers to amplify that region of the *C. segnis* genome then sequencing the amplified DNA through Sanger sequencing (a specific type of sequencing method). Through resequencing it was determined there was a sequence difference between the resequenced data and the GenBank genome. Consequently the original sequencing data was obtained in the form of Roche 454 reads (individual sequence readings that are aligned and compared to generate a single consensus sequence). In addition, the reads were available as a set of 74 compiled contigs (large sequence fragments assembled for the overlap of the original sequence reads), which were compared base by base using the

Mauve alignment software. Discrepancies in the two sequences were double checked by examining the individual reads at that position to determine the correct sequence. We found that 247 nucleotide deletions, 14 insertions, and 8 substitutions needed to be made to the publically available genome sequence. Reannotation (creation of features showing protein coding sequences) followed with the original number of pseudogenes drastically reduced from 126 to 27. All features were checked to determine whether the features created were highly probable of being protein-coding sequences. Errors in sequence also led to questioning of the genome assembly (arrangement of the contigs in the correct order). To test for correct representation, genomic DNA plugs (which allow us to look at large intact pieces of genomic DNA, which remain intact because cells are trapped inside agarose gel before opening them up to keep them encased and protected) were made and digested using 3 different restriction enzymes: *AseI*, *SpeI*, and *SwaI* (which cut the DNA into pieces of known sizes). The fragment size predictions for these enzymes were made using the current genome and were verified through pulsed field gel electrophoresis (PFGE) which allows us to look at the large segments of DNA in the DNA plugs. Thus the genome assembly appears to be correct. Additionally the absence of any true transposases (proteins that allow segments of DNA to be moved around and inserted into the genome) posed the question of whether *C. segnis* is able to accept transposons (movable DNA segments) into its genome. This was determined through horizontal gene transfer of the Tn5 transposase (a segment in one bacterial cell that can move to another) into *C. segnis* from *E. coli* strain 2992 and verified through PFGE.

A short region of the genome around the origin of *C. segnis* was compared with *Caulobacter* NA1000 and K31 to determine evolutionary patterns and relationships

between these three species. *C. segnis* is more closely related to NA1000 than it is to K31 (more DNA homology). It allowed us to determine the number of suspected insertion and deletion events between *C. segnis* and NA1000 in the region and make an estimate in differences for the entire genome.

The other reason for this thesis was to curb some of the previously published misconceptions about *C. segnis* morphologic data. The *Caulobacter* cell cycle contains two different cell types; a mobile cell and immobile cell. The mobile cell contains flagella and is immature meaning it is unable to divide as easily, it is able to swim in search of nutrient rich areas before maturing into a non-motile cell. The non-motile cell loses its flagella as it matures and develops a prosthecae which is a short appendage that allows it to adhere to surfaces or to other cells. These cells continuously divide to produce more motile cells. Through this study, we found that under low nutrient conditions *C. segnis* is stressed into developing into a fully mature stalk cell which was previously thought to be absent. Additionally it was believed that *C. segnis* possessed peritrichous flagella (many flagella on the cell with no particular ordering) but through transmission electron microscopy no cells were observed with peritrichous flagella. All examined specimens contained a single polar flagellum. Additionally it was determined that *C. segnis* has a doubling time of approximately 200 minutes at 30°C and is able to grow slowly at 4°C. In order for *C. segnis* to grow a carbon source, oxygen, nitrogen, phosphorus, water, and trace elements are required. Once these conditions are met Riboflavin is also required in order to have more rapid growth.

ABSTRACT

Caulobacter segnis is a unique species of *Caulobacter* in that it is slow growing, found in soil samples, and has other phenotypic differences from the freshwater *Caulobacter* that it is most closely related to. Initially deemed *Mycoplana segnis*, it was reclassified after a 16s rRNA analysis and found to be most closely related to the *Caulobacter* NA1000 branch. Because the annotated sequence available in GenBank contained a large number of pseudogenes (126), we compared the original sequencing data to the GenBank sequence and determined that many of the pseudogenes were due to sequence errors in the Genbank sequence. Consequently, we used multiple approaches to correct and reannotate the *C. segnis* genome sequence. In total there were 247 nucleotide deletions, 14 insertions, and 8 substitutions resulting in 233 fewer bases in our corrected sequence. The corrected sequence contains only 27 pseudogenes compared to the 126 of the original annotation. Furthermore it was found that *C. segnis* does produce prosthecae when grown in low nutrient conditions and the swarmer cells have a single, polar flagellum not peritrichous flagella. In addition we confirmed that Riboflavin is essential for growth. At 30°C *C.segnis* had a doubling time of 200 minutes and it was also able to grow slowly at 4°C.

INTRODUCTION

The cell cycle of the genus *Caulobacter* within the class of Alpha Proteobacteria is quite unique in that it relies on a dimorphic cell division that produces two different cells; a non-motile stem cell and a motile immature cell. The non-motile stem cells have prosthecae or stalks with holdfast material at the end which allows them to adhere to surfaces and form biofilms from which they continually produce swarmer cells. They also use their prosthecae to attach to other stalks and form flowery structures called rosettes consisting of clusters of *Caulobacter* cells. The swarmer cells are flagellated and reproduce on a slower timeframe than the stalked cells because they must first mature by shedding their flagellum and synthesize stalks. This delay allows the swarmer cells to search for sites with more resources. Since the aquatic environments where *Caulobacter* are found tend to be nutrient scarce, this dimorphic division allows them to spread their progeny as well as keep a firm base in the immediate region. *Caulobacter* are gram-negative bacteria with shapes that vary from rods, fusiform, or vibrioid. They vary in length from 1-2 micrometers and have widths of 0.4-0.5 micrometers. Freshwater *Caulobacter* have a low tolerance to NaCl from less than 5 g/L in certain species to 20 g/L in other species.

Caulobacter segnis was first isolated from soil samples by Takahashi and Komahara (U.S. patent 3,749,641, July 1973). It was initially classified as *Mycoplana segnis* (TK 0059) by Urakami et al. (1990) due to shared phenotypic characteristics with the genus *Mycoplana*. *Mycoplana* are soil dwelling bacteria with branching cells which have the ability to decompose aromatic compounds. The phenotypic comparison was considered

to be accurate at the time because *M. segnis* was isolated from the soil and produced 7-amino-3-methylcephem. Yet it had low levels of DNA-DNA homology with the TK0055 (13%), TK 0053(6%), and TK0051 (15%) *Mycoplana* isolates. *M. segnis* was reclassified by Abraham et al. (1999) as a *Caulobacter* through a 16S rRNA analysis which revealed that *M. segnis* was most closely related to *Caulobacter vibrioides* and more distantly related to *Caulobacter henricii*. However, *C. segnis* was thought to be morphologically unique from others in its genus. Urakami and Oyanagi (1990) claimed that *C. segnis* does not produce prosthecae and it has peritrichous flagella. Also, a lipid analysis showed that *C. segnis* does not contain the equivalent chain lengths 11:798, 15:0, 17:0, 17:1 ω 6 $\square$, and 17:1 ω 8 $\square$ which are present in all other *Caulobacter* tested.

The *C. segnis* genome available in National Center for Biotechnology Information (NCBI) database was 4.66 Mb long and had a 67% G/C content with 4139 protein coding sequences (CDS) (Brown et al. 2011). However, the *C. segnis* genome was thought to contain 126 pseudogenes where the coding region was interrupted with frameshift mutations caused by single base insertions or deletions. Since the other sequenced *Caulobacter* genomes contained few pseudogenes, we resequenced the frameshift-containing region of pseudogene Cseg_0004 after amplifying the region of concern with the polymerase chain reaction (PCR) and found that there was a small error in the Genbank sequence. When the error was corrected the gene contained a single continuous open reading frame with a deduced amino acid sequence that was nearly identical to that of the corresponding *C. crescentus* gene. This result led us to hypothesize that there were probably additional errors in the available genome sequence. Fortunately, we were able to obtain the original Illumina sequencing data from the Joint Genome Institute. When we

compared the Illumina sequencing reads with the available *C. segnis* genome sequence, we found numerous instances where the available genome sequence contained single or double base pair insertions or deletions that resulted in inappropriate reading frame interruptions. Therefore, one aim of this study was to correct the nucleotide sequence and annotation of the *Caulobacter segnis* genome.

Additionally, errors in the sequence also left speculation about the contig assembly.

Therefore, we verified the genome assembly through pulsed field gel electrophoresis of *C. segnis* DNA. We also examined the cell morphology and growth patterns of *C. segnis* and showed that it has a single polar flagellum and that it produces stalks under limiting nutrient conditions.

MATERIALS AND METHODS

Correction of the *Caulobacter segnis* Genome Sequence

The annotated *C. segnis* genome sequence downloaded from NCBI (<http://www.ncbi.nlm.nih.gov/genome/1934>) was compared to the original Roche 454 sequencing data that was obtained from the Joint Genome Institute. The original data was viewed as individual reads using the Tablet software (The James Hutton Institute). Contigs compiled from the consensus of the original reads were compared with the annotated version of the genome using the Mauve Multiple Genome Alignment software (Genome Evolution Lab). The 454 data were assumed to be correct when there were at least five reads with identical sequence. In cases where there were fewer than five reads, changes were not accepted. When correction of the sequence resulted in a combined open

reading frame, the predicted amino acid sequence of the combined reading frame was compared to those in the NCBI database using BLAST (2). If significant matches were observed, the annotation of the gene was corrected as well.

Feature Creation

After correcting the genome nucleotide sequence, the genome was submitted to the Rapid Annotation using Subsystem Technology (RAST, 5) for automated annotation. The two annotations were then compared manually at regions of sequence change and Artemis: Genome Browser and Annotation Tool (6) was used to update the *C. segnis* genome annotation downloaded from GenBank. When the differences were observed, the deduced amino acid sequences of predicted genes were compared to those in the NCBI database using BLAST. A consensus of 5 or more matches identifying the same gene function ($<1E-6$) was used as the basis for determining feature size and feature function. Finally both the RAST annotated sequence and our annotated sequence were inputted in the MICheck: Microbial genome checker (7) to verify predicted CDSs. Conflicting results in MICheck were viewed and the corresponding amino acid sequence was compared to the NCBI database using BLAST. If the predicted CDS had at least 5 matches ($<1E-6$) then it was retained as a valid gene.

Growth of *C. segnis*

A *Caulobacter segnis* culture was obtained from Dr. Yves Brun (Indiana University). *C. segnis* was grown in peptone yeast extract (PYE) medium (Johnson and Ely, 1977) or PYE supplemented with 10 mM glucose, 30 mM monosodium L-glutamate. PYE broth

consists of 0.2% Bactopeptone, 0.1% yeast extract, 0.75 mM MgSO₄, 0.5 mM CaCl₂. For PYE agar plates the broth was supplemented with 1.2% agar. *C. segnis* was also grown on minimal media (M2) plus glucose (Johnson and Ely 1977) to determine the basic growth requirements. Cultures were grown in PYE in a side-arm culture flask at 30°C and bacterial growth was measured using a Klett Colorimeter for 8 hours after initial inoculation. To select for streptomycin resistant *C. segnis*, 100 µL of a *C. segnis* culture was spread on a PYE plate containing 50 µg/ml streptomycin. The resulting colonies were purified by two rounds of single colony isolation. Transposon mutagenesis experiments were performed by mixing 1 mL of streptomycin resistant *C. segnis* with 1 mL of *Escherichia coli* strain NC2992 (Tn5+) and immediately filtering the mixture onto a 0.45 µm filter membrane. The membrane was then placed face-up on a PYE plate and incubated at 30°C for 4 hours. Next the membrane was placed in a glass tube and 1 mL of PYE was added. After resuspending the bacteria, 100 µL of mixture was spread on a PYE plate containing 50 mg/L streptomycin and 50mg/L kanamycin for selection of *C. segnis* colonies that contained a Tn5 insertion. Kanamycin and streptomycin resistant colonies were purified by two rounds of single colony isolation.

TEM Staining Protocol

To stain the *C. segnis* cell sample for transmission electron microscopy (TEM), 5 mL of liquid culture was centrifuged at 1157 x g for 10 minutes and the supernatant was discarded. The pellet was resuspended gently in 2 mL of DI water. Subsequently, 8 µL of the resuspended cells was added to the copper grid for 1 minute then carefully dried by touching filter paper to side of grid. Next 8 µL of uranium acetate (radioactive) was

added and left to sit for 2.5 minutes then carefully dried with filter paper making sure to properly dispose of any radioactive contaminants. The grid was allowed to dry for 15 minutes before use in the TEM.

DNA Isolation

To isolate, *C. segnis* DNA, 10 mL of culture was centrifuged at 4629 x g for 5 min at 4 °C and the supernatant was discarded. The pellet was resuspended in 10 mL SSC (1X Standard Saline Citrate). The centrifugation was repeated and the pellet was again resuspended in 1mL SSC. Subsequently 0.2 mL lysozyme (10 mg/mL) was added and mixed by inversion. The mixture was incubated on ice for 30 minutes and then 0.8 mL of 2% Sarkosyl solution was added and mixed by inversion. After 10 minutes at room temperature, 250 µL of phenol/chloroform/isoamyl alcohol (25:24:1) was added and mixed by inversion. The mixture was spun in a microcentrifuge for 5 minutes and the top layer was collected and transferred to a new tube. This process was repeated and the phenol/top layer collection step up to 6 times as needed until no visible precipitate between the two layers remained. To precipitate the DNA, 1/10th the volume of 3M sodium acetate (pH 5.8) and 1 mL of cold 95% ethanol were added, mixed by inversion and placed on dry ice (or -80 degree C) for 10 minutes. The mixture was then spun down in a microcentrifuge for 5 minutes, the supernatant was discarded, and the tube was blotted and allowed to dry for 15 minutes before resuspending the pellet in 200 µL TE buffer. The quality of the isolated DNA was determined by agarose gel electrophoresis.

Pulsed Field Gel Electrophoresis (PFGE)

Preparation of genomic DNA in agarose plugs

To create *C. segnis* DNA agarose plugs, 5 mL of cultured cells was centrifuged at 3214 x g for 5 min at 4°C. The supernatant was discarded and the pellet was resuspended in 2 mL of Tris-NaCl and centrifuged at 3214 x g for 5 min at 4°C. The supernatant was discarded and the pellet resuspended in 0.5 mL Tris-NaCl, then incubated at 37°C while preparing a 1% agarose solution. To make the plugs, 0.5 mL of cells was pipetted into a 1.5 mL tube and an equal volume of warm agarose (0.5mL) was added and mixed by pipetting. The mixture was quickly drawn into a 1 mL syringe and allowed to cool and harden. The end of the syringe was cut off with scissors and the hardened mixture was pushed out of the syringe and sliced at 0.1 mL increments with razor to create the plugs. The cut plugs were placed into a 15 mL tube with 2 mL of lysis buffer (1.9 mL 1% Sarkosyl-EDTA and 0.1 mL proteinase K). The tube was incubated overnight at 50°C. The liquid in the tube was removed by aspiration and then 2 mL of 1X TE buffer and 30 µL of PMSF (17.4 mg phenylmethanesulfonyl fluoride / 1 mL 100% ethanol) was added and incubated on a rotator for 1 hour at room temperature. The aspiration and PMSF addition steps were repeated once more. After the final aspiration, 2 mL of wash buffer (1X TE buffer) was added and incubated on a rotator for 1 hour at room temperature. This wash step was repeated two more times. The plugs were stored in the refrigerator (4°C) until use.

Restriction digest of genomic DNA in agarose plugs

A single 0.1 mL plug containing intact genomic DNA was placed in a 1.5mL tube containing the reagents listed in Table 1 depending on the enzyme used for digestion. The digestion mixtures were then incubated at 35°C for 4 hours and then stored at 4°C until used.

Table 1. Reagents per restriction digest of 1 DNA plug by *AseI*, *SpeI*, and *SwaI*.

	Restriction Enzymes		
Ingredients	<i>AseI</i>	<i>SpeI</i>	<i>SwaI</i>
H ₂ O	61 µL	60 µL	61 µL
Buffer 3	10 µL	--	--
Buffer 3.1	--	--	10 µL
Buffer 4	--	10 µL	--
BSA	--	1 µL	--
Restriction enzyme	3 µL	3 µL	3 µL

PFGE parameters

PFGE was carried out in a 1% agarose gel (1.5 g pulsed field gel agarose and 150 mL SBA) using the parameters described in Table 2. The running buffer was 1X SBA (35.35mM Boric Acid, 10.0mM NaOH, pH=8.5).

Table 2. Parameters for DNA fragment separation in PFGE along with the appropriate size markers.

Parameters	10Kb – 210 Kb	200Kb – 2000Kb	50Kb – 800Kb
Temperature	14°C	14°C	14°C
Initial Pulse time	10 sec	70 sec (15hours)	5 sec
Final Pulse time	20 sec	120 sec (11 hours)	120 sec
Voltage	6.0 V	6.0 V	4.5V
Time	16 hours	26 hours (15/11)	40 hours
Ladder	Lambda	Yeast	Yeast/Lambda

Origin Comparison

Mauve Multiple Genome Alignment software was used to align conserved sequences of the annotated genomes of *Caulobacter segnis*, NA1000, and K31. A region containing 226 genes downstream and 56 genes upstream of the origin of replication in *C. segnis* was compared with NA1000 to identify inversion, insertion, and deletion events since these two species diverged from K31.

RESULTS AND DISCUSSION

The presence of 126 pseudogenes in the annotated Genbank version of *Caulobacter segnis* suggested possible errors in the annotation and/or the sequencing data.

Consequently, the 74 original contigs obtained from Joint Genome Institute (JGI) were aligned with the NCBI version of the sequence and visually compared using Mauve.

When differences were observed, the original sequence reads were examined to identify

the correct sequence information. In most cases, the sequencing reads resulted in a clear consensus. When variation in the sequencing reads was observed, it usually involved differences in the number of times a base was repeated. However in all cases, most of the reads favored a particular interpretation. Therefore, the genome sequence was corrected based on the original sequencing reads. In total, we removed 247 base pairs from 157 sites and added a total of 14 base pairs at 12 sites. In addition, we corrected single base pair errors at 8 base sites (Table 3). After reannotation of the affected genes, the corrected genome sequence contained only 27 pseudogenes. An example of a change can be observed in Figure 1 in which the NCBI sequence of Cseg_3308 contains an added “AT” that was not present in the original 454 data. In this case when the extra “AT” was removed, two reading frames were merged to form a single open reading frame that corresponded to those of the homologous beta-glucuronidase-like protein genes in other *Caulobacter* species (Figure 2). Based on these results, we concluded that Cseg_3308 was not a pseudogene, and the annotation of the gene was changed to include the relevant information about the gene.

To check the corrected annotation of the *C. segnis* genome, the corrected genome nucleotide sequence was run through RAST for automated annotation. RAST features at locations of sequence change were then compared manually to the feature descriptions that we generated manually for verification of correct feature creation. Discrepancies between the two were analyzed using BLAST to determine the correct annotation. Subsequently, both the RAST annotation and our edited version were evaluated using the Microbial Genome Checker (MICheck) to check for annotation errors. MICheck suggested that the newly corrected genome contained 5 unnecessary features and the

RAST annotation contained nearly 20. These discrepancies were analyzed using BLAST and 4 features were deleted on the new annotation and 1 of the RAST annotation genes were accepted into the new annotation.

The revised *Caulobacter segnis* genome is a single 4,655,405 base pair chromosome with a 67.7% G/C content. The genome is larger than the 4 Mb genome of NA1000 but smaller than the 5.48 Mb K31 genome (Table 4). Other essential genome features such as the numbers of rRNA genes, flagellar genes, and G/C content are conserved. The *C. segnis* and NA1000 genomes have no differences in the number or identity of the tRNA genes whereas the K31 genome has differences in five tRNA genes compared to the *C. segnis* genome. These data are consistent with the idea that *C. segnis* and NA1000 are closely related.

C. segnis is unique in that it does not contain any true transposases. There is only one pseudogene reminiscent of a transposase. This observation leads to the question of whether transposons are able to be incorporated into the *C. segnis* genome. To answer this question, horizontal gene transfer was attempted between *Escherichia coli* strain NC2292, a streptomycin sensitive strain containing transposon Tn5 which codes for kanamycin resistance. Since the plasmid containing the Tn5 does not replicate in *Caulobacter*, selection for kanamycin resistance will result in selection for a transposon that has integrated into the *C. segnis* genome. When the experiment was performed, 14 single colonies were obtained and purified by two rounds of single colony isolation. Eleven of these isolates were subjected to PFGE analysis to confirm the insertion of Tn5 into the genome. Fragments containing the Tn5 insertion were identified by a shift in the

migration pattern of the bands in the 97 KB region of the gel in lanes 3, 5, & 6 (Fig. 3) proving that the *C. segnis* genome can stably harbor a transposon. Tn5 insertions are likely to be present in the remaining cases as well, but in regions of the genome that were not analyzed in this experiment.

Table 4. Basic features of the *C. segnis*, NA1000, and K31 genomes

Genome Features	<i>C. segnis</i>	NA1000	K31
Base Pairs (Kb)	4.66	4.02	5.48
G/C content	67.7%	67.2%	68.1%
GC3	88.93%	86.9%	88.4%
Genes	~4270	~3879	~5500
tRNA genes	51	51	49
rRNA genes	6	6	6
5s	2	2	2
16s	2	2	2
23s	2	2	2
other RNA	3	41	3
Flagellar genes	48	43	44
phage genes	19	15	42
transposases	1 (pseudo)	40	61
integrases	5	4	33
recombinases	2	2	7
pseudogenes	126*	1	17

Table 5. tRNA genes in *C. segnis*, NA1000, and K31

	<i>C. segnis</i>	NA1000	K31
Ala	4	4	4
Arg	4	4	4
Asn	2	2	1
Asp	2	2	3
Cys	1	1	2
Gln	1	1	1
Glu	2	2	2
Gly	4	4	4
His	1	1	1
Ile	2	2	2
Leu	5	5	4
Lys	2	2	2
Met	5	5	4
Phe	1	1	1
Pro	3	3	3
Ser	4	4	4
Thr	3	3	3
Trp	1	1	1
Tyr	1	1	1
Val	3	3	3
TOTAL	51	51	50

Table 3. Summary of sequence change events by deletions, insertions, and substitutions.

	One Base	Two Base	Three Base	Five Base	Total Bases
Base pairs removed	75	76	5	1	247
Base pairs added	10	2	0	0	14
Base pairs changed	8	0	0	0	8
Change in genome	93	78	5	1	-233

Figure 1. A comparison of NCBI sequence versus 454 contig 18 for a portion of the *Cseg_3308* gene reveals two additional base pairs in the NCBI version of the *C. segnis* genome sequence.

Figure 2. Effect of the removal of two base pairs from the Cseg_3308 nucleotide sequence. Deletion of the two nucleotides results in a single reading frame for the Cseg_3308 gene (corresponding to the blue bar in reading frame 1 of the lower panel and the red line in the plot of third codon position GC content).

When comparing 226 downstream genes and 56 upstream of the origin in *C. segnis* to the corresponding region in NA1000, we found many regions of gene homology and long stretches of conserved sequences, consistent with the close relationship of the two species (Figure 4). However, two inversion events can be observed in this region. In addition, we observed 54 genes that were present in only one of the two genomes and could be attributed to 43 insertion/deletion events when this region of *C. segnis* and NA1000 genomes were compared with that of K31 (Table 6).

Figure 4. Mauve comparison of *C. segnis* and NA1000 around the origin region.

Table 6. Origin comparison chart showing differences between *C. segnis* and NA1000 in 282 genes around the origin.

Event	Count
<i>C. segnis</i> Insertions	24
<i>C. segnis</i> Deletions	7
<i>C. segnis</i> Unique Genes	32
NA1000 Insertions	15
NA1000 Deletions	8
NA1000 Unique Genes	22
Inversions	2

Confirmation of the *C. segnis* contig assembly

PFGE analysis was used to determine if the original 76 *C. segnis* contigs were assembled in the correct order in the finished sequence. The predicted DNA fragments from restriction digests of *C. segnis* genomic DNA using *AseI*, *SpeI*, and *SwaI* are shown in Table 7. Several gels were run using different conditions to separate specific sizes shown in Figure 5 & 6. All of the predicted large bands were observed in the *AseI* and *SwaI* and all four predicted fragments were observed in the *SpeI* digests. Thus these data indicate that the assembly of the *C. segnis* genome is correct.

Table 7. Predicted fragments sizes of the *C. segnis* genome when digested with *AseI*, *SpeI*, or *SwaI*.

<i>AseI</i>	<i>SpeI</i>	<i>SwaI</i>
length (Kb)	length (Kb)	length (Kb)
0	4	337
3	8	1327
3	9	1475
15	9	1517
22	12	
69	17	
77	24	
77	34	
97	39	
110	61	
117	67	
122	87	
156	94	
178	96	
209	102	
357	114	
362	117	
458	123	
490	125	
498	126	
579	131	
655	203	
	205	
	268	
	319	
	455	
	629	
	1076

Figure 5. PFGE of the *C. segnis* genome digested with *AseI*, *SpeI*, or *SwaI* in gels A and B. (A.) was set with parameters of 70 seconds for 15 hours and 120 seconds for 11 hours at 6 Volts. (B.) was run using a 5 second initial pulse and 120 second final pulse over 40 hours at 4.5 Volts.

Figure 6. PFGE illustrating presence of cut fragments as well as insertion of Tn5 into the genome.

Phenotypic Characterization

C. segnis cells are approximately 1.7 μm to 2.2 μm by 0.6 μm to 0.7 μm and rod shaped bacteria. Colonies are white in color and tended to be mucoid probably due to extracellular polysaccharide production. In contrast to the description provided by Urakami et al., *C. segnis* divides by fission in a manner typical of other *Caulobacters* producing a mature cell that functions as a stem cell and an immature swarmer cell. The species was named *segnis* because it was slow growing. At 30°C *C.segnis* has a doubling time of approximately 200 minutes which is slower than K31 at 160 minutes and much slower than 110 minutes for NA1000 (Fig.7). Like K31, *C. segnis* is able to grow slowly at 4°C. The original strain was isolated from soil samples by Takahashi and Komahara. Initially it was phenotypically classified by its ability to produce 7-amino-3-methylcephem (Urakami) and given the identity of *Mycoplana segnis*. However, 16S rRNA analysis showed that it was in fact related most closely to *Caulobacter* and

therefore the strain was reclassified as *Caulobacter segnis* (Abraham et al). In agreement with Abraham et al. no stalks were observed in *C. segnis* grown in PYE media, however when grown in media with low levels of nutrients, we observed that *Caulobacter segnis* does in fact produce a prosthecae (Fig. 8 & 9). The prosthecae allow the fully mature cell to adhere to surfaces when nutrient conditions are optimal. The *C. segnis* prosthecae appears to be approximately 0.6 μm in length. Formation of rosettes was not observed. In contrast a rosette is observed for CB15 in figure 10. Additionally *C. segnis* swarmer cells have a single polar flagellum whereas Urakami et al stated that it had peritrichous flagella. No peritrichous flagella were observed (Fig 11). It was also shown determined that riboflavin is essential for growth.

Figure 7. Growth curve on *C. segnis* at 30°C.

Figure 8. *C. segnis* non-motile, stem cell grown in PYE with prosthecae taken using TEM. The cell in its entirety (above) and a close-up on the prosthecae (below).

4/9/2014 HT: 200 kV TEM Magnification: 6K

Figure 9. *C. segnis* with prosthecae grown in M2 plus glucose media.

2/6/2014 HT: 200 kV TEM Magnification: 7k

Figure 10. Rosette formation observed in CB15 culture.

Figure 11. *C.segnis* with single polar flagellum grown in PYE.

CONCLUSION

To conclude, *Caulobacter segnis* is a unique species of *Caulobacter* correctly classified with freshwater *Caulobacter* even though it is found in soil. More recently isolations of *Caulobacter* from the roots of plants have become routine (add references, plus Ely et al., unpublished). Multiple sequence errors in the publically available genome nucleotide sequence have been corrected and an additional table has been added identifying the errors and locations for reference (Table 8). Sequence corrections have led to a large reduction in the number of pseudogenes from 126 to 27. Moreover, information presented in previous studies by Urakami et al. was deemed unreliable. It was found that *C. segnis* contains a single polar flagellum rather than peritrichous flagella. It also has been observed to have the ability to mature into stalked cells under low nutrient conditions.

Also *C. segnis* does not grow via branching, rather it divides through binary fission.

Finally *C. segnis* is able to incorporate transposases into its genome despite having no transposases of its own.

<i>Caulobacter Segnis</i> Genome Edits					
Feature Number	Sequence	Changed Bases	Change Type	Location	Contig
3	Acc ga gagc	ga	Deletion	4228	52
4	ccgcgg cc ccctgg	cc	Deletion	5830	52
30	ggcgg ct ccggc	t	Deletion	35233	52
30	cgcg cc ccccgcgc	cc	Deletion	35565	52
46	agg cg cgcccg	cg	Deletion	54385	52
4270	Cgg cg ggtgaa	g	Deletion	56908	52
65	Ccac gg ggc	gg	Deletion	70677	52
69	Tgat ct ccccg	tc	Deletion	75713	52
71	Acc g gccgcc	g	Deletion	78749	52
128	acgca c gcgccgcggg	c	Deletion	143102	51
182	aaga ag agaaggaaga	ag	Deletion	205977	50
189-190	cccc tg tggccgcga	tg	Deletion	213199	50
198	Agca att attgg	att	Deletion	220334	50
210	tcgga g gggctgt	g	Deletion	230057	50
219	Atgcg cccc ca	c	Deletion	239494	50
231	Atgt g gggt	g	Deletion	256111	50
243	ggccttcgc ctt tagc	t	Deletion	265636	50
209	gcgt ca agggcgcgatt	a	Deletion	289461	49
291	acctc g gggcga	g	Deletion	309022	48
303	tgcc ca caacggc	ca	Deletion	322284	48
443	Attgt g ggc	g	Deletion	456093	47

485	accg cg ccacct	cg	Deletion	515481	47
497	agcgcg cccc gcg	c	Deletion	532811	47
533	aa ccgccc ccgccccgg	ccgccc	Deletion	571896	47
660	ca cccc gcgcct	cc	Deletion	689893	47
682	Gccc tg tgat	tg	Deletion	707023	47
696	agccccggc g gcgcgctt	g	Deletion	721807	47
707	Gca ct ccttt	c	Deletion	733220	46
708-709	tgg cccc tccgg	c	Deletion	734958	46
751	Ccag t gtcggc	gt	Deletion	790398	44
0851-0852	atcct cccc cgtg	c	Deletion	911645	43
0853-0854	cgga gggg gtga	g	Deletion	915063	43
917	acgacc gt gtgtgg	gt	Deletion	989589	43
917	tcgc cg cgcgccc	cg	Deletion	989986	43
981-982	tccg gggg gagga	g	Deletion	1057383	43
1055	tcga acg cgcggt	cg	Deletion	1130819	39
1065	tcacggcgcg gcccc ggat	c	Deletion	1143370	39
1128	atcgtg tt tcggcga	t	Deletion	1212448	37
1154	tcgc g gcgcgccgat	g	Deletion	1240597	36
1154	tgccggc gg ggt	gg	Deletion	1240979	36
1223	agct gg gggtgt	gg	Deletion	1310520	36
1235	ccttgatct g ggttt	g	Deletion	1326415	68
1251	ggcgc g ggccttggc	g	Deletion	1346490	35
1279	gcgcgcg ac accgg	ac	Deletion	1385177	35
1293	tgg tct tgggcgt	tc	Deletion	1400798	35
1296	tctg cccc ggccgaa	c	Deletion	1404126	35
1322-1323	gaaaagaa a aaagggt	a	Deletion	1427586	34
1363	ccgac g ggcggagagat	g	Deletion	1462908	32
1379	ttccagggcg c aca	c	Deletion	1481460	32

1387	aacgcgcgccgaa	g	Deletion	1492582	32
1404	cccc gc gcgcggaaaaa	gc	Deletion	1513682	31
1446-1447	tag cccc ctccgcg	c	Deletion	1555140	31
1531	cgtgtgcg cccc ct	c	Deletion	1628321	31
1554	gagcggctggtg ggg act	g	Deletion	1656247	31
1561	accgc gc cgagga	cg	Deletion	1664371	31
1589	accggcg cc gccgtctt	c	Deletion	1697865	31
1589	cgaaa ggggg g	g	Deletion	1698230	31
1617	gagcggct gt gaccc	gt	Deletion	1727029	31
1626	Tgc ggg cggct	g	Deletion	1736200	31
1674	tggcgggg gg ctg	gg	Deletion	1781677	31
1746	aggc ggggggg gag	g	Deletion	1864680	31
1761	ccaccgcc c acga	c	Deletion	1884980	31
1784	gagcgcg cg cgtg	cg	Deletion	1916612	31
1816	aagggg cgcc gg	c	Deletion	1957174	30
1838	tggcgt cg cgcggccc	cg	Deletion	1986821	30
1839-1840	accgcaa cg cgccga	cg	Deletion	1988958	30
1947	cggc ggg ctatt	g	Deletion	2117396	30
1950-1951	aggc ggg gagga	g	Deletion	2122567	30
1978	gccgt gt gaccaggcgta	gt	Deletion	2153187	29
1990	tgcgagaaca ggggggg g	g	Deletion	2167214	29
2009	Ttcg t atgt	t	Deletion	2191620	29
2010	ggctac ggg cgcc	g	Deletion	2193781	29
2030-2031	ttcccg cg cgcga	cg	Deletion	2221600	29
2043	tctgc ggg cccggt	g	Deletion	2235910	29
2052	Ggccct a tat	a	Deletion	2247634	29
2072-2073	tcgcg cgcc agccc	cgc	Deletion	2272936	29
2175	ttcccat a tggcc	at	Deletion	2381622	7

2208	Ttggc g cgc	g	Deletion	2418949	27
2232	aa cg gcggcggcc	cgg	Deletion	2442931	27
2306	gcctg g tcctat	tc	Deletion	2522023	26
2314	Ttcag cccc ct	c	Deletion	2530228	26
2341	gcg cccc ggggcca	c	Deletion	2554261	26
2343	cgagctggccgcgg gg gcc	gg	Deletion	2557216	26
2343	ccgc gg gggcct	gg	Deletion	2557331	26
2351	Gcc g gcgcgc	g	Deletion	2566191	26
2359	Atcg cg cacg	c	Deletion	2574246	26
2546-2547	tatacc g cgcgctat	gc	Deletion	2789088	24
2598	Tcag a cagc	a	Deletion	2844051	23
2604	actgcc cccc gga	cc	Deletion	2850943	23
2638	agcgcc cg cgcaa	cg	Deletion	2891847	23
2691	Tcgg a gcgag	g	Deletion	2941192	23
2756	Aacgc cg cgc	cg	Deletion	3015803	23
2756	Gat cccc gat	cc	Deletion	3017456	23
2773-2774	tgcccc ct ccgggggc	c,g	Deletion	3036113	23
2820	Cgt cg cgcggt	cg	Deletion	3082472	21
2872	ggg cg ccgcca	cg	Deletion	3131555	21
2876-2877	Aaca cc ctc	c	Deletion	3136290	21
2892	Gccct cccc g	cc	Deletion	3148780	21
2896	Gcg cccc gac	cc	Deletion	3152518	21
2906	Gtt cg ccgccg	cgc	Deletion	3160471	21
2929	gacc gg gagaact	g	Deletion	3183182	21
2965-2966	Acc cccc tttt	c	Deletion	3223587	21
3041	tcgacg g cggtt	gc	Deletion	3304002	21
3044-3045	Tcgaagg g cg	g	Deletion	3312667	21
3048	tactacca gt gttcct	gt	Deletion	3320376	21

3077	gcgtggg g cggccctgac	g	Deletion	3358990	20
3182	aacgcg ccc gcggccaa	c	Deletion	3474293	19
3198	Atcg cg cggg	cg	Deletion	3490390	19
3233	atggg cg cggcg	cg	Deletion	3531325	19
3233	Tgat at cct	at	Deletion	3531511	19
3251	gccgccggc cg cgcgccgt	gc	Deletion	3549581	18
3299	Ggacct ct	ct	Deletion	3615624	18
3308	Cgg ca catgc	ca	Deletion	3626943	18
3326	tcggcgggc cg gcc	cg	Deletion	3651824	17
3356	Tcag cg cgccttg	gc	Deletion	3678487	17
3453	tataggtgtg gg aa	g	Deletion	3778740	15
3456	tttcgggcggcg cg acgggggtca	cg	Deletion	3782043	15
3510-3511	tttcgatg ag acag	ag	Deletion	3843227	13
3517-3518	tcttcct ttt tgttc	t	Deletion	3847035	13
3522	Aacc ccc gct	cc	Deletion	3850996	13
3534	ggccac ccc gaga	cc	Deletion	3863944	13
3540	ttccccct ccg gccttcgg	ccg	Deletion	3872804	13
3548	acgcccc ccg ccc	c	Deletion	3882234	64
3549-3550	gcggcgagacggaggggg g cg	g	Deletion	3884640	64
3553	ccggccgg cg cccga	cg	Deletion	3888160	64
3559	cctcgac cg cgg	c	Deletion	3896360	64
3563	tcggcgt g ggccggaata	g	Deletion	3902994	64
3567	Cccgct ct ggg	ct	Deletion	3909620	64
3571	tgctgg ag acga	ag	Deletion	3914412	64
3586	tctcac ac ggcc	ac	Deletion	3933260	64
3589	taggacgc ccc ctccg	cc	Deletion	3939440	64
3765	ttcagcatat gc gcgtgg	gc	Deletion	4128365	57
3809	gggatggcca ga gggtc	ga	Deletion	4183172	57

3845-3846	agacggga g ggggtcgcgaa	g	Deletion	4221937	57
3856-3856	accggcacg cg cgcgctccat	cg	Deletion	4232625	57
3856-3857	tgtcga ct ctctgcgcct	c	Deletion	4233206	57
3868	tgaccgc cg cggtttcgcgct	cg	Deletion	4248548	57
3895	gcccacg gc gcgccc	gc	Deletion	4273387	55
3919	gagaga cc gcgatg	c	Deletion	4300904	54
3923	agcatg ac acgcggct	ac	Deletion	4307201	54
3929-3930	aaatcgccc cg cgcggtcc	cg	Deletion	4313947	54
3968	atctcgacg tc tcttcacc	tc	Deletion	4350058	54
3979	gtcgccgccc gc gggtcgca	g	Deletion	4368520	54
3987	gcgcgccct gg gagaagcg	g	Deletion	4381597	54
4002	ccggcgggca g ggga	g	Deletion	4399430	54
4003	ttacccgct cg cgacgacg	cg	Deletion	4401520	54
4013	acgctttt ga gaaacaa	ga	Deletion	4408379	54
4176	Tcga cccc cg	cc	Deletion	4566108	53
4179	agggtg tc tcggtgaa	tc	Deletion	4569751	53
4194	gggggcccga gc gcgcgagtt	gc	Deletion	4581915	53
4207	ggcgg gc gcctttcgg	gc	Deletion	4594655	53
4248	tgccgggt gg ggcgcaggaag at	g	Deletion	4638526	53
0726-0727	gcccc tc cgcgct c gcgg	c,c	Deletion/ Insertion	757034	6
3347-3348	gg cc gcgcgccct g ggggccta g	c,g	Deletion/ Insertion	3669580	17
80-81	Cgga cg ggggg	c	Insertion	89354	52
1554	ggcgagcggctggt g ggact	g	Insertion	1656260	31
2966-2967	gcctttctgccc cg cg	cg	Insertion	3224880	21
3050-3051	cgtcg gc ggctc	g	Insertion	3323843	2
3050-3051	Atct ct ccg	c	Insertion	3323898	2
3233	Gct cg gccc	c	Insertion	3532827	18

3233-3234	Ggtaa ct cc	c	Insertion	3532838	18
4120	taaggccccc c gc	c	Insertion	4516560	54
1304	ttc gg tcacg gc gcat	gg/gc	Insertion/ Deletion	1410421	35
2373	aacagggt c ggg a tcc	c a	Insertion/ Deletion	2595887	26
1072-1073	Atatcc t tttc	c→t	Substitution	1157580	38
1235	Tggga a ttgg	a→g	Substitution	1320881	68
1235	tcgtcagaca t tcagct	t→c	Substitution	1325879	68
3051	Actca a agaa	c→a	Substitution	3324429	2
3052	gaat c cgctcgg	g→t	Substitution	3325048	20
3053	Gggcca a acg	g→a	Substitution	3325452	20
3916	cggcggg g cc c gctcaacca	g→c, c→t	Substitution	4292804	54

Table 8. Sequence changes in *C.segnis* genome by feature name, change type, and location.

REFERENCES

1. 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*. Abraham WR, et al. *Int J Syst Bacteriol* 1999 Jul
2. URAKAMI, TEIZI, et al. "Recharacterization and emended description of the genus *Mycoplana* and description of two new species, *Mycoplana ramosa* and *Mycoplana segnis*." *International Journal of Systematic Bacteriology* 40.4 (1990): 434-442.
3. BLAST Scoring Parameters, E. Michael Gertz, 16 March 2005.
4. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD and Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* 14(2), 193-202.
5. 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.
6. *The RAST Server: Rapid Annotations using Subsystems Technology*. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA,

McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. *BMC Genomics*, 2008

7. Artemis: sequence visualization and annotation. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA and Barrell B. *Bioinformatics (Oxford, England)* 2000;16;10;944-5 PUBMED: 11120685
8. Cruveiller, S., J., Le Saux, D., Vallenet, A., Lajus, S., Bocs And Médigue C. :**MICheck**: A web tool for fast check of syntactic annotations of bacterial genomes. *Nucleic Acids Res.* 2005 Jul 1;33(Web Server issue):W471-9.