Spring 2019

ANTI-CRISPR vs. CRISPR: The evolutionary arms race between microorganisms

Rachael M. St. Jacques

ANTI-CRISPR vs. CRISPR: The evolutionary arms race between microorganisms

Rachael M. St. Jacques

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Science

Department of Graduate Biology

May 2019

FACULTY COMMITTEE:

Committee Chair: Steven G. Cresawn PhD

Committee Members/Readers:

James Herrick PhD

Kim Slekar PhD

Chris Berndsen PhD

Acknowledgements

I would love to acknowledge those who have had an impact on my research. My committee members Dr. Herrick, Dr. Slekar, and Dr. Berndsen have all been wonderful. Dr. Herrick is a fantastic professor and a compassionate person. He is a wonderful advisor and is always there for his students. Dr. Slekar is incredibly knowledgeable in her field, and provides wonderful insight during our conversations. Dr. Berndsen is an incredibly patient man, and will not hesitate to explain complex ideas in a kind manner.

Other professors have also been mentors to me during my time at James Madison University. Dr. Renfroe is widely known for how much he cares about his students. I have always been able to go him for both good conversation and advice. Drs. Enke, Steffen, and Parker have all taught me how to be a better scientist through their excellent courses.

Special thanks should go to my advisor, Dr. Steven Cresawn, who has all of the previously mentioned qualities and more. I consider myself incredibly lucky to have been his mentee, as a graduate student could not possibly have a better mentor.

Thank you to everyone who has had an impact on me and my research.

Table of Contents

List of Tables

List of Figures

Abstract

CRISPR arrays are a defense mechanism employed by bacteria against viral invaders. Cas proteins do the work in detecting, capturing, and integrating the viral DNA into the CRISPR array (Barrangou et al., 2007). Anti-CRISPR proteins are produced by phages, viruses that infect bacteria, to stop the bacterial host's CRISPR-Cas complex from interrupting the phage life cycle (Bondy-Denomy, et al., 2015).

SEA-PHAGES is a course-based bacteriophage research network composed of 120 colleges and known at James Madison University as Viral Discovery. JMU uses the unsequenced *Streptomyces griseus* ATCC10137 as a host for bacteriophage discovery and propagation, and in this study we report the sequencing and analysis of this strain, including a search for CRISPR-Cas arrays. To determine if the *S. griseus* ATCC 10137 encodes CRISPR-Cas arrays, next generation sequencing and bioinformatic analyses were performed.

DNA extraction and whole genome sequencing using an Illumina MiniSeq and Oxford MinION were used to obtain sequence data from *S. griseus* ATCC10137. The Illumina reads were trimmed using Trimmomatic, and the Nanopore data were filtered using Filtlong. A hybrid genome assembly using the Illumina reads and Nanopore reads was generated using Unicycler, resulting in a genome assembly that was 8,576,363 bp long. To determine if CRISPR-Cas arrays were present in the genome, the assembly fasta was uploaded to CRISPRfinder. CRISPRfinder identified 3 probable CRISPR arrays, and 10 questionable regions.

Automated annotation methods Prokka and RAST were used to predict genes in the *S. griseus* genome, but they produced substantially different output. We therefore

developed the novel bioinformatic tool, Prokkrastinator, to merge the two annotation methods. Prokkrastinator doubles as a genome browser and gene table.

To search for anti-CRISPR proteins in the genome of bacteriophage Wipeout, protein models were generated using YASARA. Of the 62.57 % of the Wipeout gene products that were not able to be modeled, 13 gene products were in the same size range as all other phage-produced anti-CRISPR proteins, 50-150 amino acids long. These 13 gene products should be further studied to determine whether or not they could be potential anti-CRISPR proteins, as they are the only proteins of the appropriate size.

Introduction

Background

Clustered regularly interspaced short palindromic repeats (CRISPR) are a defense mechanism employed by bacteria against viral and plasmid invaders (Barrangou et al., 2007). This defense is an array with a series of sequences of repeating bacterial host DNA that are interspaced with viral DNA sequences collected from prior infections (Barrangou et al., 2007; Hale et al., 2009). CRISPR-associated proteins, or Cas proteins, do the work in detecting, capturing and integrating the viral DNA into the CRISPR array (Barrangou et al., 2007). This system has been modified to work within eukaryotic cells as a genome editing tool, for example to repair mutations in the CFTR gene that causes cystic fibrosis (Firth et al., 2015).

Anti-CRISPR proteins are produced by phages, viruses that infect bacteria, to stop the bacterial host's CRISPR-Cas complex from interrupting the phage life cycle. Some of these phage anti-CRISPR proteins can manipulate the host's precursor CRISPR-Cas proteins, Csy proteins, into becoming a transcriptional repressor that blocks the host's own CRISPR-Cas transcription (Bondy-Denomy et al., 2015). Anti-CRISPR proteins have been used in cystic fibrosis studies, which used both CRISPR-Cas to edit the diseased cell genome, and anti-CRISPR proteins to halt the progression of the CRISPR-Cas complex. The use of anti-CRISPR proteins in conjunction with the CRISPR-Cas complex was necessary to inhibit unintended off-site targeting of the cell DNA by the CRISPR-Cas complex in the experimental trials (Shin et al., 2017). Because these phage

proteins are a relatively new discovery, important information such as their abundance, dispersal, structures and specific mechanisms are just beginning to be explored.

The natural purposes of CRISPR-Cas arrays are useful in a medical research setting. For example, mouse zygote genomes were edited by injecting CRISPR-Cas components into a cell during embryological development (Wang H et al., 2013). The cataract-causing allele Crygc was replaced with a non-deleterious allele in mice using CRISPR-Cas9 (Wu Y et al., 2013). While CRISPR-Cas9, also known as an RNA guided endonuclease (RGEN), has incredible capabilities and can be relatively simple to use, it is not always able to be used in a reliable and predictable manner. These particular RGENs have been shown to have off-site targeting (Cho et al., 2013).

The Cas9-single-guide RNA (Cas9-sgRNA) complex can have deleterious off-site effects when used in repairing mutated alleles or removing proviral DNA from host DNA (Shin et al., 2017). To combat these effects, the anti-CRISPR protein AcrIIA4 can be used alongside the Cas9-single-guide RNA treatment. Shin et al. used the phage protein AcrIIA4 in their experiments, as it acts as an inhibitor to the CRISPR-Cas complex by binding to the CRISPR-Cas active site. AcrIIA4 was injected into an experimental cell six hours after Cas9-sgRNA treatment. The competitive binding of anti-CRISPR left the Cas9-sgRNA complex unable to continue degrading DNA, thus alleviating some of the unintended cutting of non-target DNA (Shin et al., 2017).

CRISPR-Cas is a remarkable tool, but it is still imperfect. Anti-CRISPRs are proteins that have been shown to alleviate or negate CRISPR-Cas' deleterious effects, but there are large information gaps: how many there are, their structure, and how they

function. A deeper understanding of these proteins could lead to the development of fully functional and diverse genetic tools for further experimentation and clinical use.

CRISPR-Cas: Mechanisms, Types, and Subtypes

Before the CRISPR-Cas-guide RNA complex is fully functional, the Cas proteins and guide RNA are first transcribed from the host DNA and translated into pre-crRNA. Protein editing of the pre-crRNA creates the fully mature complex (Bondy-Denomy et al., 2015). CRISPR associated proteins, or Cas proteins, do the work of detecting, capturing and integrating the viral DNA into the CRISPR array (Brouns et al., 2008). Certain Cas proteins, such as the nuclease Cas9, bind with transcribed viral DNA that was encoded in the CRISPR array and a guide RNA protein (Hsu et al., 2014). This complex patrols the cell and cuts the invading viral genome at specific genomic sites (Rath et al., 2015). There are three main "types" of CRISPR systems (discussed below), which each share the following characteristics: a repeating phage or plasmid- derived DNA sequence and CRISPR associated proteins. These CRISPR types are categorized into subtypes.

Type I CRISPR-Cas targets and cuts foreign DNA with the sgRNA and Cas3 protein, which encodes a large helicase with DNase activity. Type I CRISPR systems also encode Cas5, Cas6, and Cas7, which are classified as being in the Repair Associated Mysterious Proteins (RAMP) superfamily and are active in the Cascade complex, which is a group of proteins that is used in lieu of Cas9 (Makarova, 2011). In the subtype I-F CRISPR-Cas system, endoribonuclease Csy4 binds to and cuts the repeat sequence in the pre-crRNA. This then complexes with the Csy1, Csy2 and Csy3 proteins that support the functional CRISPR-Cas complex (Bondy-Denomy et al., 2016).

Type II CRISPR-Cas systems employ the well-known Cas9, a large protein that processes the crRNA and cleaves target DNA. The Cas9 protein has at least two domains, with the RuvC nuclease domain residing at the N terminus and the HNH nuclease domain (likely responsible for target degradation) residing in the middle of the protein. Also included in Type II systems are the proteins Cas1 and Cas2, which are responsible for acquiring spacers for the CRISPR array (Makarova, 2011).

Type III CRISPR-Cas utilizes RAMP and Cas6 proteins where they act in a similar way to the Cascade proteins in type I. Lacking in the Type III CRISPR operon are the Cas1 and Cas2 genes; instead the type III CRISPR-Cas system utilizes bacterial trans-encoded proteins. Type III can be split into two subtypes- types III-A and III-B. Subtype III-A has been known to cleave DNA, while III-B has been shown to cleave RNA (Makarova, 2011).

In all CRISPR-Cas types, three main events transpire. The primary event is spacer acquisition, which Cas1 and Cas2 always participate in, and involves copying the invading viral DNA into the bacterial CRISPR array at the beginning, or leader end of the CRISPR array (Brouns et al., 2008). They are present in types I and II, and function in the cascade complex in type III (Makarova, 2011). These stored viral DNA sequences are used by the bacteria to provide adaptive immunity against that particular virus in the future by becoming the spacers in the array (Barrangou et al., 2007; Hale et al., 2009).

In type II systems, after the spacer is acquired, the entire CRISPR array is transcribed and processed into small CRISPR RNAs, or crRNAs (Brouns et al., 2008). Once processing of the crRNA occurs, it complexes with the tracrRNA. Once the RNA complex has been established, the Cas nuclease protein Cas9 is transcribed and

translated. The Cas9 protein then complexes with the crRNA/tracrRNA and roams the bacterial cell, hunting for Cas9's protospacer adjacent motif (PAM), which is the viral DNA (Barrangou et al., 2014; Cong et al., 2013). Because the bacterial genome is genetically altered, the changes can be passed on to daughter cells when bacteria undergo binary fission.

The third and last event of the CRISPR-Cas system is interference. Once the complex has bound to the invading DNA, the nucleases cut the target DNA and discard it (Van der Osst et al., 2014).

Anti-CRISPR: What they Inhibit and How

Of the known anti-CRISPR proteins, at least ten inhibit the type I-F CRISPR-Cas systems and are found in *Pseudomonas aeruginosa* (Pawluk et al., 2016). The more exclusively studied anti-CRISPR proteins, anti-CRISPRs 1, 2, 3 and 4 are found within the same genomic region in phage genomes. These anti-CRISPR proteins are found between two conserved capsid protein genes (Pawluk et al., 2014).

Three of these ten anti-CRISPR proteins, AcrF1, AcrF2 and AcrF4, interact with the Csy complex and directly inhibit the complex's ability to bind to the PAM site in the dsDNA target. AcrF1 and AcrF2 bind to different areas of the Csy complex, and do not inhibit each other in the process. They are able to simultaneously bind to the Csy complex and prevent it from binding to target DNA (Bondy-Denomy et al., 2015).

AcrF3 interacts with Cas3 and blocks the helicase-nuclease from binding to the Csy-DNA complex. Because Cas3 is unable to bind to Csy, the protein acts as a transcriptional repressor and blocks the transcription of the CRISPR-Cas complex (Bondy-Denomy et al., 2015).

Of the ten previously mentioned anti-CRISPR proteins, four are found in the same operon as the I-F *P. aeruginosa* anti-CRISPRs and inhibit the type I-E CRISPR system (Pawluk et al., 2014; Pawluk et al., 2016). Each of these small anti-CRISPR proteins are between 50-150 amino acids in length (Rauch et al., 2017).

There are other anti-CRISPR proteins that function against type II CRISPR-Cas9 systems. At least four of these anti-CRISPR proteins were found in *Listeria monocytogenes*. Of these four anti-CRISPR proteins, two are effective against CRISPR-Cas systems in *Streptococcus pyogenes*. AcrIIA2 and acrIIA4 have been suggested to inhibit the *S. pyogenes* Cas9 (Rauch et al., 2017).

The Arms Race

The evolutionary arms race between bacteria and their phages has likely been occurring since bacteria-phage interactions originated. In industry, the ability for bacteria to withstand phage attack is invaluable. The dairy industry struggles with phages and plasmids running rampant through their production facilities. This situation introduces selective pressure on those bacteria that are best able to fight off the infections (Barrangou et al., 2014).

Dairy industry investors had a pressing need to discover which strains of *Streptococcus thermophilus*, bacteria used in the dairy production industry, were resistant to phage infection (Barrangou et al., 2014). *S. thermophilus* CRISPR-Cas was found to encode CRISPR-Cas arrays. To determine if these bacteria were immune to common phage contaminants, *S. thermophilus* cultures introduced to *S. thermophilus* phages. Most *S. thermophilus* phages could not infect the host due to the CRISPR-Cas arrays. The

evolutionary pressures applied to *S. thermophilus* bacteria by phage infection led to the acquisition of CRISPR-Cas arrays in an industrial setting (Barrangou et al., 2014).

Pseudomonas aeruginosa also encodes at least one CRISPR-Cas array. *P. aeruginosa* phages were able to develop a work-around through evolutionary pressures on their genes, which led to the development of anti-CRISPR proteins. (Bondy-Denomy et al., 2013).

The Deleterious Effects of CRISPR-Cas, and the Anti-CRISPR Proteins that Stop Them; CRISPR-Cas Genome Editing in the Lab

CRISPR-Cas9 has recently been utilized in labs to edit genomes because of the ease of use and effective results with the molecule. Gene editing of mutations that cause disease, such as attempted manipulation of the CFTR gene, has shown promise (Firth et al., 2015). Other experiments with gene-editing and CRISPR-Cas, such as those on HIV-infected cells, have been more complicated and yielded undesired results (Wang, G et al., 2016; Wang, Z et al., 2016).

The use of CRISPR-Cas9 to remove HIV-I proviral DNA from infected cells was mildly successful, but also encouraged genetic changes in some HIV viral particles, enabling them to escape the attack. The mammalian error-prone Non-Homologous End-Joining Pathway (NHEJ Pathway) added insertions, deletions and substitutions into the genome of the HIV after being cut by the CRISPR-Cas9-guide RNA. These changes enabled the escaped particles to avoid further degradation by the pre-fabricated CRISPR-Cas9-guide RNA complex (Wang, G et al., 2016; Wang, Z et al., 2016). The natural function of the NHEJ pathway happened to benefit the virus (Wang, G et al., 2016).

Insertions and deletions (indels) can be fatal for the HIV provirus, as they can interrupt genes and prevent proper transcription of the HIV genome. Some indels are in just the right place to increase viral resistance to the host and CRISPR-Cas. This leaves the virus immune to further CRISPR-Cas9 guide RNA complex attack (Wang, Z et al. 2016; Liang et al., 2016).

Gene editing with CRISPR-Cas was also used to revert a single-base mutation in the CFTR gene, which causes cystic fibrosis. CRISPR-Mediated gene targeting showed promise as there were few mutations other than those that were intended. The few mutations that were present were found at predicted off-target nucleotide sites (Firth et al., 2015).

After cells containing the mutation was injected with the CRISPR-Cas9-guide-RNA complex to repair the CFTR gene, it was injected with AcrIIA4, an anti-CRISPR protein isolated from *Pseudomonas* phages. The phage protein bound to the CRISPR-Cas9-guide-RNA complex, keeping it from binding to the target DNA and thus inhibiting the cause of the off-site damage (Shin et al., 2017).

Anti-CRISPR: A Necessary Addition to the Medical Applications of CRISPR-Cas9

CRISPR-Cas9 has dramatically increased the genome editing abilities of medical and research science, but it is not perfect. It has been used to edit proviral genomes out of host cells (Liang et al., 2016), edit mutated genes in mouse embryonic cells (Wang H et al., 2013; Wu Y et al., 2013) and to correct disease-causing mutations in genes that affect lung epithelial cell function (Firth et al., 2015). While these medical advances in the use of CRISPR-Cas9 genetic manipulation have been useful, they are not without risk. Off-site targeting of the target DNA by the CRISPR-Cas9 complex has been known to occur,

and can cause unintended deleterious effects (Liang et al., 2016; Firth et al., 2015; Wang, G et al., 2016; Wang, Z et al., 2016).

Anti-CRISPRs have been shown to alleviate the off-site deleterious side-effects of CRISPR-Cas9 in medical applications (Shin et a.l, 2017). Through competitive inhibition and forced redirection of host protein complexes, anti-CRISPR proteins limit off-site targeting of DNA by CRISPR-Cas9-guide-RNA complexes (Shin et al., 2017). Few anti-CRISPR proteins and their mechanisms are currently known. This begs for more anti-CRISPR research due to their useful nature in conjunction with CRISPR-Cas9 gene editing.

Methods

Extracting and Precipitating DNA from *Streptomyces griseus*

DNA extraction Method 1

Cultures of *S. griseus* were grown in 50 mL of sterile 1877 liquid media for 48 hours at 28°C and 220 rpm. Once grown, 30 mL of the culture was centrifuged in a 50 mL conical tube for 5 minutes at 4,000 x g in order to pellet the cells. The supernatant was discarded and the pelleted cells were washed with 10 ml of 10% sucrose solution. The sample was centrifuged again for 5 minutes at 4,000 x g, afterwards the supernatant was discarded.

After the supernatant was discarded, the pellet was suspended in 10 mL of lysis solution, which was made up of 2 mL of 0.3 M sucrose, 2 mL of 25 mM EDTA, 2 mL of 25 mM Tris base, and 2 U of RNAse. Ten mg of lysozyme and 7 mg of achromopeptidase were added to the suspension post resuspension. The conical tube was then incubated at 37°C for 20 minutes in a standing incubator.

The conical tube was removed from the incubator, and 1 mL of 10% SDS solution and 5 mg of proteinase K were added to the sample. The sample was then incubated in a water bath for 1.5 hours at 55°C.

Once the sample was removed from the water bath, 3.6 mL of 5M NaCl and 15 mL of chloroform were added. A white clump was observed in the sample post-addition. The sample was then gently rotated end-over-end for 20 minutes. After rotation, the sample was centrifuged for 20 minutes at 5,000 x g.

The resulting aqueous phase was transferred to a sterile 50 mL conical tube, where the DNA was precipitated by adding 0.5 mL of isopropanol. The sample was then centrifuged at 4,000 x g for 5 minutes to pellet the precipitated DNA. The isopropanol was removed from the sample by pipetting.

One mL of -20°C, fresh 70% ethanol was used to rinse the sides of the sample. The ethanol and DNA mix were removed and put into a new microcentrifuge tube, where the DNA was air dried. Once air dried, the DNA was then dissolved in 100 µL of prewarmed, 60°C 10 mM Tris and 10 mM EDTA buffer. The solution sat for four days in a 4°C refrigerator to allow for the DNA resuspend. After eluting the DNA, the quality of the DNA was assessed using a BioTek Synergy H1. The Take3 Nucleic Acid Quantification program from the BioTek Synergy H1 Gen5 microplate reader software, version 2.09, was utilized.

DNA extraction Method 2

S. griseus was cultured in 50 mL of sterile 1877 liquid media for 72 hours at 28°C in a shaking incubator at 220 rpm. 30 mL of the culture was transferred to a 50 mL sterile conical tube and centrifuged for 5 minutes at 4,000 x g. The supernatant was discarded, and the pellet was washed with 10 mL of 10% sucrose. The conical tube was centrifuged for 5 minutes at 4,000 x g, after which the supernatant was discarded.

The pellet was resuspended in 10 mL of lysis solution, which was made up of 2 mL of 0.3 M sucrose, 2 mL of 25 mM EDTA, 2 mL of 25 mM Tris base, and 2 U of RNAse. After the pellet was suspended, 0.0176 g of lysozyme and 0.0077 g of achromopeptidase were added. The conical tube was then incubated in a standing incubator at 37°C for 20 minutes.

Once the sample was removed from the standing incubator, 1 mL of 10% SDS and 0.0051 g of proteinase K were added. The sample was then incubated in a water bath at 55°C for 1.5 hours. After the sample was removed from the water bath, 3.6 mL of 5 M NaCl and 15 mL of chloroform were added. The conical tube was then rotated end-over-end for 20 minutes.

The sample was then centrifuged for 5,000 x g for 20 minutes. The aqueous phase was then transferred to a sterile 50 mL conical tube, and 0.5 mL of isopropanol was added to precipitate the DNA. The sample was then centrifuged at 4,000 x g for 5 minutes. The isopropanol was removed, and 1 mL of ethanol was added to the same conical tube.

The sample was centrifuged at 4,000 x g for 5 minutes, afterwards the ethanol was removed and the DNA was allowed to air dry. The dried DNA was suspended in 4 mL of 66°C sterile double distilled water (DDI water). The sample was stored in a 4°C refrigerator for 48 hours to suspend the DNA. After suspending the DNA, the quality of the DNA was assessed using a BioTek Synergy H1. The Take3 Nucleic Acid Quantification program from the BioTek Synergy H1 Gen5 microplate reader software version 2.09 was utilized.

Precipitating DNA from Method 2

8 mL of -80°C 70% ethanol was added to 4 mL of 4°C DNA from the second extraction method to a sterile 50 mL conical tube. The sample was centrifuged at 4,000 x g for 5 minutes, after which the ethanol supernatant was removed. The “waste” conical tube was re-centrifuged at the same parameters as the original tube. Once separated, the supernatant was removed, leaving behind the full DNA pellet. Both tubes were allowed

to air dry to allow excess ethanol to evaporate overnight in a sterilized Biological Safety Cabinet.

After air-drying, 1 mL of 60°C fresh MilliQ water was added to the primary 50 mL conical tube. The sides of the tube were rinsed with 1 mL of 60° C fresh MilliQ water, and the sample was stored at 4°C for five days. After air-drying in the BSC, the 15 mL conical tube still contained ethanol. The 15 mL conical tube was centrifuged at 4,000 x g for 5 minutes, after which the supernatant was removed. Once the supernatant was removed, the DNA was air-dried in the sterilized Biological Safety Cabinet (BSC). To elute the DNA from the 15 mL conical tube, 1 mL of 60°C MilliQ water was added. After the water was added, the sample was stored at 4°C for four days to suspend the DNA. The quality of the DNA was assessed using a BioTek Synergy H1. The Take3 Nucleic Acid Quantification program from the BioTek Synergy H1 Gen5 microplate reader software, version 2.09, was utilized.

DNA Extraction Method 3

Three cultures of *S. griseus* were grown in 50 mL of 1877 liquid media for 48 hours at 28°C in a shaking incubator at 220 rpm. In order to concentrate the cells, each culture was transferred into individual 50 mL conical tubes and centrifuged at 2,000 x g for 5 minutes. After centrifugation, the supernatant was removed, leaving 10 mL of both the pellet and liquid media in the tube. The pellets were resuspended in the 10 mL of broth in the conical tubes using a vortex. Once suspended, the three concentrated cultures were combined into a new, sterile 50 mL conical tube.

The 50 mL conical tube was centrifuged for 5 minutes at 4,000 x g. After centrifugation, the supernatant was discarded. The pellet was then washed with 10 mL of

10% sucrose and centrifuged for 5 minutes at 4,000 x g. After centrifugation, the supernatant was discarded and then washed with 10% sucrose and centrifuged for 5 minutes at 4,000 x g.

The supernatant was discarded, and the pellet was suspended in 10 mL of lysis solution (made with 4 mL of 0.3 M sucrose, 4 mL 25 mM EDTA, 4 mL 25 mM Tris Base, and 6.074g RnaseA). Lysozyme (0.0176 g) and achromopeptidase (0.0077 g) were added to the suspension.

The sample was incubated at 37°C for 20 minutes. Afterwards, 1 mL of 10% SDS and 0.0051 g of proteinase K were added to the sample, which was then incubated in a 55°C water bath for 1.5 hours.

Once the 1.5 hour incubation had been complete, 3.6 mL of 5M NaCl and 15 mL of phenol-chloroform were added to the sample. The sample was then rotated end-over-end for 20 minutes at 6 rpm, and centrifuged at 5,000 x g for 20 minutes.

The aqueous phase was transferred to a new conical tube. Two syringes with Promega DNA columns were used to collect the DNA from the aqueous phase. Once each DNA column was attached to a syringe, 1.5 mL of the aqueous phase was transferred to each syringe and pushed through. The flow-through was collected in a microcentrifuge tube.

Two mL of 80% isopropanol was added to each syringe, and pushed through. This step was repeated two more times, and each time the waste was discarded. The columns were removed from the syringe, placed in new microcentrifuge tubes, and centrifuged for 6 minutes at 10,000 x g. After centrifugation, the columns were transferred from the microcentrifuge tubes to a 90°C dry bath for 1.5 minutes.

Once the columns finished incubating, they were moved to a sterile DNA-LoBind microcentrifuge tube. To elute the DNA, 50 µL of 90°C MilliQ water was added directly to each column, which were then incubated at room temperature for 1.5 minutes. The columns and tubes were then centrifuged at 10,000 x g for 1 minute. The products of both tubes were combined into one tube. The process was repeated one more time, beginning with transferring the aqueous phase to a new set of syringes and columns, and ending with the two eluates being combined into one DNA-LoBind tube. The quality of both eluate tubes were assessed using a BioTek synergy H1.

BioTek synergy analysis of DNA samples 1, 2, and 3

Each *S. griseus* DNA extraction sample was tested for quality and quantity using the BioTek synergy H1 with Gen5 and Take3 software. Nuclease-free water was used as a blank, with 2 µL samples of water being added to the glass tray in lanes A1 and A2. After the blanks were added to the glass tray, 2 µL of sample from extraction 1 was added to lane B1 and B2. The program was then run using the dsDNA setting. This process was repeated for each sample.

Qubit Quantification of *S. griseus* DNA with the dsDNA BR kit

In order to use the Qubit for quantifying DNA prior to Illumina sequencing, a working solution was made. The Qubit dye was spun down in a Qubit supplied tube, and shaken to resuspend the dye. To maintain the 1 µL:199 µL dye:buffer ratio the working solution requires for each sample and an extra, 1,791 µL of buffer and 9 µL of dye was added to two 1.5 mL eppendorf tubes. The samples were then vortexed to resuspend the dye.

Working solution was added to each of the five sample tubes, but in different quantities. One hundred ninety µL of working solution was added to the two standard sample tubes, while 198 µL of working solution was added to the three experimental sample tubes. Two µL of DNA was added to each experimental sample, while the standard sample tubes, S1 and S2, had 10 µL of DNA/blank material added. Once each tube had all reagents added, they were incubated at room temperature for 2 minutes and spun to get rid of any air bubbles present in the sample.

The outside of each tube was cleaned with a KimWipe to remove any smudges on the outside of the sample that would prevent an accurate reading from the Qubit. After the sample was put into the Qubit, the "DNA" option and " dsDNA BR Kit" option was selected. The standards were read first, and then sample tube was inserted into the instrument. The Qubit was adjusted to the volume of the sample, and the units were changed to ng/uL.

A series of DNA sample dilutions was created by adding 0.05 µL of the sample to 99.95 µL of nuclease-free, lab grade water. New working solution was made in the same ratios that were made before. To further dilute the DNA, 20.5 µL of the diluted sample was combined with 79.5 µL of nuclease-free lab-grade water.

Sequencing *S. griseus* ATCC 10137 DNA

Illumina Sequencing with the single-end 76 cycle reagent cartridge and Nextera XT DNA Library Prep Kit

An index matrix was setup to show which indices were used for each sample.

Table 1: The index matrix of the *S. griseus* and *Staphylococcus* samples run on the Illumina. Samples 1-4 of *S. griseus* are labelled SGriseus#.

	N701	**N702**	**N703**
N517	SGriseus1	SGriseus2	SGriseus3
M502	SGriseus4		

The TD buffer, ATM, and gDNA were taken from the freezer and put on ice to thaw. A 96 well PCR plate had 10 µL of TD, 5 µL of gDNA, and 5 uL of ATM added to each well. Once the reagents had been added, the PCR plate was centrifuged at 20°C for 280 rcf for 1 minute. The plate was then put in the thermocycler at 55°C for five minutes, and 10°C holding temperature. Immediately after the hold temperature was reached, 5 µL of NT was added to the wells and was centrifuged at 280 x g, at 20°C for 1 minute. After the plate was incubated at room temperature for 5 minutes, 5 µL of each index was added to each individual well. The plate was centrifuged at 280 x g for 1 minute after 15 µL of NPM was added to each well. The plate was then put in the thermal cycler at 72°C for 3 minutes, 95°C for 30 seconds, for 12 cycles of the following: 95°C for 10 seconds, 55°C for 30 seconds, 72°C for 30 seconds. The plate was then subjected to 72°C for 5 minutes, and kept at a hold temperature of 10°C.

To begin the DNA library clean up process, the PCR plate was removed from the thermal cycler and centrifuged at 280 x g at 20°C for 1 minute. The contents of each well in the plate was transferred to its respective well on a new midi plate. Twenty seven µL of fully suspended room temperature Ampure beads were added to each well. The midi plate was shaken at 1,800 rpm for 2 minutes, and was then incubated at room temperature for 5 minutes. The midi plate was put on a magnetic stand to pellet the

Ampure beads for at least 2 minutes. Once separated, the supernatant was removed from the wells. To air dry the DNA, the plate was left for 5 minutes on the bench.

Fifty two and a half µL of resuspension buffer was added to each well once the plate was removed from the magnetic stand, and was shaken at 1,800 rpm for 2 minutes. The sample was then put on the magnetic stand for 2 more minutes. After separation, 50 µL of each sample was transferred into a new hard-shell PCR plate for storage. Once part of the sample was stored, the plate was put on the magnetic stand to ensure that there were no Ampure beads left in the solution. After 2 minutes, 20 µL of each sample was transferred to a new midi plate.

In a new DNA-LoBind tube, 308 µL of LNA1 and 56 µL of LNB1 were combined. Forty-five µL of the LNA1/LNB1 mixture were added to each well of the midi plate with 20 µL of each sample. The midi plate was then shaken at 1,800 rpm for 30 minutes, and then put on the magnetic stand for 2 minutes. The supernatant was removed from the wells. 45 µL of LNW1 was added to each well, and the midi plate was shaken at 1,800 rpm for 5 minutes, and placed on the magnetic stand for 2 minutes. After being left on the magnetic stand for 2 minutes, the supernatant was removed and 45 µL of LNW1 was added to each well. The plate was again shaken at 1,800 rpm for 5 minutes, and then set on the magnetic stand for 2 minutes. Thirty µL of 0.1 M NaOH was added to each well, afterwhich the plate was shaken at 1,800 rpm for 5 minutes. The plate was checked for homogenization, and then shaken at 1,800 rpm for 5 more minutes. After being shaken, the plate was put on the magnetic stand for 2 minutes, and the supernatant was transferred to a new PCR plate labelled SGP with 30 µL of LNS1 added in each well.

The plate was then centrifuged at 1,000 x g for 1 minute. Five µL of each sample was pooled into a new LoBind tube labelled “PAL”.

995µLof hybridization buffer and 5µLof the PAL was added to a new LoBind tube, which was then vortexed and centrifuged for 1 minute. Two hundred fifty µL of the diluted library was then transferred to a new LoBind tube, with 250 µL of hybridization buffer. The new DNA-LoBind tube with the diluted library and hybridization buffer was then vortexed and centrifuged at 280 x g for 1 minute. The sample was then incubated at 98°C for 2 minutes, and then immediately cooled on ice for 5 minutes.

To dilute and denature the PhiX control, 3 µL of EBT and 2 µL of 10 mM PhiX were added to a new DNA-LoBind tube. The sample was then vortexed and pulse centrifuged. After centrifugation, 5 µL of 0.1 M NaOH was added, and vortex and centrifuged again. The diluted PhiX control was incubated for 5 minutes at room temperature, afterwhich 5 µL of 200 mM Tris-HCl was added. The sample was then vortexed and centrifuged at 280 x g for 1 minute. 985 µL of hybridization buffer was added to the sample. To dilute the PhiX control to 1.8 pM, 45 µL of denatured PhiX and 455 µL of hybridization buffer were added to a new tube. In another new tube, 5 µL of denatured and diluted PhiX control and 500 µL of denatured and diluted DNA library were combined. This final tube was then set on ice.

The reagent cartridge was removed from the freezer and placed in a shallow warm-water bath for 90 minutes. Once the cartridge was thawed and at room temperature, well #16 was punctured with a 1,000 µL pipette tip. Five hundred µL of the DNA/PhiX library was inserted into the punctured well. After loading the cartridge, the flow cell was removed from the refrigerator to be brought up to room temperature. The

flow cell was cleaned on both sides with 70% ethanol, and was loaded into the MiniSeq. Once both the flow cell and the MiniSeq were loaded, the MiniSeq was started.

MinION Sequencing with a 1D Genomic DNA by Ligation Kit (SQK-LSK109)

The DNA concentration was determined with the Qubit. To prepare the MinION flow cell for sequencing, it was warmed to room temperature. The Nanopore was then connected to a Dell Optiplex 9020, Microsoft Windows version 7 Professional. MinKNOW was opened, and "connect" was selected. The flow cell and sample IDs were entered into the form on MinKNOW. The flow cell was then slid into the Nanopore. "Platform quality control" was clicked and the product code was entered. Once the flow cell was set up, "execute" was clicked to perform a quality control check. It was determined that the flow cell had 1,281 pores active. Once the quality of the flow cell was established, the priming and barcode kits were removed from the freezer to thaw.

Seven and a half µL of DNA were added to a PCR tube with 2.5 µL of fragmentation mix, ID #RB06. The sample was flicked and stun down to pull all of the liquid in the sample to the bottom. The sample was then put in the thermal cycler, with a cycle of 30°C for 1 minute, 80°C for 1 minute, and 4°C indefinitely. The samples were then removed from the Thermal cycler, and was spun down. All of the samples were pulled into one DNA-LoBind DNA tube, with 60 µL of Ampure beads added. The LoBind tube was then flicked and allowed to incubate at room temperature to allow the DNA to bind to the beads. The sample was then spun down a second time, and then placed on a magnet for 2 minutes.

Once the Ampure beads collected in the bottom of the sample, the supernatant was removed. Two hundred µL of 70% of EtOH was used to wash the beads. The ethanol was removed 30 seconds later, and the sample was spun down. The sample was placed on the magnet again to enable the removal of the remaining alcohol. After the ethanol was removed, 11 µL of Tris-NaCl was added. This was mixed by flicking, and put back on the magnet. Once the beads had fallen out of suspension, 10 µL of the mix was transferred to a new DNA-LoBind tube. In another tube, the RAP reagent was spun down, and 1µLof RAP was added to the 10 µL of DNA mix. This new mix was incubated at room temperature for 5 minutes. Four and a half µL of water was added to the sample after incubation. Thirty-four µL of Sequencing Buffer (SQB) and 25 µL of Loading Beads (LB) were added to complete the DNA library preparation.

To prime the flow cell, 30 µL of Flush Tether reagent was added to the sample of flush buffer and mixed by pipetting. The priming port cover was moved slightly, and a P1000 pipette was set to 200 µL. This pipette was then put into the port, and turned carefully to 230 µL to draw out the liquid in the flow cell. Once the liquid was removed, 30 µL of priming mix was loaded into the priming port. To allow for full saturation of the flow cell, the priming mix was allowed to sit for five minutes after addition to the cell.

The last 200 µL of primer mix was added to the cell, with the priming port cover left open afterwards. The DNA library was resuspended, and added to the spot-on port. After the library was loaded, both port covers were closed. Finally, “sequencing run” was selected on the MinKNOW software to begin sequencing.

BaseCalling *S. griseus* ATCC 10137 Sequencing Data

Illumina MiniSeq

As the Illumina MiniSeq sequenced the *S. griseus* genome, the instrument basecalled the sequence at the same time. Those files were downloaded from BaseSpace, the cloud computing software developed by Illumina, onto a personal computer and the Virtual Machine (VM) for analysis.

Albacore-MinION

Albacore was run on a VM using Docker instance to basecall the Nanopore data. The following command was used to run the program:

```
docker run -v $PWD:$PWD -w $PWD -u $UID vera/albacore \
--recursive -t 4 -i Salmonella_DG_1_18/ -c \
/opt/albacore/r94_450bps_linear.cfg -s output
```

Trimming the Illumina data

The Illumina data was trimmed to obtain optimal data for hybrid assembly. The tool Trimmomatic was used with the Illumina reads on the command line, using the following command on the terminal:

```
java -jar /home/steve/Trimmomatic-0.36/trimmomatic-0.36.jar SE \
-phred33 Sgriseus_Cat.fastq.gz \
Sgriseus_trimmedSW420MinLen36.fastq \
ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3\
SLIDINGWINDOW:4:20 MINLEN:36
```

FastQC was used to determine the quality of the Trimmomatic output.

Filtering Nanopore data with Filtlong

The Nanopore data was filtered for 1000 bp long reads using the following Filtlong command on the server maintained by the lab:

```
./filtlong --min_length 1000 --keep_percent 90 --target_bases \
500000000 /home/nanopore/porechoppedreads/BC06.fastq | gzip > \
BC06filtered.fastq.gz
```

The output was stored on the VM.

Hybrid *S. griseus* ATCC 10137 Genome Assemblies

The bioinformatic tool Unicycler was used to produce a hybrid genome assembly of *S. griseus* reads from the Illumina sequencing and MinION sequencing data. The following command was run on the server maintained by the lab:

```
screen unicycler -s \
/home/rachael/Sgriseus_trimmedSW420MinLen36.fastq -l \
/home/rachael/BC06filtered.fastq.gz --mode bold --linear_seqs 1 \
-o \
Sgriseus_trimmedSW420MinLen36FiltlongNanoporeDataUnicyclerAssembl
y
```

The output was stored on the VM, Open Science Framework (OSF), and the laptop used.

PHASTER analysis of the *S. griseus* ATCC 10137 genome

The assembled *S. griseus* genome was uploaded to the PHASTER web application to identify potential prophages.

CRISPRfinder analysis of the *S. griseus* ATCC 10137 genome

The assembled *S. griseus* genome was uploaded to the CRISPRfinder web application to identify CRISPR-Cas arrays.

Automated *S. griseus* ATCC 10137 Genome Annotations

Prokka

To use Prokka to auto annotate the *S. griseus* genome, the following command was used:

```
prokka --outdir SgriseusUniAssT3LinAssAnnotated.fna --prefix \
Sgriseus SgriseusUniAssT3LinAss.fasta
```

The name of the file was then changed SgriseusUniAssemblyT3.

RAST

The *S. griseus* assembled genome was run through the auto annotator RAST through the WebApp. The options Genetic code 11, RASTtk, automatically fix errors, and backfill gaps were selected after the genome was uploaded to the site.

Development of Prokkrastinator, an Annotation Method combination tool

Prokkrastinator was built using JavaScript, python, HTML and CSS. JavaScript libraries, such as D3.js and Meteor.js, were implemented. Git was used as a version control software, and GitLab is the current repository for the code.

Genome Annotation of Bacteriophage Wipeout

Wipeout's genome was run through the auto annotation tool GeneMark Heuristic Approach for Gene Prediction on the WebApp. The fasta sequence for Wipeout was uploaded onto the site, and the options LST, PDF, genetic code 11, both, and heuristic 2010 were selected.

The GeneMark output file was uploaded to PECAAN, a phage genome annotation tool. PECAAN automatically generates a Glimmer annotation file and Starterator file, as well as a BLASTN, BLASTP, HHPRED, RBS site analysis, and conserved domain search.

Each auto annotated gene was analyzed using these data gathered from these programs to determine if the gene was actually present, if the proper start codon had been selected, and what the function of the gene may be. Those genes that were ~300 bp long and under were selected for protein modeling.

Bacteriophage protein modeling (potential anti-CRISPR proteins)

To convert the amino acid fastas of the genes from Wipeout, R code was used. The script was titled file_list_maker.R. These converted files were then run through YASARA to compare the gene products with homologues, and thus queried against the PDB database. The resulting model data table was converted from an HTML format into a csv file with R script titled model_table.R. The rest of the homology results were written to a csv file using R script titled prediction.R.

S. griseus ATCC 10137 Cas3 modeling

The Cas3 gene fasta was translated into amino acids using ExPASy Translate. The output option “Compact” and the Genetic Code option “ Standard” was selected. The longest open reading frame was the first reading frame in the 5’-3’ direction. Once translated to amino acids, the file was uploaded to the Phyre2 protein fold recognition server. The modeling option “normal” and the “not for profit” option were selected.

S. griseus ATCC 10137 Cas3 ligand search

The Cas3 pdb file generated by Phyre2 was uploaded to the COACH server to identify ligand interaction sites.

Results

DNA was extracted from *S. griseus* cells using a phenol-chloroform based method. As with other gram positive bacteria, DNA extraction from *Streptomyces* species is notoriously difficult (Blakely et al., 2003). Three attempts were made, each with a slightly different method. The first DNA extraction was done according to the method developed by Blakely et al. While the yield of the extraction was high, the quality of the extraction was poor (Table 2). The first DNA extraction sample yielded on average 181ng/ µL of DNA. As can be seen in table 2, The average 260/280 ratio was 1.401 and the 260/230 ratio was 2.155 on average. The second DNA extraction attempt varied slightly from the first attempt. To decrease the chances of contamination, the DNA was kept in the same conical tube that it was precipitated in. This decreased the overall yield, and had little effect on the purity of the sample. The second DNA extraction yielded an average DNA concentration of 18.35 ng/ µL. As can be seen in table 3, The 260/280 ratio was 1.433 on average and the 260/230 ratio was 2.311 on average. The third and final DNA extraction method was similar to the first two extraction methods, with the major alteration being the addition of filtering DNA through a column. The third DNA extraction yielded on average 594.5 ng/ µL of DNA. As can be seen in table 3, The 260/280 ratio was 1.8125 on average and the 260/230 ratio was 2.09 on average.

A BioTek Synergy H1 was used to collect quality and quantity data from the sample. The Take3 Nucleic Acid Quantification program from the BioTek Synergy H1 Gen5 microplate reader software, version 2.09, was used to analyze the DNA. Two µL of each DNA extraction sample was put onto each well of the microplate. Once the

microplate was inserted into the BioTek Synergy H1, the dsDNA option was selected and run.

To sequence DNA, the 260/280 ratio of the sample must be approximately 1.8, and the 260/230 must be approximately 2 (Illumina, 2018). The sample that resulted from the third DNA extraction had these parameters, so was selected for sequencing.

BioTek Synergy H1 analysis of *S. griseus* ATCC 10137 DNA extraction data

Table 2 : DNA extraction 1 analysis results. A BioTek Synergy H1 was used to collect quality and quantity data from the first DNA extraction method. The Take3 Nucleic Acid Quantification program from the BioTek Synergy H1 Gen5 microplate reader software, version 2.09, was utilized. The sample yielded an average of 181 ng/µL of DNA with a 260/280 ratio of 1.401 and a 260/230 ratio of 2.155.

	Row B2	**Row B3**
260 Raw	0.237	0.233
280 Raw	0.181	0.177
230 Raw	0.274	0.268
320 Raw	0.049	0.045
260	0.181	0.181
280	0.129	0.13
230	0.085	0.083
260/280	1.402	1.4
260/230	2.121	2.188
ng/uL	180.85	181.225

Table 3: DNA extraction 2 analysis results. A BioTek Synergy H1 was used to collect quality and quantity data from the first DNA extraction method. The Take3 Nucleic Acid Quantification program from the BioTek Synergy H1 Gen5 microplate reader software, version 2.09, was utilized. The sample yielded an average of 18.4 ng/µL of DNA with a 260/280 ratio of 1.433 and a 260/230 ratio of 2.311.

	Row B2	**Row B3**
260 Raw	0.095	0.07
280 Raw	0.089	0.062
230 Raw	0.097	0.069
320 Raw	0.073	0.046
260	0.017	0.019
280	0.013	0.013
230	0.009	0.007
260/280	1.362	1.504
260/230	2.035	2.587
ng/uL	17.3	19.4

Table 4: DNA extraction 3 analysis results. A BioTek Synergy H1 was used to collect quality and quantity data from the first DNA extraction method. The Take3 Nucleic Acid Quantification program from the BioTek Synergy H1 Gen5 microplate reader software, version 2.09, was utilized. The sample yielded an average of 594.5 ng/ µL of DNA with a 260/280 ratio of 1.8125 and a 260/230 ratio of 2.09.

	Row B2	**Row B3**
260 Raw	0.512	0.797
280 Raw	0.311	0.459
230 Raw	0.29	0.419
320 Raw	0.057	0.054
260	0.451	0.738
280	0.252	0.402
230	0.218	0.349
260/280	1.789	1.836
260/230	2.066	2.114
ng/uL	450.8	738.2

Qubit quantification of *S. griseus* ATCC 10137 DNA

The Qubit was used for quantifying DNA prior to Illumina sequencing, as required by Illumina. To test the DNA concentration of the samples, a working solution was needed. The working solution was made by spinning down the Qubit dye in a Qubit supplied tube, and shaken to resuspend the dye. A 1 µL:199 µL dye:buffer ratio was used to make the working solution, and was split between two 1.5 mL eppendorf tubes. The samples were then vortexed to resuspend the dye.

One hundred ninety µL of working solution was added to the two standard sample tubes, while 198 µL of working solution was added to the three experimental sample tubes. Two µL of DNA was added to each experimental sample, while the standard

sample tubes, S1 and S2, had 10 µL of DNA/blank material added. Once each tube had all reagents added, they were incubated at room temperature and spun to get rid of any air bubbles present in the sample.

The outside of each tube was cleaned with a KimWipe to remove any smudges on the outside of the sample that would prevent an accurate reading from the Qubit. The Qubit was adjusted to the volume of the sample, and the units were changed to ng/uL.

The original concentration of the third DNA extraction sample as determined by Qubit was 379 ng/µL. The Illumina sequencer only needed 2 ng/µL of DNA, so the sample had to be dilutes. A series of DNA sample dilutions was created by adding 0.05 µL of the sample to 99.95 µL of nuclease-free, lab grade water. New working solution was made using the same ratios as before. To further dilute the DNA, 20.5 µL of the diluted sample was combined with 79.5 µL of nuclease-free lab-grade water.

The first attempt at diluting the DNA resulted in a concentration of 0.976 ng/µL.

Table 5: Qubit quantification of *S. griseus* sample for Illumina sequencing. The DNA concentration of the *S. griseus* sample was determined to be 379 ng/µL. After dilution, the DNA concentration of the sample was determined to be 0.976 ng/µL.

Species Sample	**DNA concentration (ng/µL)**
S. griseus	379
S.griseus 1st dilution	0.976

S. griseus ATCC 10137 Sequencing Data

The Nextera XT DNA Library prep kit was used to prepare DNA libraries from the third DNA extraction sample. Once the libraries were prepared, the individual libraries were pooled to create one sample. This pooled sample was loaded into a 75-cycle reagent kit and then run on the Illumina MiniSeq. As can be seen in table 6, the Illumina sequencing run resulted in 19,773,962 total reads, and 18,982,977 identifiable reads (Table 6). *S. griseus* sample 1 had an index 1 of TAAGGCGA and index 2 of TCTTACGC, and made up 12.4625% of the total identifiable reads. *S. griseus* sample 2 had an index 1 of CGTACTAG and index 2 of TCTTACGC, and made up 10.1038% of the total identifiable reads. *S. griseus* sample 3 had an index 1 of AGGCAGAA and an index 2 of TCTTACGC, and made up 9.8552% of the total identifiable reads. *S.griseus* sample 4 had an index of TAAGGCGA and ATAGAGAG, and made up 19.0788% of the total identifiable reads.

The Illumina sequencing run resulted in 97.5% of the reads with a Q Score equal to or above 30 (Figure 1). The beginning of the sequencing run produced reads with a Q Score between 25-32, while the majority of the run produced reads with a Q Score equal to or above 30 (Figure 2).

Illumina data

Table 6: Results from the Illumina sequencing run with *S. griseus* and *Staph* samples. The sequencing resulted in 19,773,962 total reads, and 18,982,977 identifiable reads. Of the reads generated, 92.9043% were able to be identified. Each *S. griseus* sample was given a unique name and index. *S. griseus* sample 1 had an index 1 of TAAGGCGA and index 2 of TCTTACGC, and made up 12.4625% of the total identifiable reads. *S. griseus* sample 2 had an index 1 of CGTACTAG and index 2 of TCTTACGC, and made up 10.1038% of the total identifiable reads. *S. griseus* sample 3 had an index 1 of AGGCAGAA and an index 2 of TCTTACGC, and made up 9.8552% of the total identifiable reads. *S.griseus* sample 4 had an index of TAAGGCGA and ATAGAGAG, and made up 19.0788% of the total identifiable reads.

TOTAL READS	PF READS	% READS IDENTIFIED (PF)	CV	MIN	MAX
19,773,962	18,982,977	92.9043	0.3410	9.8552	21.7553

INDEX	SAMPLE ID	PROJECT	INDEX 1 (I7)	INDEX 2 (I5)	% READS IDENTIFIED (PF)
1	Sgriseus1-2002	NA	TAAGGCGA	TCTTACGC	12.4625
2	Sgriseus2-2002	NA	CGTACTAG	TCTTACGC	10.1038
3	Sgriseus3-2002	NA	AGGCAGAA	TCTTACGC	9.8552
4	Sgriseus4-2002	NA	TAAGGCGA	ATAGAGAG	19.0788
5	StaphSciuri1-2002	NA	CGTACTAG	ATAGAGAG	21.7553
6	StaphSciuri2-2002	NA	AGGCAGAA	ATAGAGAG	19.6486

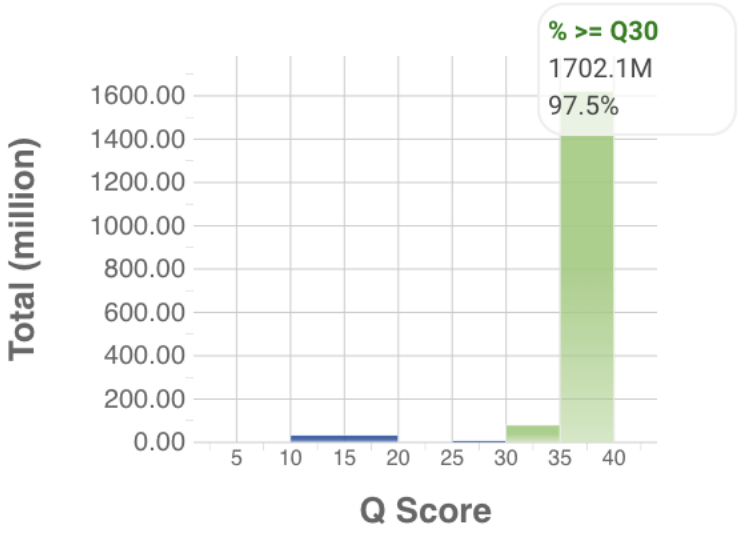


Figure 1: Q Score distribution of the reads from Illumina sequencing. A Q Score of 30 means that 1 base in every 1,000 bases called was a mistake. The Illumina sequencing run resulted in 97.5% of the 19,773,962 total reads with a QScore equal to or above 30.

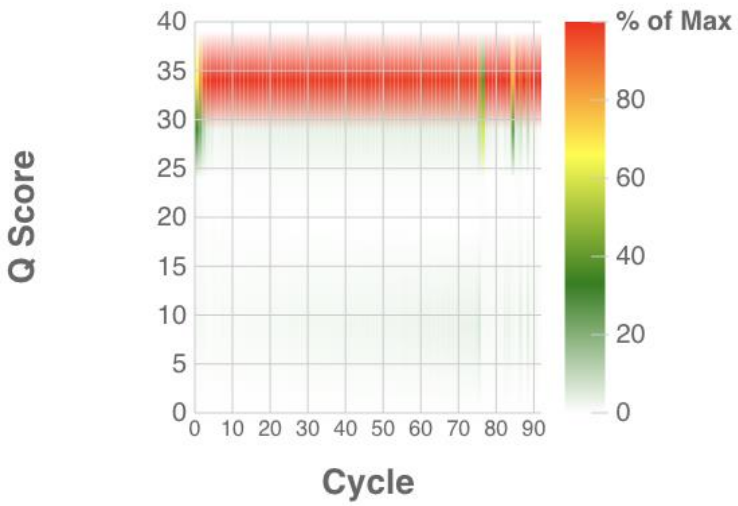


Figure 2: Q Score heat map of the reads and their quality during the sequencing cycles from the Illumina sequencing run. During the start of the Illumina sequencing run, the Q Score of the reads was between 25 and 35. A majority of the reads generated had a Q Score above 30, while the quality of the reads began to drop near the end of the sequencing run.

Nanopore data

The 1D Genomic DNA Ligation Kit (SQK-LSK109) was used to create DNA libraries from the third DNA extraction samples. Once the libraries were prepared and pooled, a P1000 pipette was used to load the sample onto the flow cell. The MinION was then plugged into a Dell Optiplex 9020, Microsoft Windows version 7 Professional. After the MinION was connected, the sequencing run was started.

The MinION Nanopore sequencing run generated 1,688,493 reads and 7,601,206,820 reads with a read length N50 of 7,576.0 bases and 140,000 reads with the mean read length of 4,501.8 bases (Figure 5). The run lasted for approximately 50 hours, and generated a cumulative yield of over 7 gigabases of data (Figure 3) and approximately 1,750,000 bases (Figure 4).

The read lengths remained similar over the course of the run (Figure 6), however the quality of the reads decreased over the 50 hours that the sequencer was collecting data (Figure 7).

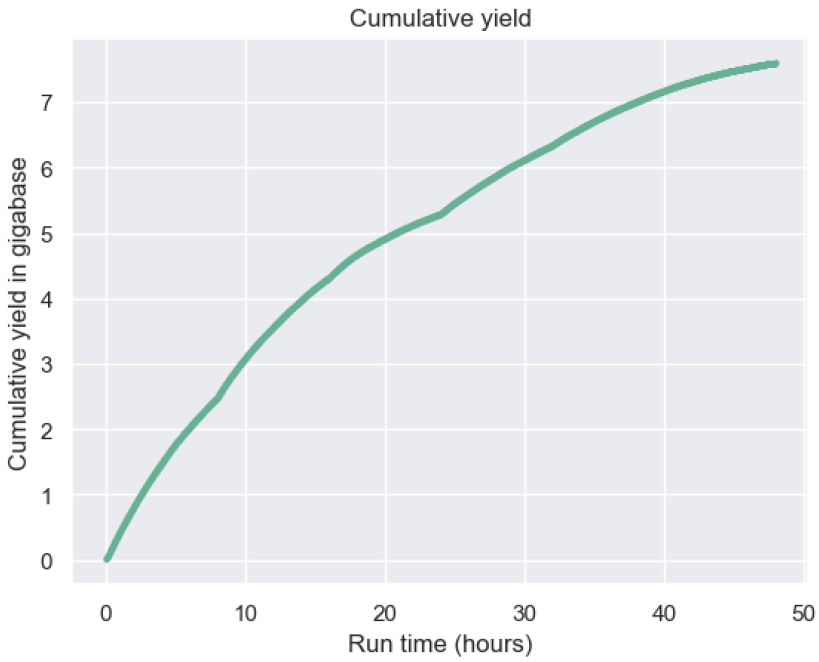


Figure 3: The cumulative yield (gigabases) during the total run time (hours) of the Nanopore. Over seven gigabases of data were generated during the Nanopore Min ION sequencing run. The bioinformatic tool Nanoplot was used to generate the graphic.

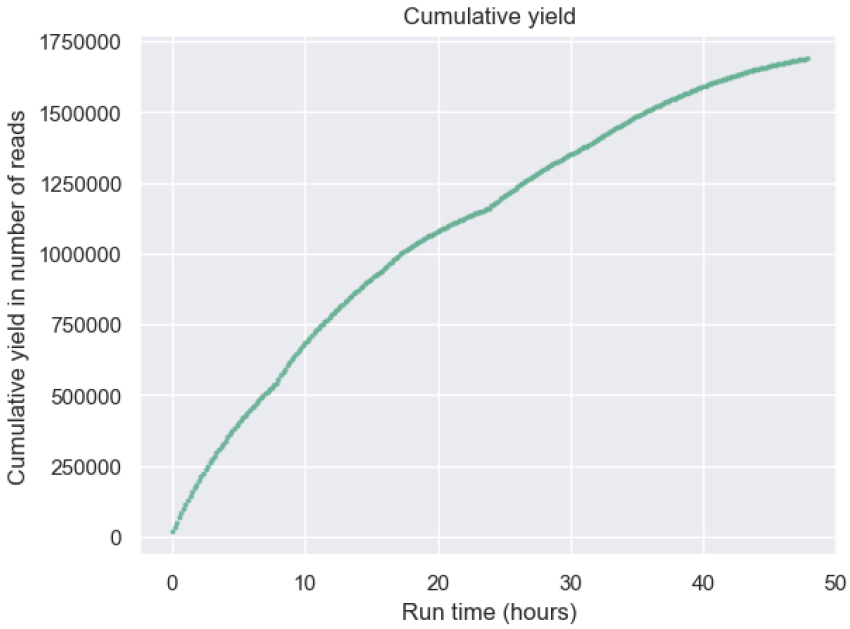


Figure 4: The cumulative yield (number of bases) during the total run time (hours) of the Nanopore. Over 1.5 million bases were called during the entire 48 hour Nanopore MinION sequencing run. The bioinformatic tool Nanoplot was used to generate the graphic.

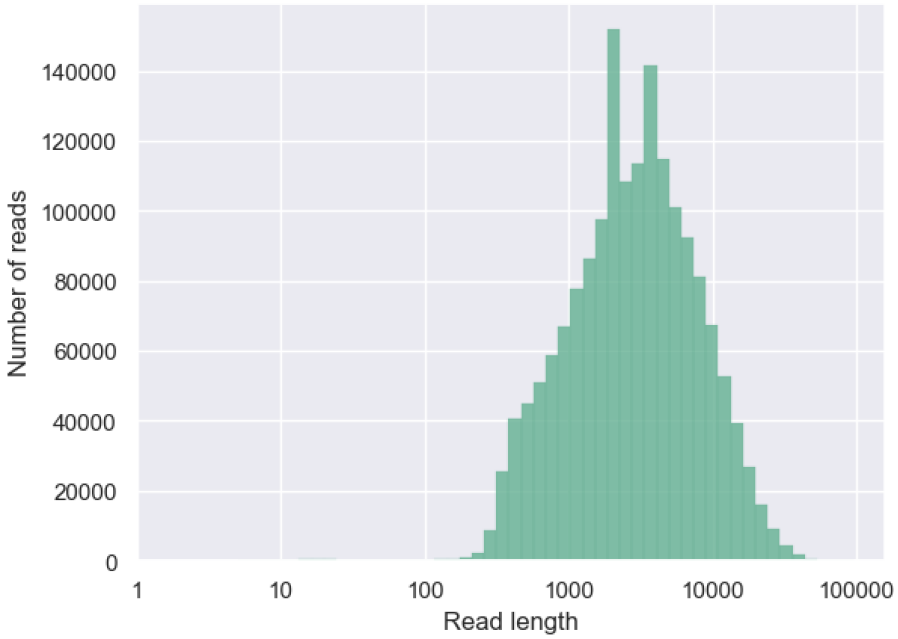


Figure 5: The read length of the reads from the Nanopore run compared to the number or reads generated from the run after log transformation. The majority of the reads generated from the Nanopore MinION sequencing run were between 1,000 and 10,000 base pairs in length. The bioinformatic tool Nanoplot was used to generate the graphic.

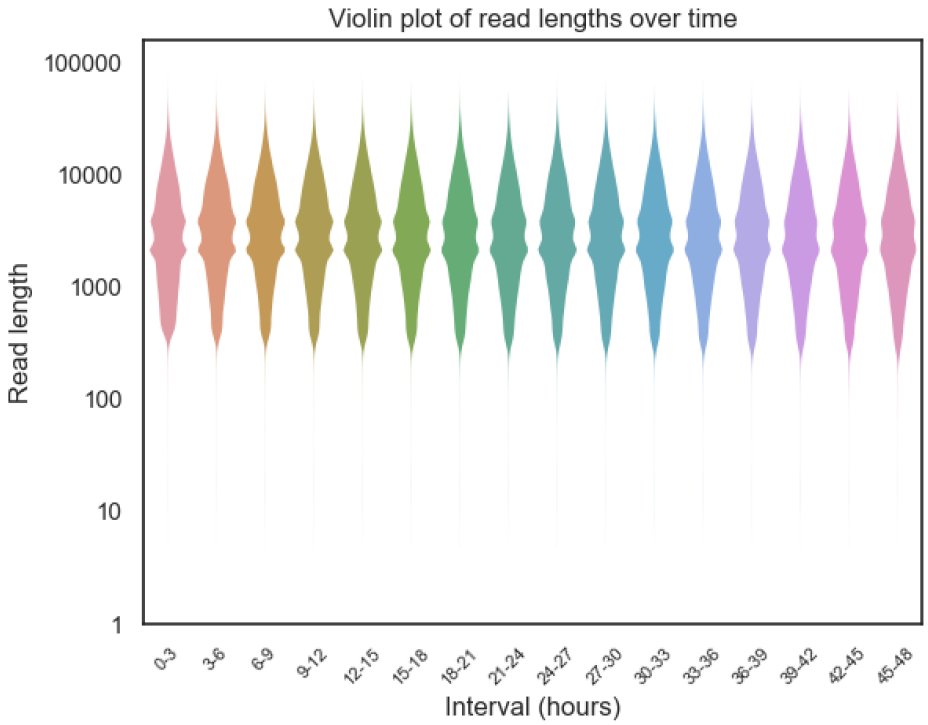


Figure 6: The distribution of Nanopore read lengths over time. The read lengths over each 3 hour period of the Nanopore MinION sequencing run were similar to each other for the entire 48 hour sequencing run. The bioinformatic tool Nanoplot was used to generate the graphic.

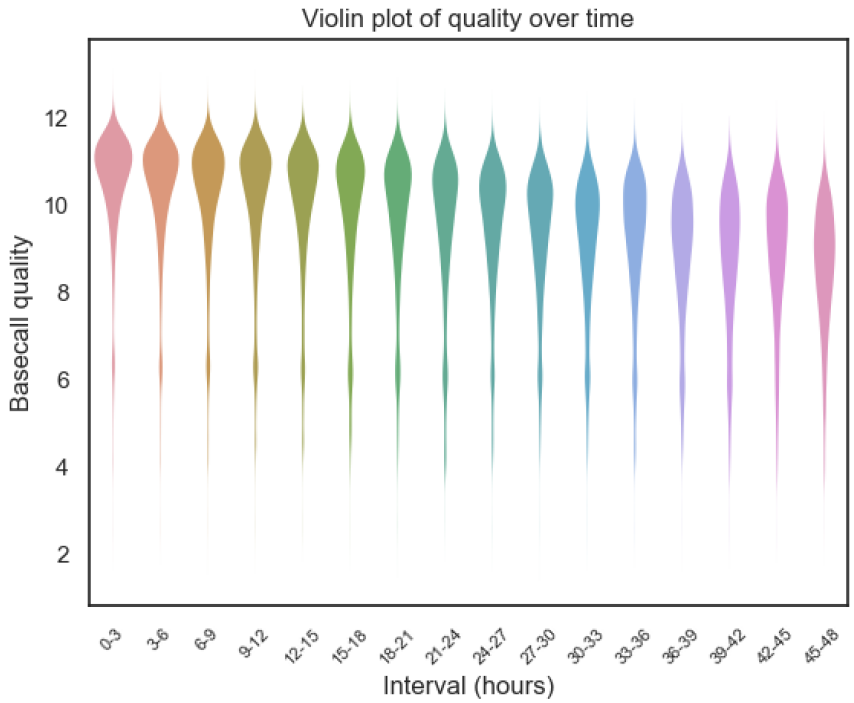


Figure 7: The quality of the basecalls over time. The quality of the Nanopore MinION bases fell between basecall quality 10-12 in the first three hours. As the sequencing run continued, the basecall quality of the reads steadily decreased. The quality of Nanopore MinION sequencing reads tends to decrease over time due to pore degradation in the flow cell. The bioinformatic tool Nanoplot was used to generate the graphic.

Hybrid *S. griseus* ATCC 10137 genome assembly data

The bioinformatic tool Unicycler was used to generate *S. griseus* ATCC 10137 genome assemblies. The first genome assembly was generated using just the Nanopore data that was over 1,000 bp long. With the filtered Nanopore data, Unicycler generated an assembly of four contigs; 6,410,491 bp long, 1,358,778 bp, 780,416 bp, and 29,804 bp (figure 8). The 1,000 bp filtered Nanopore data genome assembly results were visualized using the bioinformatic tool Bandage.

The 1,000 bp filtered Nanopore data and the 4:20 sliding window trimming approach, 36 minimum length trimmed illumina data generated a hybrid genome assembly with 9 contigs and 1 unitig, which is 8,576,363 bp long (figure 9). The 1,000 bp filtered Nanopore data genome assembly results were visualized using Bandage.

Nanopore only assembly

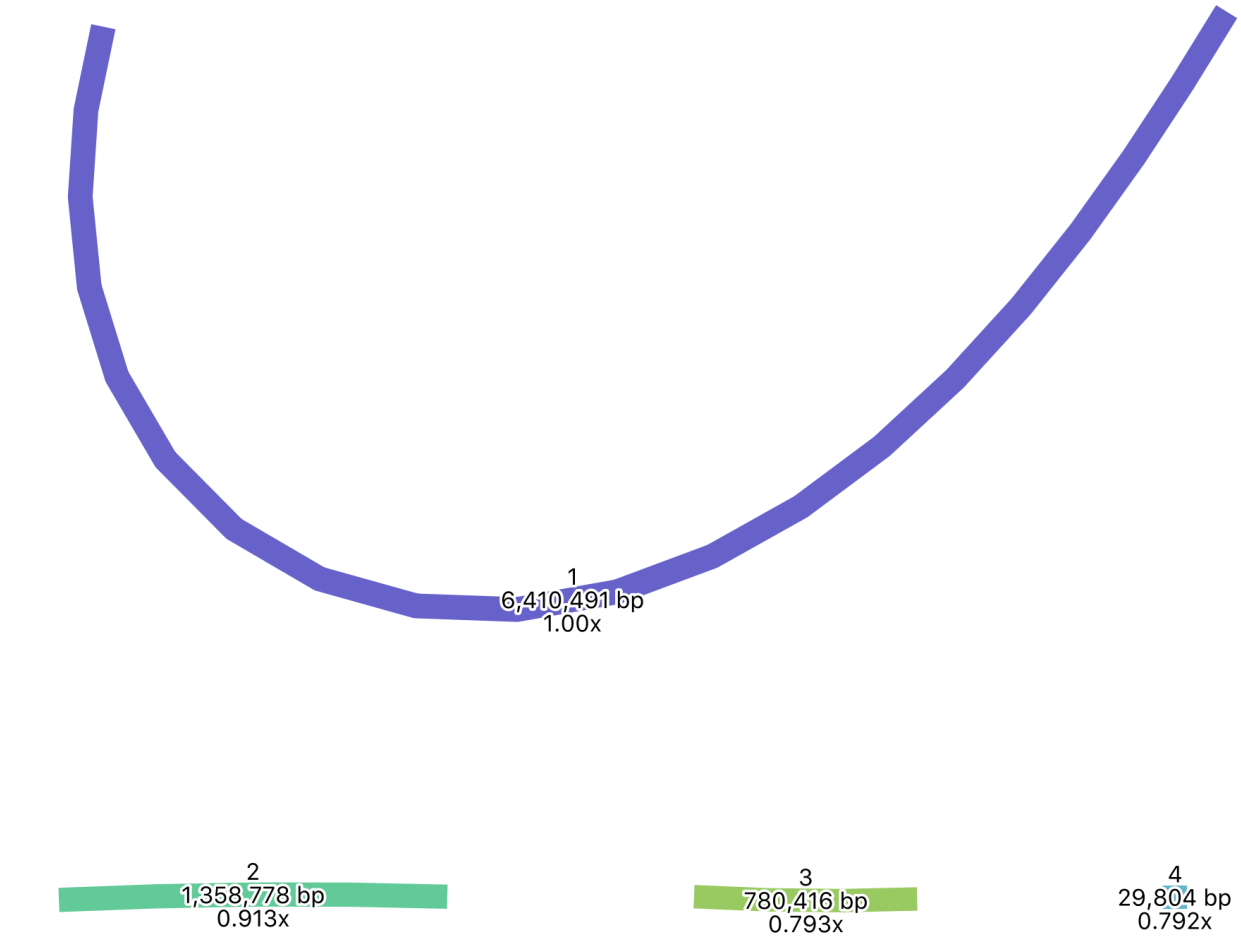


Figure 8: Unicycler assembly with the 1,000 bp filtered Nanopore MinION sequencing run data. Unicycler produced four contigs from the 1,000 bp filtered Nanopore MinION sequencing data. The first contig was 6,410,491 bp, the second contig was 1,358,778 bp, the third contig was 780,416 bp, and the fourth contig was 29,804 bp long. The assembly was visualized using Bandage, with contigs being drawn in proportion to their length.

Nanopore and Illumina Unicycler Assembly

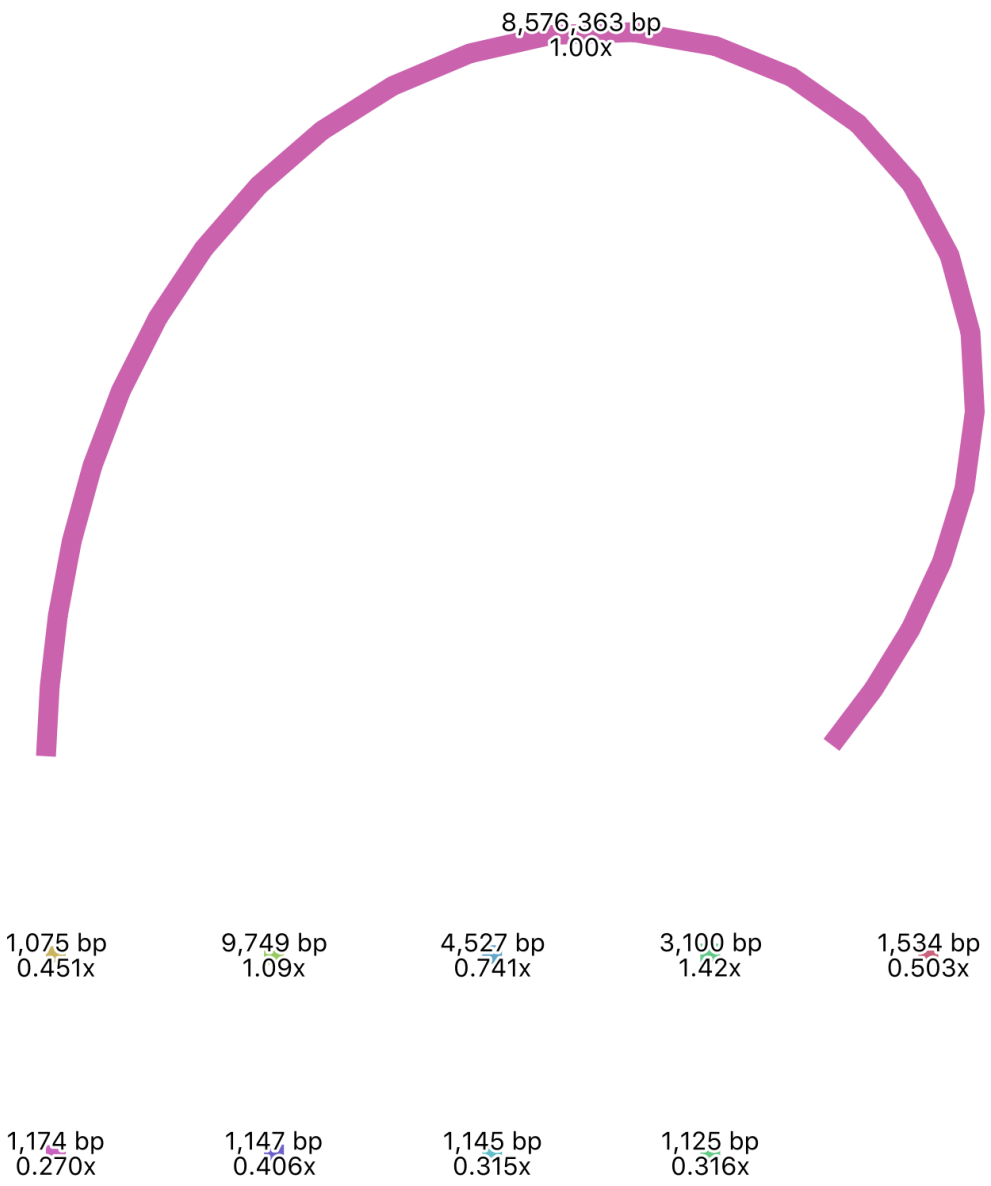


Figure 9: Unicycler assembly with the filtered Nanopore MinION sequencing run data and trimmed Illumina MiniSeq sequencing run data. The 1,000 bp filtered Nanopore MinION data and 4:20 sliding window and 36 minimum length trimmed Illumina MiniSeq data were assembled using Unicycler, settings linear and bold. Unicycler generated an assembly that produced 10 contigs. The first contig is 8,576,363 bp long, and is thought to be the *S. griseus* ATCC 10137 genome. The other contigs are all under 10,000 bp long, and did not return any plasmid matches when analyzed using BLAST. The assembly was visualized using Bandage, with contigs being drawn in proportion to their length.

Analysis of the *S. griseus* ATCC 10137 genome

Prophages identified in the *S. griseus* ATCC 10137 genome using PHASTER

The *S. griseus* genome assembly generated by Unicycler was uploaded to PHASTER to scan the *S. griseus* genome to identify potential prophages. One incomplete prophage was identified in the *S. griseus* genome. The region length was 22.7Kb, coordinates 3,217,534-3,240,311 bp. The incomplete prophage had a PHASTER score of 60, had a GC content of 70.95 %, and contained 22 potential proteins. Three tail proteins were found on the 5'-3' strand, coordinates 3,223,067-3,224,266 bp, 3,224,263-3,225,546 bp, and 3,227,818-3,228,924 bp. A fiber protein was identified at position 3,229,705-3,231,171 bp.

CRISPR-Cas arrays in the *S. griseus* ATCC 10137 genome using CRISPRfinder

The hybrid *S. griseus* ATCC 10137 genome assembly produced by Unicycler was uploaded to the web server www.crispr.i2bc.paris-saclay.fr, or CRISPRfinder, to identify potential CRISPR-Cas arrays in the genome (Grissa et al., 2007).

CRISPRfinder identified 3 confirmed CRISPR arrays, and 10 questionable regions in the *S. griseus* ATCC 10137 genome. The first confirmed CRISPR array begins at genome position 4,000,169 and ends at 4,001,234, for a total length of 1,065 base pairs. In the first array, there are 17 spacers. The second confirmed CRISPR array begins at genome position 4,004,378, ends at 4,005,445, and contains 17 spacers. The final confirmed CRISPR array begins at position 4,015,427, ends at 4,016,611, and has 19

spacers (Table 7). The number of questionable CRISPR arrays and the number of spacers in each of those arrays can be found in Table 8.

The NCBI BLASTn with default parameteres was used to identify Confirmed CRISPR-Cas spacer regions that are also found in phage genomes (Altschul et al, 1990). The first confirmed CRISPR-Cas region of *S. griseus* ATCC 10137 had 17 spacers, but BLAST only identified spacer 11 in CRISPR-Cas region 1 as matching other bacteriophage DNA. Bacteriophages Biscum, Ididsumtinwrong, Austintatious, PapayaSalad, and Darolandstone were found to have similar sequence as spacer 11. CRISPR-Cas region 2, spacer 2 encoded sequence that was similar to bacteriophage Rando14. CRISPR-Cas region 2, spacer 9 encoded sequence that was similar to bacteriophage Menlow. CRISPR-Cas region 3, spacer 6 encoded sequence that was similar to bacteriophage ToastyFinz. CRISPR-Cas region 3, spacer 7 encoded sequences that were similar to bacteriophages SV1, Mojorita, Picard, and ToastyFinz (Table 9).

Table 7: Data from the confirmed CRISPR-Cas arrays in *S. griseus*. CRISPRfinder was used to identify the 3 confirmed CRISPR-Cas arrays in the *S. griseus* genome. The first two CRISPR regions contained 17 spacers each, while the third CRISPR region contained 19 spacers.

CRISPR Region	Start	End	Number of Spacers	DR Consensus
1	4000169	4001234	17	GTGGTCCCCGCGCGTGCGGGGGTGTTCCC
2	4004378	4005445	17	GTGGTCCCCGCGCGTGCGGGGGTGTTCCC
3	4015427	4016611	19	GTGGTCCCCGCGCGTGCGGGGTTGTTCC

Table 8: Data from the questionable CRISPR-Cas arrays in *S. griseus*.CRISPRfinder was used to identify the 10 questionable CRISPR-Cas arrays in the *S. griseus* genome. The first 4 CRISPR regions each encoded 1 spacer, the fifth region encoded 3 spacers, the sixth region encoded 1 spacer, the seventh region encoded 5 spacers, and regions eight through ten each encoded 1 spacer.

CRISPR Region	Start	End	Number of Spacers	DR Consensus
1	9090	9248	1	GGGCGGGGGCGCCGGTCAGGTCGGCGGCTTCG CGCAGGAGGGCGGCGGCCTGGGC
2	1580588	1580696	1	CCCGGAGCCGAACCCGGTGTCGCTGGA
3	2854405	2854508	1	AGCGGTTACGCGCATACATCTGCGGTGG
4	3550255	3550340	1	GGTTCCGTCCTCAAACGCCGGACGGGCTG
5	4049500	4049704	3	CAACCCCGCACGCGCGGGGACCAC
6	4154117	4154184	1	CCGCCGCCCTGCTGGCCTCCGCC
7	4346586	4346950	5	GGAGTCGGTGTACCGGTTGTCGT
8	4578862	4578927	1	CCGTCCGTCGGTCCCGCCGTGCC
9	5522346	5522450	1	CTACAACGGCCCCGTCTTCAACCAG
10	6045450	6045534	1	CTCGCCCGCCCCACGGCGGCCGGCC

Table 9: BLAST results from comparing *S. griseus* ATCC 10137 CRISPR spacers in the genome to other genomes in the NCBI BLAST database. Confirmed CRISPR region 1, spacer 11 encoded sequence that was found in bacteriophages Biscum, Ididsumtinwrong, Austintatious, PapayaSalad, and Darolandstone. Confirmed CRISPR region 2, spacer 2 encoded sequence that was found in bacteriophage Rando14. Confirmed CRISPR region 2, spacer 9 encoded sequence that was found in bacteriophage Menlow. Confirmed CRISPR region 3, spacer 6 encoded sequence that was found in bacteriophage ToastyFinz. Confirmed CRISPR region 3, spacer 7 encoded sequence that was found in bacteriophages SV1, Mojorita, Picard, and ToastyFinz.

CRISPR region	**Spacer**	**Phage**	**Query coverage**	**E-value**	**% identity**
1	11	Biscum	75 %	0.023	100.00
1	11	Ididsumtinwrong	75%	0.023	100.00
1	11	Austintatious	75%	0.99	95.83
1	11	PapayaSalad	75	0.99	95.83
1	11	Darolandstone	93	3.5	86.67
2	2	Rando14	78	0.28	96.00
2	9	Menlow	78	3.5	96.00
3	6	ToastyFinz	100	0.007	90.91
3	7	SV1	100	1e-05	96.97
3	7	Mojorita	100	2e-04	93.94
3	7	Picard	100	2e-04	93.94
3	7	ToastyFinz	90	0.007	93.33

Automated genome annotation of *S. griseus* ATCC 10137

The auto annotation methods Prokka and RAST were used to predict genes in the *S. griseus* ATCC 10137 genome. Prokka was run with a terminal command using the hybrid *S. griseus* ATCC 10137 genome assembly, while the the hybrid *S. griseus* ATCC 10137 genome assembly was uploaded onto the web application-based RAST. Analysis of the *S. griseus* genome with auto annotation methods Prokka and RAST resulted in different numbers of potential genes. Prokka predicted 7,438 genes and RAST predicted 7,775 genes (Figure 10).

There are CRISPR arrays present in the genome, as well as CRISPR-associated protein coding genes such as CRISPR-associated endoribonuclease *Cas1*, CRISPR-associated endoribonuclease *Cas2*, CRISPR-associated endoribonuclease *Cse3*, CRISPR system Cascade unit *casD*, CRISPR system Cascade unit *casC*, and CRISPR-associated helicase/nuclease *Cas3*.

The genome also encodes the antiseptic resistance genes *qacA_2*, *qacA_3*, *qacA_5*, and *qacA_6*. The multidrug resistance gene *MdtH_1* and the Vancomycin B-type resistance coding gene *VanW* were found in the genome. Also present are genes that belong to the SMR, or small multidrug resistance protein family, such as the quaternary ammonium compound-resistance coding genes *sugE_1* and *sugE_2*. Tetracycline resistance is conferred to the bacteria by the presence of the gene *TetR_2*. Soluble epoxide hydrolase and multidrug ATP/binding permease protein are also encoded by genes found in the genome.

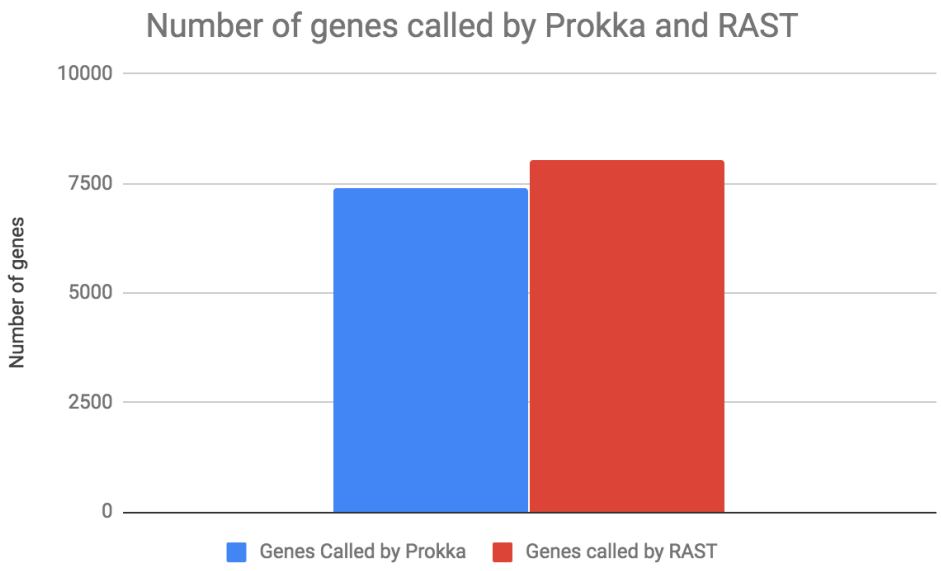


Figure 10: Number of genes called by Prokka and RAST. Prokka version predicted 7,438 genes to be present in the *S. griseus* genome, while RAST predicted 7,775 genes to be present in the *S. griseus* genome.

Prokkrastinator data

Analysis of the *S. griseus* genome with the auto annotation methods Prokka and RAST resulted in different numbers of potential genes. Prokkrastinator was written in the coding languages HTML, CSS, and JavaScript. JavaScript libraries such as Meteor.js and D3.js were used to draw interactive and reactive depictions of data, respectively. This novel bioinformatic tool accepts the uploaded gff output files of RAST and Prokka. Once uploaded, the gene data were visualized on a genome ruler. Gene data such as start, stop, and coding strand were included in the genome browser, as well as the gene table below the graphic.

Analysis of the auto annotation method's gff files with Prokkrastinator revealed that Prokka predicted 7,381 genes and RAST predicted 8,045 genes (Figure 11). Both Prokka and RAST called 6,675 genes that had the same start coordinates, stop coordinates, and coding strand. These were considered shared genes. There were 376 variations of genes between Prokka and RAST, where the genes had different start coordinates, but the same stop coordinates and coding strands. Prokka called 432 genes that were unique to that auto annotation method, or that were not called by RAST. RAST called 1,012 genes that were unique to that auto annotation method, or that were not called by Prokka (Figure 17).

Of the 8,045 genes called by RAST, 17.3 % were not predicted by Prokka, while 82.7 % were. Of the 7,381 genes predicted by Prokka, 9.8 % were not called by RAST, while 90.2 % were (Figure 18).

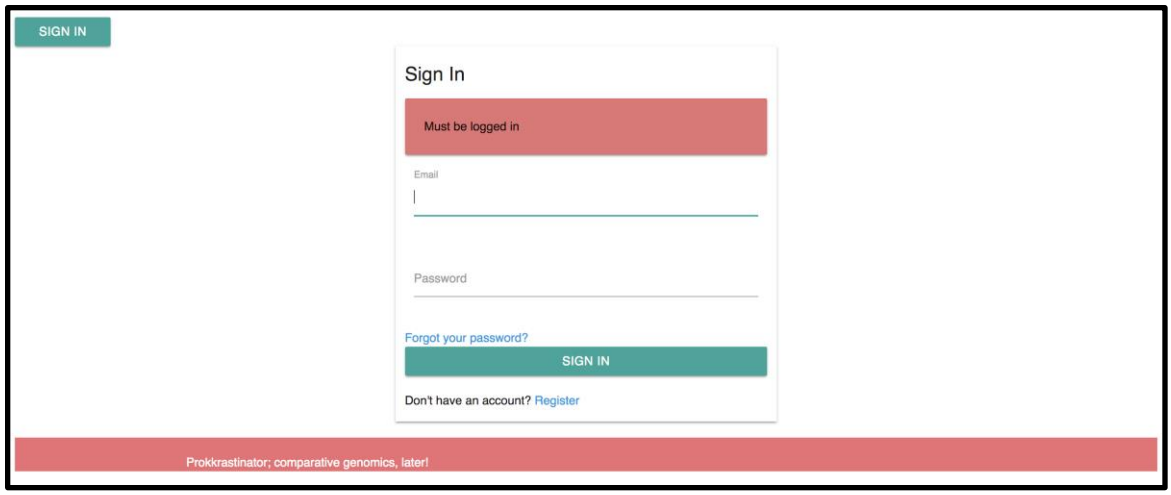


Figure 11: Opening screen of Prokkrastinator. To make a new account, click Register and fill out the form. Click Forgot your password? to reset your password. Enter your email and password to log in.

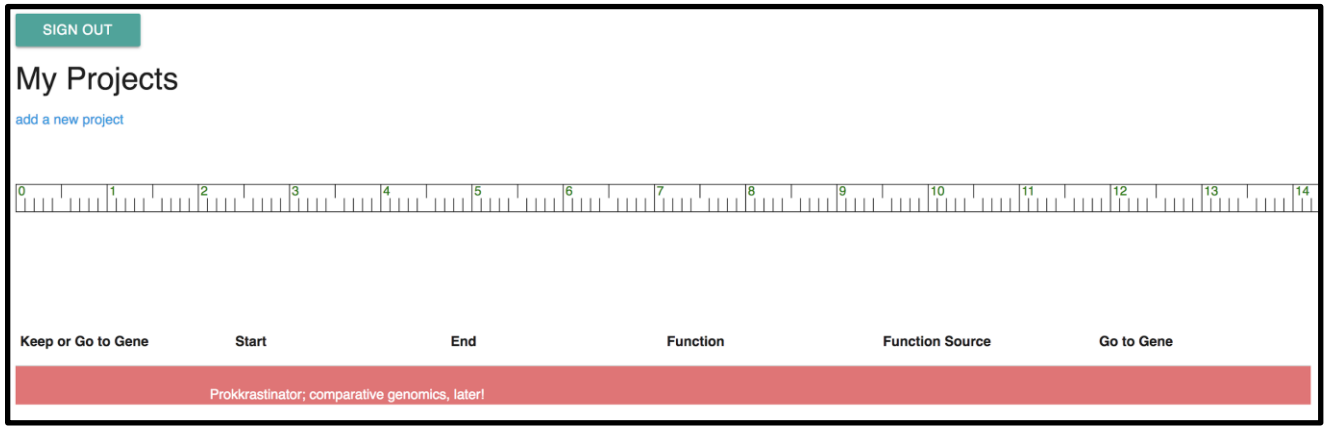


Figure 12: The homepage of Prokkrastinator. Located here is the genome ruler, template for the gene box, and an empty list of all of your projects. To create a new project, click add a new project.

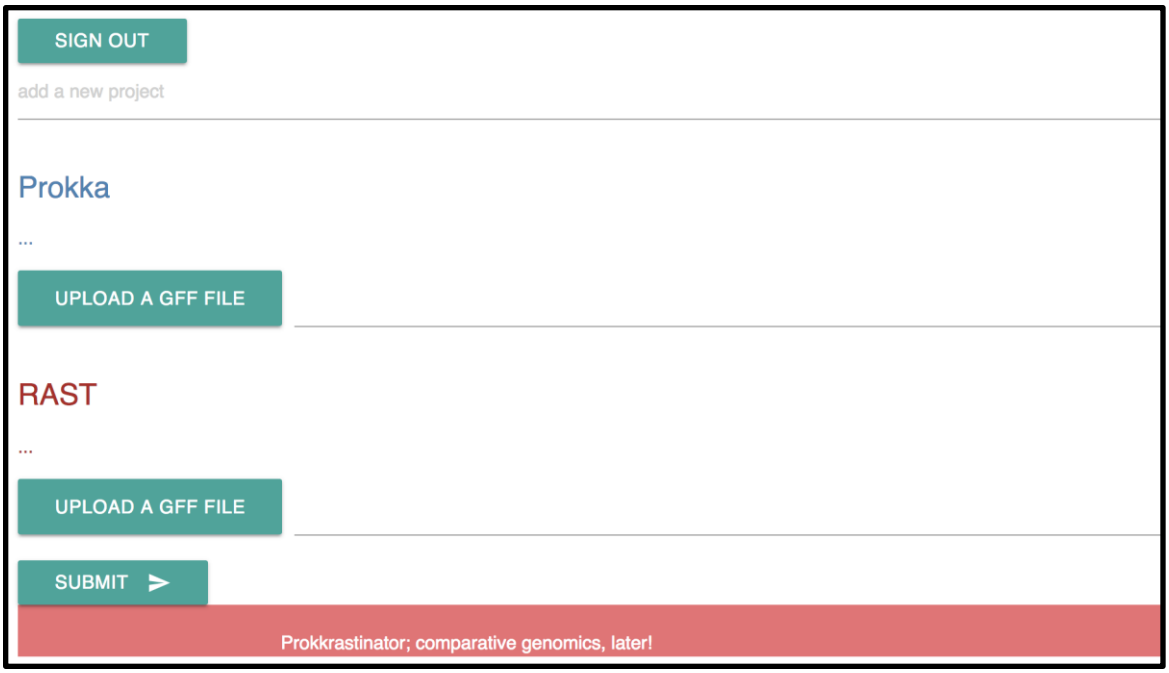


Figure 13: The upload page of Prokkrastinator. To upload the Prokka.gff file, click UPLOAD A GFF FILE under the Prokka title. Parsing the data takes approximately one minute.

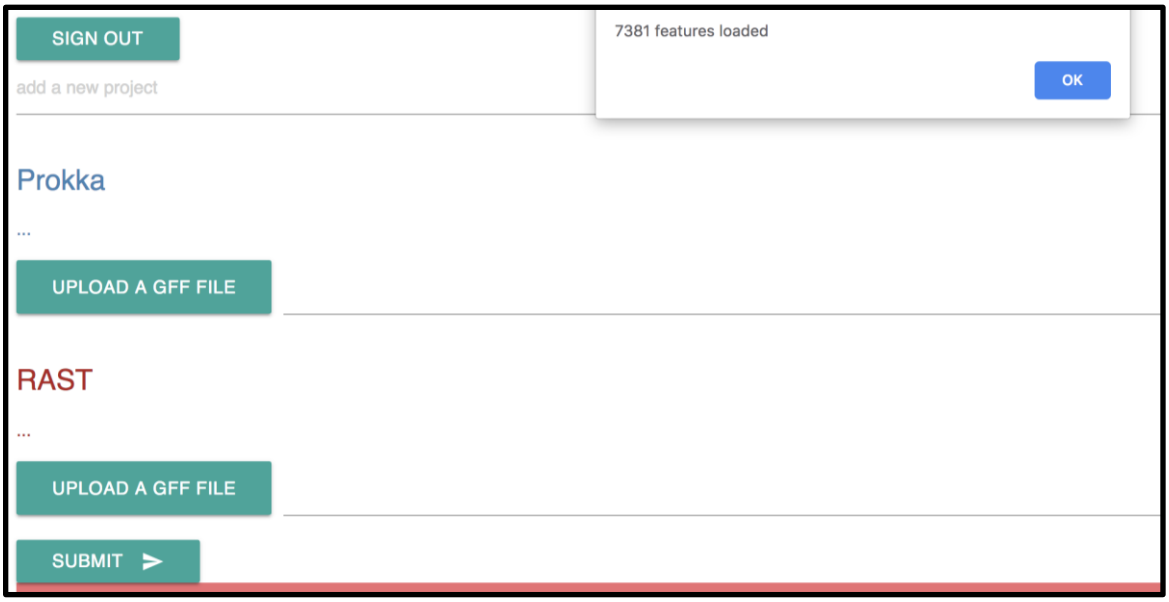


Figure 14: Parsing data completion. The alert message shows how many features, or genes, were uploaded with the gff upload. Click OK to close the alert.

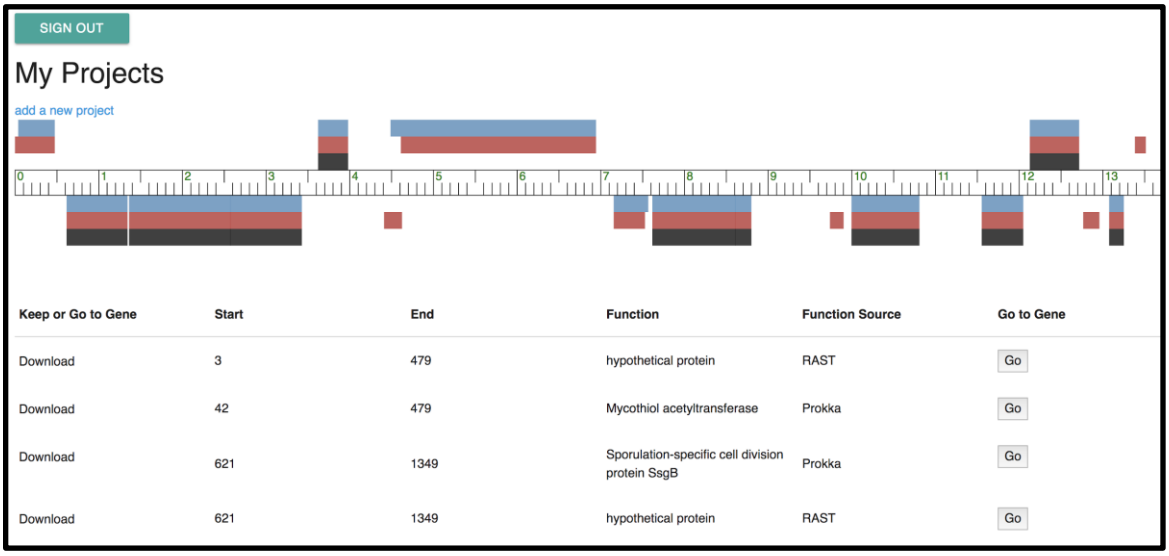


Figure 15: The Prokkrastinator homepage with data loaded. The blue boxes are the Prokka genes, the red boxes are the RAST genes, and the black boxes are what was in common between the two annotation methods. Underneath the genome ruler is the gene box, which provides a list of genes, their genomic coordinates, their potential function, and the program that predicted them.

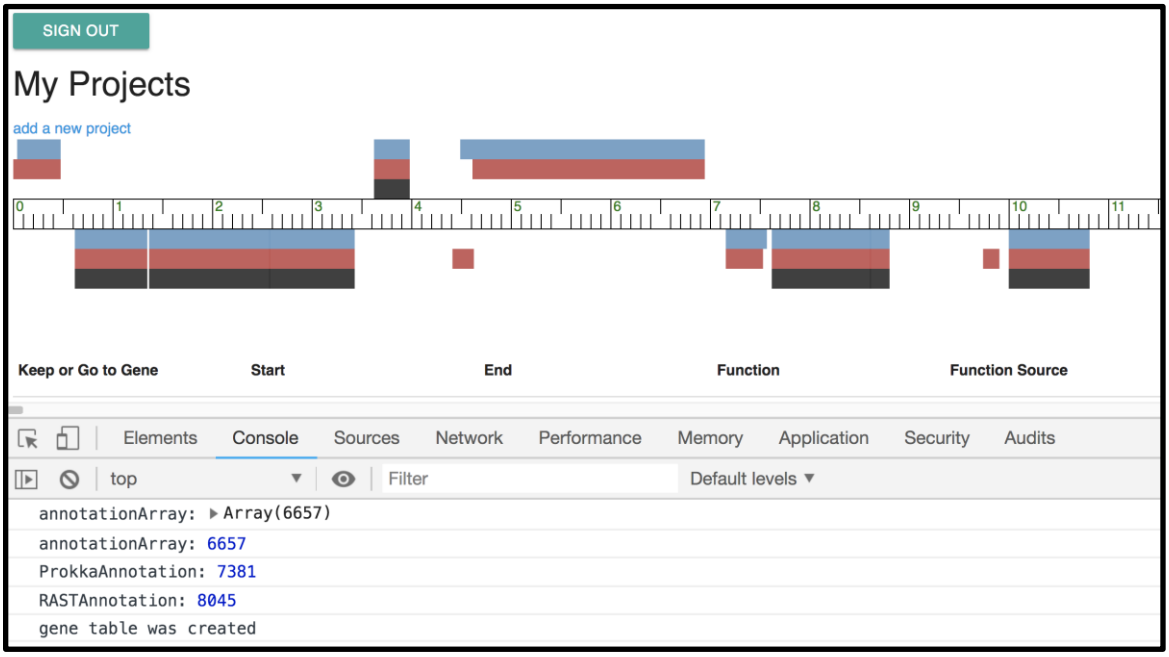


Figure 16: The Prokkrastinator homepage with the annotation method number of genes data loaded. To view this screen, click More Tools, then Developer Tools. Here the number of genes called by each annotation method can be viewed, as well as the number of genes both annotation methods have in common.

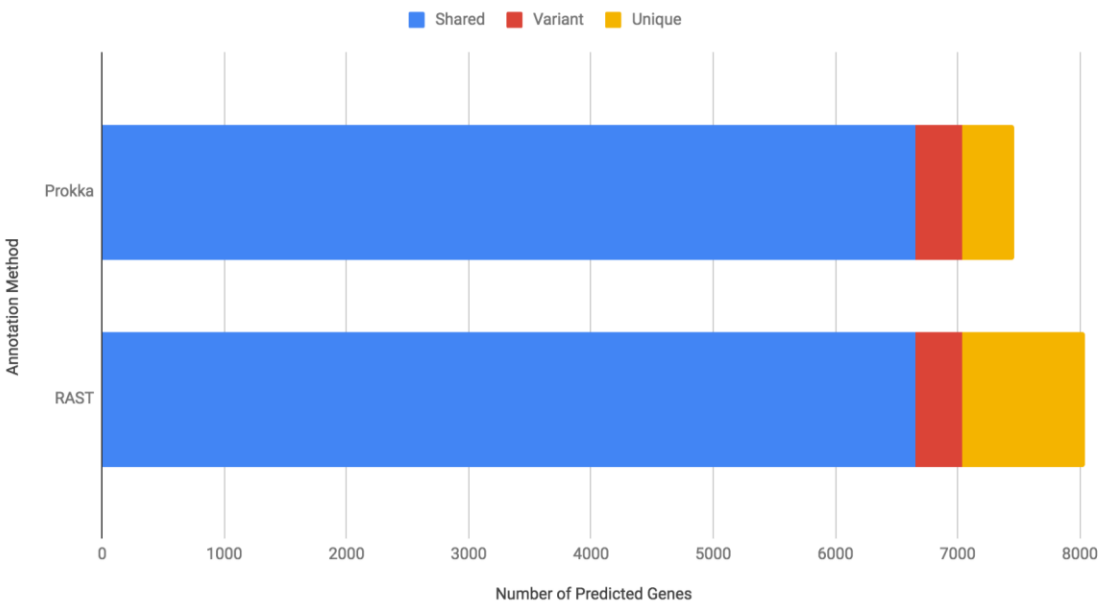
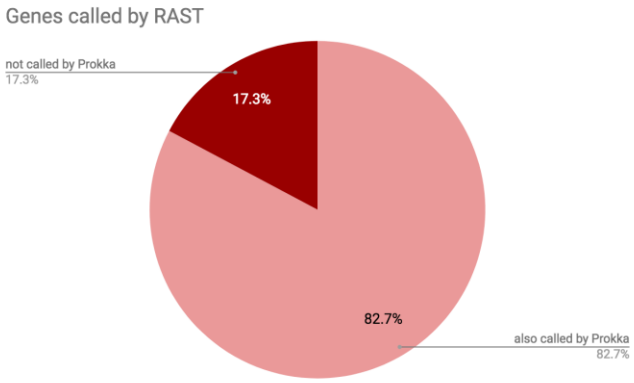


Figure 17: A comparison of the results from Prokka and RAST with the genes unique to Prokka and Rast. Prokka predicted 7,381 genes and RAST called 8,045 genes. Of these, 6,675 genes were called identically by the two programs and 376 genes varied between Prokka and RAST only by the position of their start codon. The remaining genes, 432 in Prokka and 1,012 in RAST, were unique to the program that called them.

18a



18b

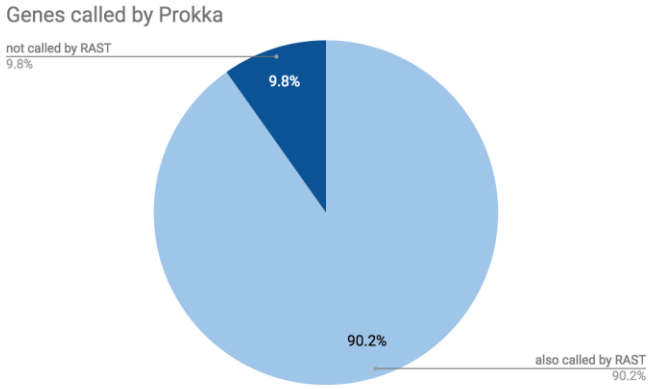


Figure 18: Prokka was used to analyze genes called by RAST that were also called by Prokka (a) and genes predicted by Prokka that were also predicted by RAST (b). Of the genes that were predicted by RAST, 17.3 % were not predicted by Prokka, while 82.7 % were predicted by Prokka. Of the genes that were predicted by Prokka, 9.8 % were not predicted by RAST, while 90.2 % were predicted by RAST.

Bacteriophage protein modeling data

Overall results

To model the gene products from Bacteriophage Wipeout, the bioinformatic tool YASARA was used. DNA Fasta data was uploaded to YASARA after being converted to amino acid fastas with R code. These converted files were then run through YASARA to compare the gene products with homologues, and thus queried against the PDB database.

Modeling the 258 gene products from bacteriophage Wipeout using Yasara yieled 155 protein models. Of these models, 2 were Disgusting (z-score <= -4), 18 were Terrible (z-score between -4.9 and -4.0), 47 were Bad (z-score between -3.9 and -3), 47 were Poor (z-score between -2.9 and -2) , 25 were Satisfactory (z-score between -1.9 to -1) , 9 were Good (z-score between -0.9 to -0.05), and 7 were Optimal(z-score >=0.0). Some gene products, 103 of them, were unable to be modeled at all (Figure 13).

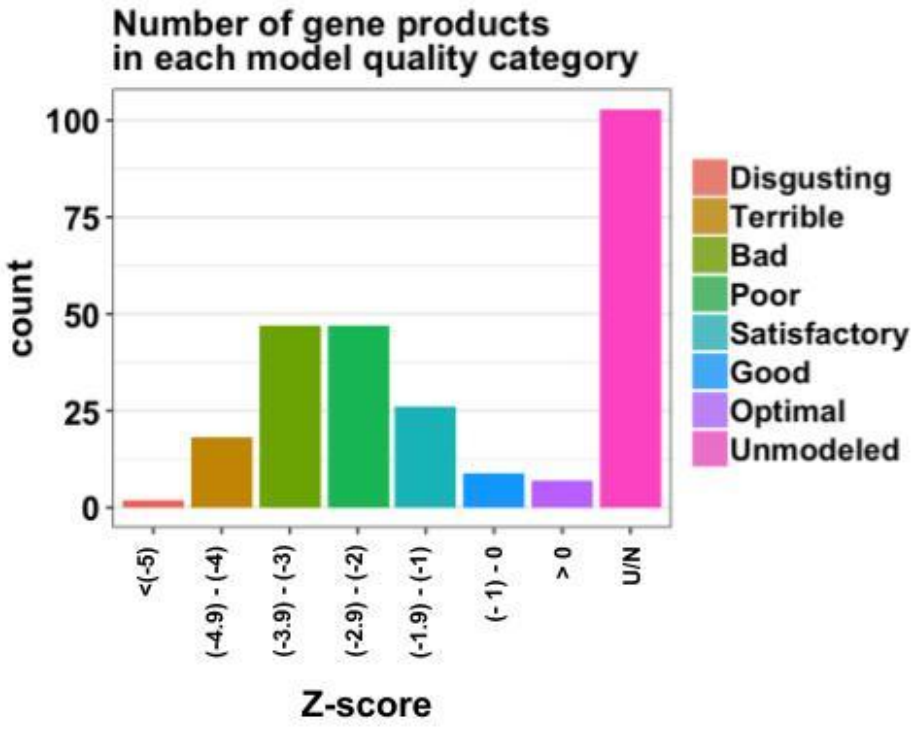


Figure 19: The number of Phage Wipeout gene products in each model quality category according to YASARA. Protein models with a z-score of over 0.0 were considered "optimal" models, models with a z-score of -0.9 to 0 is considered a "good" model, models with a z-score of -1.9 to -1 is considered a "satisfactory" model, models with a z-score between -2.9 and -2 is is considered "poor" quality, models with a z-score between -3.9 and -3 are considered "bad" quality, models with a z-score between -4.9 -4 were considered "terrible" quality, and a model with a score below -5 is considered a "disgusting" model. Out of the 258 genes encoded by phage Wipeout, 103 gene products were unable to be modeled by YASARA.

Essential proteins required for infection

Two gene products were predicted to be structurally functional proteins. Gene product 129 was predicted to be a capsid protein. The model had a BLAST e-value of 7.5e-1, an alignment score of 45, 93% query coverage, and a z-score of -3.65. The capsid protein prediction fell in the Bad quality category.

Gene product 110 was predicted to be a portal protein. The model had a BLAST e-value of 7.8e-1, an alignment score of 52, 66% query coverage, and a z-score of -2.141. This model fell into the Poor quality category.

Proteins of interest

Other proteins of interest from Wipeout include gene products 49, 265, 94, 62, 57, 79, and 93. Gene product 49 fell into the Poor quality category, and was predicted to be a Bcl-2 protein from Human herpesvirus 8. The model had a BLAST e-value of 9.8e-1, an alignment score of 53, 53 % query coverage, and a z-score of -2.724. Wipeout gene product 265 was predicted to be a swarming motility protein, and had a BLAST e-value of 7e-17, an alignment score of 170, 90 % query coverage, and a z-score of -2.35.

Wipeout gene product 94 was in the Bad quality category, and was predicted to be a CRISPR-associated exonuclease Cas4 protein. The model had a BLAST e-value of 1e-3, an alignment score of 66, 79 % query coverage, and a z-score of -3.303. Gene product 66 was predicted to be a CdiI immunity protein, and had a BLAST e-value of 8.9e-1. The model also had an alignment score of 47, 77 % query coverage and a z-score of -3.01.

Gene product 57 resembles a penicillin G acylase protein, and was in the Poor quality category. The model had a BLAST e-value of 1.8, an alignment score of 45, 76 % query coverage, and a z-score of -2.401. Wipeout gene product 79 was identified as a potential pertussis toxin-like subunit ArtA protein. The BLAST e-value for the pertussis toxin-like subunit was 0.9, the alignment score was 42, had 67 % query coverage, and had a z-score of -2.586. Gene product 93 was identified as a potential tetanus toxin-like protein. This model had a BLAST e-value of 1.6, an alignment score of 45, 90 % query coverage, and a z-score of -2.352.

Potential anti-CRISPR proteins

All bacteriophage produced anti-CRISPR proteins are between 50- 150 amino acids long, or 150-450 base pairs long (Dong et al., 2017). Of the 103 unmodeled gene products from phage Wipeout, 13 gene products fell between 150-450 base pairs long. Gene products 2, 36, and 42 are 150, 291, and 157 bp long, respectively. Gene product 56 is 332 base pairs long. Gene products 59 and 262 are 164 and 262 base pairs long, respectively. Gene product 60 is 262 base pairs long, and gene product 61 is 167 base pairs long. Gene products 101 and 106 are 207 and 206 base pairs long, respectively. Gene product 144 is 179 base pairs long. Gene products 183 and 266 are 264 and 160 and base pairs long, respectively. Gene product 266 is 160 base pairs long. Finally, gene product 275 is 150 base pairs long (Table 9).

Table 10: Unmodeled phage Wipeout gene products with nucleotide lengths of 50-450 bp. Listed with the length of the genes are the amino acid sequences that correspond to each gene product.

Gene Product	Length (base pairs)	Amino Acid sequence
2	150	MRHTQTIEGRMAAAREALNITFAYADEFEREMNELFRTECTKNQFDELI KTIYGDRPAENVKGSQVKWDGKRDLLMGIFTDTGDGPKTTQSLAGTM AGALNALTERLDWYRMPRKGEVDNLFAAASGFDPVINSEKNKIRKAVL SMALAA
36	291	MKKPDLKRLMDWRGAVSFGSVLIVALALSWWSLYSMAVEFYGVPKEL AFGVSIAFDGAALFVADLASKYARTEDSGLAPKLATYAFVGTSVYLNV EHAALLNYGTPGKVLFGAPPVIAGVLFELYLRFVHRSEMRSNGLVAKR MPVFGKVSWMIFPGKTFKGFKNVVFFRLNEVVTNVTGESLSRKRDKPV TPKKMSRDKRDKTDDKSVTKPVMTKDDTPVKTNAPIVTTVTEIPRRDV TDDKQKSVSRLVKELWDDGTRDKAELRKLISDIKGRDIPANTITVALSR MSP
42	157	MKTDEDRCREYGVPYDPTITRPAVYQQSGWKCHLCGKRVRKSLKYPH PRSASLDHIVPLSWRKDSPGHVWGNVALAHLRCNQSKGARFAGSTKPA PRKPGIVTPLWKLRLTLFAGTAALFYFNAEPTVLTIAAVLCILSVVPQRK VRRRRRRAWWKL
56	332	MSDELIEKVIRTTEVASGGGGLLNAEQSNRFIDYMWEATVLGTQVRTIR MRADTVDIDKLGIGERLMRVATEAVDDGVNAGATFSKISLTTKKLRLD WEISTESLEDNIEGDALEDHIARLMATQAGNDLEDLAINGDTALTGNPL LKAFDGWRKRALAGGHVIDHGGNGVDRSVFNKALKAMPRKYMQRRN GLKFFTGSNVIQDYLFSLQNTSADYVTPEALAAAGINSGVRTEGPAGFT TGNAFGIPVQEVPLFEETLDGDYSGASGDHADVWLTFPNNMLWGVKR EVQIFTEFKPKKDTTEYTMYCRVGTQIENADAFVVVKNVKIAS
59	164	MSCLISSRFNMRATVLRQAGTNPQENPGGHWETVQDPDSGAIERVWVP DEDSGTPGDQTLVIKCMVRGVTNGGIRAAGTTQRFSEIYENVDWAIIQF PASVVLSKYDRVTNISNSKGQLIWREEEINQAPATVFQVMGSTPVIDPFG NHIENTALLQRAQVQSG
60	262	MAKGKAFVGVTADTTEVSALTGFLSTLSTQIKADVNMAPVLDYAHSA MSSKFDAYMSAIAPTAPNQFHHVYDWGRIGVPQNQLWKNVIRGRGAN KYASFEFRASVLPVPLPEGNKKPFMRKHRFIYKAMIMEYNIGVTVKPKR AKMLAFPNERGDIIFTKGPVFVQNPGGMGTTGAFTAAWSNWWGGAGA DQVFKSEIQRVLERDLGGSALSRFMRRFKKAKFRTGRIAVADSKSAMA DGSRLALEFLQERNRKNEARRKAQ
61	167	MTYDITATHALNKFLVDQLSGENLIDLTKYNGLSPFIPAQQQPELTNLPS GVPFFVYNYASNGEYEDWWLEHEQAAYIIYSDNEKQIRQISNYLNQLL KRYDWAADDINDYLREFGTTEQKKFEFKYTRVISMASIEPATEEGGRY AGSITVNLCFTSQLNQEGMRV
		MIKIGMTGAQGTGKTSMAQAMINSPAFKDFVLVPSTARQIKDYGYPINR

101	207	EATELSQLLVPALRMVDEWETMNSPQNTLYKQGIISDRTLIDSLAYTYY QNEHVWENGALVENVTRRLTEMHMQSYHFVLYFPVYWEAEDDGVRD ADESYRTRIDDYVIQALEMLNVPYLTVPDVSPEERVEWFIGEIQELYAE AERQYLRRFGNDLL
106	206	MSTTTAMLLIVDRSGSMSSIQMEAEQALNDFLTKQKAVDGRCGVKVV QFDGEVEDLFGPVALAQAPEIKIVPRGMTALLDAIGKSVTEFRETMDHM PVDKRLVVILTDGLENVSQEWKLDAVNKLISESRELGYEFIFLAANQDA IATGAQMGIPTASSLTFAASPAGVRGMGMTMDMYVTNYREGKAAEFT DQDRATASGLTDED
144	179	MTDRTDKFTGRPVTGDRPENTSMVEQANTQLFIDAMDELLAHPLVKA VRWTQYTPGFNDGEPCTFDGHTPEVCLTGLELGEEGIETERDHYEDGD EVWLGEYDMYEYPRKENGQIDWDQPKIYKVGGVDTTEINELLSKFAGC VEGGRHDVWMNRTFGDPAEVIATPEGFEVSFYDCGY
183	264	MGIFSRKVKPAFVPAQNLSAKKTVTLTPIVSGQPAANVSLIKQNGVSFE KKVESAVKLQKDAGVANRFDVIGLIDESYSMDYLFANGTVQTISERVL AWTAGVDADGMAPVGGFANGFQWHGEIDLTNVMGCASGWGCWGGT DLAAGLREAFEVAKGADNPVYLFIVTDGAPNDRQAVIDLIAQMSEYPIF IKIVLVGDDPQGKKFVEYLDDLEKHEPGRRLFDNTDAQHIRDASRVSDD DFNKAMTEEVPSAIDAMRQAGLVV
266	160	MTDLARMTDAEISSKFKGLVEGYAKDLAKSNKSKGDPGTASIASVRVG KIKRKKGQLVPSQILLAMRAAIWKMGWAQGVLRDEEGRLCLRGAMV WLYRKGYFHQEDGDIAAQWLHDEVRRKNGDAQKYNFIYWNNDPRRT LAEILKILEDAGNRAKLAGE
275	150	MRHTQTIEGRMAAAREALNITFAYADEFEREMNELFRTECTKNQFDELI KTIYGDRPAENVKGSQVKWDGKRDLLMGIFTDTGDGPKTTQSLAGTM AGALNALTERLDWYRMPRKGEVDNLFAAASGFDPVINSEKNKIRKAVL SMALAA

Cas3 protein modeling data

The Cas3 gene from *S. griseus* ATCC 10137 was obtained using Prokkrastinator. Once the Cas3 DNA fasta file was obtained, Expasy was used to translate that DNA fasta data into an amino acid fasta file. Once translated, the amino acid file was uploaded to the Phyre2 server to compare the sequence to potential homologues. Phyre2 outputs a PDB file, which was uploaded to YASARA to further analyze the protein. The Cas3 protein from *S. griseus* is 1,559 amino acids long, of which 95 % Yasara could be modeled at >90 % confidence. Of the 1,559 residues in the protein, 900 residues (58 % of the protein) could be modeled with 100% confidence.

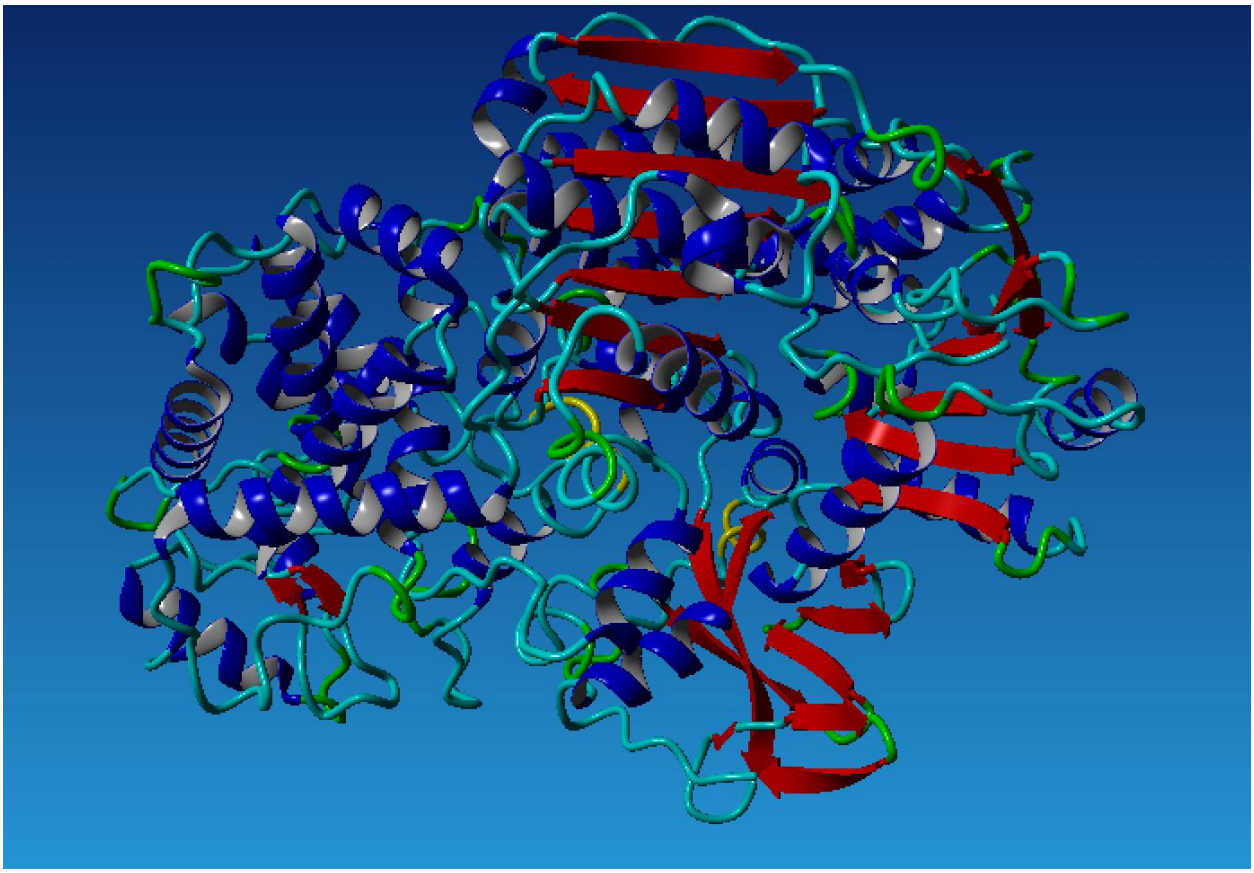


Figure 20: Cas3 protein from *S. griseus*. The fasta sequence of the *S. griseus* gene was obtained using Prokkrastinator. The DNA sequence was translated and modeled in using Phyre2, and then modeled using YASARA.

Cas3 ligand search data

COACH modeling

After the potential structure of the *S. griseus* ATCC 10137 Cas3 protein was predicted, the PDB file was uploaded to the COACH server to predict potential ligand binding sites and active residues. COACH predicted Mg, ATP, DTP, and nucleic acid to be potential ligands of the *S. griseus* Cas3 protein. The magnesium ligand was predicted to interact with Cas3 residues 314 and 465, as seen in Figure 56. This prediction had a C-score of 0.25, and had a total of 10 clusters (Figure 57).

The ATP ligand prediction potentially interacts with Cas3 residues 280, 282, 283, 284, 287, 309, 310, 311, 312, 313, 314, 315, 349, and 465. The C-score for this prediction was 0.18, and had 9 clusters (Figure 58). The ligand DTP was predicted to interact with Cas3 residues 280, 284, 287, 309, 310, 312, 313, 314, 315, 349, 656, 674, 678, and 681. The C-score for this prediction was 0.08 and had 3 clusters (Figure 59).

COACH predicted nucleic acid to be a possible ligand for *S. griseus* Cas3. The nucleic acid is predicted to interact with Cas3 residues 21, 68, 71, 109, 143, 217, 218, 223, 339, 340 ,372, 436, 438, 439 ,590, 613, 648, 649, 650, 653, 712, 716, 717, 823, 827, 828, and 880. This prediction had a C-score of 0.02, and had 1 cluster (Figure 60).

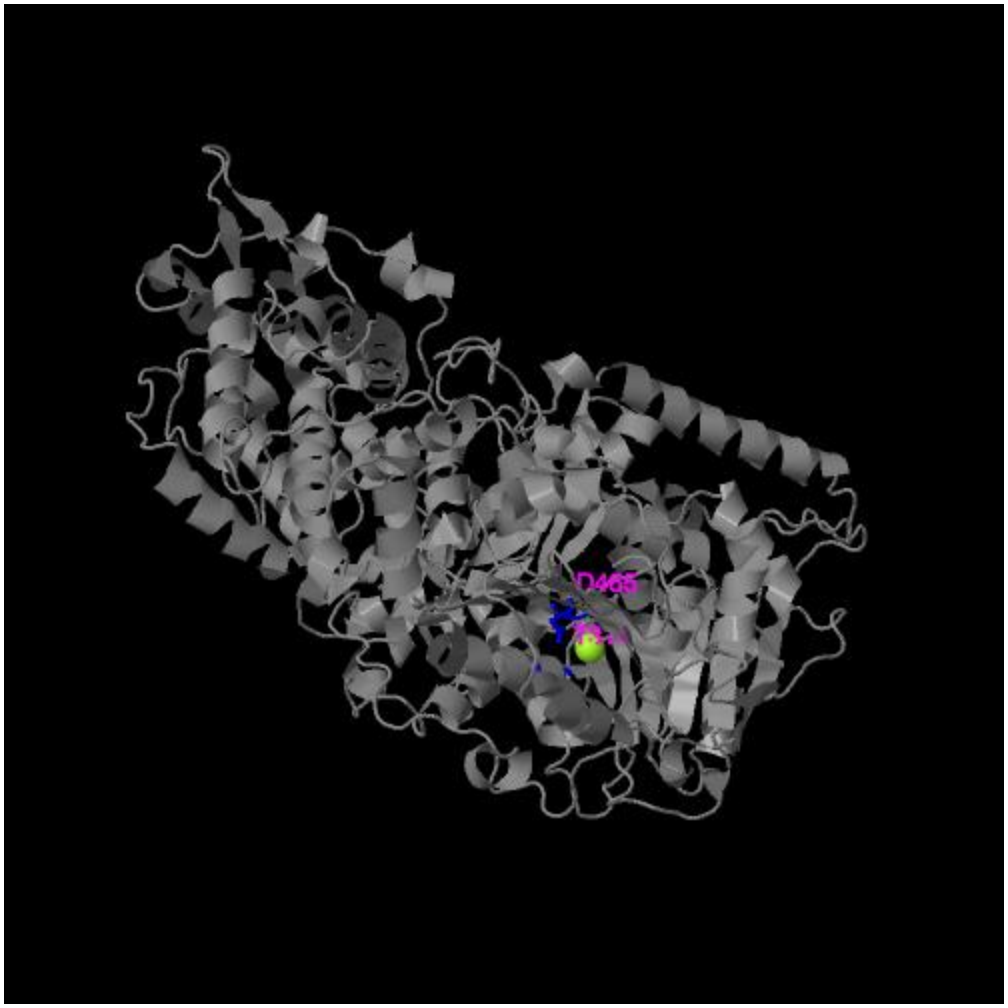


Figure 21: The COACH ligand MG ligand prediction is highlighted in yellow. The consensus binding residues 314 and 465 are highlighted in blue and annotated in purple. The c-score for this ligand prediction was 0.25 and had 10 clusters.

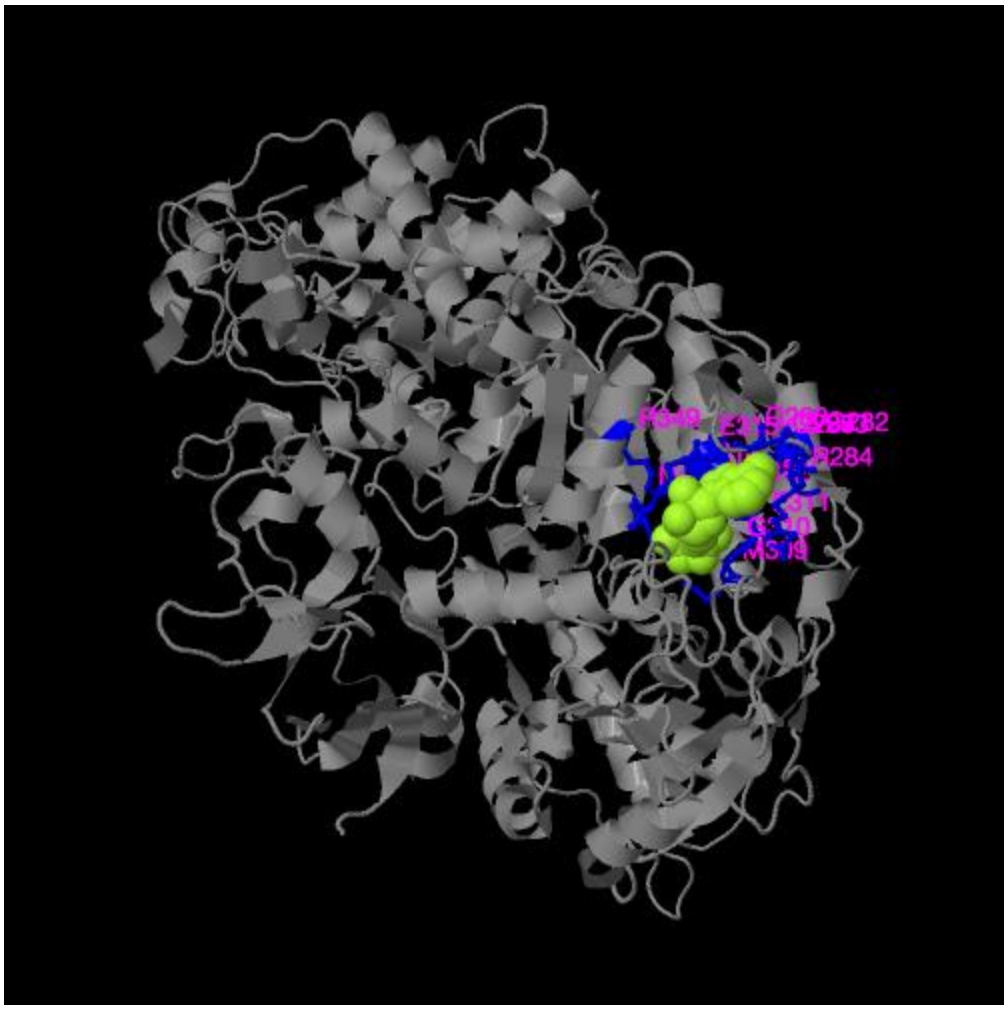


Figure 22: The COACH ligand ATP prediction is highlighted in yellow. The consensus binding residues 280, 282, 283, 284, 287, 309, 310, 311, 312, 313, 314, 315, 349, and 465 are highlighted in blue and annotated in purple. The c-score for this ligand prediction was 0.18 and had 9 clusters.

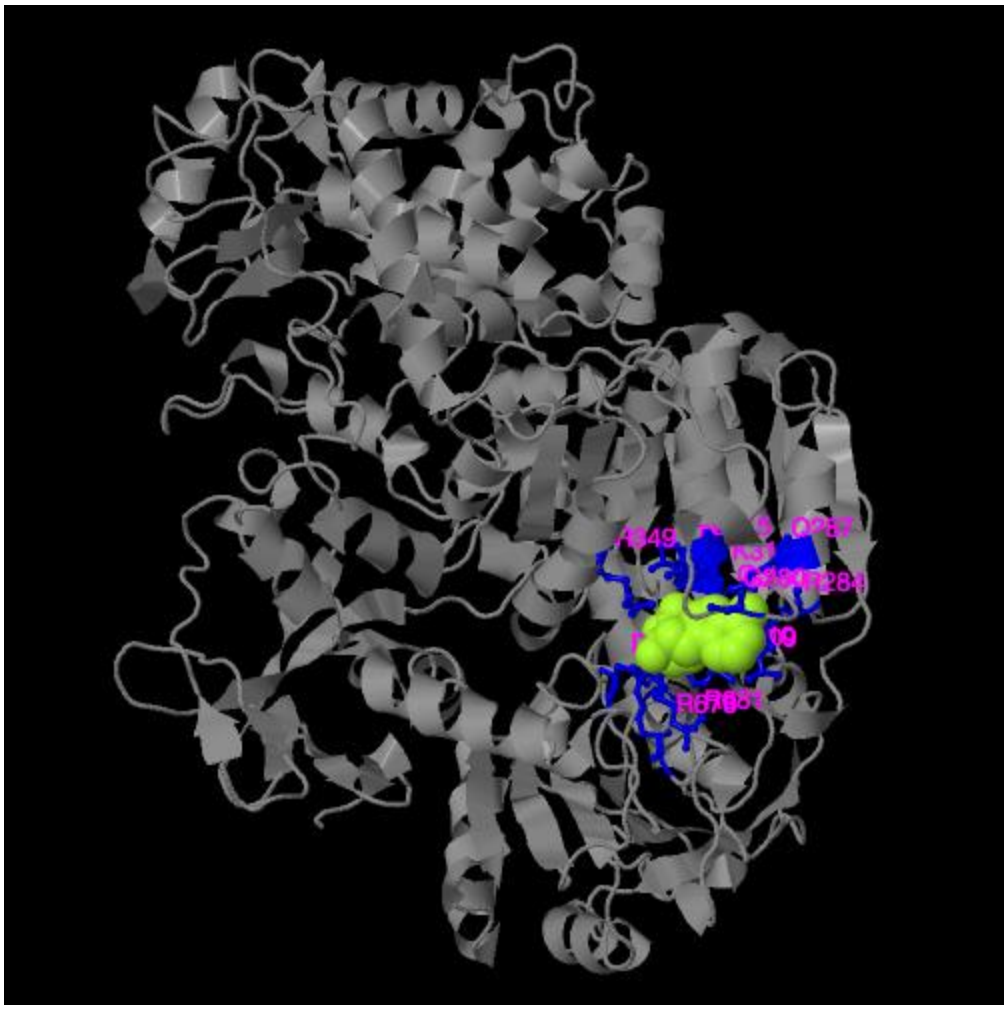


Figure 23: The COACH ligand DTP prediction is highlighted in yellow. The consensus binding residues 280, 284, 287, 309, 310, 312, 313, 314, 315, 349, 656, 674, 678, and 681 are highlighted in blue and annotated in purple. The c-score for this ligand prediction was 0.08 and had 3 clusters.

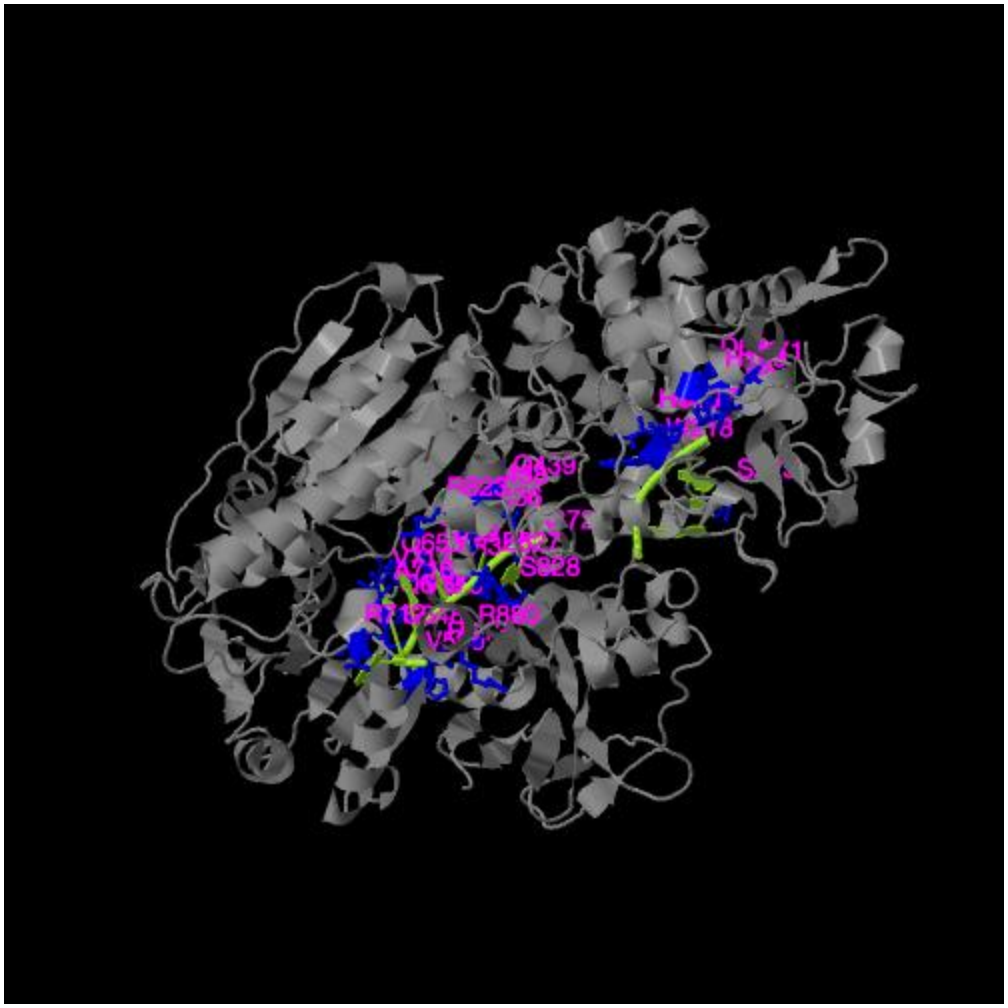


Figure 24: The COACH ligand nucleic acid prediction is highlighted in yellow. The consensus binding residues 21, 68, 71, 109, 143, 217, 218, 223, 339, 340 ,372, 436, 438, 439 ,590, 613, 648, 649, 650, 653, 712, 716, 717, 823, 827, 828, and 880 are highlighted in blue, and annotated in purple. The c-score for this ligand prediction was 0.02, and had 1 cluster.

TM Site Results

The COACH server also ran TM Site predictions for the *S. griseus* Cas3 protein. The TM Site prediction identified DTP and nucleic acid as potential ligands. DTP is predicted to interact with Cas3 residues 280, 284, 287, 309, 310, 312, 313, 314, 315, 349, 656, 674, 678, and 681. This prediction had a c-score of 0.31 and 1 cluster (Figure 25). Nucleic acid is predicted to interact with Cas3 residues 21, 68, 71, 109, 143, 217, 218, 223, 339, 340, 372, 436, 438, 439, 590, 613, 648, 649, 650, 653, 712, 716, 717, 823, 827, 828, and 880. This prediction had a c-score of 0.25 and 1 cluster (Figure 26).

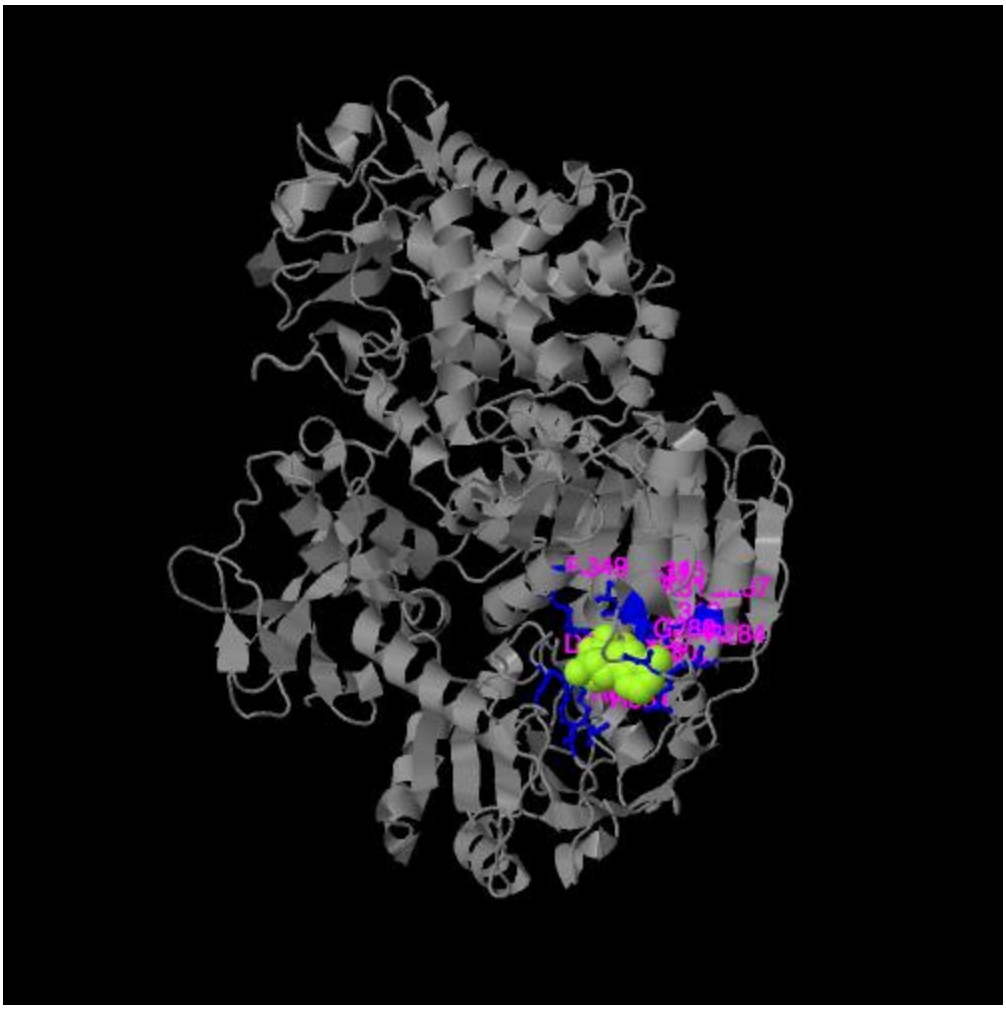


Figure 25: The TM Site prediction of the ligand DTP is highlighted in yellow. The predicted binding residues 280, 284, 287, 309, 310, 312, 313, 314, 315, 349, 656, 674, 678, and 681 are highlighted and annotated in blue and purple, respectively. The c-score for this prediction was 0.31, and had 1 cluster.

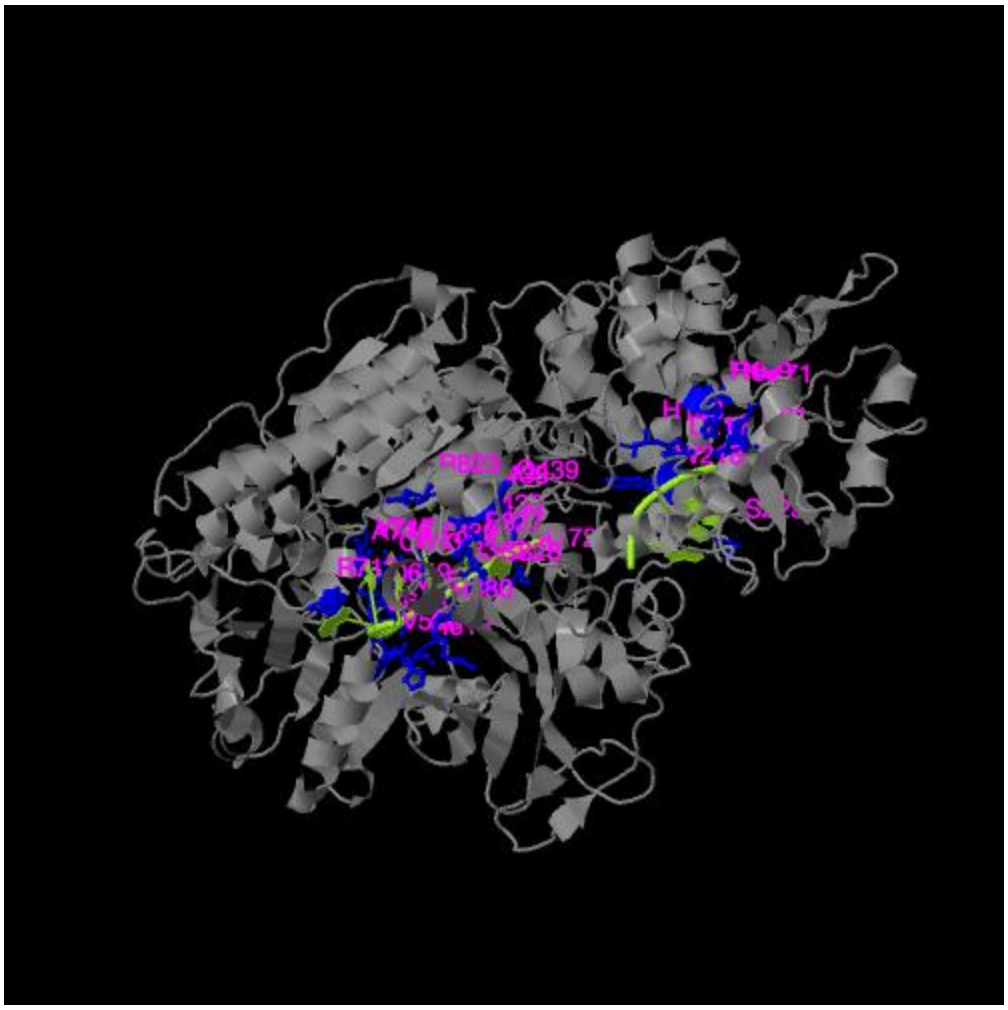


Figure 26: The TM Site prediction of the nucleic acid ligand is highlighted in yellow. The predicted binding site residues 21, 68, 71, 109, 143, 217, 218, 223, 339, 340, 372, 436, 438, 439, 590, 613, 648, 649, 650, 653, 712, 716, 717, 823, 827, 828, and 880 are highlighted in blue and annotated in purple. The c-score for the prediction was 0.25 and had 1 cluster.

Discussion

The evolutionary arms race between viruses that infect bacteria, bacteriophages, and their hosts has resulted in ingenious strategies to promote and defend against viral infection and propagation. For example, during phage infection the virus injects its DNA into the host cell. In response bacteria encode and express exonucleases, enzymes that specifically degrade linear phage DNA while leaving untouched the typically circular chromosomes and plasmids of the bacterial cell (Berg et al., 2002). This is problematic for the phage; without a way to avoid the bacterial exonucleases, the phage fails to replicate within the bacterium. The evolutionary pressure of bacterial exonucleases on phages selected for the circularization of the phage genome once injected into the cell (Lehman et al., 1964).

The circularization of phage DNA to avoid degradation by bacterial exonucleases is not new knowledge, unlike the discovery of bacterial CRISPR-Cas systems (Barrangou et al., 2007). CRISPR-Cas is a bacterial adaptive immune system. As the bacterial cell interacts with infecting phages, CRISPR-associated proteins (Cas proteins) like Cas9 and Cas3 cut the phage DNA. While both Cas9 and Cas3 cut phage DNA, the results are quite different. Cas9 is known in popular culture as the gene editing protein. The DNA cuts that are made with Cas9 are precise, enabling it to be combined with a guide RNA (gRNA) and pre-designed DNA on a separate plasmid to introduce targeted substitutions, insertions, or deletions (Wu Y et al., 2013). Cas3, however, does not make a clean cut in the genome. The Cas3 protein shreds the phage DNA beyond repair (Bondy-Denomy et al., 2016)

Interacting with either the Cas9 and Cas3 proteins would be detrimental to the phage, as a cut phage genome is no longer infectious (Bondy-Denomy et al., 2016). It has recently been discovered that phages have evolved a defense against CRISPR. These phage-encoded anti-CRISPR, or Acr, proteins are the latest advance in an ancient arms race.

A total of ten Acr proteins have been described in the literature, four of which are well characterized (Pawluk et al., 2014). Three of these ten anti-CRISPR proteins, AcrF1, AcrF2, and AcrF4, interact with the Csy complex, which is a group of proteins that binds the target DNA through hybridization of the target and crRNA. These proteins directly inhibit the complex's ability to bind to the PAM site in the dsDNA target. AcrF1 and AcrF2 bind to different areas of the Csy complex, and do not inhibit each other in the process. They are able to simultaneously bind to the Csy complex and prevent it from binding to target DNA (Bondy-Denomy et al., 2015).

AcrF3 interacts with Cas3 and blocks the helicase-nuclease from binding to the Csy-DNA complex. Because Cas3 is unable to bind to Csy, Csy acts as a transcriptional repressor and blocks the transcription of the CRISPR-Cas complex (Bondy-Denomy et al., 2015).

Acr proteins are not only useful for ensuring the success of phage infections; these proteins have been used in tandem with CRISPR-Cas genome engineering mechanisms to attempt to decrease the occurrence of off-target genomic changes. For instance, one such study used CRISPR-Cas editing to correct a mutation responsible for cystic fibrosis, followed by an *Acr* protein treatment to halt the progression of the CRISPR-Cas complex. The use of Acr proteins was effective in preventing deleterious

effects caused by CRISPR-Cas complexes (Shin et al., 2017). Acr proteins have been shown to be effective tools for CRISPR-Cas genome editing, and are already designed and implemented by phages. There are an estimated 10^{31} phages on the planet (Hatful, 2013), and those phages that infect bacteria that encode CRISPR-Cas arrays require some mechanism to circumvent CRISPR-Cas-mediated immunity.

The SEA-PHAGES program is an international consortium of universities, tribal colleges, and community colleges that offers a course-based undergraduate research experience focused on the discovery and genomic characterization of novel actinobacteriophages, phages that infect members of the phylum Actinobacteria. The program has stored these phage data (genomic and amino acid data) on Phamerator.org.

Most of the bacterial species and strains used in this program have been completely sequenced, but this is not true of the *S. griseus* ATCC 10137 strain that is used at JMU. A related strain, *S. griseus* 13350, has been sequenced and published. Interestingly, this strain has been shown to encode CRISPR-Cas arrays, implying that phages which can infect it use a mechanism that allows them to circumvent CRISPR-Cas-mediated immunity.

If *S. griseus* 13350 encodes CRISPR-Cas arrays, the *S. griseus* strain used by JMU and the SEA-PHAGES consortium would probably contain CRISPR-Cas arrays as well. If *S. griseus* 10137 encodes CRISPR-Cas arrays, then the phages that infect them may have anti-CRISPR proteins. If those anti-CRISPR proteins are similar to other anti-CRISPR proteins, models of optimal or satisfactory quality will be generated from the phage amino acid fastas. If those proteins are unique anti-CRISPR proteins, they will not be able to be modeled, but will fall in the same size range (50-150 amino acids long) as

other known Acrs. Thus, the overall goals of this study were to determine if *S. griseus* ATCC 10137 encoded CRISPR-Cas arrays, and to discover the mechanism phages were using to work around those host defense systems.

To determine if *S. griseus* ATCC 10137 encodes CRISPR-Cas arrays, *S. griseus* ATCC 10137 DNA was extracted using a phenol chloroform based method. The DNA was then assessed for quality and quantity using both a BioTek Synergy H1 and Qubit fluorometer. Analysis with the BioTek synergy concluded that the third DNA extraction yielded on average 594.5 ng/ µLof DNA. As can be seen in table 3, The 260/280 ratio for the extraction was reported to be 1.8125 on average, and the 260/230 ratio was 2.09 on average. Analysis with the Qubit showed that the concentration of the DNA was 379 ng/uL. This difference in quantity could be due to the individual accuracy of the different instruments.

Once the concentration and quality of the DNA was assessed, DNA sequencing was possible. The Illumina MiniSeq and Nanopore MinION owned and operated by the James Madison CGEMS facility were utilized for the project. As can be seen in Table 6, sequencing with the MiniSeq yielded 9,773,962 total reads, and 18,982,977 identifiable reads. *S. griseus* samples 1, 2, 3, and 4 combined made up a total of 51.5% of the total identifiable reads from the sequencing run. 97.5 % of the total reads had a QScore equal to or above 30 (Figure 1), with a majority of those reads being generated during each cycle (Figure 2).

The Nanopore MinION sequencing run yielded 1,688,493.0 total reads, and 7,601,206,820.0 reads with a read length N50 of 7,576.0 bases. The sequencing run also generated 140,000 reads with the mean read length of 4,501.8 bases (Figure 5). The run

lasted for approximately 50 hours, and generated a cumulative yield of over 7 gigabases of data (Figure 3) and approximately 1,750,000 bases (Figure 4).

The read lengths remained similar over the course of the run (Figure 6), however the quality of the reads decreased over the 50 hours that the sequencer was collecting data (Figure 7). This is not unusual with the Nanopore MinION, as it is known to have lower quality reads compared to the Illumina MiniSeq. The Nanopore MinION has an accuracy rate of 97 %, while the Illumina MiniSeq has an accuracy rate of 99.9 % (Tyler et al., 2018 & Glenn, 2011). Nanopore reads are also longer than the Illumina reads generated, making them useful for hybrid assemblies.

By combining both the Nanopore and Illumina reads with the bioinformatic tool Unicycler, a hybrid genome assembly was built. The final product resulted in one unitig which was 8,576,363 base pairs long, and 9 smaller contigs (Figure 9). An assembly with just the filtered Nanopore data provided evidence that the smaller contigs were missasemblies due to the lower quality of Nanopore reads. This leads to the conclusion that the unitig from the hybrid assembly is the full genome of *S. griseus* ATCC 10137.

Once the assembly was generated and assessed for completion, the *S. griseus* genome was run through CRISPRfinder to determine if CRISPR-Cas arrays were present in the genome. There were a total of 3 confirmed CRISPR-Cas arrays present, with a total of 53 spacers shared between them (Table 7). There were also 10 questionable regions, or regions where the CRISPRfinder program was not certain enough about the presence of a true CRISPR-Cas array. Between the 10 questionable regions there were 16 spacers (Table 8). This provided evidence that there were CRISPR-Cas arrays encoded in the *S. griseus* ATCC 10137 genome, completing one of the objectives of the study.

To identify a *Cas9* or *Cas3* gene and determine the rest of the contents of the *S. griseus* genome, the *S. griseus* ATCC 10137 fasta file was analyzed using Prokka and RAST. The Prokka program, version 1.13, was run on the command line and called 7,381 genes. Out of curiosity, the genome was run through the web app genome annotator RAST. RAST predicted 8,045 genes (Figure 11). The disparity between the number of genes called by Prokka and the number of genes called by RAST was alarming. However, there was no tool to be able to view both the Prokka and RAST annotations side by side to gauge which annotation method was correct.

This led to the development of Prokkrastinator, a genome browser that combines both Prokka and RAST annotations. The majority of the Prokkrastinator code is written in JavaScript and uses libraries such as D3.js and Meteor.js. Python is used for the general feature format file parser, while GitLab is used to store the code and for version control. To use Prokkrastinator, an account was made on prokkrastinator.org (Figure 11). Once the account was made, "add a new project" was clicked (Figure 12). Once at that form, the project name was entered. Prokka and RAST annotations were individually exported as GFF files and submitted to Prokkrastinator for analysis and visualization. After the necessary information is added, the "submit" button was clicked (Figure 13). Only 6,657 genes were called by both Prokka and RAST (Figures 14 & 15). Of the 8,045 genes called by RAST, 17.3% were not predicted by Prokka and 82.7% were. Of the 7,381 genes predicted by Prokka, 9.8% were not called by RAST, while 90.2% were (Figure 18).

S. griseus ATCC 10137 was also found to encode genes such as CRISPR-associated endoribonuclease *Cas1*, CRISPR-associated endoribonuclease *Cas2*, CRISPR-

associated endoribonuclease *Cse3*, CRISPR system Cascade unit *casD*, CRISPR system Cascade unit *casC*, and CRISPR-associated helicase/nuclease Cas3. Prokkrastinator was used to obtain the nucleotide sequence corresponding to the *Cas3* gene for subsequent protein modeling. Modeling the Cas3 protein would provide further evidence as to its structure and function.

Once the fasta file for the *Cas3* protein was obtained, it was translated into amino acids using the ExPASy DNA translator. The Phyre2 server was used to generate a potential structural model of the Cas3 protein. Phyre2 also predicted this protein to have the same structure as other *Cas3* homologues, which was further evidence to show that *S. griseus* has and uses CRISPR-Cas arrays.

Once the evidence had been gathered that supported *S. griseus* ATCC 10137 encodes CRISPR-Cas arrays, the next step of research was to begin modeling bacteriophage proteins. The bacteriophage Wipeout was selected for gene modeling because it was identified and purified at James Madison University, and is in a small cluster of just 14 bacteriophages. This would have been useful if Acr proteins were identified, as there were a limited number of other, related bacteriophage genomes to search through.

To identify potential Acr proteins in the Wipeout phage genome, the annotation of the genome was modeled in Yasara. This generated 155 protein models, 2 were Disgusting, 18 were Terrible, 47 were Bad, 47 were Poor, 25 were Satisfactory, 9 were Good, and 7 were Optimal. These quality score categories are generated by the Yasara software. Of the entire Wipeout genome, 37.43 % of gene products were able to be modeled (Figure 13).

Of the gene products that were able to be modeled, gene products 49, 265, 94, 62, 57, 79, and 93 were of particular interest. Gene product 49 was predicted to be a Bcl-2 protein from Human herpesvirus 8. Wipeout gene product 265 was predicted to be a swarming motility protein. Wipeout gene product 94 was predicted to be a CRISPR-associated exonuclease Cas4 protein. This is not unusual, as other phages have been shown to encode their own CRISPR-Cas arrays to avoid degradation by the host (Seed et al., 2013). However, Wipeout has no confirmed CRISPR-Cas arrays. Future studies should be done to determine how many and which actinobacteriophages contain CRISPR-Cas arrays. Gene product 57 resembles a penicillin G acylase protein, which is one of the most prevalently used biocatalysts utilized for the production of the antibiotic beta-lactam (Srirangan et al., 2013). Wipeout gene product 79 was identified as a potential pertussis toxin-like subunit ArtA protein, and gene product 93 was identified as a potential tetanus toxin-like protein. These proteins are of obvious interest to the medical field, as bacteriophages have been known to transfer genes to and from bacteria through transduction (Wagner & Waldor, 2002).

YASARA was able to predict two gene products that are proteins for the phage to be infectious. Gene product 129 was predicted to be a capsid protein, and gene product 110 was predicted to be a portal protein.

While there were genes that able to be modeled, none of the gene products that were modeled were Acr proteins. However, 39.9 % of the gene products were unable to be modeled. These are proteins with no known structure, and as such are novel proteins. All known phage-produced Acr genes are 150-450 base pairs long (Table 9). Of the 39.9 % of unmodeled Wipeout gene products, 13 fell in this size range. Because of the length

of the gene that codes for them, these gene products are the most likely candidates for potential Wipeout Acr proteins.

These Acr proteins are potentially useful in the bioengineering field. Other Acr proteins have been shown to help negate the deleterious effects that genetic engineering with CRISPR-Cas9 can have. While knowledge of these proteins is growing, there are still unanswered questions as to how their mechanisms work. With the vast resources provided by the SEA-PHAGES consortium, these proteins can be further studied in a classroom setting to provide information to the medical field (Shin et al., 2017).

Literature Cited

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

2. Arndt D, Marcu A, Liang Y, Wishart DS, Wishart D. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes.

3. 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. 2008. The RAST Server: Rapid Annotations using Subsystems Technology.

4. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science (80-) 315:1709–1712.

5. Barrangou R, Marraffini LA. 2014. CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. Mol Cell 54:234–244.

6. Blakley R, Benkovic S, Chemistry P, of Folates John Wiley B, York N, Ryan T, Chave K, Rhee M, Yao R, Yin D. 2003. High yield preparation of genomic DNA from Streptomyces.

7. Bolger AM, Lohse M, Usadel B. 2014. Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data 30:2114–2120.

8. Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. 2012. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature 493:429–432.

9. Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, Wiedenheft B, Maxwell KL, Davidson AR. 2015. Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. Nature 526:136–139.

10. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin E V., van der Oost J. 2008. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science (80-) 321:960–964.

11. Cho SW, Kim S, Kim Y, Kweon J, Kim HS, Bae S, Kim J-S. 2014. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res 24:132–41.

12. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science (80-) 339:819–823.

13. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol 190:1390–400.

14. Dong C, Hao G-F, Hua H-L, Liu S, Labena AA, Chai G, Huang J, Rao N, Guo F-B. 2017. Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins. Nucleic Acids Res 46:393–398.

15. Edgar R, Qimron U. 2010. The Escherichia coli CRISPR System Protects from Lysogenization, Lysogens, and Prophage Induction. J Bacteriol 192:6291–6294.

16. Firth AL, Menon T, Parker GS, Qualls SJ, Lewis BM, Ke E, Dargitz CT, Wright R, Khanna A, Gage FH, Verma IM. 2015. Functional Gene Correction for Cystic Fibrosis in Lung Epithelial Cells Generated from Patient iPSCs. Cell Rep 12:1385–1390.

17. GLENN TC. 2011. Field guide to next-generation DNA sequencers. Mol Ecol Resour 11:759–769.

18. Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35.

19. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. 2009. RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. Cell 139:945–956.

20. Hatfull GF. 2015. Dark Matter of the Biosphere: The Amazing World of Bacteriophage Diversity. J Virol 89:8107–8110.

21. Hsu PD, Lander ES, Zhang F. 2014. Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell 157:1262–1278.

22. Illumina. 2018. Nextera XT DNA Library Prep Kit Reference Guide For Research Use Only. Not for use in diagnostic procedures.

23. Jansen R, Embden JDA van, Gaastra W, Schouls LM. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43:1565–75.

24. Lehman IR, Nussbaumt AL. 1964. The Deoxyribonucleases of Escherichia coli V. ON THE SPECIFICITY OF EXONUCLEASE I (PHOSPHODIESTERASE)*THE JOURNAL OF BIOLOGICAL CHEMISTRY.

25. Liang C, Wainberg MA, Das AT, Berkhout B. 2016. CRISPR/Cas9: a double-edged sword when used to combat HIV infection. Retrovirology 13:37.

26. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J, Koonin E V. 2011. Evolution and classification of the CRISPR–Cas systems. Nat Rev Microbiol 9:467–477.

27. Maxwell KL. 2016. Phages Fight Back: Inactivation of the CRISPR-Cas Bacterial Immune System by Anti-CRISPR Proteins. PLOS Pathog 12:e1005282.

28. Rath D, Amlinger L, Rath A, Lundgren M. 2015. The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie 117:119–128.

29. Rauch BJ, Silvis MR, Hultquist JF, Waters CS, McGregor MJ, Krogan NJ, Bondy-Denomy J. 2017. Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. Cell 168:150–158.e10.

30. Seed KD, Lazinski DW, Calderwood SB, Camilli A. 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494:489–491.

31. Seemann T. 2014. Genome analysis Prokka: rapid prokaryotic genome annotation 30:2068–2069.

32. Shin J, Jiang F, Liu J-J, Bray NL, Rauch BJ, Baik SH, Nogales E, Bondy-Denomy J, Corn JE, Doudna JA. 2017. Disabling Cas9 by an anti-CRISPR DNA mimic. Sci Adv 3:e1701620.

33. Srirangan K, Orr V, Akawi L, Westbrook A, Moo-Young M, Chou CP. 2013. Biotechnological advances on Penicillin G acylase: Pharmaceutical implications, unique expression mechanism and production strategies. Biotechnol Adv 31:1319–1332.

34. Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, Corbett CR. 2018. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci Rep 8:10931.

35. van der Oost J, Westra ER, Jackson RN, Wiedenheft B. 2014. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nat Rev Microbiol 12:479–492.

36. Wagner PL, Waldor MK. 2002. Bacteriophage control of bacterial virulence. Infect Immun 70:3985–93.

37. Wang G, Zhao N, Berkhout B, Das AT. 2016. CRISPR-Cas9 Can Inhibit HIV-1 Replication but NHEJ Repair Facilitates Virus Escape. Mol Ther 24:522–526.

38. Wang G, Zhao N, Berkhout B, Das AT. 2016. CRISPR-Cas9 Can Inhibit HIV-1 Replication but NHEJ Repair Facilitates Virus Escape. Mol Ther 24:522–526.

39. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R. 2013. One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. Cell 153:910–918.

40. Wang Z, Pan Q, Gendron P, Zhu W, Guo F, Cen S, Wainberg MA, Liang C. 2016. CRISPR/Cas9-Derived Mutations Both Inhibit HIV-1 Replication and Accelerate Viral Escape. Cell Rep 15:481–489.

41. Wick RR, Judd LM, Gorrie CL HK. 2017. Unicycler. GitHub.

42. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.

43. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies.

44. Wu Y, Zhou H, Fan X, Zhang Y, Zhang M, Wang Y, Xie Z, Bai M, Yin Q, Liang D, Tang W, Liao J, Zhou C, Liu W, Zhu P, Guo H, Pan H, Wu C, Shi H, Wu L, Tang F, Li J. 2015. Correction of a genetic disease by CRISPR-Cas9-mediated gene editing in mouse spermatogonial stem cells. Cell Res 25:67–79.