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

Derrick C. Scott

Background

- *Alphaproteobacteria*
  - large and metabolically diverse group that includes the genus *Caulobacter*

Background

- *Alphaproteobacteria*
  - large and metabolically diverse group that includes the genus *Caulobacter*
  - found in essentially all habitats

Background

- *Alphaproteobacteria*
  - large and metabolically diverse group that includes the genus *Caulobacter*
  - found in essentially all habitats

- *Caulobacter* thrive in low nutrient conditions and generally share the same phenotypic properties.

Background

- *Alphaproteobacteria*
  - large and metabolically diverse group that includes the genus *Caulobacter*
  - found in essentially all habitats

- *Caulobacter* thrive in low nutrient conditions and generally share the same phenotypic properties.
  - Rod shaped and usually curved

Background

- *Alphaproteobacteria*
  - large and metabolically diverse group that includes the genus *Caulobacter*
  - found in essentially all habitats

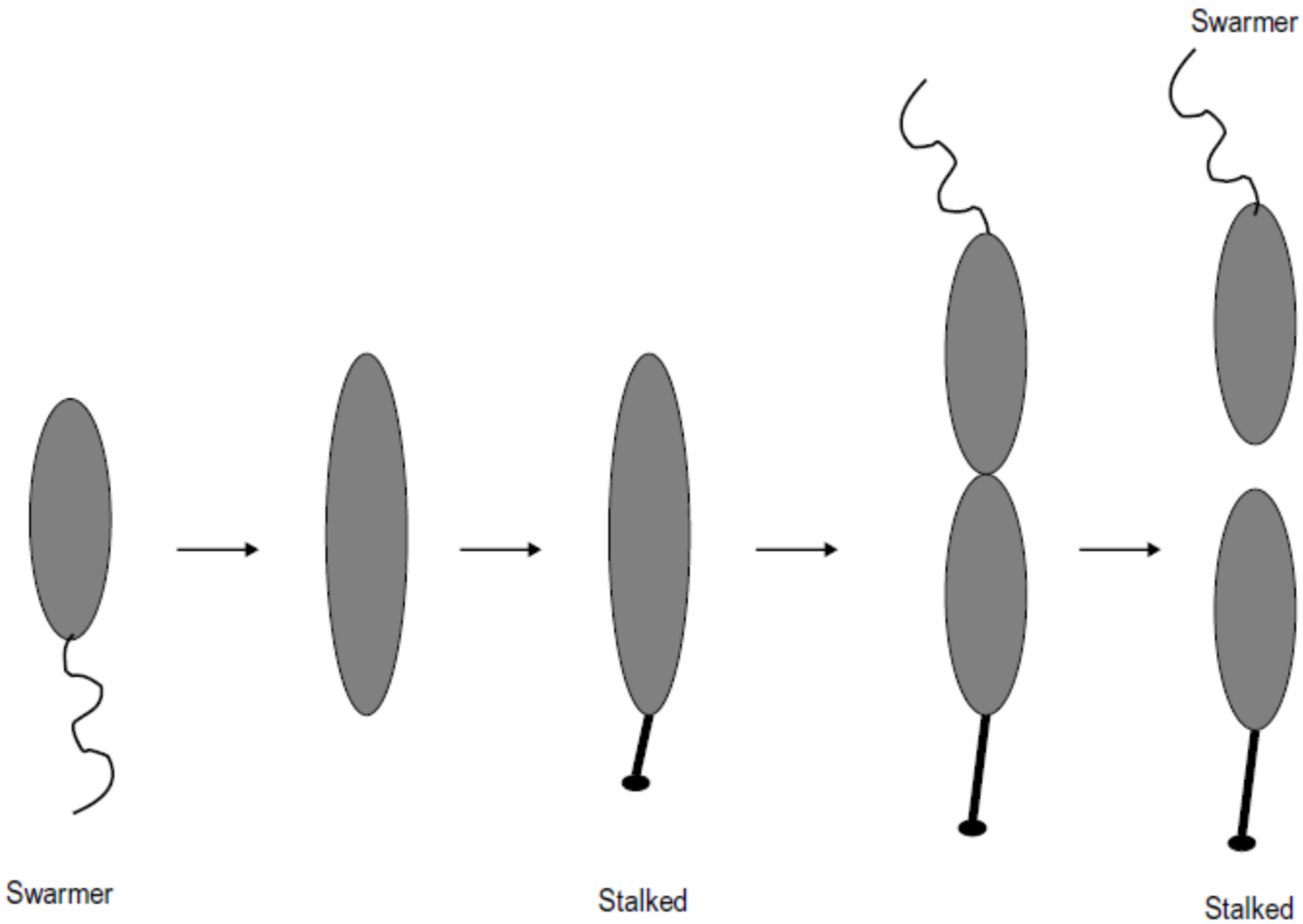
- *Caulobacter* thrive in low nutrient conditions and generally share the same phenotypic properties.
  - Rod shaped and usually curved
  - Gram negative

Background

- *Alphaproteobacteria*
  - large and metabolically diverse group that includes the genus *Caulobacter*
  - found in essentially all habitats

- *Caulobacter* thrive in low nutrient conditions and generally share the same phenotypic properties.
  - Rod shaped and usually curved
  - Gram negative
  - Display rare dimorphic phenotype

Background



Background

- Wealth of information available to support cell cycle research
- Study of evolutionary biology of Caulobacters is minimal

Introduction

- Despite ground breaking advances in the field of prokaryotic biology, there are many unanswered questions left to be studied that require the assembly of high quality bacterial genome sequences.
  - Extensive Evolutionary Studies

Introduction

- Despite ground breaking advances in the field of prokaryotic biology, there are many unanswered questions left to be studied that require the assembly of high quality bacterial genome sequences.
  - Extensive Evolutionary Studies
  - Comparison of Genomes

Introduction

- Despite ground breaking advances in the field of prokaryotic biology, there are many unanswered questions left to be studied that require the assembly of high quality bacterial genome sequences.
  - Extensive Evolutionary Studies
  - Comparison of Genomes
  - Proteomics

Introduction

- Despite ground breaking advances in the field of prokaryotic biology, there are many unanswered questions left to be studied that require the assembly of high quality bacterial genome sequences.
  - Extensive Evolutionary Studies
  - Comparison of Genomes
  - Proteomics

→ Cannot be done without a high quality genome

Introduction

- Despite ground breaking advances in the field of prokaryotic biology, there are many unanswered questions left to be studied that require the assembly of high quality bacterial genome sequences.
  - Extensive Evolutionary Studies
  - Comparison of Genomes
  - Proteomics

  → Cannot be done without a high quality genome
  - Most genomes are in permanent draft status
  - Traditionally has been labor intensive to sequence and finish assembling a genome

Objectives

- PART 1: To find a way to quickly and reliably sequence and assemble a bacterial genome
- PART2: Use our new sequences to do evolutionary studies, genome comparisons, and gain insights into “genome scrambling”
- PART 3: Follow up with additional strains of Caulobacter using data from PART 1 and PART 2

Comparison of Genome Sequencing Technology and Assembly Software For the Analysis of a GC-Rich Bacterial Genome

Derrick C. Scott

Methods

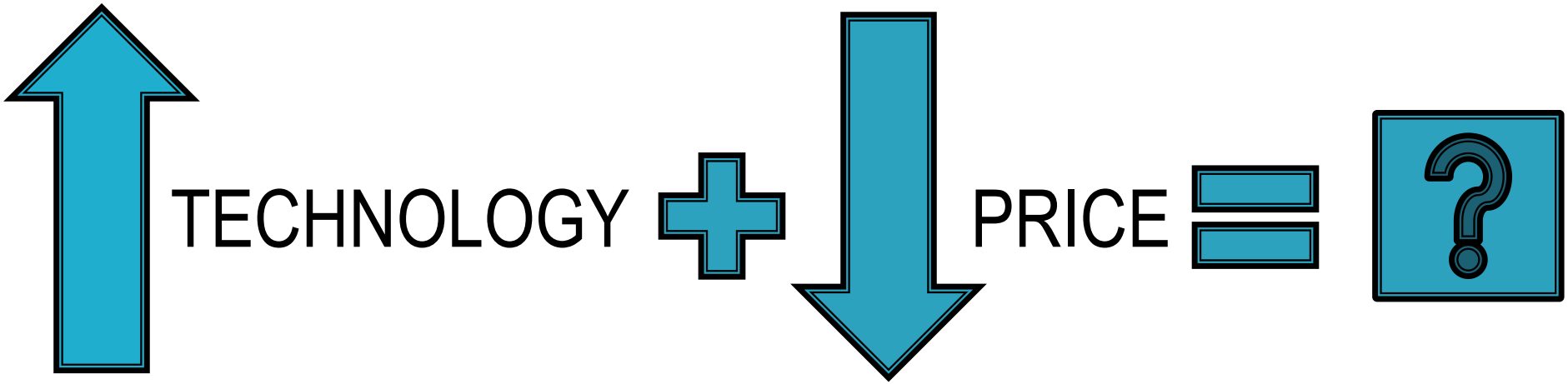
- First Publicly Funded Human Genome Project
  - 13 years and $3,000,000,000 (3 Billion) to complete

Methods

- First Publicly Funded Human Genome Project
  - 13 years and $3,000,000,000 (3 Billion) to complete

- Currently in Development
  - Less than 24 hours and $1,000 to complete

Methods



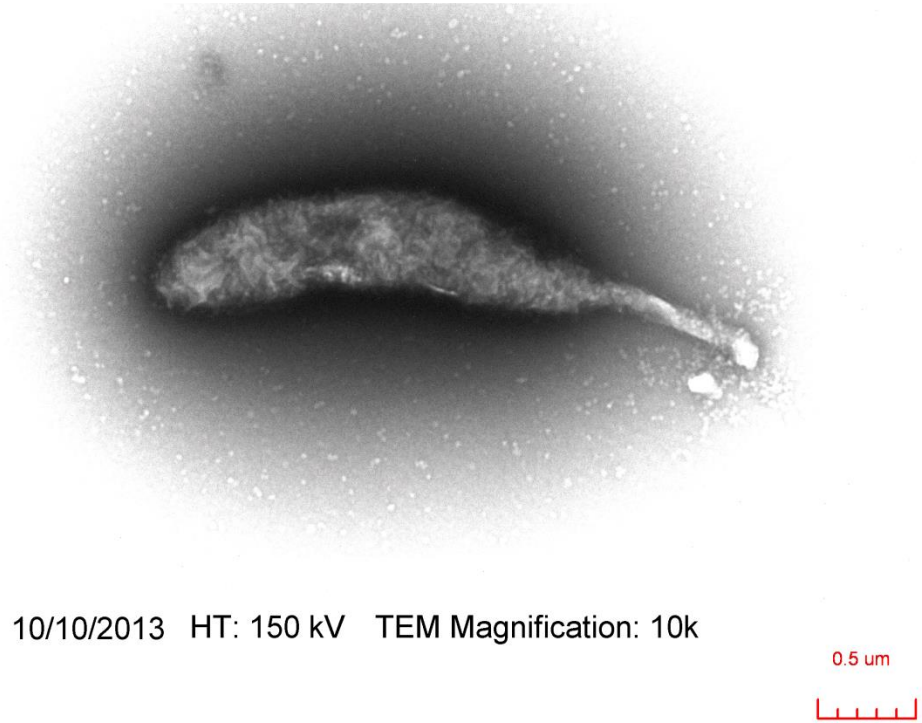
Methods

- Advantages and disadvantages associated with each individual technology
- No one size fits all approach to a quality genome assembly.
- Researchers with no experience in bioinformatics will be attempting the process of genome assembly.
  - ME!

Methods

- 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*

Methods



- We obtained a sample of the *Caulobacter henricii* bacterium from the American Type Culture Collection and extracted genomic DNA.

Methods

The Sequencers

Three genome sequencers were used in this study:

- Roche 454 GS FLX
- Illumina Miseq
- Pacific Biosciences RSII

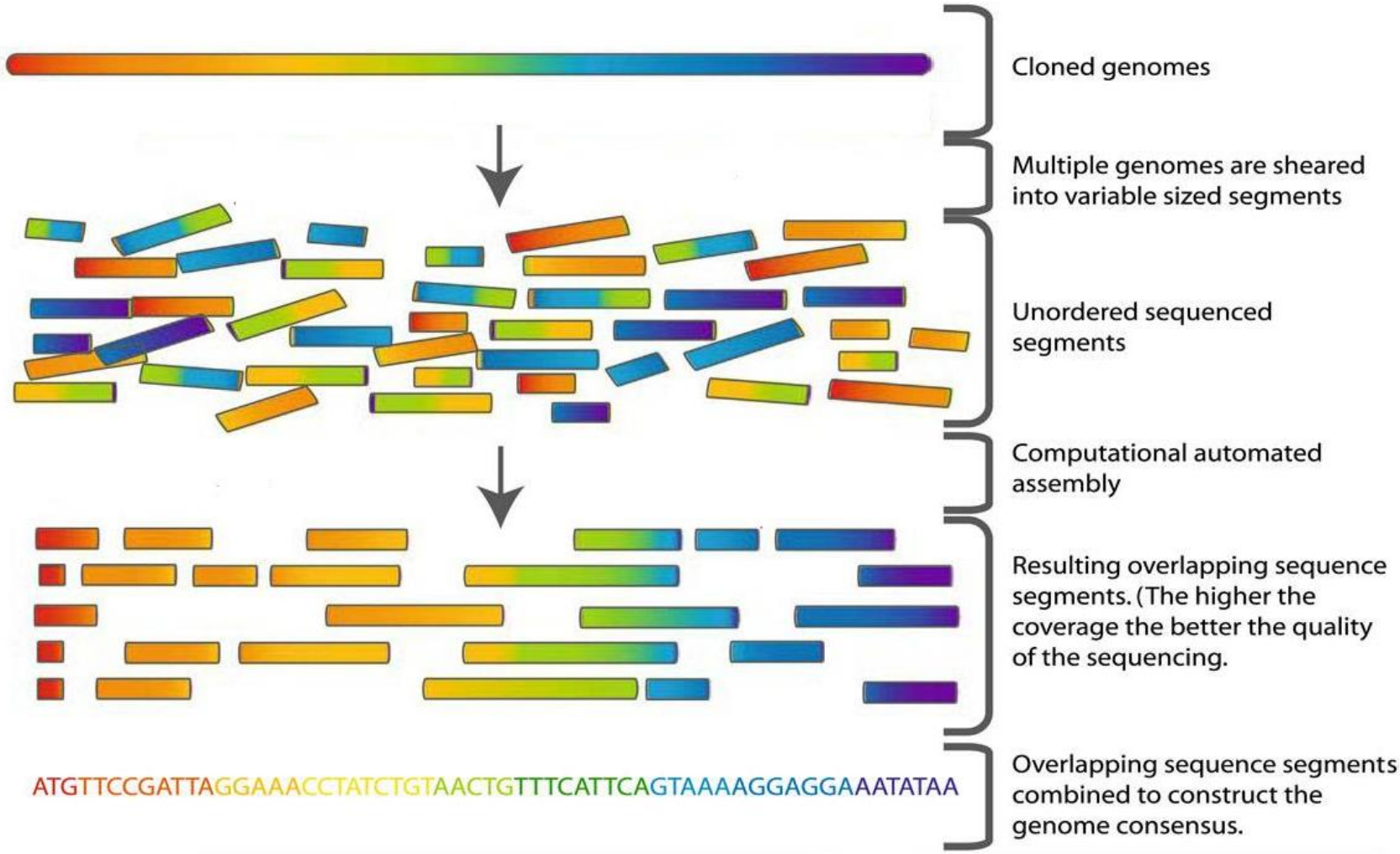
Methods

The Assemblers

Eight genome assemblers were used in this study:

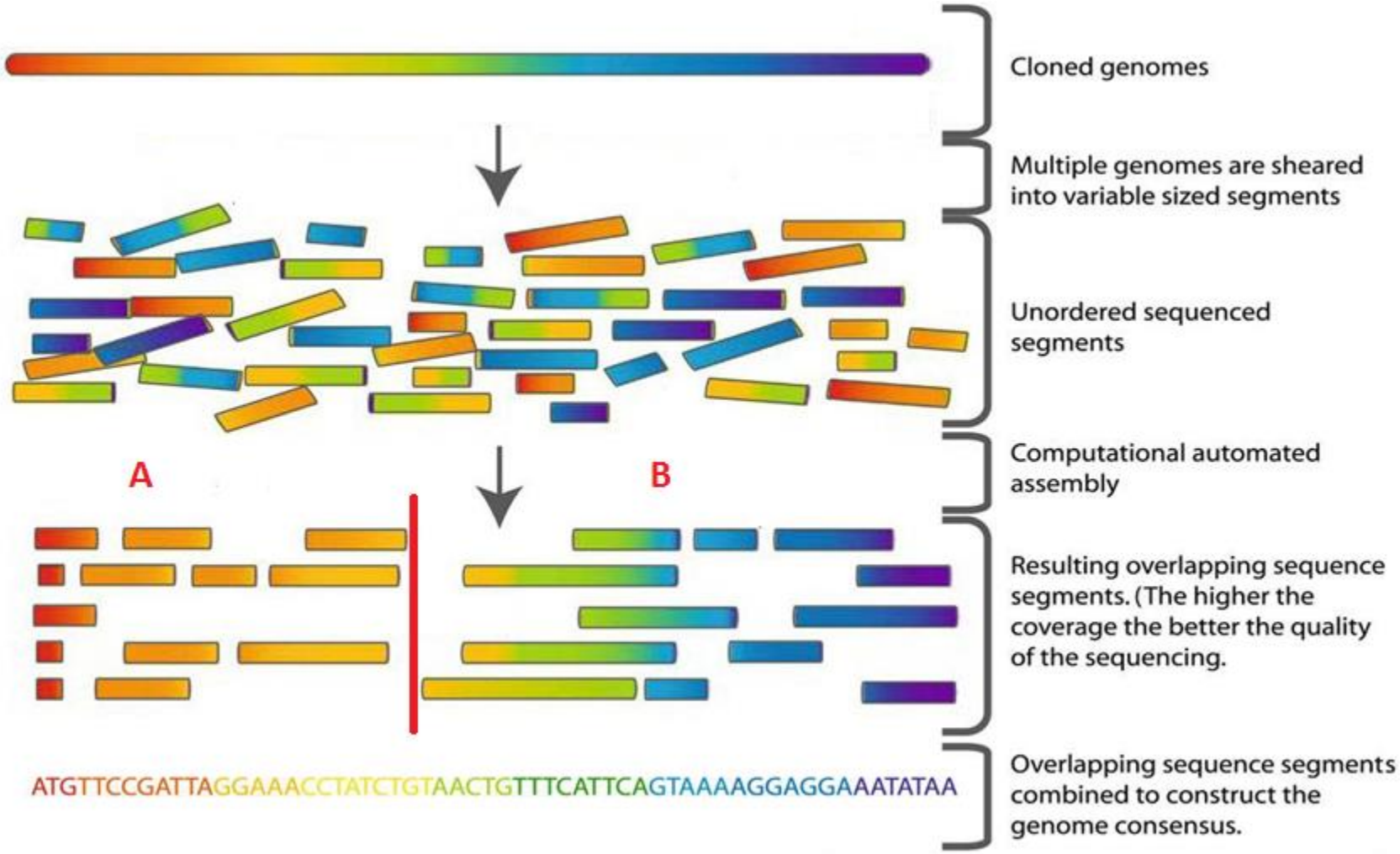
- Celera Assembler 8.0
- CLC Genomics Workbench 6
- HGAP 2.0
- MaSuRCA v2.1.0
- Newbler v2.6
- PANDAseq
- DNAStar SeqMan NGen 11.2.1
- SPAdes v2.5.1

Whole genome shotgun sequencing



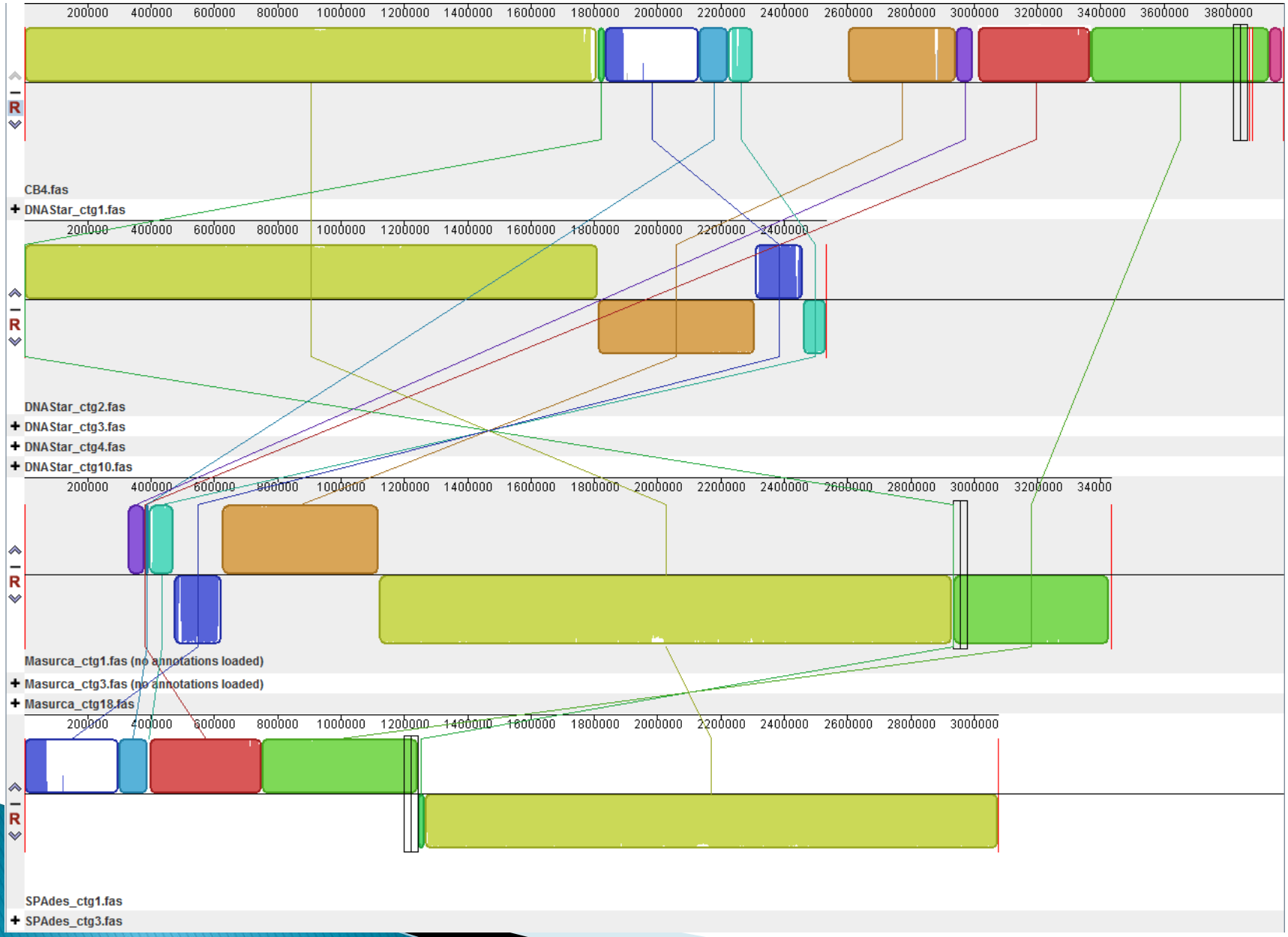
Adapted from Commins et al.

Whole genome shotgun sequencing



Adapted from Commins et al.

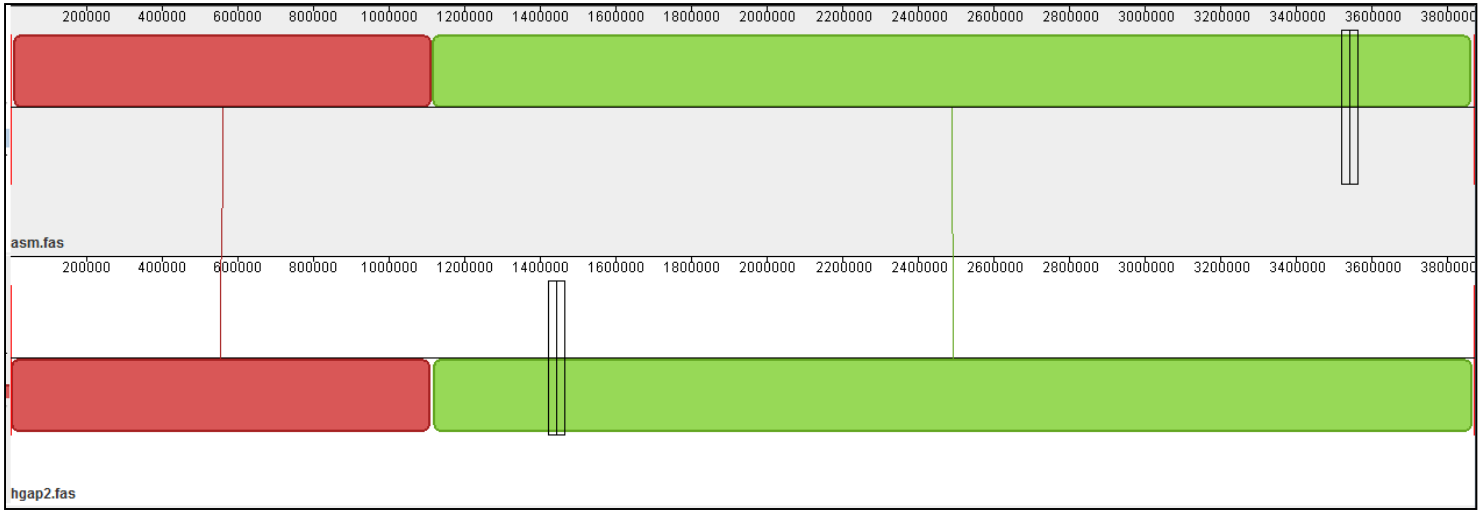
Assembler Builds



Comparison of assembler builds

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

Results



- Independently reached the same consensus build using two separate assembly algorithms.

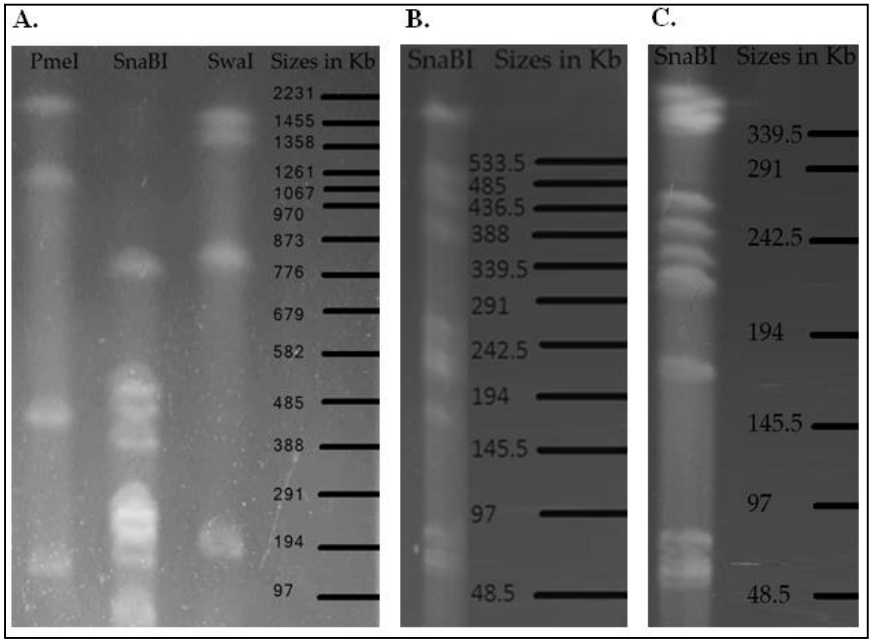
Results

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

- Positions reported of HGAP2 cut sites after Webcutter 2.0 analysis

Results

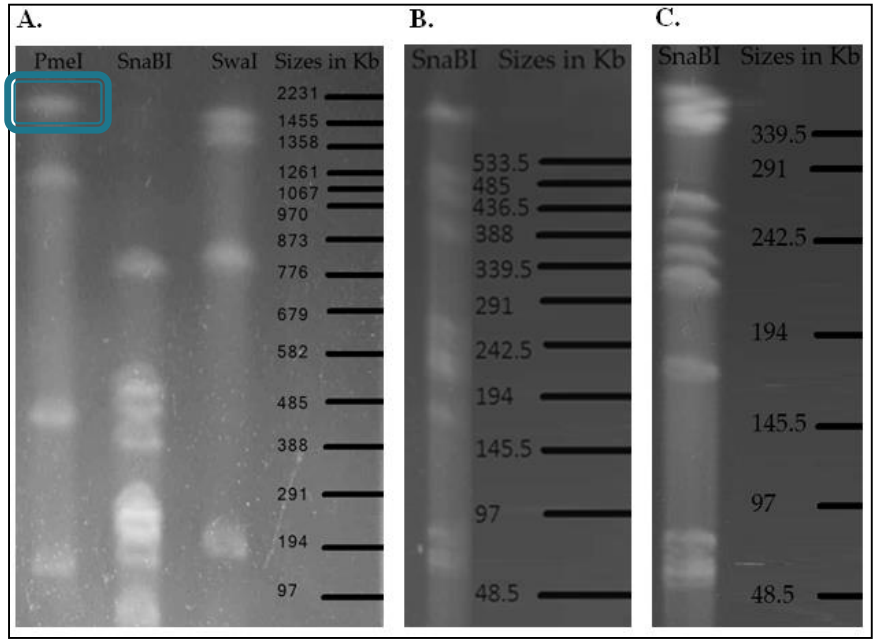
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	**3,868,732**	**3,868,732**	**3,868,732**



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

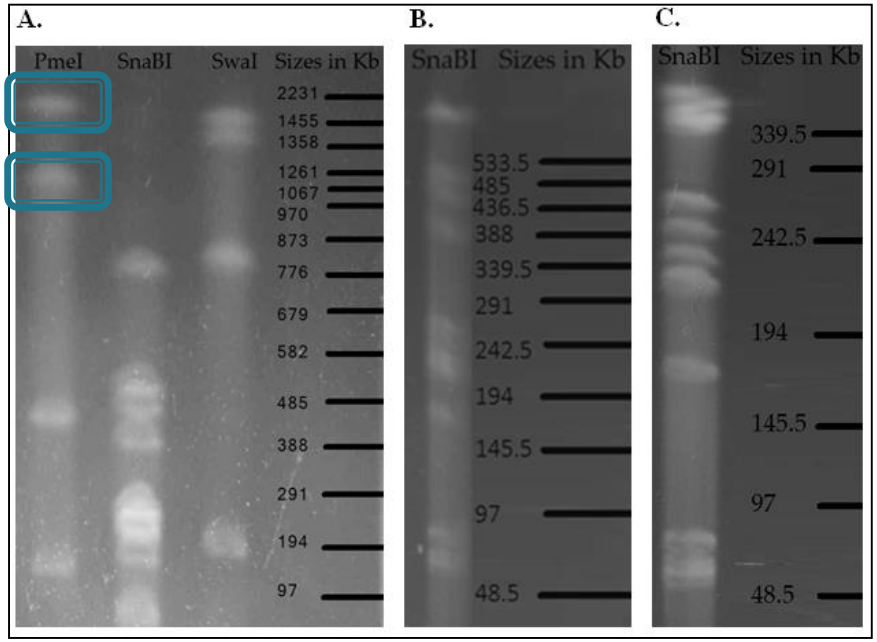
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	3,868,732	3,868,732	3,868,732



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

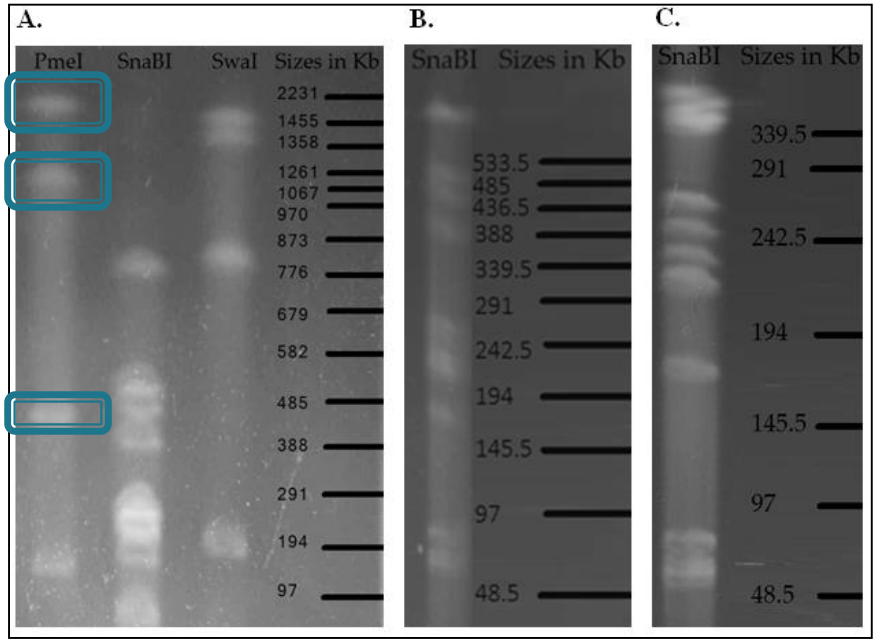
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	**3,868,732**	**3,868,732**	**3,868,732**



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

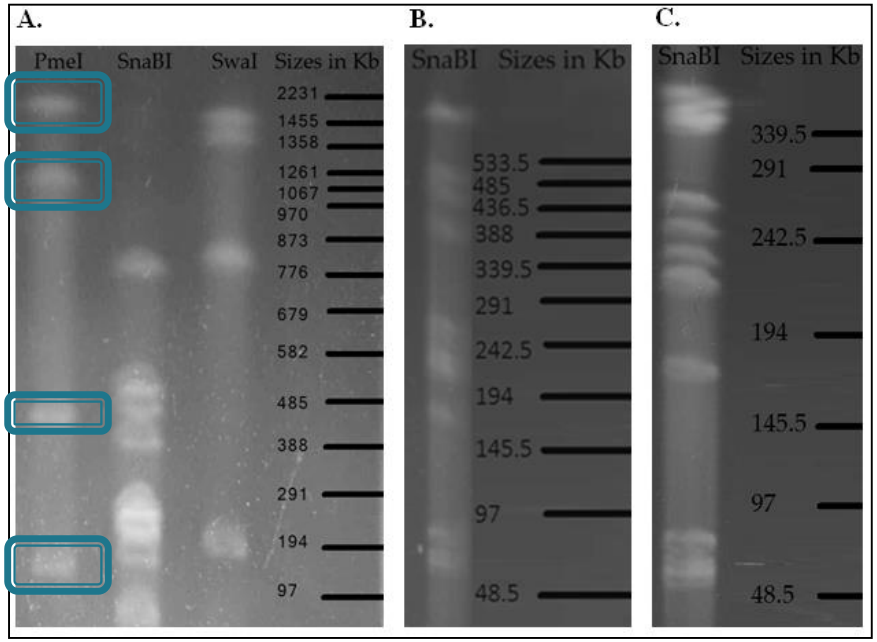
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	**3,868,732**	**3,868,732**	**3,868,732**



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

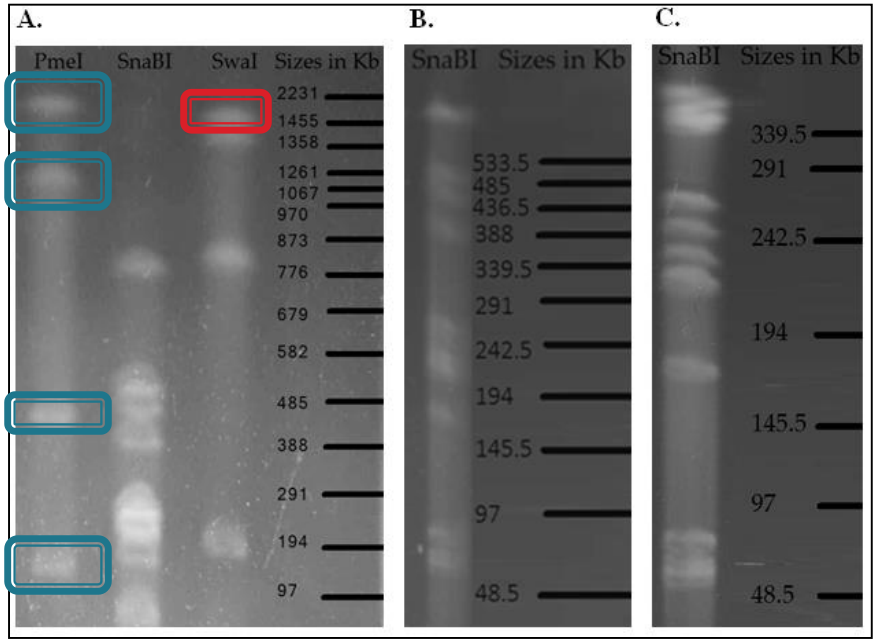
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	**3,868,732**	**3,868,732**	**3,868,732**



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

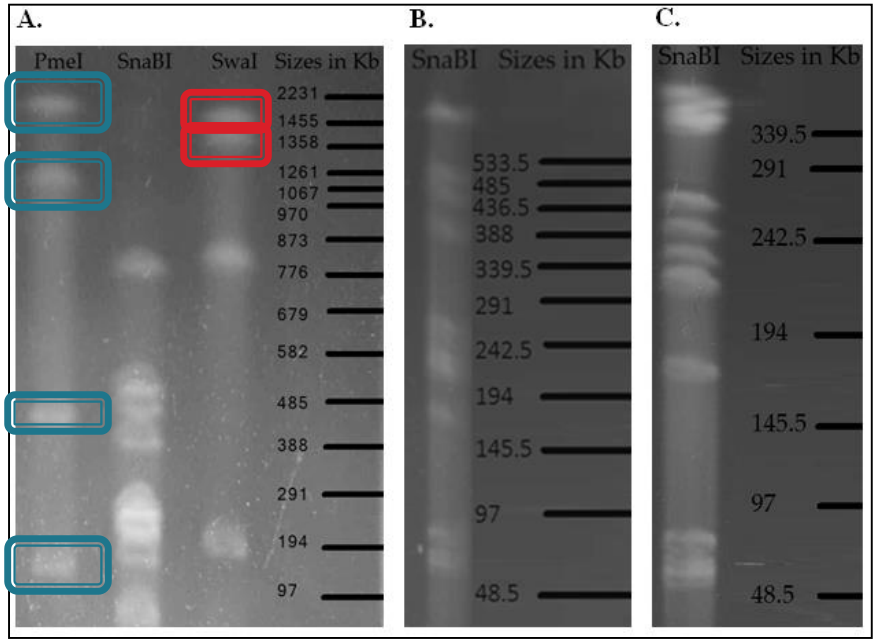
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	3,868,732	3,868,732	3,868,732



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

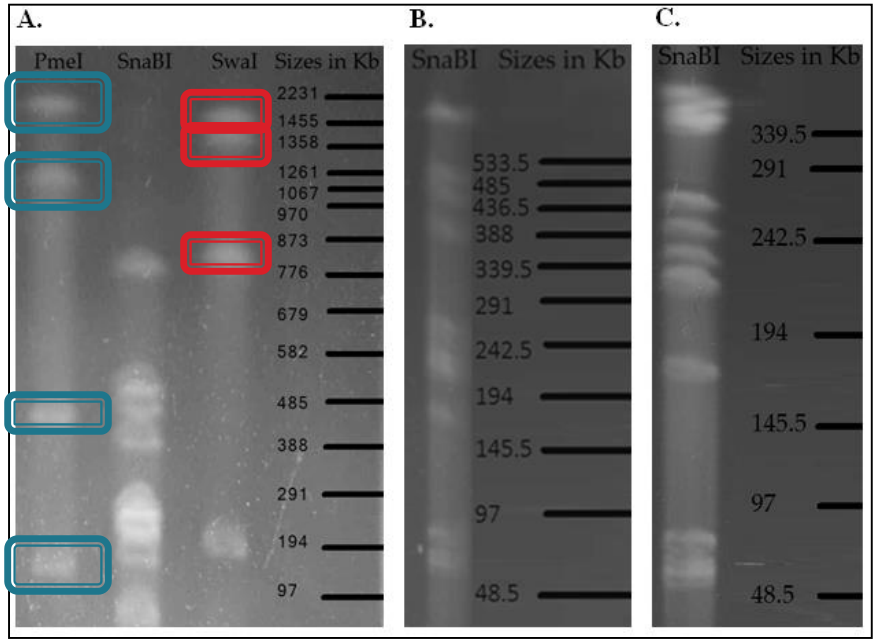
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	3,868,732	3,868,732	3,868,732



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

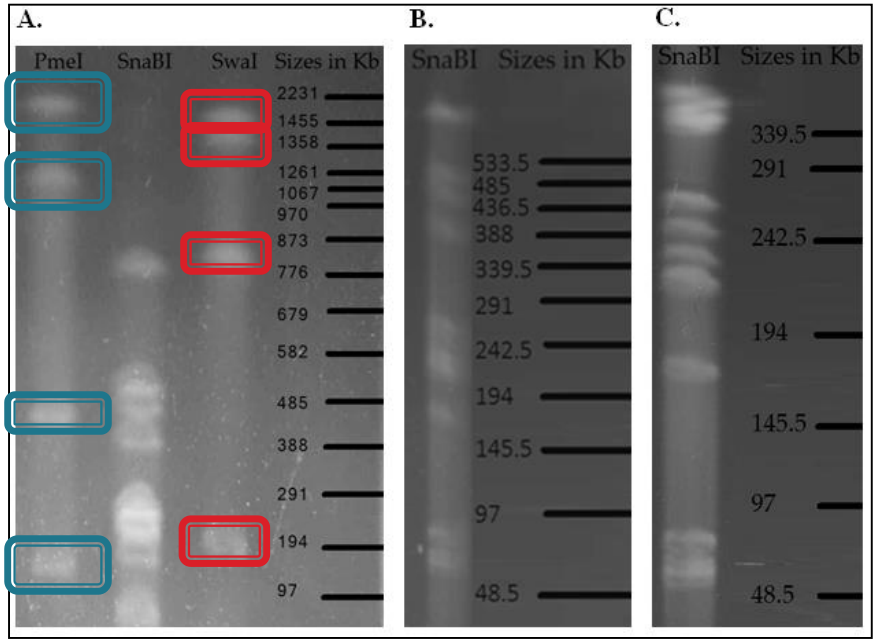
	PmeI	SnaBI	SwaI
	2,213,704	813,832	1,489,762
	1,072,969	514,225	1,312,723
	430,867	471,842	847,235
	151,192	404,820	219,012
		271,615	
		252,161	
		236,788	
		223,579	
		223,151	
		183,196	
		90,832	
		77,552	
		71,350	
		28,879	
		4910	
Total Bases	3,868,732	3,868,732	3,868,732



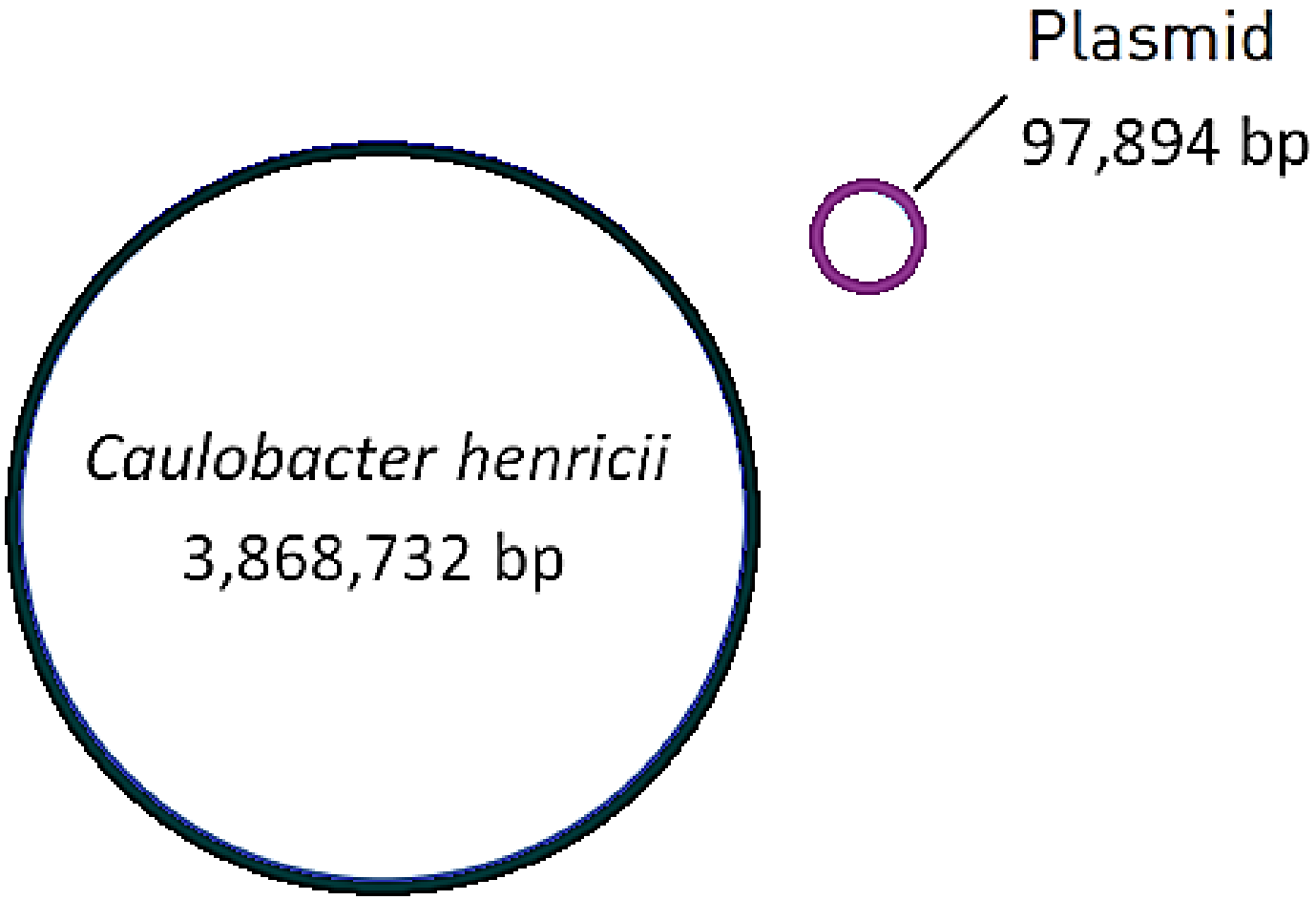
- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion

Results

	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



- Predicted fragment sizes of HGAP2 build after enzymatic digestion VS Observed fragments after enzymatic digestion



De novo, finished assembly of *Caulobacter henricii* with predicted accuracy of >99.987% (QV39)

Conclusions

- We found that software programs using only the MiSiq/454 data provided accurate yet numerous contigs that did not result in a complete assembly.
- The HGAP 2.0 assembler generated an accurate and complete *de novo* genome assembly of *Caulobacter henricii* using Pacific Biosciences RS II data.
- So did the Celera 8.0 assembler by error correcting the PacBio RS II long reads with Illumina short reads (PBcR).

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

Derrick C. Scott and Bert Ely

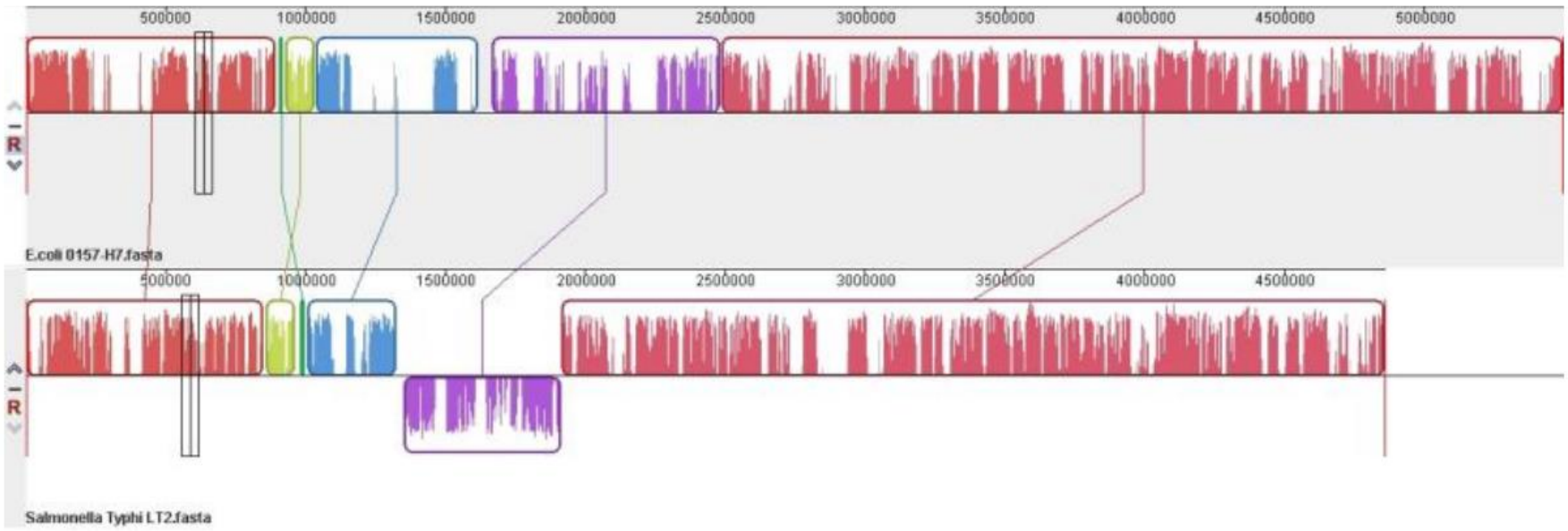
Department of Biological Sciences,

University of South Carolina, Columbia, SC, USA

Background

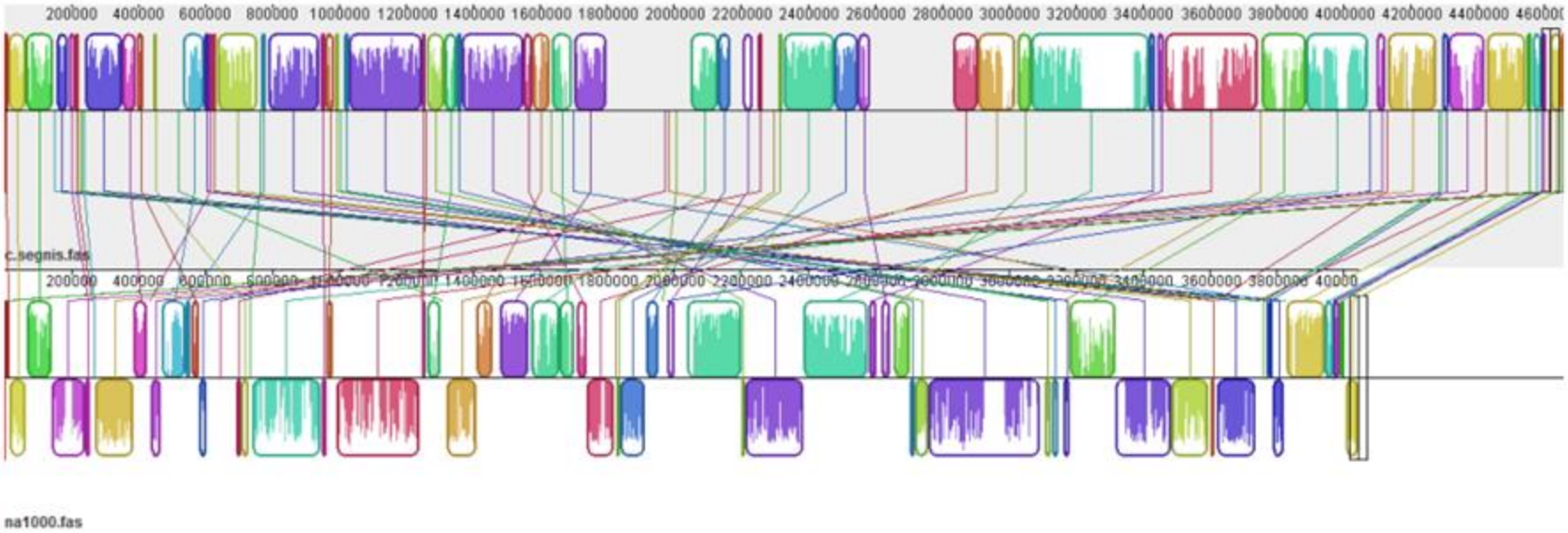
- *Caulobacter sp.* K31
  - Novel *Caulobacter* which was isolated from a research station in Finland.
- *C. crescentus* NA1000
  - Laboratory strain derived from *C. crescentus* CB15 *C. crescentus* NA1000
- *C. segnis* strain TK0059
  - Genome published in 2011
- *C. henricii CB4*
  - Newly Sequenced for this Study
- *Brevundimonas subvibrioides* strain CB81
  - Genome published in 2010
- *Brevundimonas DS20*
  - Newly Sequenced for this Study

Results



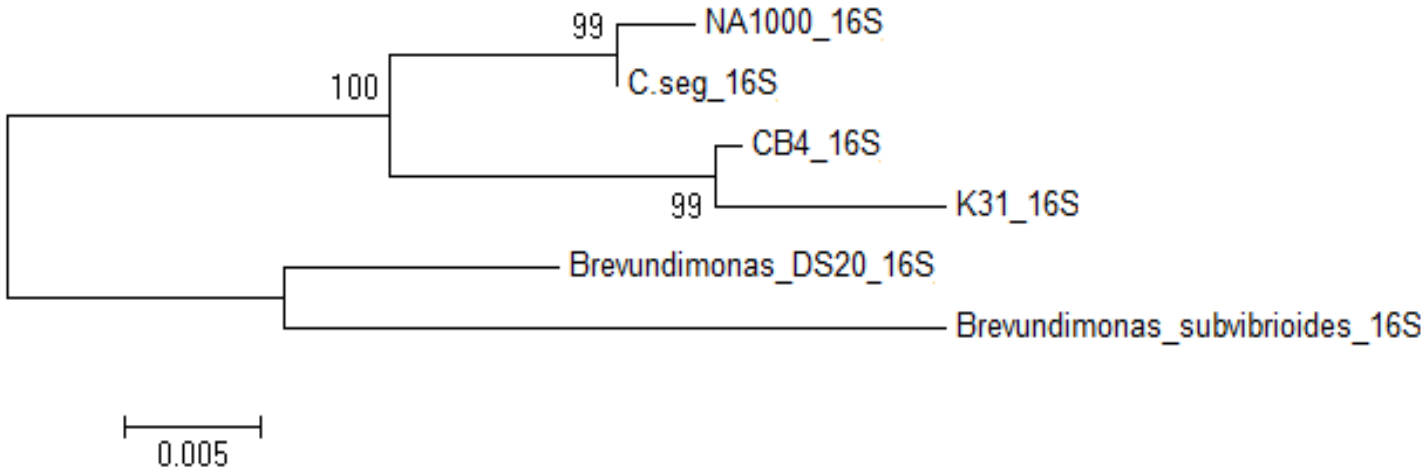
A comparison of the *Escherichia coli* and *Salmonella typhi* genomes. Ash and Ely, unpublished. Each line represents a rearrangement event.

Results



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

Results



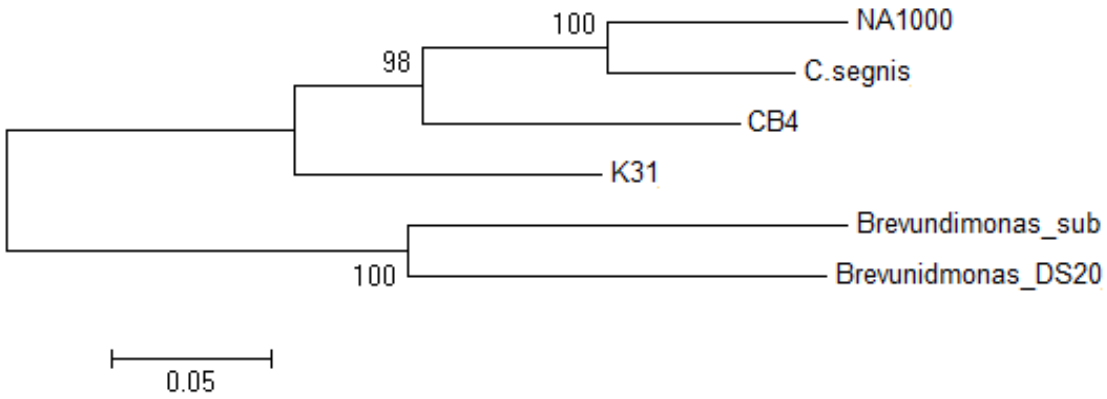
A comparison of 16S rRNA nucleotide sequences among the species included in this study (percent identity).

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

Disadvantages of 16s

- 16s region is relatively short
- Bacterial species often have multiple copies
- Most 16s in databanks are truncated
- 16s tree sometimes not congruent with actual gene-gene homology

Results



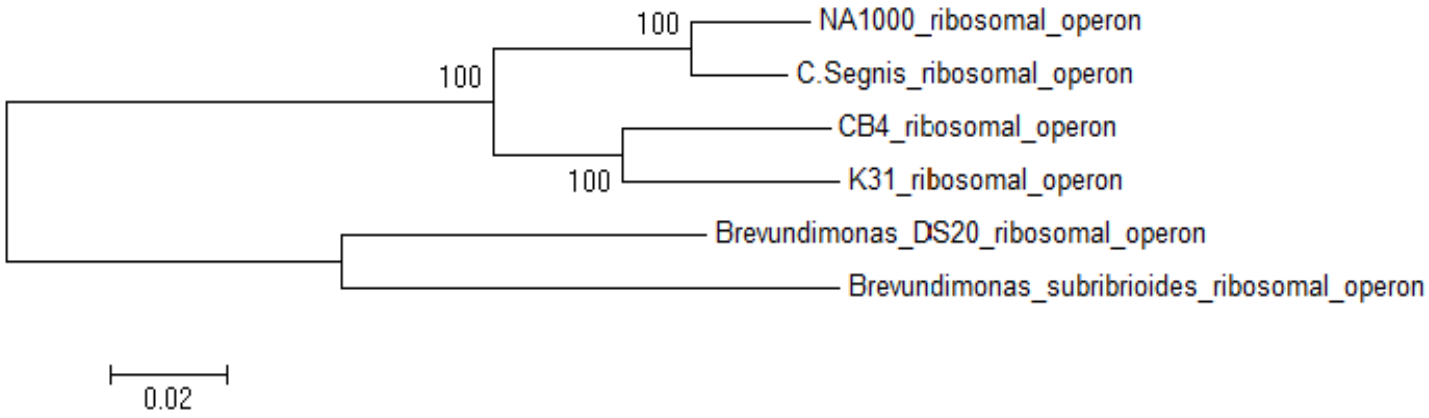
A comparison of dcw cluster nucleotide sequences among the species included in this study (percent identity). 26 gene operon.

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

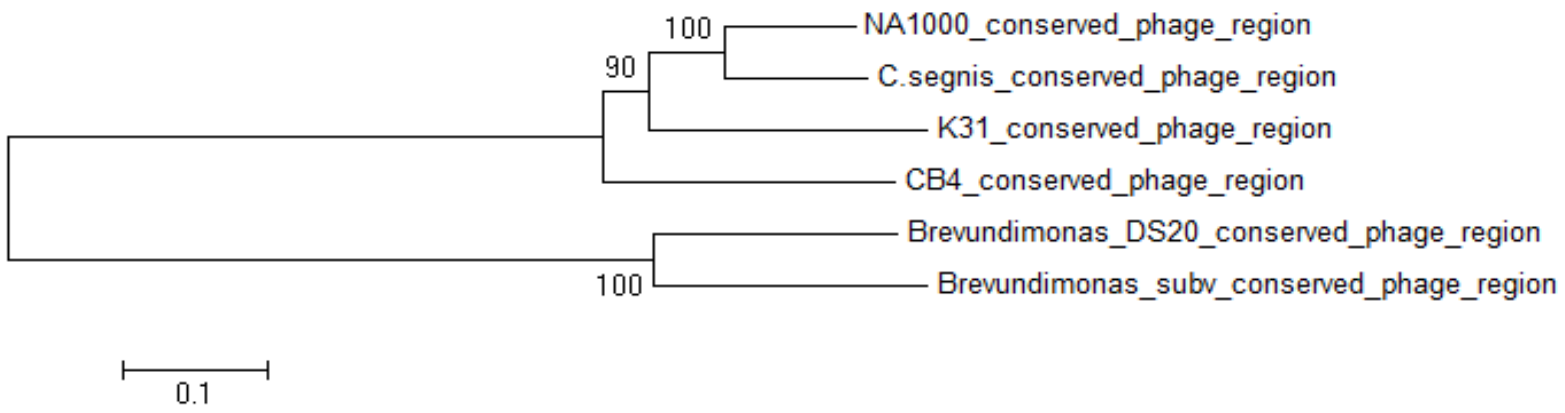
Results

A comparison of ribosomal protein operon nucleotide sequences among the species included in this study (percent identity). 28 gene operon.

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



Results



A comparison of conserved phage region nucleotide sequences among the species included in this study (percent identity). 20 gene operon.

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

Background

- The nucleotide sequence differs by as much 17% in pairwise
- No significant identity in Brevundimonads
- Upon closer inspection, we found that there was significant amino acid identity among the genes in this region in all six genomes.

Background

- *Caulobacter* phage regions
  - Codon usage bias for CTG (Leucine), GGG (Glycine), GCG (Alanine), and CGG (Arginine)

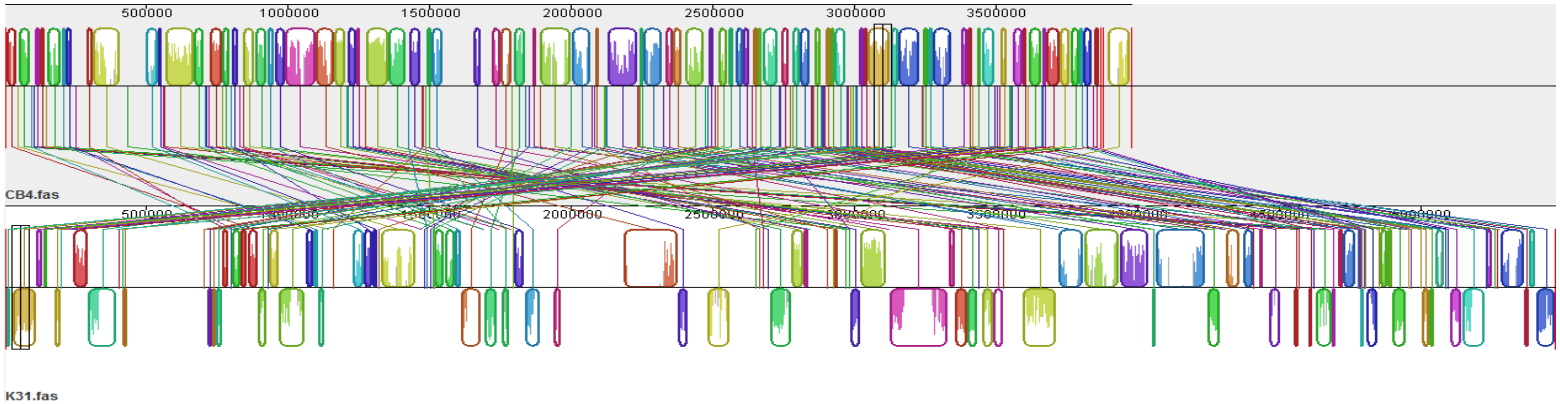
- *Brevundimonas* phage region
  - Bias towards CTC(Leucine), CGC (Glycine), GCC (Alanine), and CGC(Arginine)

- We were also able to locate an inversion event in the *Brevundimonads* that was absent in the *Caulobacters.*

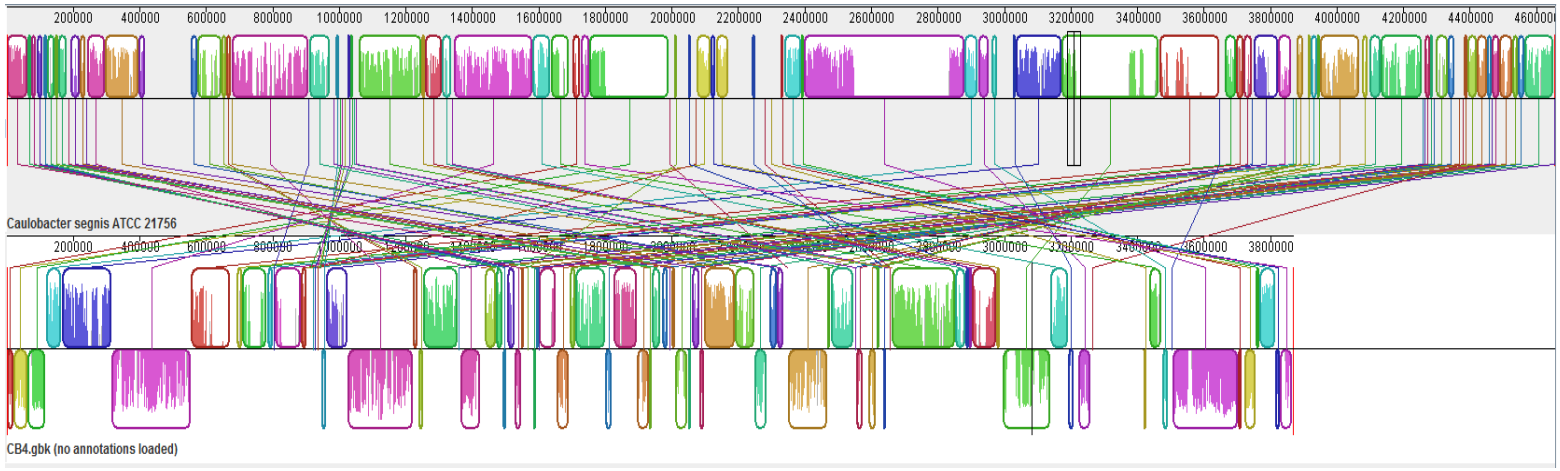
- Only found codon bias in phage region

Results

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



MAUVE alignment of CB4 (top) and K31 (bottom).



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

Results

- Previous studies have shown that *Caulobacters* exhibit an extremely high rate of genome rearrangement when compared to similarly related bacteria.
- We found no correlation between relatedness and genome scrambling
- Scrambling did not disrupt the conservation of the essential genome
- More studies are needed to determine exactly what is responsible for the organized chaos that is genome scrambling in *Caulobacters*.