

A Pathogenomic Examination of Virulent and Avirulent
Aerococcus viridans var. *homari*

By
Emma S. Garlock

A thesis submitted to the
Department of Chemistry and Biochemistry
Mount Allison University
In partial fulfillment of the requirements for the
Bachelor of Science degree with Honours
April 9th, 2019

Abstract

The American lobster (*Homarus americanus*) is the most successful commercial fishery in Canada. The industry creates 7800 jobs in the Atlantic region, making it a pivotal part of the local economy in rural Atlantic Canada. When lobsters are harvested, they are either sent directly to market, or kept in storage to ensure there is a supply of live lobster available for consumers year-round. The lobster fishery experiences periodic post-harvest loss due to an infection called gaffkemia. Gaffkemia is caused by the bacterium *Aerococcus viridans* var. *homari*. Large live lobster storage facilities have *A. viridans* screening procedures in place to limit post-harvest loss, but they are not sensitive enough to differentiate between naturally occurring virulent and avirulent strains. When a shipment tests positive for the bacterium, this creates a financial burden on the facility as they must treat the shipment and it must be sold at a lower value. This project hopes to characterize genomes of virulent and avirulent strains of *A. viridans* var. *homari* to identify the factors responsible for virulence in this bacterium. This project uses genomic approaches to identify variations in the genomes of virulent and avirulent strains of *A. viridans* var. *homari*. Genomic polymorphisms were compared between the phenotypes and putative virulence factors and impacted metabolic pathways were identified. No pathways had a significantly different number of polymorphisms between the phenotypes. However, the results of this study do indicate that particular regions of the genome are more prone to mutation than others.

Acknowledgements

I would like to sincerely thank my supervisor Dr. K. Fraser Clark for his knowledge, encouragement and patience during this project. Your focus on a positive, supportive and friendly working environment for all of your students made completing this work much easier. I would also like to thank Dr. Jeffery Waller for being my second reader. This project would not be possible without the support of my lab mates, Lizzy, Kristena, Shae, Brian and Gabby. An additional thank you to the members of the Waller and Cockshutt lab who generously shared their lab space with us. Finally, I would like to give my deepest thanks to my wonderful friends for always letting me bounce ideas off them even though they had no idea what I was talking about.

Table of Contents

ABSTRACT	2
ACKNOWLEDGEMENTS.....	3
LIST OF TABLES AND FIGURES.....	6
1. INTRODUCTION.....	7
1.1 LOBSTER INDUSTRY	7
1.2 BACTERIAL PATHOGENESIS.....	7
1.2.1 Common Mechanisms.....	7
1.2.2 Bacterial Immune Evasion.....	9
1.2.3 Virulence Factors	10
1.2.4 Common Medical Bacterial Pathogens.....	10
1.2.5 Common Veterinary Bacterial Pathogens.....	11
1.3 AQUATIC PATHOGENS	11
1.3.1 Vertebrate Hosts	12
1.3.2 Invertebrate Hosts	13
1.4 AEROCOCCUS PATHOGENS.....	15
1.4.1 Virulence Factors Associated with the Aerococcus Genus	16
1.4.2 Aerococcus viridans.....	16
1.4.3 Aerococcus viridans var. homari- Crab and Lobster Infections	17
1.4.5 Prevalence of Infection	19
1.5 EXPERIMENTAL OBJECTIVES	19
2. MATERIALS AND METHODS.....	19
2.1 STRAIN SEQUENCING AND IDENTIFICATION	19
2.2 ALIGNMENT AND VARIANT CALLING.....	20
2.3 PRIMARY VARIANT FILTERING.....	20
2.4 VARIANT ANNOTATION.....	21
2.5 SECONDARY VARIANT FILTERING.....	21
2.6 GROUPING OF VARIANTS.....	21
2.7 VARIANT VISUALIZATION	22
2.9 VIRULENCE FACTOR IDENTIFICATION	22
2.10 PHYLOGENETICS.....	23
2.11 DETERMINATION OF MOST VARIANT GENOME REGION	23
2.12 ALIGNMENT OF VIRULENCE FACTORS ACROSS AEROCOCCUS SPECIES.....	24
2.13 ASSIGNMENT OF VIRULENCE FACTORS TO KEGG PATHWAYS.....	24
3. RESULTS.....	25
3.1 CHARACTERIZATION OF GENOMES.....	25
3.2 ALIGNMENT, VARIANT CALLING AND VARIANT FILTRATION	27
3.3 VIRULENCE FACTOR IDENTIFICATION	28
3.4 PHYLOGENETICS.....	28
3.5 METABOLIC PATHWAYS.....	29
3.6 STATISTICAL ANALYSIS	31
4. DISCUSSION	32
4.1 ALIGNMENT, VARIANT CALLING AND PRIMARY FILTRATION	32
4.2 VARIANT ANNOTATION AND FILTERING	32
4.3 PHYLOGENY	33
4.4 STATISTICAL ANALYSIS	33
4.5 VIRULENCE FACTOR IDENTIFICATION	34
4.5.1 CptV.....	34

4.5.2 <i>Tuf</i>	35
4.5.3 <i>SugABC</i>	35
4.5.4 <i>hsdM2</i>	36
4.6 METABOLIC PATHWAYS.....	36
4.6.1 <i>Purine and Pyrimidine Metabolism</i>	36
4.6.2 <i>Starch and Sucrose Metabolism</i>	37
4.6.3 <i>Quorum Sensing</i>	37
5. CONCLUSIONS AND FUTURE DIRECTIONS	38
6. REFERENCES	39
7. APPENDIX	46

List of Tables and Figures

Table 1: Virulent Genome Characteristics.....	25
Table 2: Avirulent Genome Characteristics.....	26
Table 3: Genes Unique to Avirulent Strains.....	26
Figure 1: Variant Composition at Each Stage of Filtration.....	27
Figure 2: Average Variant Frequency per 100 base pairs.....	28
Figure 3: Phylogenetic Tree of GroEL sequences.....	29
Table 4: Metabolic Pathways.....	30
Figure 4: Compact letter displays of Variant Frequency analysis.....	31

1. Introduction

1.1 Lobster Industry

In 2017, American lobster (*Homarus americanus*) landings in Canada were valued at \$1,461,661,000, the largest commercial landing for any fishery in the country. A large volume of lobster will also come from the USA, be processed, and shipped out, bringing the total export value up to \$2 billion annually [1]. Lobster fisheries are an essential element of many coastal economies. For many of the fishers, lobster landings account for 90% of their total landings. In Nova Scotia alone, 7800 jobs directly rely on this industry via the harvesting and processing sectors [2]. For this industry to be sustainable, lobsters can only be caught after they have reached a minimum carapace length, which roughly equates to a certain weight and one reproductive cycle.

1.2 Bacterial Pathogenesis

The shapes, clustering, motility, and surface molecules of bacterial pathogens allows for a wide range of infections and host immune responses. For a bacterium to be considered pathogenic, it must meet the five characteristics outlined by Smith, 1984 [3]. It must: (1) infect the mucosal surfaces of the respiratory, alimentary or urogenital tracts; (2) enter the host usually by penetration of the mucosal surfaces; (3) multiply in the environment of the host's tissues; (4) resist or interfere with host-defense mechanisms that try to remove or destroy them; (5) cause damage to the tissues of the host. However, if the bacteria are delivered directly to host tissues, only the final three qualifications must be met.

1.2.1 Common Mechanisms

In order for bacteria to cause infection, they must first avoid the immune system of the host. Quorum sensing and biofilm formation are two mechanisms commonly employed by bacterial pathogens to aid infection. Both mechanisms allow for large colonies to form in ways that either combat or subvert the host's immune system. Quorum sensing is a type of cell-cell communication that initiates specific gene expression when signaling molecules reach a threshold concentration. The genus *Staphylococcus* uses quorum sensing to determine the timing for cell surface molecule and virulence factor presentation. This helps the bacteria minimize detection by the host immune system until the colony is large enough to effectively compromise

the host's immune defense. *Staphylococcus aureus* (*S. aureus*) uses quorum sensing to achieve apoptosis of epithelial cells during infection [4].

Biofilm-forming bacteria are often the cause of chronic bacterial infections, the most common is *Staphylococcus epidermis* (*S. epidermis*) [5]. Bacterial biofilms do not always have immediate pathogenic impacts. This is due to both quorum sensing and the fact that biofilms sometimes induce pathogenicity by a triggered release of mobile planktonic cells. Biofilms are difficult to deal with, as a polysaccharide matrix surrounds them. The matrix facilitates the development of separate channels within them for water and nutrient delivery. This creates a closed system of exopolysaccharides that are resistant to antibiotics [6]. Research has shown that biofilms can be composed of many types of bacteria, complicating drug delivery and increasing the rate of horizontal gene transfer [7]. Not all bacteria are resistant to the same antibiotics, so selecting an effective antibiotic for many bacteria can be complicated, while using a cocktail of antibiotics can be a risk as well. Overuse of antibiotics can increase the rates of antibiotic resistance gene development [8].

Antibiotic resistance genes can be inherent or acquired. Efflux pumps in the cell's membrane designed to removed antibiotics from the cell are a form of inherent resistance. Similarly, the bacteria could also have enzymes capable of destroying the antibiotics as they enter the cell. For example, β -lactamases present in bacteria such as *Escherichia coli* will hydrolyze most penicillin-type drugs rendering them inactive. In comparison, acquired resistance is typically seen in the form of target modification. This process takes place when the cellular target for the antibiotic is modified to the point where the antibiotic has a decreased affinity [4, 9].

Bacterial translocation is an infection mechanism often seen in the gut of those who are chronically ill, such as cancer patients [10]. Bacterial translocation is the movement of the bacteria from the lumen to the extraintestinal site. While this is a normal process that should result in an immune response from the mesenteric lymph nodes, chronically ill and immunocompromised patients often lack this ability. Therefore, bacterial translocation serves as a mechanism to spread infection throughout the body via the lymph and then to the mesenteric lymph nodes [11, 12].

1.2.2 Bacterial Immune Evasion

Bacteria such as *S. aureus* have an enhanced ability to multiply in host tissues via immune evasion. This ability is broken down into four stages or mechanisms: avoiding reactive oxygen species (ROS) attacks, then defensin, complement, and neutrophil evasion.

Staphylococcus is the most well-studied bacterium showing immune evasion; however, *Yersinia*, *Bordetella*, and *Borrelia* also exhibit similar functions [13].

In the innate immune system, phagocytic cells produce toxic hydrogen peroxide and free radicals after pathogen ingestion. However, as a catalase-positive bacterium, *S. aureus* can inactivate the ROS and hydrogen peroxide to survive longer. *S. aureus* also expresses an extracellular adherence protein (Eap), which limits leukocyte movement. Inhibition of leukocytes lowers the overall number of neutrophils present by inhibiting the neutrophil's ability to bind and transverse the endothelial surface [14]. There is also an impact on the expression of chemoattractants, as the bacterium can block formylated peptide receptors, a unique and potent identifier of bacterial pathogens. This results in less stimulation and taxis of phagocytic leukocytes. The combination of these factors limits phagocytosis of the invading pathogen and reduces the communication to the adaptive immune system.

Another mechanism of defense is secretion of α -defensins, which disrupt the bacterial cell membrane. However, inactivation of α -defensins can occur by bacterial secretion of staphylokinase (SAK) [15]. Similarly, aureolysin cleaves antimicrobial peptides, rendering them ineffective. Certain bacteria also have the *oatA* gene, which makes the bacterial membranes resistant to lysozyme activity [16].

Staphylococcus aureus can also block the complement pathway, a crucial part of the immune response. All three pathways involve the cleavage of precursor molecules to activate downstream effectors. The classical pathway is the only complement pathway where the initial activation can get blocked with immune evasion strategies. The majority of cleavage steps in the complement pathway have a corresponding stabilizing protein from *S. aureus* that will inhibit any of the principal complement outcomes which are the formation of membrane attack complexes, chemotaxis for phagocyte activation, phagocytic killing, and antigen presentation. SAK also works in conjunction with protein A to inhibit the formation of C1q, C1r, and C1s, which are cleavage products that have the ability to activate downstream immune functions[13].

1.2.3 Virulence Factors

Gene products that allow bacteria to colonize host tissues and cause disease are referred to as virulence factors. Similarly, how pathogenic the bacteria are by producing these factors is referred to as virulence [17]. Virulence factors can be broken down into four functional categories; host cell attachment and invasion, bacterial secretion systems and effectors, toxins, and iron acquisition systems. The overall virulence of a bacterium is multifactorial, as many VFs work together to achieve the necessary five qualifications of pathogenicity: (1) infect the mucosal surfaces of the respiratory, alimentary, or urogenital tracts; (2) enter the host usually by penetration of the mucosal surfaces; (3) multiply in the environment of the host's tissues; (4) resist or interfere with host-defense mechanisms that try to remove or destroy them; (5) cause damage to the tissues of the host.

All virulence factors aim to help the bacteria avoid or combat the host immune system in order to assure survival. Conserved virulence factors for plant, nematode, and mammalian bacterial pathogens include factors for quorum sensing, oligosaccharide production, and multidrug efflux pumps [18]. Conservation of these factors is expected as quorum sensing is a common infection strategy, and oligosaccharides or efflux pumps are characteristic of bacterial membranes. The oligosaccharides help to specify adherence targets and efflux pumps are critical cellular components for antibiotic resistance in biofilms [19, 20]. There is a close relationship between the standard mechanisms of bacterial infections and frequently occurring virulence factors.

1.2.4 Common Medical Bacterial Pathogens

Streptococci, *Staphylococci*, and *Enterococci* spp. continue to be some of the most prevalent bacterial infections in clinical settings [21]. *Staphylococcus* spp. are found frequently in clinical settings as a hospital-acquired infection infecting surgical wounds and other damaged skin. Not only are methicillin-resistant *Staphylococcus* infections (MRSA) increasing in prevalence in clinical settings, but they are a multi-drug resistant version as well. *Streptococci* spp. are most well known as the cause of acute pharyngitis, or strep throat. There are many strains of *Streptococcus* and subsequently, types of infection. *Enterococcus* spp. are known for UTIs, soft-tissue infections, and endocarditis. However, it is noteworthy that *Enterococcus* spp. can also be considered commensal and can be found on the skin and in the digestive tract of all humans.

Many bacterial pathogens, such as *Pseudomonas aeruginosa*, take advantage of suppressed immune systems; this can be a result of immunosuppressive drugs, recent surgery, chronically compromised immune systems or any combination of the three [21]. It is for that reason that hospitals provide an excellent environment for bacterial infections and the spread of antibiotic resistance since there are many immunocompromised people in a comparatively small environment.

1.2.5 Common Veterinary Bacterial Pathogens

Veterinary microbiologists have categorized commonly identified bacterial pathogens into four categories: *Staphylococci*, *Streptococci*, Gram-negative bacilli, and anaerobes. Since studies found the majority of *Staphylococcal* and *Streptococcal* infections in farm and livestock animals, they suspect that many of the *Staphylococci* spp. infections are contact infections from interacting with humans [22]. Veterinary pathogens are capable of forming biofilms and employing quorum sensing, providing the same challenges found in human infections. Significant biofilm producing infections include mastitis, enteritis, and pneumonia [23]. *Staphylococcus* spp., *Streptococcus* spp., and *Pseudomonas* spp. bacteria all employ this mechanism in all of their animal hosts.

The increase in household pets, and the increased use of broad-spectrum antibiotics in the veterinary practice, has led to a rise in antimicrobial-resistant bacteria. This is explicitly seen with *Staphylococcal* and *Streptococcal* canine infections. The overuse of antibiotics in livestock feed could also have this result. Screenings of the bacteria present in pets have shown that they can act as a reservoir for human pathogens, which poses the risk that more human pathogens could develop antibiotic resistance [24].

1.3 Aquatic Pathogens

Common aquatic pathogens are Gram-positive fecal coliforms such as *Enterococcus*, *Staphylococcus*, and *Streptococcus*. These pathogens originate in the digestive tracts of warm-blooded animals and can enter aquatic ecosystems through direct delivery via defecation or sewage runoff. These bacteria are the standard indicators used to test water quality for both fresh and saltwater systems [25]. While Gram-positive bacteria are capable of infecting marine species, Gram-negative bacteria have much higher infection rates due to their characteristic cell

membranes containing lipopolysaccharides (LPS). Some theorize that the LPS allow for better survival in the harsh marine environment, and also allow for better infiltration to potential hosts and induction of varied immune responses [26].

The Gram-negative bacterial pathogens of the genus *Vibrio* are some of the most frequent and damaging types of infections for both marine vertebrates and invertebrates. The *Vibrio* bacteria cause a wide range of infections with varying mortalities and morbidities. Their extracellular properties help to increase strong proteolytic, phospholipase, and hemolytic properties, resulting in tissue degradation in infected animals. There is also evidence that the bacteria are capable of quorum sensing and biofilm formation, which would also increase the pathogenicity [27].

While marine pathogens would use the same pathogenicity mechanisms to infect a host, the rates of infection differ. The migration rates of many aquatic vertebrates create higher incidences of infection. The rates of marine infections are similar to what one would expect from a human disease with an insect vector. Dispersal barriers such as political or geographic borders are also lacking in the ocean; allowing for a much wider spread of pathogens than what one would typically find on land [28].

1.3.1 Vertebrate Hosts

Vertebrate hosts combat pathogens using their innate and adaptive immune systems. In the innate system, circulating phagocytes, granulocytes, and other cells work together to remove pathogens. At this time, cellular signals are being sent out, and adaptive immunity is activated as degraded pathogens are presented to lymphocytes. This system can be highly specific and allows for memory and tolerance upon re-exposure to the same pathogen [29]. This response is activated by pathogen-associated molecular patterns specific to bacteria, such as LPS on cellular membranes, or other bacteria specific gene products [30].

Vibrio spp. are significant pathogens for many marine vertebrates; normally causing dermal lesions and other tissue degradation. Sharks in captivity have shown to be particularly susceptible to *Vibrio* spp. infections with death being caused by an increased blood pH due to urea metabolization by the bacteria [31]. This pathogen impacts multiple species in multiple geographic regions, making it difficult to study and understand since the strains and infection mechanisms change between hosts [27].

Farmed cold water fish are susceptible to a wide range of bacterial pathogens. Different species of *Vibrio* can cause a septicemic infection. *Streptococcus* spp. and *Aeromonas* spp. bacteria cause dermal lesions that vary in severity and location, with infections being found on the tail and inside gills. A large-scale survey off the east coast of the USA found that *Brucella* spp, *Giardia* spp, *Cryptosporidium* spp., and *Leptospira* spp., are the most common bacterial pathogens for marine vertebrates and seabirds. Antibiotic testing revealed that these strains were resistant to the majority of antibiotics frequently used to treat human infections. Further investigation into the *Brucella* strain showed similarity to a *Brucella* isolate found in humans, implying that infection rates increase with human contact. This confirms that marine vertebrates are also vectors for human pathogens [32].

1.3.2 Invertebrate Hosts

The major difference between invertebrate and vertebrate immune systems is the adaptive immune system. There is little evidence supporting that invertebrates are capable of forming memory cells that will be used upon re-exposure to a pathogen, whereas a vertebrate immune response will prompt the creation of memory cells that can be reactivated during subsequent infections of the same pathogen [33]. However, invertebrates do have the same innate abilities as vertebrates; as they can phagocytize, opsonize, and generate antimicrobial peptides [34].

1.3.2.1 Crab

Crabs and other decapods will generally employ bactericidal, agglutinin, and phagocytic methods to combat bacterial pathogens in their hemocyte. Following phagocytosis, studies have shown an overall decrease in hemocyte counts, leading to an inhibition of clotting factors, an essential part of the immune response [35]. After clearing an infection, the agglutinin process will cause cell aggregates to be found in many tissues, principally the gills. In times of stress, like those created during fishing, handling, and storage, the crab's ability to combat pathogens is reduced. This would indicate crabs held in storage for an extended period would have increased chances of infecting others in the population or other consumers. Another study confirmed this by showing that the majority of infected blue crabs in storage contracted the pathogen as a result of the handling and storage process [36].

Crabs have a wide range of bacterial pathogens that vary in physical localization and symptomatic manifestation. Additionally, certain bacterial infections do not appear to be location specific; with isolates obtained from crabs collected from various geographic regions. The most prominent infections concerning researchers, and the aquaculture industry, are shell disease and bacteremia, commonly caused by members of the *Vibrionaceae* family. Since this bacterium can also be pathogenic to humans, there is a concern of increased aquaculture infections increasing *Vibrio* spp. infection rates in humans. This would happen through direct consumption of raw meat or as it gets transferred up the food chain via lobsters or larger commercial fish [37].

There is particular interest in the relationship between crab and lobster infections. Crabs acting as bacterial reservoirs is a concern because they are food source for lobsters. Red crabs (*Geryon quinquedens*) and snow crabs (*Chionoectes opilio*) have been shown to be relatively tolerant to *Aerococcus viridans* (*A. viridans*), a significant lobster pathogen. One theory for the resistance is the presence of *A. viridans* var. *homari* agglutinins which are absent in lobster [38].

1.3.2.2 Lobster

Lobsters are immune to most bacterial infections as the hemolymph of lobsters has been shown to have substantial nonspecific bactericidal activity [39]. There are only three known bacterial infections that can circumvent the innate immune defense system of lobsters, limp lobster disease, shell disease, and gaffkemia [40]. Degradation of the carapace due to chitinovorous *Vibrio* bacteria characterizes shell disease [41]. Even though the infection only impacts the carapace, it is still classified as an opportunistic infection as preexisting lesions predispose the organism to infection [42]. Lobsters in pre-molt exhibit higher instances of shell disease, which again supports the theory that a loss of carapace integrity is needed for shell disease to occur [43]. The immune response resulting from this infection is characterized by inflammation and melanization of the underlying tissues. This immune response usually blocks the bacteria from entering the shell and prevents septicemic infections. This results in the granular and burnt appearance of the carapace associated with shell disease [44]. Limp lobster disease is caused by a *Vibrio fluvialis*-like bacteria. This disease manifests as lethargy and reduced response times, eventually leading to death. When first characterized, it was thought this disease was unique to captive lobsters, but it was later identified in fresh harvests [45].

Gaffkemia is the most concerning septicemic infection impacting lobster fisheries. *Aerococcus viridans* var. *homari*, formerly *Gaffkya homari*, is the cause of this infection.

Aerococcus viridans is exceptionally pathogenic for lobsters, 10 virulent bacteria per kilogram of lobster can be fatal [46]. With the bacteria lacking mechanisms to penetrate the shell, infection is predicated upon lesions or other deformities in the carapace. For this reason, high incidences of infection get reported after fishing, handling, and storage. Infection rates in captivity can increase quickly due to the pathogenicity of the bacteria and also the tendency for lobsters to fight and damage their carapaces [47]. Lobster hemolymph is an ideal growth medium for the bacteria, with bacteria showing increased growth rates when compared to other media. This may be due to lobsters' lack of agglutinin for *A. viridans*. Bacteria specific agglutinins would typically be a part of the animal's humoral defense mechanisms. Cellular defense mechanisms also prove ineffective as the capsule surrounding the bacteria prevents ingestion by phagocytes [48]. Currently, the exact mechanism of infection and associated virulence factors are unknown.

1.4 *Aerococcus* Pathogens

The genus *Aerococcus* was identified in 1953 [49]. Growth in tetrads, the lack of catalase production, and positive Gram test defines the strain as *Aerococcus viridans*. Further studies showed the species *A. urinae*, *A. sanguinicola*, *A. christensenii*, *A. urinaehominis*, and *A. urinaeequi* are often identified from human infections. However, *Aerococcus* are frequently misidentified as *Streptococcus* due to their similar appearance in blood cultures. The workflow followed by the clinician determines whether the correct diagnosis will be made. Standard commercial identification tests include morphological identification, biochemical and physiological traits, and antibiotic susceptibility. Since *Aerococcus* shares many similarities with other bacteria, these commercial tests are sometimes ineffective as differentiating between *Streptococcus* spp. and *Aerococcus*. 16S sequencing is the ideal method for identification of *Aerococcus* species [50].

While the *Aerococcus* genus shares similar antibiotic susceptibilities with the *Enterococci* genus, the different mean inhibitory concentration (MIC) can help to differentiate between species. β -lactam antibiotics affect most species of *Aerococci*. Penicillin has a lower MIC for *A. urinae* and *A. sanguinicola* when compared to *A. viridans*, which is useful since *A. urinae* and *A. sanguinicola* are the two most common causative agents of urinary tract infections (UTIs). Other antibiotics such as tetracycline, erythromycin, and gentamicin have been shown to be effective on *Aerococcus* *in vitro*, but the impact are unknown *in vivo* [50].

The challenge associated with identification of *Aerococci* allows for the possibility that there is a misrepresentation of the prevalence of human *Aerococcal* infections. When *Aerococcus* were known to be present, they are associated with UTIs, sepsis, and infective endocarditis. The diagnosis rate is especially high in older males with other urological conditions [51]. This same trend follows with *A. viridans* infections in humans, with patients having been previously ill or immunocompromised. In addition to UTIs, *A. viridans* has been isolated from patients suffering from either septic arthritis or sepsis, and in the cerebrospinal fluid of those with meningitis [52].

1.4.1 Virulence Factors Associated with the *Aerococcus* Genus

Carkaci et al., identified virulence factors in *Aerococcus sanguinicola* and *Aerococcus urinae* using whole genome sequencing. This study showed that *Aerococcus* virulence factors fit into three main categories: anti-phagocytosis, adherence, and biofilm formation. The majority of antiphagocytic genes are related to the polysaccharide capsule surrounding the bacteria. For adhesion, fibronectin binding proteins and more generic surface proteins were identified. The only gene identified for biofilm formation was *bopD*, a sugar binding transcriptional regulator [53].

When avirulent strains of *A. viridans* var. *homari* were identified, Clark and Greenwood identified differentially expressed proteins between the virulent and avirulent strains. Of the proteins expressed exclusively in the virulent strain, the only one identified was Cpn60. Cpn60 is a chaperonin, assisting in proper protein folding, however, studies have shown it also functions as a cell surface protein, enabling entrance into host phagocytes as a method to avoid the host immune system [54]. This avoidance mechanism is common in many zoonotic bacteria and has been shown to be essential in many different bacterial infections.

1.4.2 *Aerococcus viridans*

Few studies have been done to characterize *A. viridans* isolated from other mammals. The 58 porcine *A. viridans* isolates from the joints of arthritic pigs (30), brains of pigs with meningitis (14) and lungs of pigs with pneumonia (14) help to characterize the pathogenicity of the bacteria. Identification with pulse field gel electrophoresis (PFGE) revealed many strains within the pig herds. This supports theories suggesting the bacteria is part of the commensal gut flora but also widespread in the environment, as this would provide an opportunity for separate

strains to develop. Bacteria other than *A. viridans* cause the primary infections noted in the pigs. All pigs used for isolation had a previous infection of porcine respiratory and reproductive syndrome virus (PRRSV), the subsequent immunosuppression would also contribute to the pig's vulnerability to an opportunistic pathogen [55]. The presence of *A. viridans* in those samples supports the theory of its opportunistic nature, with the bacteria taking advantage of previously damaged tissues or an immunosuppressed host.

Opportunistic infections are also characteristic of *A. viridans* infections found in humans. A 28-day-old patient tested positive for a UTI caused by *A. viridans*. However, this was a secondary UTI with the primary UTI being caused by *Enterococcus faecalis*. The patient also presented with many symptoms causing failure to thrive which compromised the immune system. This was the first case of an *A. viridans* infection in an infant. However, little progress was made in clarifying the best practice for treatment as a mix of antibiotics was administered and the infection was cleared within three days without identifying the most effective antibiotic [56].

However, *A. viridans* has also been shown to cause UTIs without a known previous bacterium causing a weakened immune system. The bacteria were isolated from a pregnant woman, and an antibiotic susceptibility test was performed. In this case, the isolate was susceptible to vancomycin, cefoperazone-sulbactam, imipenem, ampicillin, and intermediately resistant to amoxicillin-clavulanate, and resistant to cefotaxime, cefazolin, cefuroxime, ciprofloxacin, gentamicin. However, this study remarked that previous isolates of *A. viridans* found in differently infected tissues had displayed varying susceptibilities to different antibiotics.

1.4.3 *Aerococcus viridans* var. *homari*- Crab and Lobster Infections

When identified in 1947, *A. viridans* var. *homari* was initially called *Gaffkya homari*, since it was recognized as the causative agent of gaffkemia. Rabin (1965), Cornick and Stewart defined gaffkemia as a non-localized septicemic infection as no enzymes were found that would contribute to the degradation of host tissue [57, 58]. As a Gram-positive bacterium, it is unusual that this would be causing a bacterial infection in marine invertebrates, as most marine bacteria are Gram-negative. The severity of *A. viridans* infections have been shown to vary based on the integrity of the carapace and efficiency of the animal's immune system. Furthermore, *in vitro* studies have shown bacterial growth levels to plateau once the carbohydrates in the hemolymph have been depleted.

A comparative study completed by Deibel and Niven (1960) looked at *G. homari* and *A. viridans* isolated from meat curing brines. Based on the results of growth rate, hydrolysis and fermentation tests, Deibel and Niven concluded the two were similar enough to comprise their own species in the genus of *Pediococcus* [59]. With the work of Williams et al., (1953) on defining the genus *Aerococcus* and further research in *G. homari/P. homari* resulted in the consensus that what was initially identified as *G. homari* was *A. viridans* var. *homari* [49]. Identification of avirulent *A. viridans* strains required further studies to determine if the avirulent strain differed significantly enough to require a new species. The serological analysis concluded that the two were not different and should remain the same species [60].

When comparing virulent and avirulent strains, a significant difference is the alcianophilic acidic polysaccharide capsule that is absent in the avirulent strain [61]. This trait acts as a defense against host immune response, as the capsule prevents the bacteria from being agglutinated and rendered ineffective. Stewart et al., (2004) found that growth in heated serum could attenuate virulence, showing that the principal virulence factors are heat sensitive [48]. Clark and Greenwood determined that Cpn60, a surface protein homologous to GroEL, was the only upregulated gene that was identifiable in the virulent strain during their comparative study. They suggest that this protein attracts and mediates entry into phagocytic cells to proliferate and avoid the innate immune system [62].

The exact source of infection has not been established but Red Crab (*Geryon quinquedens*) and Snow Crab (*Chionoecetes opilio*), two food sources for lobster, have been shown to be carriers of gaffkemia. Cornick and Stewart showed that the bacteria could survive up to 100 days in the crabs without causing infection. Additionally, the time spent in the crabs does not alter the virulence, and the bacteria were still pathogenetic when re-isolated after incubation in the crabs [38].

Gallager et al. found 37% of tests were presumptive positive, and 1.5% were confirmed positive for *A. viridans* in *Cancer irroratus* and *Cancer borealis* [63]. This is similar to Stewart et al., who found 24% were presumptive positive and 4.7% were confirmed positive for *Homarus americanus* using bacterial property characterization looking specifically at gram staining, beta-hemolytic activities and colony formation [64]. The reported prevalence of *A. viridans* in crabs supports the theory that crabs are acting as a carrier.

1.4.5 Prevalence of Infection

In 1966, Stewart et al. found an overall incidence of *A. viridans* infections of 4.7% in lobsters fresh caught from the Atlantic coast of Nova Scotia [64]. In 2001, Lavallee et al. found *A. viridans* infection rates from 5.8-6.9% when investigating the prevalence of freshly caught lobsters before entering containment in the spring, summer, and fall [65]. For lobsters dying in storage, *A. viridans* accounted for 10%-75% of deaths [47]. However, the total mortality for the stock was determined by the park owners and not the research team.

1.5 Experimental Objectives

The aims of this project are to:

1. Identify influential nucleotide variations in the genomes of 10 *Aerococcus viridans* var. *homari* strains
2. Use the nucleotide variations to identify potential proteins and metabolic pathways that could account for the different pathogenic phenotypes found in *Aerococcus viridans* var. *homari*.

It is expected that a causative agent can be identified in the genome that will explain the differing pathogenicities. This study looks to use a combination of bioinformatic techniques to identify virulence factors by investigating the genome difference between virulent and avirulent strains. Ultimately the results will allow for more sophisticated screening for *A. viridans* in the lobster fishery to increase harvest yield and prevent post-harvest economic losses.

2. Materials and Methods

2.1 Strain Sequencing and Identification

All bacterial cultures, DNA extraction, and sequencing preparation were performed by K.F. Clark. Illumina HiSeq 2000 PE100 reads with coverage of 3.5-4.1Gb for each of the ten strains were obtained from Genome Quebec. Based on the results of Greenwood *et al.*, strains AVC, 10400, 88R, 37R and 700406 were placed in the avirulent group with 1030, 1032, 29838, Rabin's and Nfld in the virulent group [66].

2.2 Alignment and Variant Calling

The reference genome assembled as contigs used was *Aerococcus viridans* ATCC 11563 (ASM17843v1) from NCBI. All alignment and variant calling took place in Galaxy Europe [67]. FastQC tests determined read quality for Illumina HiSeq outputs [68]. Trim Galore! v0.4.30 trimmed each raw Illumina HiSeq output to improve quality and subsequent alignment, a second FastQC verified the quality improvement. Each strain had two fastqsanger read files resulting from Trim Galore!, a forward and reverse. Each alignment tool made one .bam file from the paired forward and reverse strands. HISAT2 v2.1.0 required "Paired-End" options, with strand information set to "Forward (F)" and all other settings as defaults [69]. Bowtie2 v2.3.4.2 only required one change to "set paired-end" with all others as defaults [70]. Map with BWA v0.7.17.4 also used paired input settings [71]. The Stats v2.0.1 ran on three sets of alignments in order to determine the optimal alignment program [72]. The properly paired reads (PPR) and error rate (ER) outputted by the Stats v2.0.1 determined the best alignment files to be used going forward. BWA files showed the highest PPR with lowest ER. Additionally, BWA with FreeBayes as the variant caller improves SNP calling [73].

The PPR and ER of all alignments were also compared between strains. An additional fully assembled *Aerococcus viridans* genome (ASM 154328v) was aligned to the contig genome in order to confirm the proper order of the contigs[53]. The fully assembled genome could not be used as the reference for this study as downstream bioinformatic programs required the genome to be assembled in contigs.

The variant caller FreeBayes v1.1.0.46-0 was run on the BWA alignments for the *Aerococcus viridans* var. *homari* strains. With the full list of options, under "Population Model options", changing the ploidy to 1 removed the diploid defaults [74].

2.3 Primary Variant Filtering

Variant filtration required VCFfilter v.0.03[75]. The command "-f "QUAL >50"" removed calls with quality scores less than 50. The commands "-f "MQM>35"" and "-f MQMR>35" removed calls with mapping quality scores less than 35 for the alternate and reference alleles respectively. After the final filtration, VCFtools_Merge v0.1.1 combined all ten VCF files.

2.4 Variant Annotation

The command line `snpEff v.4.3t` program allowed for functional annotation [76]. The program already contained an `Aerococcus_viridans_atcc_11563_ccug_1143` database. Adding the verbose, `"-v"`, argument to the command line of a standard `snpEff` input showed what the program would recognize as contig names along with their corresponding lengths. Note that the annotation program failed since the contig names generated by `FreeBayes` were not recognized by `SnpEff`. However, each assigned contig from `FreeBayes` has the full name and length specified in the info section of the metadata.

A Microsoft Excel file containing the lengths and names of the contigs from each source was made. Using a `VLOOKUP` function, the contig names were matched based on their lengths. When properly matched, the file was saved after removing the contig lengths. A `VLOOKUP` function with the `FreeBayes` contig names as the lookup value and the file with matched up contig names as the table array replaced the `FreeBayes` generated contig names with `SnpEff` accepted ones.

To output an additional statistics file, the standard `snpEff.jar` command used a `"-stats"` argument along with an `.html` file name. The arguments `"-canon"` and `"-strict"` ensured assignment of only canonical and validated transcripts in the annotation. The `SnpEff` command annotated all ten individual strains as well as the combined VCF file. Running the `SnpEff` program generated an annotated VCF, a `gene.txt` file and a `stats.html` file.

2.5 Secondary Variant Filtering

The command line program, `SnpSift` allowed for VCF filtering based on the `IMPACT` field of the `SnpEff` annotation [77]. To filter, a standard `snpSift.jar` filter command used the argument `"ANN[0]. IMPACT has 'impact of interest'"`. Filtering made four new VCF files for high, moderate, modifier and low impact changes. `SnpSift` split with the `"-j"` argument combined the VCF files for high, moderate and modified impact.

2.6 Grouping of Variants

Looking at the grouping of variants used the VCF file generated from the `snpSift` split command, now known as the working VCF, the program did remove the sample names for each

strain. However, the names were replaced manually by looking at each FreeBayes to see where the first variant call happened and matching it with the column in the working VCF.

A COUNTIF function in Excel determined how many strains in each group were not expressing the variant at a certain position. When the strain did not express a variant, the cell contained “.”, which was the criteria for COUNTIF. Using a COUNTIF function to count strains expressing a variant was not possible due to the variability of the format IDs in the cells of interest. The COUNTIF results allowed for further sorting.

The determined threshold for a variant being more present in avirulent vs virulent was three of five expressing when less than three expressed in the other group. If variants were expressed in more than eight of ten strains, or present in less than three of 10 strains, they were removed. Calls with equal variance, which were 4:4, 3:3 and 2:2 were also removed. This left only variants that met the threshold. For this project, the threshold was determined as presence in more than 3 strains of one phenotype, and less than 3 strains of the other.

2.7 Variant Visualization

Visualization of variant tracks required Genome Savant [78]. All tracks required tabix indexing before visualization. The command line tabix package in with Samtools-1.8 provided the code to index the ten individual VCF files by following the standard protocol [79]. The reference genome in the Genome Savant browser was the same FASTA file used for alignment and variant calling.

2.9 Virulence Factor Identification

The snpEff gene output file uses Regulatory Sequence Analysis Tools (RAST) transcript IDs to identify the genes present. The *A. viridans* RAST database was downloaded as a tabular file with the amino acid sequences and associated IDs. All the transcript IDs were copied from the gene file and put in a new Excel document. In the new file, a VLOOKUP using the transcript ID as the lookup value and the RAST database tabular file as the table array created a file with all the amino acid sequences for the proteins present. Biological annotation required a local BLAST using the Virulence Factor Database (VFDB) with a tabular to FASTA converted file [17].

The genes.txt files were combined into virulent and avirulent groups using R. v 3.5.1. They were then filtered to have only the unique genes for each group. Following this, the virulent and avirulent groups of unique genes were compared to find any that existed within only one phenotype.

2.10 Phylogenetics

All genomes were aligned to the reference *Aerococcus_viridans_atcc_11563_ccug_4311* using BWA to create 10 BAM files that were converted to SAM files in order to be opened in Excel. The location of GroEL was determined based on the corresponding .GFF file. Using the data from the SAM file, the portion of each respective query genome corresponding to GroEL was extracted. All ten GroEL sequences and the references sequence were aligned with ClustalW in Mega X. Once aligned, a TN93+G maximum likelihood tree with bootstrap values of 500 was created for the nucleotide sequences.

To attempt a whole genome phylogenetic analysis, all trimmed reads were run through Unicycler to get a fully assembled genome. To finish the genome and identify sequence overlaps, the Unicycler file was uploaded into MegaX and ClustalW alignment was attempted. The computer available did not have enough power to compete the tasks. Similarly, an alignment of the largest regions from each partially assembled genome failed due to lack of computing power.

2.11 Determination of most Variant Genome Region

A Friedman test, the non-parametric equivalent of a 2-way ANOVA was done to determine whether or not there was a significant difference in the SNP pattern across the genomes, and whether or not that pattern differed between virulent and avirulent strains. The genome was divided into seven approximately balanced groups based on the total SNP sites present in that region. The test was run using the `friedman()` function from the `agricolae` package available in R v3.5.1 using strain as the blocking variable [80]. Fisher LSD post-hoc comparisons are done automatically for significant ($p < 0.05$) p-value, however a Bonferroni correction was applied manually to account for multiple comparisons.

Based on the results of the Friedman's test, a parametric plot model was run to determine whether or not there was justification to incorporate the pathogenicity factor into the section of

the genome. A second Friedman test was run on the data with added pathogenicity factor. The same procedure was followed for a significant ($p < 0.05$) p-value.

2.12 Alignment of Virulence factors across *Aerococcus* species

A sample of available protein sequences from different *Aerococcus* spp. were collected from NCBI and aligned with NCBI COBALT. Species collected were *A. sanguinicola*, *A. urinae*, *A. urinaehominis*, *A. christensenii*, *A. suis*, *A. urinaeequi*, and *A. viridans* var. *homo sapiens*.

An R code was written to identify the genes with the largest number of high impact SNPs (supplemental materials). Combined with the SnpEff output, the locations of these SNPs within the genes were identified in the alignments. This allowed us to see whether or not the SNPs occurred in a region that was highly variable across the *Aerococcus* spp.

2.13 Assignment of Virulence factors to KEGG pathways

To identify potential metabolic pathways that may be impacted by the SNPs present in the genome, a KEGG with KAAS SBH assignment method search was done with the same set of sequences used for the BLAST search. The genes dataset used can be found in the supplemental materials. For each pathway, the number of times a gene had a SNP were totaled for each of the ten strains. This was formatted as a matrix and an EdgeR was run using a separate file to specify the virulent and avirulent strains [81]. EdgeR was run without gene annotations, using a virulent-avirulent contrast with a minimum count of 5, and TMM normalization with robust settings, all other settings were default for edgeR Galaxy version 3.24.1. A separate file was used to identify which of the strains were virulent or avirulent. Differential expression was determined by ($p < 0.05$) for adjusted p-values determined by the Benjamini and Hochberg (1995) method.

3. Results

3.1 Characterization of Genomes

The genes.txt files generated from SnpEff gave a list of all the genes present. As seen in Tables 1 and 2, there is a higher average number of genes in the avirulent strains than in the virulent. Many of the genes identified were uncharacterized, but all characterized genes unique to the avirulent strains can be seen in Table 3. Conversely, there was one gene, *hsdM2*, that was found in virulent strains 1032, 29838, and Rabin's that was not present in any avirulent strains.

Table 1. Virulent Genome Characteristics. Genome characteristics for the 5 virulent strains of *Aerococcus viridans* var. *homari*. Genome total length, genome effective length, variant rate, genes and number of effects determined by SnpEff. GC% determined by FASTQC.

	Nfld	Rabin's	29838	1032	1030
Genome total length (bp)	2,006,060	2,006,060	2,006,060	2,006,060	2,006,060
Genome effective length (bp)	1,855,735	1,862,892	1,854,572	1,855,598	1,831,270
Variant rate /100 bp	0.309	0.389	0.334	0.379	0.33
Genes	1780	1798	1780	1788	1781
Number of Effects	53,234	66,512	57,697	65,267	55,799
GC%	40	39	39	39	39

Table 2. Avirulent Genome Characteristics. Genome characteristics for the 5 avirulent strains of *Aerococcus viridans* var. *homari*. Genome total length, genome effective length, variant rate, genes and number of effects determined by SnpEff. GC% determined by FASTQC in Galaxy Europe.

	10400	700406	AVC	37R	88R
Genome total length (bp)	2,006,060	2,006,060	2,006,060	2,006,060	2,006,060
Genome effective length (bp)	1,843,122	1,851,331	1,855,055	1,874,256	1,877,399
Variant rate /100 bp	0.391	0.342	0.323	0.251	0.418
Genes	1780	1818	1783	1829	1833
Number of Effects	66,024	58,759	56,242	43,309	71,894
GC%	39	40	39	39	40

Table 3. Unique Genes to Avirulent Strains. Using the output from SnpEff, unique genes for avirulent strains were identified. Since no unique genes were found in 10400 or AVC, they were omitted from the table. The only genes included in this table are the characterized genes, all others were omitted.

	700406	88R	37R
<i>actP1</i>	Yes	Yes	Yes
<i>agcS</i>	Yes	Yes	No
<i>copB</i>	Yes	Yes	Yes
<i>copZ</i>	Yes	Yes	Yes
<i>copZ2</i>	Yes	Yes	Yes
<i>gloA</i>	Yes	No	Yes
<i>kdgT</i>	Yes	Yes	Yes
<i>mphB</i>	Yes	No	Yes
<i>rluC</i>	Yes	No	Yes
<i>pseI</i>	No	Yes	No

3.2 Alignment, Variant Calling and Variant Filtration

The statistics output show BWA files having the highest PPR with lowest ER. Running the stats tool on all BWA alignments verified that there was no significant difference ($p < 0.05$) between the number of unmapped reads in the virulent versus avirulent strains. This confirms that there are no large genomic regions present in one strain that are not present in another. After filtering the vcf files by phred scaled quality scores and mapping quality scores, 9.66-13.48% of the total SNPs remained. Following the SnpEff program, filtering out all low impact SNPs only removed about 7% of the total variants for each strain, which would include synonymous but not silent mutations, which made up 80% of variants.

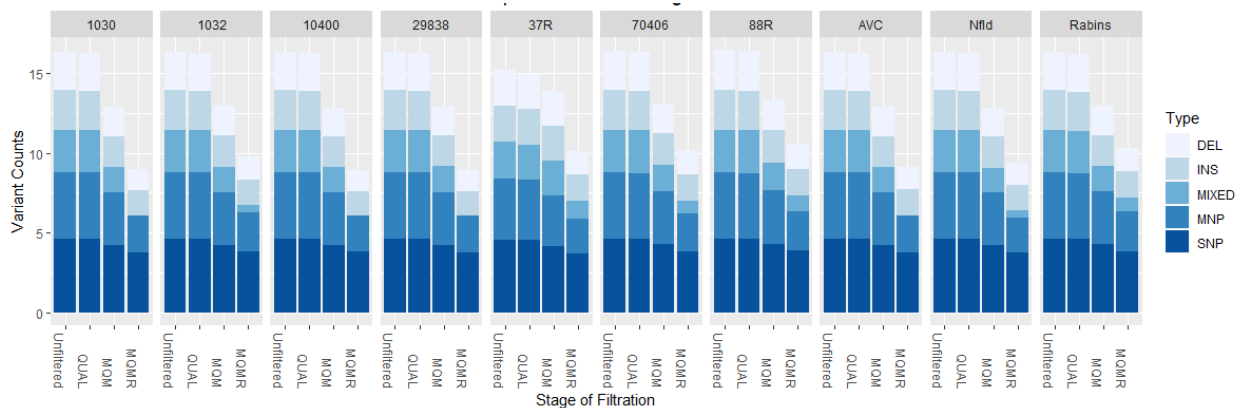


Figure 1. Variant Composition at Each Stage of Filtration. Impact of variant filtration on the total variant calls and variant composition. Variant counts shown on log scale. Total variants and variant types determined by SnpEff. Plot created using ggplot2 packages in Rv3.5.1

The SnpEff output was also able to calculate the variant frequency per 100 base pairs for each contig. The proper order of the reference contigs was determined based on the alignment to another fully assembled *A. viridans* var. *Homo sapiens* genome. To view the overall pattern which can be seen in Figure 2, the average frequency at each position was taken for the virulent and avirulent groups. However, no confidence intervals are present for ease of visualization.

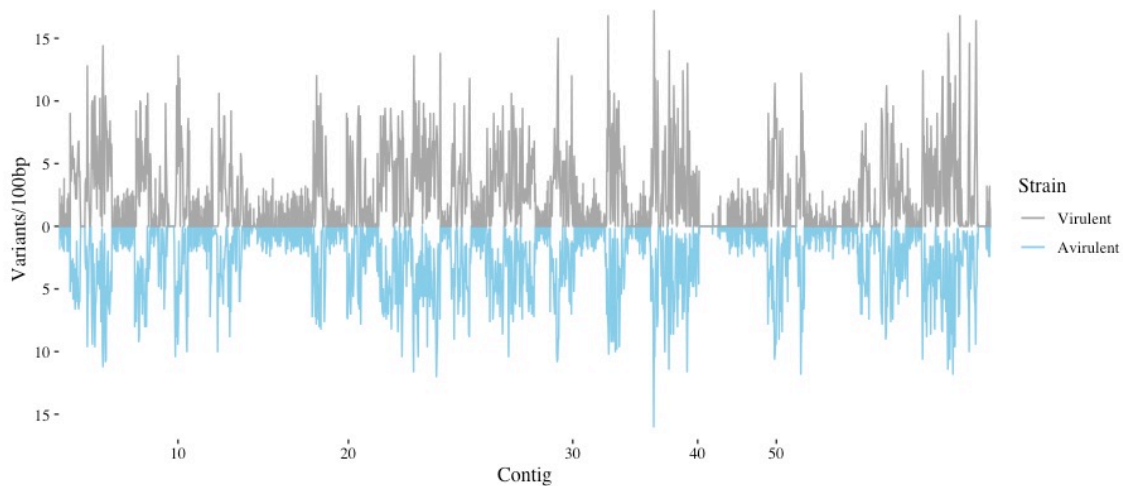


Figure 2. Average Variant Frequency per 100 Base Pairs. Variant frequency determined by SnpEff stats output with *Aerococcus viridans* atcc_11563_ccug_4311 as reference genome. Virulent strains depicted in grey, avirulent in blue. Plot constructed using ggplot2 in R v3.5.1.

The pattern of variants appears consistent between the virulent and avirulent strains. However, the magnitude of the frequencies is larger for the virulent strains.

3.3 Virulence Factor Identification

The local BLAST using the virulence factor database highlighted 131 proteins as potential factors present in the *A. viridans* var. *homari* genome. Frequently occurring ontologies include: cation transport, nucleotide binding, ATP binding, ATPase activity, plasma membrane and transmembrane transport. Looking at the recurrence of SNPs at different sites within those genes, which were annotated from SnpEff, the putative metal cation transporter P-type ATPase (*ctpV*), Elongation Factor Tu (*tuf*), and the sugar ABC transport system (SugABC) proteins were highlighted as the probable virulence factors; in addition to the *hsdM2* gene identified earlier. *CptV* contained 12 SNPs, *tuf* contained 21 SNPs and the sugar ABC transport system had 24 SNPs.

3.4 Phylogenetics

Figure 3 shows the phylogenetic relationship of the 11 different GroEL sequences investigated in this study. This shows the majority of the GroEL sequences in the avirulent

strains are more closely related to the sequences found in the human strains than the virulent strains are to the same reference strain. The virulent strains also have identical GroEL sequences.

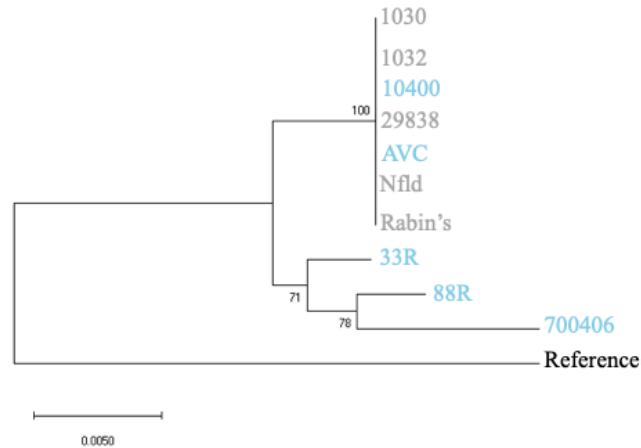


Figure 3. Phylogenetic Tree of GroEL Sequences. A TN93+G maximum likelihood tree with bootstrap values of 500 was created for the nucleotide sequences created using MegaX. The reference strain used was *Aerococcus viridans*_atcc_11563_ccug_4311. Avirulent strains are shown in blue text while virulent strains are shown in grey.

3.5 Metabolic Pathways

Using the genes identified via SnpEff, sequences were gathered and grouped into metabolic pathways using KEGG. Since all genes were found using SnpEff, all of the genes identified in the pathways contain genomic polymorphisms. A summary of impacted pathways can be seen in Table 4.

Table 4. Metabolic Pathways. KEGG generated list of metabolic pathways with the number of proteins impacted by SNPs. Inputted proteins were determined based on a BLAST hit from the virulence factor database.

Pathway	Genes	Pathway	Genes
ABC transporters	56	Nicotinate and nicotinamide metabolism	6
Ribosome	53	Sulfur metabolism	6
Purine metabolism	48	Vancomycin resistance	6
Pyrimidine metabolism	38	Glucagon signaling pathway	6
Quorum sensing	29	Central carbon metabolism in cancer	6
Glycolysis / Gluconeogenesis	25	Tyrosine metabolism	5
Starch and sucrose metabolism	25	Benzoate degradation	5
Amino sugar and nucleotide sugar metabolism	24	Tryptophan metabolism	5
Aminoacyl-tRNA biosynthesis	24	Selenocompound metabolism	5
Pyruvate metabolism	23	Antifolate resistance	5
Two-component system	20	RNA polymerase	5
Peptidoglycan biosynthesis	19	Longevity regulating pathway - worm	5
Cysteine and methionine metabolism	18	Monobactam biosynthesis	4
Phenylalanine, tyrosine and tryptophan biosynthesis	18	Valine, leucine and isoleucine biosynthesis	4
Homologous recombination	18	Phenylalanine metabolism	4
Phosphotransferase system (PTS)	17	Vitamin B6 metabolism	4
Pentose phosphate pathway	16	Biotin metabolism	4
Fructose and mannose metabolism	15	Nitrogen metabolism	4
Alanine, aspartate and glutamate metabolism	15	HIF-1 signaling pathway	4
Glycine, serine and threonine metabolism	15	Peroxisome	4
Propanoate metabolism	14	Biofilm formation - Vibrio cholera	4
Methane metabolism	14	Pentose and glucuronate interconversion	3
Mismatch repair	14	Ascorbate and aldarate metabolism	3
Galactose metabolism	13	Lysine degradation	3
Carbon fixation pathways in prokaryotes	13	Novobiocin biosynthesis	3
RNA degradation	13	Taurine and hypotaurine metabolism	3
Fatty acid biosynthesis	12	D-Glutamine and D-glutamate metabolism	3
Lysine biosynthesis	12	Streptomycin biosynthesis	3
Glycerolipid metabolism	12	Chloroalkane and chloroalkene degradation	3
DNA replication	12	Aminobenzoate degradation	3
Oxidative phosphorylation	11	Styrene degradation	3
One carbon pool by folate	11	MAPK signaling pathway - plant	3
Cell cycle - Caulobacter	11	Sulfur relay system	3
Arginine biosynthesis	10	Longevity regulating pathway - multiple species	3
Glycerophospholipid metabolism	10	Tuberculosis	3
Butanoate metabolism	10	Synthesis and degradation of ketone bodies	2
Thiamine metabolism	10	Ubiquinone and other terpenoid-quinone biosynthesis	2
Terpenoid backbone biosynthesis	10	Carbapenem biosynthesis	2
Carbon fixation in photosynthetic organism	9	Prodigiosin biosynthesis	2
Folate biosynthesis	9	Cyanoamino acid metabolism	2
β -Lactam resistance	9	D-Alanine metabolism	2
Protein export	9	Inositol phosphate metabolism	2
Citrate cycle (TCA cycle)	8	Naphthalene degradation	2
Fatty acid degradation	8	C5-Branched dibasic acid metabolism	2
Histidine metabolism	8	Fluid shear stress and atherosclerosis	2
Glyoxylate and dicarboxylate metabolism	8	Tropane, piperidine and pyridine alkaloid biosynthesis	2
Pantothenate and CoA biosynthesis	8	Cationic antimicrobial peptide (CAMP) resistance	2
Base excision repair	8	Biofilm formation – Pseudomonas aeruginosa	2
Arginine and proline metabolism	7	RNA transport	2
Drug metabolism – other enzyme	7	FoxO signaling pathway	2
Biofilm formation – Escherichia coli	7	Phosphatidylinositol signaling system	2
Bacterial secretion system	7	Longevity regulating pathway	2
Nucleotide excision repair	7	Necroptosis	2
Valine, leucine and isoleucine degradation	6	Thyroid hormone synthesis	2
Glutathione metabolism	6	Insulin resistance	2
Riboflavin metabolism	6	Legionellosis	2

3.6 Statistical Analysis

When looking at different regions of the genome, it was found there was a significant difference in variant frequency per 100 base pairs between regions of the genome, but not when comparing the same regions on different phenotypes. This can be seen in Figure 4, using a compact letter display. To interpret, if there are any letters in common between the two groups, then they cannot be considered significantly different. However, it is important to note that the lack of significance does not increase with the number of letters two groups have in common.

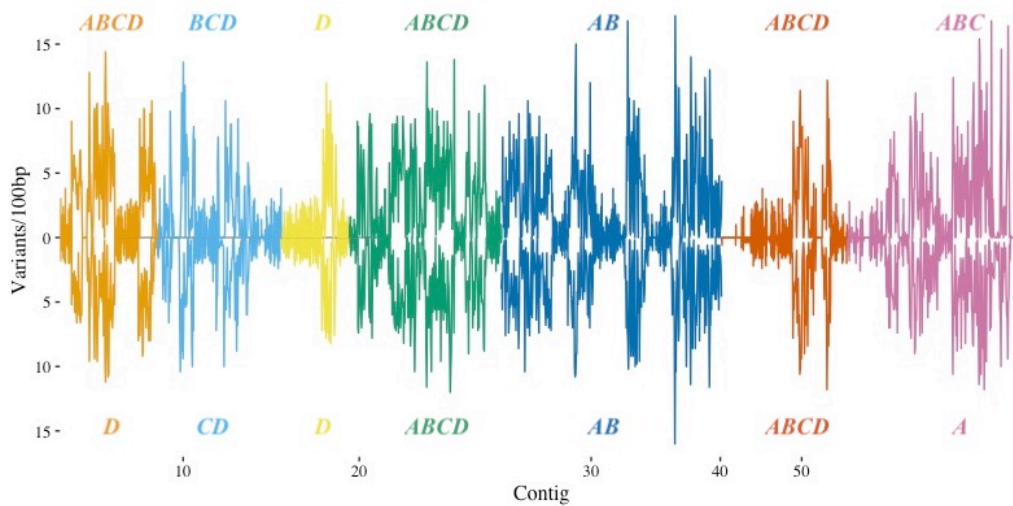


Figure 4. Results of Statistical Analysis. Variant frequencies for both virulent and avirulent phenotypes coloured based on level for analysis. Level A, B, C, D, E, F and G, are shown by grey, yellow, light blue, green, dark blue, red and pink respectively. The positive Y axis values are the averages for the virulent strain while the values on the negative side of the axis are for avirulent. Plot constructed using ggplot2 in R v3.5.1.

4. Discussion

4.1 Alignment, Variant Calling and Primary Filtration

Evaluating the results of the BAM stats (full table available in the appendix) make it clear there is no significant difference between the rates of properly paired reads or error during the mapping. This shows that there are no larger sections of genome that are present in one phenotype that are not present in the other, therefore virulence cannot be attributed to a large, detectable genome insertion. Another alignment was performed aligning a different *A. viridans* strain (ASM 154328v) to the reference strain used in this procedure. This also did not reveal a significant difference ($p < 0.05$). Therefore, the differing pathogenicity for this bacterium is not caused by a large, detectable genome insertion or deletion.

We were unable to extract protein sequences from the query genomes to align and compare to protein sequences of other *Aerococcus* species to see whether variants existed in regions of high variability. This technique could also determine whether or not the variant regions were species or genus specific. While all genomes were aligned and mapped, it is much more technically complicated to extract single protein sequences and as such this was determined to be beyond the scope of this project.

4.2 Variant Annotation and Filtering

Due to the presence of multiple reading frames, it is difficult to predict the impact of one specific polymorphism on virulence when there are many possible encoded proteins. The same polymorphism may have varying impacts depending on the protein it encodes for a certain reading frame. It is not within the scope of this project to predict the proteins most likely expressed in the disease states. SnpEff has no functionality to determine which possible reading frame is the most likely to be expressed. There are many combinations of expressed genes when there are polycistronic genes across an entire genome. It is possible that there are particular combinations that get expressed in certain strains that cause them to be virulent. It is for this reason that the genes.txt file was used to identify the sequences for BLAST and not the first entry in each SnpEff row.

A FASTA file containing all available amino acid sequences for *A. viridans* proteins containing variants was searched against a the VFDB (downloaded July 7th, 2018)[17]. The *A.*

viridans proteins searched all met the qualification of having more variants present in one phenotype than the other. All the factors available in this database are predominantly virulence factors of importance to humans, therefore it is possible that the most prominent virulence factor is not included in this list if it is exclusive to crustaceans.

4.3 Phylogeny

The goal of this project was to identify a set of “smoking gun” polymorphisms or variants that confer virulence to *A. viridans* var. *homari*. While this was not identified, several trends were revealed. The pattern of variants between the 10 genomes aligned with what would be expected based on the 16S phylogeny determined by Greenwood et al., [66]. This is also shown in the phylogenetic tree constructed for the GroEL sequences which can be seen in Figure 3. GroEL was of interest as it was identified as a differentially expressed protein in virulent versus avirulent strains of *Aerococcus viridans* var. *homari* by Clark and Greenwood [54]. The characterization of virulent versus avirulent may be too broad of a comparison to make. While all strains may still be considered the same species, it is possible that they differ enough that when comparing variants, one cannot be sure what constitutes a difference in virulence or a difference in phylogeny.

4.4 Statistical Analysis

From the initial Friedman test not including the pathogenicity and the ANOVA, both agree that there are significant differences in the variant frequencies between different regions of the genome. This demonstrates that the polymorphisms shown in Figure 2 are not due to a consistent rate of random mutation throughout the genome, but that there are some areas more prone to variation than others. However, it is important to note that this analysis only shows where differences in frequency lie, not the magnitudes of those differences. Therefore, it does not statistically determine which sections of the genome have the highest frequencies of variants.

Inclusion of the pathogenicity in the group factor allowed for the assessment of pathogenicity on variant frequencies. This was justified based on the significant interaction of Strain: Group from the parametric ANOVA model. This test found no significant differences between any of the groups when comparing virulent to avirulent strains, which can be seen in

Figure 4. Biologically, this does not provide any indication for which section of the genome may contain probable virulence factors since there is no statistically significant differences in polymorphism frequencies when comparing the same genome section between the two phenotypes.

4.5 Virulence Factor Identification

Another question raised was whether or not virulence in a crustacean host was the result of a gain or loss of function. When looking at the number of variants and how impacted the metabolic pathways are, it supports the theory that virulence was gained with respect to lobsters. Using the *A. viridans* var. *homo sapiens* genome as the reference, the avirulent group showed a lower number of polymorphism and metabolic pathway changes when compared to the virulent group. This implies that the bacteria may have been introduced to the environment as the human variant, then mutated into the var. *homari* strain in order to optimize its survival in a marine environment. This is further supported when looking at the roles of the impacted genes identified as potential virulence factors. For example, variants in genes utilized in trehalose metabolism, a sugar not found in mammals, were identified[82, 83].

In addition to different host environments, vertebrates and invertebrates have very different immune systems. Different virulence factors make pathogens successful in different immune systems, therefore it is expected that a different infection mechanism would be required by *A. viridans* var. *homari* in order to successfully colonize a crustacean. This is the driving factor behind the mutations and is exemplified when looking at the genes that have been impacted.

4.5.1 *CptV*

In order to optimize survival in a crustacean host, pathogens would require modifications to the copper regulation system. There were 12 variant locations identified in the ten strains for *ctpV*. For seven of locations, the variant was more common in the virulent phenotype. Crustaceans have a copper-coordinated oxygen carrier whereas the human host uses iron. While copper is required for normal cellular function as an enzyme cofactor, it induces toxicity at excess concentrations. First characterized in *Mycobacterium tuberculosis*, the probable copper-exporting P-type ATPase, or *ctpV*, shows a link between copper regulation and levels of

virulence. In murine models, *ctpV* is required in order for *M. tuberculosis* to display full virulence. *CtpV* mutants lack the ability to export copper resulting in protein denaturation, and other toxicity impacts [84].

4.5.2 *Tuf*

Elongation factor Tu (*tuf*) is commonly known as a transcription factor, however studies show a moonlighting role as an immune molecule that binds to fibronectin [85]. Fibronectin is a common mammalian host molecule that is often targeted by pathogens. Binding of fibronectin initiates the process of microbial adhesion and invasion. For mammals, fibronectin can also function as a coagulation factor [86].

For crustaceans, the analogous protein would be coagulogen, which is the second most prominent protein in crustacean haemolymph by concentration [87]. This protein is responsible for the coagulation and melanization of carapace and tissue injury in crustaceans [88]. As an opportunistic bacterium, *A. viridans* var. *homari* is only able to enter the host via shell damage. If the bacterium was able to block clotting factors, this would increase the chances of *A. viridans* var. *homari* colonization in a compromised host. For 17 of the 20 variant locations on *Tuf*, the variant was more common for virulent phenotype. If the bacterium is mammalian in origin, virulence in a crustacean would require *tuf* be modified in order to bind the coagulogen protein to exert a function similar to the one in the mammalian host.

4.5.3 *SugABC*

The *SugABC* system is responsible for trehalose recycling in bacterium. Trehalose can be used as a glycolipid in cell membranes and has been previously implicated as a crucial aspect of *Mycobacterium smegmatis* virulence [89]. While bacteria have the tools to deal with trehalose, they do not normally have a host capable of trehalose production when infecting a mammal. Crustaceans possess the enzyme trehalose 6-phosphate synthase in order to create trehalose. Animals found in cooler environments produce trehalose as a preventative mechanism for internal ice formation, the peculiar carbohydrate structure facilitating the process of vitrification as opposed to freezing [82]. Mirroring the situation with copper, in order to survive in the crustacean host, *A. viridans* var. *homari* would require modifications to its trehalose regulating proteins in order to maintain homeostasis and propagate in the host. Of the 23 variant sites

located in this system, 20 of them had the variant more common in the virulent phenotype. In addition, the variant was present in 4/5 or 5/5 strains. However, the edgeR results show that none of the pathways had a statistically significant difference ($p < 0.05$) in SNPs present in their genes.

4.5.4 *hsdM2*

This gene was identified as the one gene found in the virulent strains 1032, 29838, and Rabin's that was not found in any of the avirulent strains. For the purposes of this project, any gene or set of SNPs present in 3 or more of one phenotype and less than 3 in the other is of interest. The *hsdm2* gene encodes an DNA adenosine methyltransferase (Dam).

Methyltransferases are an essential component of epigenetic regulation, helping to control transcriptional regulation, mismatch repair and other genetic functions [90]. The methylation patterns caused by the adenine methyltransferases can result in differential gene expression as regulatory proteins will have different affinities for methylated regions. A study using *Salmonella typhimurium* found that *Dam*⁻ mutants were avirulent [91]. Since this gene is not found in any of the avirulent strains, it is possible the causative mechanism of virulence is *via* epigenetic regulation of transcription.

4.6 Metabolic Pathways

4.6.1 Purine and Pyrimidine Metabolism

At its most basic level, bacteria must be able to take up molecules and match their metabolism to the nutrients available. Included in the list of necessary molecules are purine and pyrimidines. It was found in *E. coli* that deletions in genes responsible for purine and pyrimidine metabolism can lower bacteria levels in blood anywhere from 20 to 1000 times. This same study determined the purine and pyrimidine pathway was the most critical pathway for ensuring infection persistence [92]. In *Salmonella*, *S. pneumoniae*, and *S. aureus* purine and pyrimidine metabolism is essential, and a lack thereof can attenuate virulence in a murine model. All of these bacteria are commonly the source of septicemic infections in humans [93]. As the cause of gaffkemia, a septicemic crustacean infection, it would make sense that purine and pyrimidine metabolism would be an important regulatory pathway for virulence in *Aerococcus viridans* var.

homari, this aligns with the fact that the purine and pyrimidine metabolic pathways were found to have more proteins containing variants in them for the virulent phenotype.

4.6.2 Starch and Sucrose Metabolism

Biofilm and polysaccharide capsule formation rely on the effective utilization of starch and sucrose. These same compounds have also been found to impact the adherence, with higher levels of starch being linked to better adherence [94]. It is possible this is because of a higher proportion of insoluble polysaccharides present in the capsule due to the presence of starch. There has also been research done looking into how the composition of different polysaccharide capsules and the stage of capsule formation may influence the gene expression [95]. For *A. viridans* var. *homari*, capsule formation has been identified as a key immune avoidance mechanism and GroEL identified as an upregulated gene responsible for its formation [54]. This pathway had more modifications for the virulent phenotype. It is possible the alternation in this pathway causes differential gene expression that creates the virulent phenotype.

Adhesion factors play a large part in the mechanism of bacterial infections. For the bacteria, the rate of adherence can be critical in invading the target cells. In a murine model, virulent *A. viridans* var. *bovis* were found to have increased levels of cAMP, a known virulence factor for *Streptococcus* that causes pores in the target cell [96]. Carkaci et al., also identified several gene families in the *Aerococcus* genus that would be responsible for controlling adhesion and cellular invasion [53].

4.6.3 Quorum Sensing

Quorum sensing is a staple of bacterial infection mechanisms [97]. This pathway was found to have a high number of polymorphism containing proteins in its pathways, and the majority of these were in the virulent phenotype. In this case, the virulence strain was found to have more than the avirulent. This may suggest that there has been some kind of mutation that increases the effectiveness of this signaling in the virulent strains. This would give the bacteria an advantage as quorum sensing is a major factor in immune evasion. If the ability to quorum sense was lost, this would make it possible for the host to eliminate the infection before it became systemic, thus making it avirulent.

5. Conclusions and Future Directions

The mentality of a “smoking gun” initially adopted during this project ended up being too simplistic. However, a lot of data was accumulated to help elucidate the virulence mechanism present in *A. viridans* var. *homari*. Overall, the number of variants present in virulent and avirulent species suggest that the avirulent strains are more similar to the var. *Homo sapiens* strain used for reference. Using gene files and identified polymorphisms, several potential virulence factors and mechanisms conferring virulence were identified. Among these are cation regulation and sugar regulation proteins. Both of which would require a shift for bacteria to survive in a crustacean host. Another gene found only in virulent strains encoded an adenosine methyltransferase, which would be an important epigenetic regulator and has been found to increase virulence in other bacterial species. GroEL phylogeny did not reveal any differences in the DNA sequence that would account for its differential expression. This further suggests that part of the differing phenotypes is caused by differential gene regulation and not purely mutations in the genome.

As of January 2019, there is a new analysis pipeline available for virulence factor identification called VFAnalyzer, based on the same database used for virulence factor identification in this project. The motivation behind this platform is to provide a pipeline to those with limited bioinformatics experience to cluster orthologues, identify gene clusters, and get draft assemblies of bacterial genomes [98]. The ability to identify gene clusters would be incredibly beneficial to this study as it would allow one to examine potential gene interactions as opposed to the isolated context used in this project. This pipeline would also utilize whole genome sequencing technology instead of curated FASTA sequences. This would reduce the risk that some factors may be overlooked due to the filtering criteria employed by the researcher. This may also eliminate the issue of lacking computer power when it comes to whole genome phylogeny analysis and alignment which was encountered with this project.

6. References

1. Government of Canada Fisheries and Oceans (2019) 2017 Value of Atlantic Landings. <http://dfo-mpo.gc.ca/stats/commercial/land-debarq/sea-maritimes/s2017av-eng.htm>. Accessed 16 Mar 2019
2. Scotia CN (2014) Government of Nova Scotia. <https://novascotia.ca/fish/commercial-fisheries/economic-impact/>. Accessed 16 Mar 2019
3. Smith H (1984) The biochemical challenge of microbial pathogenicity. *J App Bact* 57:395–404.
4. Stewart PS, William Costerton J (2001) Antibiotic resistance of bacteria in biofilms. *The Lancet* 358:135–138.
5. Vuong C, Dürr M, Carmody AB, et al (2004) Regulated expression of pathogen-associated molecular pattern molecules in *Staphylococcus epidermidis*: quorum-sensing determines pro-inflammatory capacity and production of phenol-soluble modulins. *Cell Micro* 6:753–759.
6. Vuong C, Otto M (2002) *Staphylococcus epidermidis* infections. *Microbes and Inf* 4:481–489.
7. Martínez JL, Baquero F (2002) Interactions among Strategies Associated with Bacterial Infection: Pathogenicity, Epidemicity, and Antibiotic Resistance. *Clin Micro Rev* 15:647–679.
8. Fair RJ, Tor Y (2014) Antibiotics and Bacterial Resistance in the 21st Century. *Perspect Medicin Chem* 6:25–64.
9. Cooksey R, Swenson J, Clark N, et al (1990) Patterns and mechanisms of beta-lactam resistance among isolates of *Escherichia coli* from hospitals in the United States. *Antimicrob Agents Chemother* 34:739–745
10. Deitch EA, Berg R (1987) Bacterial translocation from the gut: a mechanism of infection. *J Burn Care Rehabil* 8:475–482
11. Tandon P, Garcia-Tsao G (2008) Bacterial Infections, Sepsis, and Multiorgan Failure in Cirrhosis. *Semin Liver Dis* 28:26–42.
12. Berg RD (1999) Bacterial translocation from the gastrointestinal tract. *Adv Exp Med Biol* 473:11–30
13. Rooijackers SHM, van Kessel KPM, van Strijp JAG (2005) *Staphylococcal* innate immune evasion. *Trends Micro* 13:596–601.
14. Haggar A, Ehrnfelt C, Holgersson J, Flock J-I (2004) The Extracellular Adherence Protein from *Staphylococcus aureus* Inhibits Neutrophil Binding to Endothelial Cells. *Infect Immun* 72:6164–6167.

15. Jin T, Bokarewa M, Foster T, et al (2004) *Staphylococcus aureus* Resists Human Defensins by Production of Staphylokinase, a Novel Bacterial Evasion Mechanism. *J Imm* 172:1169–1176.
16. Aubry C, Goulard C, Nahori M-A, et al (2011) OatA, a Peptidoglycan O-Acetyltransferase Involved in *Listeria monocytogenes* Immune Escape, Is Critical for Virulence. *J Infect Dis* 204:731–740.
17. Chen L, Zheng D, Liu B, et al (2016) VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res* 44:D694-697.
18. Rahme LG, Ausubel FM, Cao H, et al (2000) Plants and animals share functionally common bacterial virulence factors. *Proc Natl Acad Sci USA* 97:8815–8821
19. Soto SM (2013) Role of efflux pumps in the antibiotic resistance of bacteria embedded in a biofilm. *Virulence* 4:223–229.
20. Jacques M (1996) Role of lipo-oligosaccharides and lipopolysaccharides in bacterial adherence. *Trends Micro* 4:408–410
21. Srinivasan R, Karaoz U, Volegova M, et al (2015) Use of 16S rRNA Gene for Identification of a Broad Range of Clinically Relevant Bacterial Pathogens. *PLOS ONE* 10:e0117617.
22. Watts JL, Yancey RJ (1994) Identification of veterinary pathogens by use of commercial identification systems and new trends in antimicrobial susceptibility testing of veterinary pathogens. *Clin Microbiol Rev* 7:346–356
23. Olson ME, Ceri H, Morck DW, et al (2002) Biofilm bacteria: formation and comparative susceptibility to antibiotics. *Can J Vet Res* 66:86–92
24. Guardabassi L, Schwarz S, Lloyd DH (2004) Pet animals as reservoirs of antimicrobial-resistant bacteria. *J Antimicrob Chemother* 54:321–332.
25. Lleò M del M, Signoretto C, Canepari P (2005) Gram-Positive Bacteria in the Marine Environment. In: Belkin S, Colwell RR (eds) *Oceans and Health: Pathogens in the Marine Environment*. Springer US, Boston, MA, pp 307–330
26. Anwar MA, Choi S (2014) Gram-Negative Marine Bacteria: Structural Features of Lipopolysaccharides and Their Relevance for Economically Important Diseases. *Mar Drugs* 12:2485–2514.
27. Austin B, Zhang X-H (2006) *Vibrio harveyi*: a significant pathogen of marine vertebrates and invertebrates. *Let App Micro* 43:119–124.
28. McCallum H, Harvell D, Dobson A (2003) Rates of spread of marine pathogens. *Ecology Letters* 6:1062–1067.
29. Cunningham AJ (1978) A comparison of the immune strategy of vertebrates and invertebrates. *Dev Comp Imm* 2:243–252.

30. Silhavy TJ, Kahne D, Walker S (2010) The Bacterial Cell Envelope. Cold Spring Harb Perspect Biol 2:.
31. Grimes DJ, Stemmler J, Hada H, et al (1984) *Vibrio* species associated with mortality of sharks held in captivity. Microb Ecol 10:271–282.
32. Bogomolni AL, Gast RJ, Ellis JC, et al (2008) Victims or vectors: a survey of marine vertebrate zoonoses from coastal waters of the Northwest Atlantic. Dis Aquat Organ 81:13–38.
33. Boehm T (2012) Evolution of Vertebrate Immunity. Curr Biol 22:R722–R732.
34. Rowley AF, Powell A (2007) Invertebrate Immune Systems—Specific, Quasi-Specific, or Nonspecific? J Imm 179:7209–7214.
35. Stewart JE, Arie B, Zwicker BM, Dingle JR (1969) Gaffkemia, a bacterial disease of the lobster, *Homarus americanus*: effects of the pathogen, *Gaffkya homari*, on the physiology of the host. Can J Microbiol 15:925–932.
36. Johnson PT (1976) Bacterial infection in the blue crab, *Callinectes sapidus*: course of infection and histopathology. Journal of Invertebrate Pathology 28:25–36.
37. Davis JW, Sizemore RK (1982) Incidence of *Vibrio* species associated with blue crabs (*Callinectes sapidus*) collected from Galveston Bay, Texas. Appl Environ Microbiol 43:1092–1097
38. Cornick JW, Stewart JE (1975) Red Crab (*Geryon quinqueedens*) and Snow Crab (*Chionoecetes opilio*) Resistance to Infection by the Lobster Pathogen *Aerococcus viridans* (var.) *homari*. J Fish Res Bd Can 32:702–706
39. Mori K, Stewart JE (1978) Natural and induced bactericidal activities of the hepatopancreas of the American lobster, *Homarus americanus*. J Invert Path 32:171–176.
40. Paterson WD, Stewart JE, Zwicker BM (1976) Phagocytosis as a cellular immune response mechanism in the American lobster, *Homarus americanus*. J Invert Path 27:95–104.
41. Malloy SC (1978) Bacteria induced shell disease of lobsters (*Homarus americanus*). J Wildl Dis 14:2–10
42. Castro KM, Factor JR, Angell T, Donald F, Landers J (2006) The Conceptual Approach to Lobster Shell Disease Revisited. J Crust Biol 26:646–660
43. Glenn RP, Pugh TL (2006) Epizootic Shell Disease in American Lobster (*Homarus Americanus*) in Massachusetts Coastal Waters: Interactions of Temperature, Maturity, and Intermolt Duration. J Crust Biol 26:639–645.
44. Smolowitz R, Chistoserdov AY, Hsu A (2005) A description of the pathology of epizootic shell disease in the american lobster, *homarus americanus*, h. milne edwards 1837. J Shell Res 24:749–756.

45. Tall BD, Fall S, Pereira MR, et al (2003) Characterization of *Vibrio fluvialis*-Like Strains Implicated in Limp Lobster Disease. *Appl Environ Microbiol* 69:7435–7446.
46. Stewart JE (1984) Lobster diseases. *Helgoländer Meeresuntersuchungen*, Volume 37, Issue 1-4, pp 243-254 37:243–254.
47. Gjerde J Occurrence and characterization of *Aerococcus viridans* from lobsters, *Homarus gammarus* L., dying in captivity. *J Fish Dis* 7:355–362.
48. Stewart JE, Cornick JW, Zwicker BM, Arie B (2004) Studies on the virulence of *Aerococcus viridans* (var.) *homari*, the causative agent of gaffkemia, a fatal disease of homarid lobsters. *Dis Aquat Org* 60:149–155.
49. Williams REO, Hirsch A, Cowan ST (1953) *Aerococcus*, a New Bacterial Genus. *Microbiology* 8:475–480.
50. Rasmussen M (2013) *Aerococci* and *aerococcal* infections. *J Infect* 66:467–474.
51. Rasmussen M (2016) *Aerococcus*: an increasingly acknowledged human pathogen. *Clin Micro Inf* 22:22–27.
52. Gopalachar A, Akins R, R Davis W, Siddiqui A (2004) Urinary tract infection caused by *Aerococcus viridans*, a case report. *Medical science monitor : international medical journal of experimental and clinical research* 10:CS73-5
53. Carkaci D, Dargis R, Nielsen XC, et al (2016) Complete Genome Sequences of *Aerococcus christensenii* CCUG 28831T, *Aerococcus sanguinicola* CCUG 43001T, *Aerococcus urinae* CCUG 36881T, *Aerococcus urinaeequi* CCUG 28094T, *Aerococcus urinaehominis* CCUG 42038 BT, and *Aerococcus viridans* CCUG 4311T. *Genome Announc* 4:.
54. Clark KF, Greenwood SJ (2011) *Aerococcus viridans* expression of Cpn60 is associated with virulence during infection of the American lobster, *Homarus americanus* Milne Edwards. *J Fish Dis* 34:831–843.
55. Martín V, Vela AI, Gilbert M, et al (2007) Characterization of *Aerococcus viridans* Isolates from Swine Clinical Specimens. *J Clin Micro* 45:3053–3057.
56. Leite A, Vinhas-Da-Silva A, Felício L, et al (2010) *Aerococcus viridans* urinary tract infection in a pediatric patient with secondary pseudohypoaldosteronism. *Revista argentina de microbiología* 42:269–270
57. Cornick JW, Stewart JE (1968) Interaction of the Pathogen *Gaffkya homari* with Natural Defense Mechanisms of *Homarus americanus*. *J Fish Res Bd Can* 25:695–709.
58. Rabin H (1965) Studies on gaffkemia, a bacterial disease of the American lobster, *Homarus americanus* (Milne-Edwards). *J Invert Path* 7:391–397.
59. Deibel RH, Niven CF (1960) Comparative Study Of *Gaffkya Homari*, *Aerococcus Viridans*, Tetrad-Forming Cocci From Meat Curing Brines, And The Genus *Pediococcus*1. *J Bacteriol* 79:175–180

60. Steenbergen JF, Kimball HS, Low DA, et al (1977) Serological grouping of virulent and avirulent strains of the lobster pathogen *Aerococcus viridans*. J Gen Microbiol 99:425–430.
61. Johnson PT, Stewart JE, Arie B (1981) Histopathology of *Aerococcus viridans* var. *homari* infection (Gaffkemia) in the lobster, *Homarus americanus*, and a comparison with histological reactions to a gram-negative species, *Pseudomonas perolens*. J Invert Path 38:127–148.
62. Clark KF, Wadowska D, Greenwood SJ (2016) *Aerococcus viridans* var. *homari*: The presence of capsule and the relationship to virulence in American lobster (*Homarus americanus*). J Invert Path 133:20–26.
63. Gallagher ML, Rittenburg JH, Bayer RC, Leavitt DF (1979) Incidence of *Aerococcus viridans* (Var.) *Homari* in Natural Crab (*Cancer Irroratus*, *Cancer Borealis*) Populations From Maine Coastal Waters. Crustaceana 37:316–317.
64. Stewart JE, Cornick JW, Spears DI, McLeese DW (1966) Incidence of *Gaffkya homari* in Natural Lobster (*Homarus americanus*) Populations of the Atlantic Region of Canada. J Fish Res Bd Can 23:1325–1330.
65. Lavallée J, Hammell KL, Spangler ES, Cawthorn RJ (2001) Estimated prevalence of *Aerococcus viridans* and *Anophryoides haemophila* in American lobsters *Homarus americanus* freshly captured in the waters of Prince Edward Island, Canada. Dis Aquat Org 46:231–236.
66. Greenwood SJ, Keith IR, Després BM, Cawthorn RJ (2005) Genetic characterization of the lobster pathogen *Aerococcus viridans* var. *homari* by 16S rRNA gene sequence and RAPD. Dis Aquat Org 63:237–246.
67. Afgan E, Baker D, van den Beek M, et al (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44:W3–W10.
68. Andrews, S.. FastQC A Quality Control tool for High Throughput Sequence Data.
69. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nature Methods 12:357–360.
70. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.
71. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.
72. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.
73. Cornish A, Guda C (2015) A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int 2015:.

74. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv:12073907
75. Erik Garrison (2018) vcflib: a simple C++ library for parsing and manipulating VCF files, + many command-line utilities. vcflib
76. Cingolani P, Platts A, Wang LL, et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6:80–92.
77. Ruden DM, Cingolani P, Patel VM, et al (2012) Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet 3:.
78. Fiume M, Smith EJM, Brook A, et al (2012) Savant Genome Browser 2: visualization and analysis for population-scale genomics. Nucleic Acids Res 40:W615–W621.
79. (2018) samtools: Tools (written in C using htlib) for manipulating next-generation sequencing data. samtools
80. Mendiburu F de, Simon R (2015) Agricolae - Ten years of an open source statistical tool for experiments in breeding, agriculture and biology. PeerJ PrePrints.
81. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.
82. Chung JS (2008) A trehalose 6-phosphate synthase gene of the hemocytes of the blue crab, *Callinectes sapidus*: cloning, the expression, its enzyme activity and relationship to hemolymph trehalose levels. Saline Systems 4:18.
83. Argüelles J-C (2014) Why can't vertebrates synthesize trehalose? J Mol Evol 79:111–116.
84. Ward SK, Abomoelak B, Hoye EA, et al (2010) CtpV: a putative copper exporter required for full virulence of *Mycobacterium tuberculosis*. Molecular Microbiology 77:1096–1110.
85. Kunert A, Losse J, Gruszin C, et al (2007) Immune Evasion of the Human Pathogen *Pseudomonas aeruginosa*: Elongation Factor Tuf Is a Factor H and Plasminogen Binding Protein. The Journal of Immunology 179:2979–2988.
86. Parkin J, Cohen B (2001) An overview of the immune system. The Lancet 357:1777–1789.
87. Depledge MH, Bjerregaard P (1989) Haemolymph protein composition and copper levels in decapod crustaceans. Helgoländer Meeresuntersuchungen 43:207.
88. Holmblad T, Söderhäll K (1999) Cell adhesion molecules and antioxidative enzymes in a crustacean, possible role in immunity. Aquaculture 172:111–123.
89. Wolber JM, Urbanek BL, Meints LM, et al (2017) The trehalose-specific transporter LpqY-SugABC is required for antimicrobial and anti-biofilm activity of trehalose analogues in *Mycobacterium smegmatis*. Carb Res 450:60–66.

90. Heusipp G, Fälker S, Alexander Schmidt M (2007) DNA adenine methylation and bacterial pathogenesis. *International J Med Micro* 297:1–7.
91. Heithoff DM, Sinsheimer RL, Low DA, Mahan MJ (1999) An Essential Role for DNA Adenine Methylation in Bacterial Virulence. *Science* 284:967–970.
92. Samant S, Lee H, Ghassemi M, et al (2008) Nucleotide Biosynthesis Is Critical for Growth of Bacteria in Human Blood. *PLOS Pathogens* 4:e37.
93. Rajagopal L, Vo A, Silvestroni A, Rubens CE (2005) Regulation of purine biosynthesis by a eukaryotic-type kinase in *Streptococcus agalactiae*. *Molec Micro* 56:1329–1346.
94. Vacca-Smith AM, Venkitaraman AR, Quivey RG, Bowen WH (1996) Interactions of streptococcal glucosyltransferases with α -amylase and starch on the surface of saliva-coated hydroxyapatite. *Arch Oral Biol* 41:291–298.
95. Duarte S, Klein MI, Aires CP, et al (2008) Influences of starch and sucrose on *Streptococcus mutans* biofilms. *Oral Micro Imm* 23:206–212.
96. Liu G, Yin J, Han B, et al (2019) Adherent/invasive capacities of bovine-associated *Aerococcus viridans* contribute to pathogenesis of acute mastitis in a murine model. *Vet Micro* 230:202–211.
97. Yarwood JM, Schlievert PM (2003) Quorum sensing in *Staphylococcus* infections. *J Clin Invest* 112:1620–1625.
98. Liu B, Zheng D, Jin Q, et al (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* 47:D687–D692.
99. Manhattan plot in R: a review. https://www.r-graph-gallery.com/wp-content/uploads/2018/02/Manhattan_plot_in_R.html. Accessed 8 Apr 2019

7. Appendix

Many supplemental materials could not be included due to file size. If you require any data not provided here, contact Dr. K. Fraser Clark at fraser.clark@dal.ca.

Table 5. Avirulent BAM alignment Stats. BWA v.0.7.17.4 mapping statistics using BAM stats v. 2.01 in Galaxy Europe. First column shows parameter tested while each column is for a different avirulent strain of *A. viridans* var. *homari*.

Parameter	37R	88R	10400	AVC	700406
1st fragments:	18460419	19040604	19396884	20293637	17117070
average length:	98	98	98	98	98
average quality:	36.5	36.4	36.4	36.5	36.4
bases duplicated:	0	0	0	0	0
bases mapped (cigar):	1782789111	2359681659	2133127426	2176726414	2011973031
bases mapped:	1782789121	2419581438	2184187737	2233300650	2056172704
bases trimmed:	0	0	0	0	0
error rate:	0.0231123	0.0314	0.03355852	0.03276333	0.0315
filtered sequences:	0	0	0	0	0
insert size average:	0	256.9	277.5	263.2	282.5
insert size standard deviation:	0	74.8	73.8	73.8	74.3
inward oriented pairs:	0	11981107	10863115	11077709	10180427
is sorted:	1	1	1	1	1
last fragments:	18460419	19040604	19396884	20293637	17117070
maximum length:	100	100	100	100	100
mismatches:	41204360	74071440	71584593	71316814	63376392
non-primary alignments:	0	0	0	0	0
non-primary alignments:	0	0	0	0	34234140
outward oriented pairs:	0	587	460	511	378
pairs on different chromosomes:	1788	40428	33262	32481	48343
pairs with other orientation:	9017446	1955	3245	2542	2226
reads duplicated:	0	0	0	0	0
reads mapped and paired:	18038473	24115315	21823632	22272808	20491066
reads mapped:	18038473	24557921	22194241	22672051	20895505
reads MQ0:	11588	9508	1829	1906	3801
reads paired:	36920838	38081208	38793768	40587274	34234140
reads properly paired:	1	23962947	21727238	22157502	20360018
reads QC failed:	0	0	0	0	0
reads unmapped:	18882365	13523287	16599527	17915223	13338635
sequences:	36920838	38081208	38793768	40587274	34234140
total length:	3650593380	3752211211	3818875166	3999582932	3370131748

Table 6. Virulent BAM alignment Stats. BWA v.0.7.17.4 mapping statistics using BAM stats v. 2.01 in Galaxy Europe. First column shows parameter tested while each column is for a different virulent strain of *A. viridans* var. *homari*.

Parameter	1030	1032	29838	Rabin's	Nfld
1st fragments:	18768915	20331624	19582185	18183088	17237761
average length:	98	98	98	98	98
average quality:	36.5	36.4	36.4	36.5	36.3
bases duplicated:	0	0	0	0	0
bases mapped (cigar):	2010129555	2222616102	2120271934	1833818925	1919123330
bases mapped:	2061210556	2272908018	2170724394	1905455141	1960039101
bases trimmed:	0	0	0	0	0
error rate:	0.03295099	0.0333	0.03314526	0.03128259	0.03351993
filtered sequences:	0	0	0	0	0
insert size average:	263.1	282.9	283.5	207.5	298.1
insert size standard deviation:	70.7	71.6	77.7	53.6	72.5
inward oriented pairs:	10247575	11290935	10763674	9424697	9729515
is sorted:	1	1	1	1	1
last fragments:	18768915	20331624	19582185	18183088	17237761
maximum length:	100	100	100	100	100
mismatches:	66235752	73936220	70276952	57366604	64328876
non-primary alignments:	0	0	0	0	0
non-primary alignments:	0	0	0	0	0
outward oriented pairs:	454	500	482	742	402
pairs on different chromosomes:	26350	40262	44619	11288	39822
pairs with other orientation:	3677	1211	2833	8075	1824
reads duplicated:	0	0	0	0	0
reads mapped and paired:	20582900	22691419	21648506	18979523	19569227
reads mapped:	20926106	23092644	22060882	19335213	19929365
reads MQ0:	1612	1876	1803	1403	1901
reads paired:	37537830	40663248	39164370	36366176	34475522
reads properly paired:	20495682	22583290	21529286	18846746	19459821
reads QC failed:	0	0	0	0	0
reads unmapped:	16611724	17570604	17103488	17030963	14546157
sequences:	37537830	40663248	39164370	36366176	34475522
total length:	3698682828	4003983575	3855181926	3582495427	3391843468

Table 7. Matchup of Contig Names between programs. Original File names assigned by Galaxy Europe, SnpEff Names assigned according to names from *Aerococcus viridans* SnpEff database, Ordered names come from an alignment of the reference genome to a fully assembled genome of *Aerococcus viridans* var. *homo sapiens*.

Original File	SnpEff	Ordered	Length	Original File	SnpEff	Ordered	Length
ADNT01000001.1	contig00001	contig_77	4025	ADNT01000076.1	contig00090	contig_13	1464
ADNT01000002.1	contig00002	contig_24	25766	ADNT01000077.1	contig00091	contig_33	30928
ADNT01000003.1	contig00003	contig_47	6267	ADNT01000078.1	contig00092	contig_83	98012
ADNT01000004.1	contig00004	contig_119	12887	ADNT01000079.1	contig00094	contig_56	7792
ADNT01000005.1	contig00005	contig_76	1609	ADNT01000080.1	contig00096	contig_89	4733
ADNT01000006.1	contig00006	contig_85	21618	ADNT01000081.1	contig00060	contig_84	834
ADNT01000007.1	contig00007	contig_55	6886	ADNT01000081.1	contig00097	contig_84	834
ADNT01000008.1	contig00008	contig_21	6202	ADNT01000082.1	contig00098	contig_4	1356
ADNT01000009.1	contig00009	contig_135	2027	ADNT01000083.1	contig00099	contig_16	2115
ADNT01000010.1	contig00010	contig_71	34245	ADNT01000084.1	contig00100	contig_31	1613
ADNT01000011.1	contig00011	contig_32	7658	ADNT01000085.1	contig00102	contig_7	83054
ADNT01000012.1	contig00012	contig_75	41947	ADNT01000086.1	contig00103	contig_3	1344
ADNT01000013.1	contig00013	contig_66	4327	ADNT01000087.1	contig00111	contig_137	1431
ADNT01000014.1	contig00014	contig_111	1676	ADNT01000088.1	contig00116	contig_118	74517
ADNT01000015.1	contig00016	contig_74	12420	ADNT01000089.1	contig00117	contig_78	22089
ADNT01000016.1	contig00017	contig_143	11273	ADNT01000090.1	contig00118	contig_128	5848
ADNT01000017.1	contig00018	contig_138	20754	ADNT01000091.1	contig00119	contig_114	1417
ADNT01000018.1	contig00019	contig_126	946	ADNT01000092.1	contig00120	contig_40	64016
ADNT01000019.1	contig00020	contig_131	1816	ADNT01000093.1	contig00121	contig_38	19774
ADNT01000020.1	contig00021	contig_94	18293	ADNT01000094.1	contig00122	contig_147	74319
ADNT01000021.1	contig00022	contig_51	16451	ADNT01000095.1	contig00123	contig_146	26939
ADNT01000022.1	contig00023	contig_88	48865	ADNT01000096.1	contig00124	contig_125	3801
ADNT01000023.1	contig00024	contig_27	6310	ADNT01000097.1	contig00125	contig_108	21041
ADNT01000024.1	contig00025	contig_26	1603	ADNT01000098.1	contig00126	contig_120	4852
ADNT01000025.1	contig00026	contig_44	2823	ADNT01000099.1	contig00127	contig_10	5798
ADNT01000026.1	contig00027	contig_60	22273	ADNT01000100.1	contig00128	contig_145	58141
ADNT01000027.1	contig00028	contig_22	7843	ADNT01000101.1	contig00129	contig_117	12504
ADNT01000028.1	contig00029	contig_18	2075	ADNT01000102.1	contig00131	contig_69	55174
ADNT01000029.1	contig00030	contig_105	11765	ADNT01000103.1	contig00132	contig_35	8455
ADNT01000030.1	contig00031	contig_101	3923	ADNT01000104.1	contig00133	contig_65	5099
ADNT01000031.1	contig00032	contig_141	31678	ADNT01000105.1	contig00135	contig_70	1017
ADNT01000032.1	contig00034	contig_29	4641	ADNT01000106.1	contig00136	contig_90	562
ADNT01000033.1	contig00035	contig_139	11107	ADNT01000107.1	contig00137	contig_2	583
ADNT01000034.1	contig00036	contig_87	7592	ADNT01000108.1	contig00138	contig_132	2072
ADNT01000035.1	contig00038	contig_98	5243	ADNT01000109.1	contig00139	contig_123	14782
ADNT01000036.1	contig00041	contig_58	19292	ADNT01000110.1	contig00140	contig_63	39978
ADNT01000037.1	contig00042	contig_121	1113	ADNT01000111.1	contig00141	contig_140	11287
ADNT01000038.1	contig00043	contig_34	6991	ADNT01000112.1	contig00142	contig_62	15131
ADNT01000039.1	contig00048	contig_102	27757	ADNT01000113.1	contig00144	contig_112	2229
ADNT01000040.1	contig00050	contig_124	21003	ADNT01000114.1	contig00146	contig_5	6765
ADNT01000041.1	contig00051	contig_6	1711	ADNT01000115.1	contig00148	contig_81	4925
ADNT01000042.1	contig00052	contig_99	56907	ADNT01000116.1	contig00149	contig_54	1160
ADNT01000043.1	contig00053	contig_41	2089	ADNT01000117.1	contig00150	contig_133	1104
ADNT01000044.1	contig00054	contig_49	16907	ADNT01000118.1	contig00152	contig_46	1917
ADNT01000045.1	contig00055	contig_115	19800	ADNT01000119.1	contig00153	contig_9	1327
ADNT01000046.1	contig00056	contig_97	1522	ADNT01000120.1	contig00156	contig_42	6929
ADNT01000047.1	contig00057	contig_93	41782	ADNT01000121.1	contig00157	contig_106	2067
ADNT01000048.1	contig00058	contig_30	3143	ADNT01000122.1	contig00158	contig_12	2661
ADNT01000049.1	contig00059	contig_107	2933	ADNT01000123.1	contig00160	contig_28	1321
ADNT01000051.1	contig00061	contig_59	30377	ADNT01000124.1	contig00162	contig_96	551
ADNT01000052.1	contig00062	contig_148	39596	ADNT01000125.1	contig00163	contig_110	14192
ADNT01000053.1	contig00063	contig_150	23413	ADNT01000126.1	contig00164	contig_8	518
ADNT01000054.1	contig00064	contig_68	9275	ADNT01000127.1	contig00165	contig_25	542
ADNT01000055.1	contig00065	contig_73	7735	ADNT01000128.1	contig00173	contig_122	3002
ADNT01000056.1	contig00066	contig_43	25879	ADNT01000129.1	contig00176	contig_11	1213
ADNT01000057.1	contig00067	contig_15	3958	ADNT01000130.1	contig00178	contig_50	1463
ADNT01000058.1	contig00068	contig_23	5443	ADNT01000131.1	contig00180	contig_116	4747
ADNT01000059.1	contig00071	contig_36	60871	ADNT01000132.1	contig00181	contig_37	3125
ADNT01000060.1	contig00073	contig_61	3724	ADNT01000133.1	contig00184	contig_103	1009
ADNT01000061.1	contig00074	contig_136	8312	ADNT01000134.1	contig00186	contig_17	4928
ADNT01000062.1	contig00075	contig_91	26947	ADNT01000135.1	contig00187	contig_104	1010
ADNT01000063.1	contig00076	contig_57	19300	ADNT01000136.1	contig00188	contig_113	6288
ADNT01000064.1	contig00077	contig_1	10692	ADNT01000137.1	contig00190	contig_53	956
ADNT01000065.1	contig00078	contig_92	29391	ADNT01000138.1	contig00193	contig_52	2600



Figure 5. Snapshot of Genome Savant. Region shown is from 4228-12680 on contig ADNT0000044.1. From top to bottom: 10400, 37R, AVC, 1030,1032, 29838, Nfld, Rabin's, 700406, 88R. Lines in individual gene tracks represent a SNP based on the reference genome *Aerococcus_viridans_atcc_11563_ccug_4311*.

Genome Savant File Set Up

```
$ cat snpEff.config
```

```
$ java -Xmx4g -jar snpEff.jar -v -stats ex2.html Aerococcus_viridans_atcc_11563_ccug_4311
FreeBayes_700406.vcf >700406.ann.vcf
```

```
$ tar jxvf tabix-0.2.6.tar.bz2
```

```
User~/samtools-1.8
```

```
#set the directory to be tabix so we can work within that
```

```
$ cd tabix-0.2.6
```

```
User ~/samtools-1.8/tabix-0.2.6
```

```
$ make
```

```
User ~/samtools-1.8/tabix-0.2.6
```

```
#this is where it started working, just typed "bgzip" to see if the options would come up, indicated
the tool was loaded
```

```
$ bgzip
```

```
Usage: bgzip [options] [file] ...
```

```
Options: -c write on standard output, keep original files unchanged
```

```
-d decompress
```

```
-f overwrite files without asking
```

```
-b INT decompress at virtual file pointer INT
```

```
-s INT decompress INT bytes in the uncompressed file
```

```
-h give this help
```

```
User ~/samtools-1.8/tabix-0.2.6
#same check for tabix
$ tabix
```

```
#each file needs to go through the bgzip before it can be tabixed
User ~/samtools-1.8/tabix-0.2.6
$ bgzip FreeBayes_700406.vcf
```

```
User ~/samtools-1.8/tabix-0.2.6
$ tabix -p vcf FreeBayes_700406.vcf.gz
```

R Code to determine unique genes between Virulent and Avirulent Genomes

Take out the first column of Gene Names for each strain

```
just_37R=Genes_37R[,1]
just_88R=Genes_88R[,1]
just_AVC=Genes_AVC[,1]
just_10400=Genes_10400[,1]
just_700406=Genes_700406[,1]
just_1030=Genes_1030[,1]
just_1032=Genes_1032[,1]
just_29838=Genes_29838[,1]
just_Nfld=Genes_Nfld[,1]
just_Rabins=Genes_Rabins[,1]
```

Set all of the factors to characters in order to compare them properly

```
c_just_37R=as.character(just_37R)
c_just_88R=as.character(just_88R)
c_just_AVC=as.character(just_AVC)
c_just_10400=as.character(just_10400)
c_just_700406=as.character(just_700406)
c_just_1030=as.character(just_1030)
c_just_1032=as.character(just_1032)
c_just_29838=as.character(just_29838)
c_just_NFLD=as.character(just_Nfld)
c_just_Rabins=as.character(just_Rabins)
```

Make dataframes for all of the virulent and then all of the avirulent, set all the lists to be in the same column so we only have to compare two columns later. Also use the discrete() thing to make sure you are looking at only unique characters

```
Virulent=distinct(data.frame(x=c(c_just_1030,c_just_1032,c_just_29838,c_just_NFLD,c_just_Rabins)))
Avirulent=distinct(data.frame(x=c(c_just_37R,c_just_88R,c_just_10400,c_just_AVC,c_just_700406)))
Avirulent$x=as.character(Avirulent$x)
Virulent$x=as.character(Virulent$x)
```

Figure out which genes are unique to Avirulent and Virulent phenotypes

```
Avirulent$presence=is.element(Avirulent$x,Virulent$x)
```

```
Virulent$presence=is.element(Virulent$x,Avirulent$x)
```

Make new dataset of all the genes that do not appear in the virulent strains

```
Diff=subset(Avirulent,presence=="FALSE")
```

Make new dataset of all the genes that do not appear in the avirulent strains

```
V_diff=subset(Virulent,presence=="FALSE")
```

Check to see which avirulent strains have the unique genes

```
colnames(Diff)=c("X.GeneName","presence")
```

```
diff_10400=merge(Diff,Genes_10400,by="X.GeneName")
```

```
diff_AVC=merge(Diff,Genes_AVC,by="X.GeneName")
```

```
diff_37R=merge(Diff,Genes_37R,by="X.GeneName")
```

```
diff_88R=merge(Diff,Genes_88R,by="X.GeneName")
```

```
diff_700406=merge(Diff,Genes_700406,by="X.GeneName")
```

check which virulent strains have the unique genes

```
test2=subset(Avirulent,x=="hsdM2")
```

```
hsd_1030=subset(Genes_1030,X.GeneName=="hsdM2")
```

```
hsd_1032=subset(Genes_1032,X.GeneName=="hsdM2")
```

```
hsd_29838=subset(Genes_29838,X.GeneName=="hsdM2")
```

```
hsd_Nfld=subset(Genes_Nfld,X.GeneName=="hsdM2")
```

```
hsd_Rabins=subset(Genes_Rabins,X.GeneName=="hsdM2")
```

Organisms specified for KEGG search

Escherichia coli K-12 MG1655

Salmonella enterica subsp. *enterica* serovar *Typhi* CT18

Shigella flexneri 301 (serotype 2a)

Enterobacter sp. 638

Enterobacter cloacae subsp. *cloacae* ATCC 13047

Yersinia pestis CO92 (biovar *Orientalis*)

Haemophilus influenzae Rd KW20 (serotype d)

Vibrio cholerae O1 El Tor N16961

Photobacterium profundum

Pseudomonas aeruginosa PAO1

Legionella pneumophila subsp. *pneumophila* *Philadelphia 1* (serogroup 1)

Nitrosococcus oceani

Neisseria meningitidis MC58 (serogroup B)

Bordetella pertussis Tohama I

Helicobacter pylori 26695

Rickettsia prowazekii Madrid E

Wolbachia wMel (*Drosophila melanogaster*)

Bacillus subtilis subsp. *subtilis* 168

Oceanobacillus iheyensis

Staphylococcus aureus subsp. *aureus* N315 (MRSA/VSSA)

Listeria monocytogenes EGD-e (serotype 1/2a)

Lactococcus lactis subsp. *lactis* II1403

Streptococcus pyogenes M1 GAS (serotype M1)

Enterococcus faecalis V583

Pediococcus pentosaceus ATCC 25745

Oenococcus oeni

Aerococcus urinae ACS-120-V-Col10a

Mycoplasma genitalium G37

Mycobacterium tuberculosis H37Rv

Chlamydia trachomatis D/UW-3/CX

Table 8. EdgeR output. Statistical analysis done of KEGG identified pathways using avirulent strains as a reference point with TMM Normalization. EdgeR test run in Galaxy Europe.

Pathway	PValue	Pathway	Pvalue
ko02020 Two-component system†	0.0021	ko02026 Biofilm formation - Escherichia coli†	0.4112
ko01100 Metabolic pathways†	0.0182	ko03410 Base excision repair†	0.4240
ko01120 Microbial metabolism in diverse environments†	0.0300	ko00720 Carbon fixation pathways in prokaryotes†	0.4333
ko01130 Biosynthesis of antibiotics†	0.0359	ko00450 Selenocompound metabolism†	0.4413
ko00030 Pentose phosphate pathway†	0.0362	ko00061 Fatty acid biosynthesis†	0.4415
ko00410 beta-Alanine metabolism†	0.0382	ko01212 Fatty acid metabolism†	0.4415
ko00660 C5-Branched dibasic acid metabolism†	0.0386	ko01524 Platinum drug resistance†	0.4601
ko00780 Biotin metabolism†	0.0386	ko00230 Purine metabolism†	0.4622
ko00830 Retinol metabolism†	0.0386	ko00053 Ascorbate and aldarate metabolism†	0.4669
ko00300 Lysine biosynthesis†	0.0390	ko00903 Limonene and pinene degradation†	0.4669
ko00362 Benzoate degradation†	0.0613	ko00981 Insect hormone biosynthesis†	0.4669
ko00071 Fatty acid degradation†	0.0747	ko00982 Drug metabolism - cytochrome P450†	0.4669
ko00380 Tryptophan metabolism†	0.0800	ko04122 Sulfur relay system†	0.4669
ko00521 Streptomycin biosynthesis†	0.0861	ko05204 Chemical carcinogenesis†	0.4669
ko00650 Butanoate metabolism†	0.1162	ko00980 Metabolism of xenobiotics by cytochrome P450†	0.4669
ko02060 Phosphotransferase system	0.1366	ko00906 Carotenoid biosynthesis†	0.4669
ko01230 Biosynthesis of amino acids†	0.1452	ko03010 Ribosome†	0.4770
ko01200 Carbon metabolism†	0.1488	ko04922 Glucagon signaling pathway†	0.4863
ko00333 Prodigiosin biosynthesis†	0.1500	ko00290 Valine, leucine and isoleucine biosynthesis†	0.4988
ko00790 Folate biosynthesis†	0.1756	ko00240 Pyrimidine metabolism†	0.5005
ko05230 Central carbon metabolism in cancer†	0.1756	ko00620 Pyruvate metabolism†	0.5005
ko00190 Oxidative phosphorylation†	0.1797	ko03013 RNA transport†	0.5334
ko01502 Vancomycin resistance†	0.1896	ko03060 Protein export†	0.5334
ko04918 Thyroid hormone synthesis†	0.1896	ko04213 Longevity regulating pathway - multiple species†	0.5334
ko00680 Methane metabolism†	0.1897	ko04016 MAPK signaling pathway - plant†	0.5528
ko00710 Carbon fixation in photosynthetic organisms†	0.1897	ko00260 Glycine, serine and threonine metabolism†	0.5896
ko01210 2-Oxocarboxylic acid metabolism†	0.1941	ko00400 Phenylalanine, tyrosine and tryptophan biosynthesis†	0.5965
ko00310 Lysine degradation†	0.1941	ko03030 DNA replication†	0.5995
ko04930 Type II diabetes mellitus†	0.2044	ko03440 Homologous recombination†	0.6057
ko04068 FoxO signaling pathway†	0.2044	ko00130 Ubiquinone and other terpenoid-quinone biosynthesis†	0.6187
ko05165 Human papillomavirus infection†	0.2044	ko04212 Longevity regulating pathway - worm†	0.6225
ko04070 Phosphatidylinositol signaling system†	0.2044	ko04931 Insulin resistance†	0.6225
ko05203 Viral carcinogenesis†	0.2044	ko00561 Glycerolipid metabolism†	0.6345
ko03420 Nucleotide excision repair†	0.2044	ko00500 Starch and sucrose metabolism†	0.6607
ko00550 Peptidoglycan biosynthesis†	0.2058	ko00640 Propanoate metabolism†	0.6723
ko00625 Chloroalkane and chloroalkene degradation†	0.2076	ko00480 Glutathione metabolism†	0.6743
ko03020 RNA polymerase†	0.2076	ko04146 Peroxisome†	0.6761
ko05010 Alzheimer disease†	0.2076	ko00220 Arginine biosynthesis†	0.7249
ko00010 Glycolysis / Gluconeogenesis†	0.2181	ko00430 Taurine and hypotaurine metabolism†	0.7249
ko00330 Arginine and proline metabolism†	0.2220	ko04917 Prolactin signaling pathway†	0.7576
ko00670 One carbon pool by folate†	0.2423	ko00920 Sulfur metabolism†	0.7576
ko01501 β-Lactam resistance†	0.2631	ko00730 Thiamine metabolism†	0.7669
ko00250 Alanine, aspartate and glutamate metabolism†	0.2631	ko00760 Nicotinate and nicotinamide metabolism†	0.7669
ko00350 Tyrosine metabolism†	0.2751	ko00520 Amino sugar and nucleotide sugar metabolism†	0.8449
ko00261 Monobactam biosynthesis†	0.3001	ko00020 Citrate cycle	0.8650
ko00052 Galactose metabolism†	0.3029	ko00270 Cysteine and methionine metabolism†	0.8705
ko00473 D-Alanine metabolism†	0.3036	ko04066 HIF-1 signaling pathway†	0.8791
ko00340 Histidine metabolism†	0.3036	ko02010 ABC transporters†	0.8887
ko00405 Phenazine biosynthesis†	0.3036	ko03018 RNA degradation†	0.8998
ko05134 Legionellosis†	0.3036	ko00040 Pentose and glucuronate interconversions†	0.9043
ko00626 Naphthalene degradation†	0.3036	ko05111 Biofilm formation - Vibrio cholerae†	0.9140
ko00361 Chlorocyclohexane and chlorobenzene degradation†	0.3217	ko01523 Antifolate resistance†	0.9140
ko01110 Biosynthesis of secondary metabolites†	0.3553	ko03430 Mismatch repair†	0.9162
ko00970 Aminoacyl-tRNA biosynthesis†	0.3931	ko02024 Quorum sensing†	0.9329
ko01220 Degradation of aromatic compounds†	0.3975	ko04973 Carbohydrate digestion and absorption†	0.9394
ko00750 Vitamin B6 metabolism†	0.4109	ko00401 Novobiocin biosynthesis†	0.9394
ko04211 Longevity regulating pathway†	0.4109	ko00622 Xylene degradation†	0.9394
ko00770 Pantothenate and CoA biosynthesis†	0.4109	ko00627 Aminobenzoate degradation†	0.9394
ko00280 Valine, leucine and isoleucine degradation†	0.4109	ko00643 Styrene degradation†	0.9394
ko00562 Inositol phosphate metabolism†	0.4109	ko05340 Primary immunodeficiency†	0.9394
ko00590 Arachidonic acid metabolism†	0.4109	ko02025 Biofilm formation - Pseudomonas aeruginosa†	0.9433
ko04013 MAPK signaling pathway - fly†	0.4109	ko03070 Bacterial secretion system†	0.9433
ko05016 Huntington disease†	0.4109	ko00051 Fructose and mannose metabolism†	0.9646
ko00983 Drug metabolism - other enzymes†	0.4109	ko00564 Glycerophospholipid metabolism†	0.9666

FASTA formatted GroEL sequences

>1030_GroEL

TTAATACATACCTGCACCTGGATCCATTGCTGGCATTGTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGGTCAACGATACC
TTCTTCAATCATGTTTACCCATTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCAGCGATTTGACGGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTTTTAGCATTGATAAAGGCAGTACC
ACCACCCGCTACAATACCTTGGTCAACAGCCGCACGTGTTGCGTTTAAAGCATCTTCGATACGTAATTTACGTTCTT
TTAATTTCTGATTAGCTGCACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAA
TTTCTCACGGTCATATTCAGAAGTTGTTTCTTCAATTTGACTACGGATTTGGGCAACACGTTGTGCTAATTGCTCTTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACGCGGTTAGCAGAACCTAATTGATCAATTG
TCGTATCTTTTAATCAAGTCCTAAATCTTCAGTAATTTACTTGACCACCTGTTAGGATAGATAAATCTTCTAACATT
GCTTTACGACGATCACAAGGCTGGCGTTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAATCTTTTGTCTGTTAATAAAAATATAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCAATTGTCAGTTACCATATATTGAGATAGGTAACCACGGTCAAATTGCATACCCTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTTCTCAATGGTGATGACGCCATCTTGGCCAACCTTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTACCTGATGAAATTGATGCAACGTTAGCGCTTACGCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTCATACCTTACGTTACAATTGCTTGAAGTATAGAACAGTGGCAGTTGTTGTACCGTCACCGGCAATATCA
TTTGTTTTTGAAGCTACTTACGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCTAATTTCAATTTCTTTAGCG
ATAGTTACACCATCGTTAGTGATTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTTGGTCCTA
AAGTTACTTTACCGTATCAGCTAATATATCAATCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGGCATTATTCATTCACCTCATAAAATTTCAATTTGTTATTAATCTGTAAATATTGATAAATTGACTAATCT
ACAATTTGCGATAATATCGCTTTCTTTAATAAATTAATAATTTCTTGACCATCATATTCAAATGTTGAACCTGCATATTT
TCAAATAGAACGCGGTGCGCTTCTTTCAATTTGATGTGAAATCTTCCGTATTCTCAGCAACTGCTTCAACAACACC
TACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCAC
AACTGCGCGTCCGTTTAATGGTTTAATCATTTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA

>37R_GroEL

TTAATACATACCTGCACCTGGATCCATTGCTGGCATTGTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGATCAACAATACC
TTCTTCAATCATGTTTACCCATTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCAGCGATTTGACGTACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTCTTAGCATTGATAAAGGCAGTACC
ACCACGACTACGATACCTTGGTCAACAGCCGCACGTGTTGCGTTTAAAGCATCTTCAATACGTAATTTACGTTCTT
TTAATTTCTGATTAGCTGCACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAGTAA
TTTTTACGCGTATATTCAGAAGTTGTTTCTTCAATTTGACTACGCTTTGGGCAACACGTTGTGCTAATTGCTCTTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACGCGGTTAGCAGAACCTAATTGATCAATTG
TCGTATCTTTTAATCAAGTCCTAAATCTTCAGTAATTTACTTGACCACCTGTTAGGATAGATAAATCTTCTAACATT
GCTTTACGACGGTACCAAAGGCTGGCGTTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAATCTTTTATCTGTTAATAAAAATATAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCAATTGTCAGTTACCATATATTGAGATAGGTAACCAGTCAAATTGCATACCCTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTTCTCAATAGTGATGACGCCATCTTGGCCAACCTTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTACCTGATGAAATTGATGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTCATACCTTACGTTACAATTGCTTGAAGTATAGAACAGTGGCAGTTGTTGTACCGTCACCGGCAATATCA
TTTTTTTTGAAGCTACTTACGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCTAATTTCAATTTCTTTAGCG
ATAGTTACACCATCGTTAGTGATTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTTGGTCCTA
AAGTTACTTTACCGTATCAGCTAATATATCAATCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGGCATTATTCATTCACCTCATAAAATTTCAATTTGTTATTAATCTGTAAATATTGATAAATTGACTAATCT
ACAATTTGCGATAATATCGCTTTCTTTAATAAATTAATAATTTCTTGACCATCATATTCAAATGTTGAACCTGCATATTT
TTCAAATAGAACGCGGTGCGCTTCTTTCAATTTGCGATGTGAAATCTTCCGTATTCTCAGCAACTGCTTCAACAACAC
CTACTTGTGTTTCTTGTGGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCA
CAACTGCGCGTCCGTTTAATGGTTTAATCATTTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTA
A

>88R_GroEL

TTAATACATACCTGCACCTGGATCCATTGCTGGCATTGTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGATCAACAATACC
TTCTTCAATCATGTTTACCCATTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCAGCGATTTGACGTACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTTTTAGCATTGATAAAGGCAGTACC
ACCACGACTACGATACCTTGGTCAACAGCCGCACGTGTTGCGTTTAAAGCATCTTCGATACGTAATTTACGTTCTT

TTAATTCTGATTAGCTGCACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAGTAA
TTTTTCACGGTCATATTCAGAAGTTGTTTCTCAATTTGACTACGGATTTGGGCAACACGTTGTGCTAATTGCTCTTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACCGCGTTAGCAGAACCTAATTGATCAATTG
TCGTATCTTTAATTCAAGTCCTAAATCTTCAGTAATTAAGTACTTGACCACCTGTTAGGATAGATAAATCTTCAACATT
GCTTTACGACGGTCACCAAAGGCTGGCGCTTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAACACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAATCTTTTGTCTGTTAATAAAAATAAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCATTGTCAGTTACCATATATTGAGATAGGTAACCGCGGTCAAATTGCATACCTTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTTTCAATAGTGATGACGCCATCTTGCCAACTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTCGCTGATGAAATGATGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTTCATACCTTCAGTTACAATTGCTTGAGTTAGAACAGTGGCAGTTGTTGTACCGTCACCAGCGATATCA
TTTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCAATTTCAATTTCTTTAGCG
ATGGTTACACCATCGTTAGTGATTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTACCGTATCAGCTAATATATCAATCCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAAATTTCAATTTGTTATTAATCTGTAATAATTGATAAATTTGACTAATCT
ACAATTTGCATAATACCTTTCGCTTTCTTTAATAAATAAATAATTTCTTGACCTCATATTTCAAATGTTGAACCTGCATATTT
TTCAAATAGAACGCGGTGCGCTTCTTTCAATTTGTGATGTGAAATCTTCCGTAATTTCTCAGCAACTGCTTCAACAACAC
CTACTTGTGTTTCTTTAGCAGCTGACGGTAATACAATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCA
CAACTGCGCGTCCGTTAATGGTTAATCATTTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTA
A

>1032_GroEL

TTAATACATACTGCACCTGGATCCATTGCTGGCATTGTTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGGTCAACGATACC
TTCTTCAATCATGTTTACCATTTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCCTCAGCGATTTGACGGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTTTTAGCATTGATAAAGGCAGTACC
ACCACCCGCTACAATACCTTGGTCAACAGCCGCACGTTGCGTTTTAAAGCATCTTCGATAACGTAATTTACGTTCTT
TTAATTCTGATTAGCTGCACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAAA
TTTCTCACGGTCATATTCAGAAGTTGTTTCTCAATTTGACTACGGATTTGGGCAACACGTTGTGCTAATTGCTCTTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACCGCGTTAGCAGAACCTAATTGATCAATTG
TCGTATCTTTAATTCAAGTCCTAAATCTTCAGTAATTAAGTACTTGACCACCTGTTAGGATAGATAAATCTTCAACATT
GCTTTACGACGATACCAAAGGCTGGCGCTTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAATCTTTTGTCTGTTAATAAAAATAAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCATTGTCAGTTACCATATATTGAGATAGGTAACCACGGTCAAATTGCATACCTTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTTCAATGGTGATGACGCCATCTTGCCAACTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTCACCTGATGAAATGATGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTTCATACCTTCAGTTACAATTGCTTAGATTAAGTACTTGACCTTCAATTTCAATGTTGAACCTGCATATTT
TTTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCAATTTCAATTTCTTTAGCG
ATAGTTACACCATCGTTAGTGATTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTACCGTATCAGCTAATATATCAATCCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAAATTTCAATTTGTTATTAATCTGTAATAATTGATAAATTTGACTAATCT
ACAATTTGCATAATACCTTTCGCTTTCTTTAATAAATAAATAATTTGACCTTCAATTTCAAATGTTGAACCTGCATATTT
TCAAATAGAACGCGGTGCGCTTCTTTCAATTTGTGATTGTAATACTTCCGTAATTTCTCAGCAACTGCTTCAACAACAC
TACTTGTGTTTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCAC
AACTGCGCGTCCGTTAATGGTTAATCATTTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA

>10400_GroEL

TTAATACATACTGCACCTGGATCCATTGCTGGCATTGTTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGGTCAACGATACC
TTCTTCAATCATGTTTACCATTTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCCTCAGCGATTTGACGGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTTTTAGCATTGATAAAGGCAGTACC
ACCACCCGCTACAATACCTTGGTCAACAGCCGCACGTTGCGTTTTAAAGCATCTTCGATAACGTAATTTACGTTCTT
TTAATTCTGATTAGCTGCACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAAA
TTTCTCAGGTCATATTCAGAAGTTGTTTCTCAATTTGACTACGATTTGGGCAACACGTTGTGCTAATTTGCTCTTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACCGCGTTAGCAGAACCTAATTGATCAATTG
TCGTATCTTTAATTCAAGTCCTAAATCTTCAGTAATTAAGTACTTGACCACCTGTTAGGATAGATAAATCTTCAACATT
GCTTTACGACGATACCAAAGGCTGGCGCTTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAATCTTTTGTCTGTTAATAAAAATAAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCATTGTCAGTTACCATATATTGAGATAGGTAACCACCGTCAAATTGCATACCTTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTTCAATGGTGATGACGCCATCTTGCCAACTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTCACCTGATGAAATGATGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAACCTGAAATT

TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTCATACCTTCAGTTACAATTGCTTGAGTTAGAACAGTGGCAGTTGTTGTACCGTCACCGCAATATCA
TTTTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCAATTTCAATTTCTTTAGCG
ATAGTTACACCATCGTTAGTATTAGCGGTGAACCAATACTTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTTACCCTATCAGTAAATATATCAATCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAAATTTCAATTTGTTATTAATCTGTAAATATTGATAAATTGACTAATCT
ACAATTGCGATAATATCGCTTTCTTTAATAATTAATAATTCTTGACCTTCATATTCAAATGTTGAACCTGCATATTTT
TCAAATAGAACGCGGTGCGCTTCTTTCAATTGTGATGTGAAATCTCCGTATTCTCAGCAACTGCTTCAACAACACC
TACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCAC
AACTGCGCGTCCGTTAATGGTTAATCATTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA
>29838_GroEL
TTAATACATACTGCCTGGATCCATTGCTGGCATTGTGGTTGAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGGTCAACGATACC
TTCTTCAATCATGTTTACCCATTGAGAAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCAGCGATTTGACGGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAAGTTGTTTATCACCAGCTAATGTTTACGTAATTCAGCAACTGCTTTTTAGCATTGATAAAGCCAGTACC
ACCACCGCTACAATACTGGTCAACAGCCGACGTGTTGCGTTTTAAAGCATCTTCGATACGTAATTTACGTTCTT
TTAATTCTGATTAGTGTGACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAA
TTTCTCACGGTCATATTCAGAAGTTGTTTCTTCAATTTGACTACGGATTTGGGCAACACGTTGTGCTAATTGCTCTTT
ATCGCCCGCACCTCAACAATAGTTGTTTCACTTTTGTGATTACAACGCGGTTAGCAGAACCCTAATTGATCAATTG
TCGTATCTTTTAAATCAAGTCTAAATCTTCAGTAATTACTTGACCACCTGTTAGGATAGATAAAATCTTCAACATT
GCTTTACGACGATCAACAAAGGCTGGTGTCTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAATCTTTTTGTCTGTTAATAAAAATATAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCAATTGTCAGTTACCATATATTGAGATAGGTAACCACGGTCAAATTGCATACCCTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTCTTCAATGGTGATGACGCCATCTTGCCCAACTTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTACCTGATGAAATGTGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAAGTAAATTT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTTCATACCTTCAGTTACAATTGCTTGAGTTAGAACAGTGGCAGTTGTTGTACCGTCACCGGCAATATCA
TTTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCAATTTCAATTTCTTTAGCG
ATAGTTACACCATCGTTAGTGTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTTACCCTATCAGCTAATATATCAATCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAAATTTCAATTTGTTATTAATCTGTAAATATTGATAAATTGACTAATCT
ACAATTGCGATAATATCGCTTTCTTTAATAATTAATAATTCTTGACCTTCATATTCAAATGTTGAACCTGCATATTTT
TCAAATAGAACGCGGTGCGCTTCTTTCAATTGTGATGTGAAATCTCCGTATTCTCAGCAACTGCTTCAACAACACC
TACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCAC
AACTGCGCGTCCGTTAATGGTTAATCATTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA
>700406_GroEL
TTAATACATACTGCCTGGATCCATTGCTGGCATTGTGGTTGAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCATTTTGCATGCTGAGCGAGATACTTTAGTTGGATCAACAATACC
TTCTTCAATCATGTTTACCCATTGAGAAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCGCGATTTGACGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAAGTTGCTCATCACCAGCTAATGTTTACGCTAATTCAGCAACTGCTTCTTAGCATTGATAAAGGCAGTACC
ACCACCGTACGATACCTTGGTCAACAGCCGACGTGTTGCGTTTTAAAGCATCTTCGATACGTAATTTACGTTCTT
TTAATTCTGATTAGTGTGACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAGTAA
TTTATCAGGTCATATTGAGAAAGTTGTTTCTTCAATTTGACTACGGAATTTGGGCAACACCTGTTGCTAATTTGCTCTT
ATCACCCGACCCCTCAACAATAGTTGTTTCACTTTTGTGATTACTACGCGGTTAGCAGAACCCTAATTGATCAATTG
TCGTATCTTTTAAATCAAGTCTAAATCTTCAGTAATTACTTGACCACCTGTTAGGATAGATAAAATCTTCAACATT
GCTTTACGACGGTCACCAAGGCTGGTGTCTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTGAACAACCTGTTTCTAAA
ACTGGTAAGATGCTTGGATATTAGAAATCTTTTTGTCTGTTAATAAAAATATAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCAATTATCAGTTACCATATATTGAGATAGGTAACCGCGGTCAAATTGCATACCTTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTCTTCAATAGTGATGACGCCATCTTGCCCAACTTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTGCGCTGATGAAATGTGCAACGTTAGCGATTGCACCTTTAGAATCTACAGGAAGTAAATTT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTTCATACCTTCAGTTACAATTGCTTGAGTTAGAACAGTGGCAGTTGTTGTACCGTCACCGGCAATATCA
TTTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCAATTTCAATTTCTTTGCGG
ATGGTTACACCATCGTTAGTGTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTTACCCTATCAGCTAATATATCAATCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAAATTTCAATTTGTTATTAATCTGTAAATATTGATAAATTGACTAATCT
ACAATTGCGATAATATCGCTTTCTTTAATAATTAATAATTCTTGACCATCATATTCAAATGTTGAACCTGCATATTT
TTCAAATAGAACGCGGTGCGCTTCTTTCAATTGTGATGTGAAATCTCCGTATTCTCAGCAACTGCTTCAACAACAC
CTACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCA
CAACTGCGCGTCCGTTAATGGTTAATCATTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA
A

>AVC_GroEL

TTAATACATACCTGCACTTGGATCCATTGCTGGCATTGTTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGGTCAACGATACC
TTCTTCAATCATGTTTACCCATTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCAGCGATTTGACGGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTTTTAGCATTGATAAAGGCAGTACC
ACCACCCGCTACAATACCTTGGTCAACAGCCGCACGTGTTGCGTTTAAAGCATCTTCGATACGTAATTTACGTTCTT
TTAATTTCTGATTAGCTGCACCAACTTTAAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAA
TTTCTCACGGTCATATTCAGAAGTTGTTTCTCAATTTGACTACGGATTTGGGCAACACGTTGTGCTAATTTGCTCTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACGCGTTAGCAGAACCCTAATTGATCAATTG
TCGTATCTTTTAATTCAAGTCCTAAATCTTCAGTAATTACTTGACCACCTGTTAGGATAGATAAATCTTCAACATT
GCTTTACGACGATCACCAAAGGCTGGCGCTTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAAATCTTTTGTCTGTTAATAAAAATAAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCATTGTGCTAGTACCATATATTGAGATAGGTAACCACGGTCAAATTCATACCCTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTCTTCAATGGTGATGACGCCATCTTGCCCAACTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTACCTGATGAAAATTGATGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTTCATACCTTCAGTTACAATTGCTTGAGTTAGAACAGTGGCAGTTGTTGTACCGTCACCGGCAATATCA
TTTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCTAATTTCAATTTCTTTAGCG
ATAGTTACACCATCGTTAGTGATTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTTACCAGTATCAGCTAATATATCAATCCCAACGAACTTGGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAAATTTCAATTTGTTATTAATCTGTAAAATATTGATAAATTGACTAATCT
ACAATTGCGATAAATATCGCTTTCTTTAATAATTAATAATTCTTGACCTTCATATTCAAATGTTGAACCTGCATATTTT
TCAAATAGAACGCGGTTCGCCTTCTTTCAATTGTGATGTGAAATCTCCGTATTCTCAGCAACTGCTTCAACAACACC
TACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCAC
AACTGCGCGTCCGTTAATGGTTAATCATTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA

>Nfld_GroEL

TTAATACATACCTGCACTTGGATCCATTGCTGGCATTGTTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGGTCAACGATACC
TTCTTCAATCATGTTTACCCATTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCAGCGATTTGACGGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTTTTAGCATTGATAAAGGCAGTACC
ACCACCCGCTACAATACCTTGGTCAACAGCCGCACGTGTTGCGTTTAAAGCATCTTCGATACGTAATTTACGTTCTT
TTAATTTCTGATTAGCTGCACCAACTTTAAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAA
TTTCTCACGGTCATATTCAGAAGTTGTTTCTCAATTTGACTACGGATTTGGGCAACACGTTGTGCTAATTTGCTCTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACGCGGTTAGCAGAACCCTAATTGATCAATTG
TCGTATCTTTTAATTCAAGTCCTAAATCTTCAGTAATTACTTGACCACCTGTTAGGATAGATAAATCTTCAACATT
GCTTTACGAGATCACCAAAGGCTGGCGCTTTTACCCCAACGACGTTGAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTTGAACAACCTGTTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAAATCTTTTGTCTGTTAATAAAAATAAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCATTGTGCTAGTACCATATATTGAGATAGGTAACCACGGTCAAATTCATACCCTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTCTTCAATGGTGATGACGCCATCTTGCCCAACTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTACCTGATGAAAATTGATGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTTAGCACCAGCAG
TAACGTTTTTTCATACCTTCAGTTACAATTGCTTGAGTTAGAACAGTGGCAGTTGTTGTACCGTCACCGGCAATATCA
TTTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCTAATTTCAATTTCTTTAGCG
ATAGTTACACCATCGTTAGTGATTAGCGGTGAACCATAACTTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTTACCAGTATCAGCTAATATATCAATCCCAACGAACTTGGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAAATTTCAATTTGTTATTAATCTGTAAAATATTGATAAATTGACTAATCT
ACAATTGCGATAAATATCGCTTTCTTTAATAATTAATAATTCTTGACCTTCATATTCAAATGTTGAACCTGCATATTTT
TCAAATAGAACGCGGTTCGCCTTCTTTCAATTGTGATGTGAAATCTCCGTATTCTCAGCAACTGCTTCAACAACACC
TACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCAC
AACTGCGCGTCCGTTAATGGTTAATCATTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA

>Rabins_GroEL

TTAATACATACCTGCACTTGGATCCATTGCTGGCATTGTTGGTTTCAGGTTGCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCGTTTTGCAATGCTGAGCGAGATACTTTAGTTGGGTCAACGATACC
TTCTTCAATCATGTTTACCCATTCAGAAGTTGCTGCATTAATCCGACACCTTCTGCTTGTCTTTCAATTTCTCAAC
AATTACAGAACCTTCTAAACCAGCGTTCTCAGCGATTTGACGGACTGGAGCTTCTAAAGCACGTAAGACGATTTGT
GCACCAGTTTGTTCATCACCAGCTAATGTTTCAGCTAATTCAGCAACTGCTTTTTTAGCATTGATAAAGGCAGTACC
ACCACCCGCTACAATACCTTGGTCAACAGCCGCACGTGTTGCGTTTAAAGCATCTTCGATACGTAATTTACGTTCTT
TTAATTTCTGATTAGCTGCACCAACTTTAAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAA
TTTCTCACGGTCATATTCAGAAGTTGTTTCTCAATTTGACTACGGATTTGGGCAACACGTTGTGCTAATTTGCTCTT
ATCGCCCGCACCTCAACAATAGTTGTTTCATCTTTTGTGATTACAACGCGGTTAGCAGAACCCTAATTGATCAATTG

TCGTATCTTTTAATTCAAGTCCTAAATCTTCAGTAATTACTTGACCACCTGTTAGGATAGATAAAATCTTCTAACATT
GCTTTACGACGATCACCAAAGGCTGGCGCTTTTACCCCAACGACGTTGAAAAGTACCACGGATTTTATTCAAGACTA
ATGTAGGTAATGCTTCACCATCAATATCATCAGCTACTAATAATAATGCACGACCTTCTTGAACAACCTTGTCTAGA
ACTGGCAAGATGTCTTGGATATTAGAAATCTTTTGTCTGTTAATAAAAATAAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCAATTGTCAGTTACCATATATTAGATAGGTAACCACGGTCAAATTGCATACCCTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTCTTCAATGGTGATGACGCCATCTTGGCCAACCTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTACCTGATGAAAATTGATGCAACGTTAGCGATTGCGCCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGAAGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTAGCACCAGCAG
TAACGTTTTTCATACCTTCAGTTACAATTGCTTGAGTTAGAACAGTGGCAGTTGTTGTACCGTCACCGGCAATATCA
TTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCAATTCATTTCTTTAGCG
ATAGTTACACCATCGTTAGTGATTAGCGGTGAACCATAAATTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTACCGTATCAGCTAATATATCAATCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAATTTCAATTTGTTATTAATCTGTAAATATTGATAAATTGACTAATCT
ACAATTGCGATAATATCGCTTTCTTTAATAATTAATAATTCTTGACCTTCATATTCAAATGTTGAACCTGCATATTTT
TCAAATAGAACGCGGTTCGCCTTCTTTCAATTGTGATGTGAAATCTTCCGTATTCTCAGCAACTGCTTCAACAACAC
TACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCAC
AACTGCGCGTCCGTTAATGGTTAATCATTTACATGTCCCTCCATTTATATAAAAATATCTGTCAAACCCAAGCTAA
>reference GroEL
TTAATACATACCTGCACCTGGATCCATTGCTGGCATTGTGGTTTTCGGTTGGCTTGCAACAGCTGCTTCAG
TAGATAAGATTAATCCAGCTACTGAACCTGCGTTTTGCAATGCTGAGCGAGAACTTTTGTGGGTCAACAATTCC
TTCTTCAATCATGTTTACCATTGAGAAGTTGCTGCATTAACAACCAATGCCGTCTGCTTGTCTTTTAATTTTCAAC
AATAACAGATCTTCTAAACACGACTTCTCAGCGATTGACGTAGTCTTAAAGCAGCTGAAGCAGTATTGTG
GCACCACTTCGCTTCAACCACTAATGTTTTCAGTAATTCAGCAACTGCTTTTTAGCATTGACGAAGGCAGTACC
ACCACCAGCTACGATACCTTGATCAACAGCCGCACGTGTGGCATTAAAGCATCTTCGATACGTAATTTACGTTCTT
TTAATTCTGACTCAGTAGCTGCACCAACTTTAACAACAGCAACACCACCAGCTAATTTAGCTAAGCGTTCTAATAA
TTTCTCACGGTCATATTCAGAAGTTGTTTCTTCAATTTGAGTACGGATTTGGGCAACACGTTGTGCTAATTGCTCTT
ATCGCCCGCACCTCAACAATAGTTGTTTCACTTTTGTGATTACAACGCGGTTAGCAGAACCTAATTGGTCAATTG
TCGTATCTTTTAATTCAAGTCTAAGTCTTCAAGTAATAAATTTGACCACCTGTTAGGATAGATAGGTCTTCTAACAT
GCTTTACGACGGTCACCAAAGGCTGGTGTCTTTACCCCAACAACGTTGAAAAGTACCACGGATTTTATTTAAAACTA
GTGTTGGTAATGCTTACCGTCAATATCGTCAGCTACCAATAGTAATGCACGACCTTCTTGAACAACCTTGTCTAGA
ACTGGCAAGATGCTTGGATATTAGAAATCTTTTATCTGTTAATAAAAATAAAGGATTTTCTAAAACGGCTTCCAT
CTTGTCAATTGTCAGTTACCATATATTAGATAGGTAACCACGGTCAAATTGCATACCCTCAACAACATCTAGAGAG
GTATCAATTGATTGAGATTCTTCAATAGTGATGACGCCATCTTGGCCAACCTTTTCCATTGCTTCTGCAATTAAGTC
ACCAACTTCGCTGTTCGCTGATGAAATTGATGCAACGTTAGCGATAGCGCCTTTAGAATCTACAGGAACCTGAAATT
TCAGATAAACGATCTGACGCTACGCGAATTGCTTTGTCAATCCCGCGACGAACACCTACTGGGTAGCACCAGCAG
TAACGTTTTTCATACCTTCAGTTACAATTGCTTGTGTTAAAACAGTTGCAGTTGTTGTACCGTCACCGGCAATATCA
TTGTTTTTGAAGCTACTTCAGCAACTAGTTTTGCACCCATATCTTCAAATTTATCTTCAATTCATGCTTTAGCG
ATAGTTACACCATCGTTAGTGATTAGCGGTGAACCATAAATTTGGTCTAAAACAACATTACGACCTTTTGGTCCTA
AAGTTACTTTACCGTATCAGCTAATATCAATCCACGAACCATTGATTGACGTGCATCTTCTGAAAATTTAATA
TCTTTTGCCATTTATTCATTCACCTCATAAAATTTATATTTTATTAATCTGTAAATATTGATAAATTGACTAATCT
ACAATTGCGATAATATCGCTTTCTTTAATAATTAATAATTCTTGACCATCATATTCAAATGTTGAACCTGCATATTT
TTCAAATAGAACGCGGTTCGCCTTCTTTCAATTGTGATGTGAAATCTTCCGTATTCTCAGCAACTGCTTCAACAACAC
CTACTTGTGTTTCTCTTTAGCAGCTGACGGTAATACGATACCTGAAACTGTTTGTCTTTAGCCTCTTCAACTCTCA
CAACTGCGCGTCCGTTAATGGTTAATCATTTACATGTCCCTCCATTTAGATAAAAATTTTGTCAAACCCAAGCTA
A

R code for Figure 1

call up the data set required and assign x and y for the figure, also say what column you want to base the colours on. Also setting to log scale to make for better visualization

```
snp_plot=ggplot(snp_numbers,aes(x=Stage,y=log10(Variant_Count),fill=Type))+
  geom_bar(stat="identity")+
  Facet by strain to compare
  facet_grid(~Strain)+
  Set the grey background
  theme_gray()+
  centre the title and move the x axis labels
  theme(axis.text.x = element_text(angle = 270, hjust = 1),plot.title = element_text(hjust = 0.5))+
  set the title
  ggtitle("Variant Composition at Each Stage of Filtration")+
  Set axis labels
  xlab("Stage of Filtration")+
  ylab("Variant Counts")+
```

Set the colours to be used for each type of SNP
 scale_fill_brewer(type="seq",palette = "Blues",direction=1)

Code to Create Figure 2

Set up is based off of code written by Yan Holtz [99]

Load Data

```
don3 <- Avirulent_plot %>%
```

Compute chromosome size

```
group_by(contig_num) %>%
summarise(chr_len3=max(pos)) %>%
```

Calculate cumulative position of each chromosome

```
mutate(tot3=cumsum(chr_len3)-chr_len3) %>%
select(-chr_len3) %>%
```

Add this info to the initial dataset

```
left_join(Avirulent_plot, ., by=c("contig_num"="contig_num")) %>%
```

Add a cumulative position of each SNP

```
arrange(contig_num, pos) %>%
mutate( poscum3=pos+tot3)
```

```
axisdf3 = don3 %>% group_by(contig_num) %>% summarize(center=( max(poscum3) + min(poscum3)/2 ))
```

Repeat for virulent

```
don4 <- Virulent_plot %>%
```

```
# Compute chromosome size
group_by(contig_num) %>%
summarise(chr_len4=max(pos)) %>%
```

```
# Calculate cumulative position of each chromosome
mutate(tot4=cumsum(chr_len4)-chr_len4) %>%
select(-chr_len4) %>%
```

Add this info to the initial dataset

```
left_join(Virulent_plot, ., by=c("contig_num"="contig_num")) %>%
```

Add a cumulative position of each SNP

```
arrange(contig_num, pos) %>%
mutate( poscum4=pos+tot4)
```

```
axisdf4 = don4 %>% group_by(contig_num) %>% summarize(center=( max(poscum4) + min(poscum4)/2 ))
```

Plotting

```
ggplot()+
```

```
geom_line(data=don3, aes(x=poscum3, y=(-Average_Counts),colour="skyblue"))+
```

```
geom_line(data=don4,aes(x=poscum4,y=Average_Counts,colour="darkgrey"))+
```

```
#add in the Virulent's frequencies and set it's colour and transparency
```

```
# custom X axis:
```

```
#to figure out what axis labels match up with what number, you need to look at the axisdf dataframe
```

```
scale_x_continuous(labels=c("10","20","30","40","50"),breaks=c(235600,572410,1015860,1262730,1418370),expand=c(0,0))
```

```
+
```

```
scale_y_continuous(labels=c("20","15","10","5","0","5","10","15","20"),breaks=c(-20,-15,-10,-5,0,5,10,15,20),expand = c(0,
```

```
1)) + # remove space between plot area and x axis and also change the axis label so that there is no negative sign
```

```
#do this in order to get a legend for what colour is what strain
```

```
scale_colour_manual(name = 'Strain', values =c('skyblue','darkgrey'='darkgrey'), labels = c('Virulent','Avirulent'))+
```

```
guides(colour=guide_legend(override.aes = list(size = 2)))+
```

```
xlab("Contig")+
```

```
ylab("Variants/100bp")+
```

```
ggtitle("Variant Frequency/100bp")+
```

```

# Custom the theme:
theme_bw() +
theme(panel.grid.minor.y = element_blank(),#remove some more gridlines
axis.text.x=element_text(colour="black"), #change the colour of the x axis text to black
axis.text.y=element_text(colour="black"), #change the colour of the y axis text to black
plot.title = element_text(hjust = 0.5),# centre the plot title(specified by ggtitle)
text=element_text(size=12,family="serif"),#make the front times new roman
panel.border = element_blank(),#no outline around plot
panel.grid.major.x = element_blank(), #remove some gridlines
panel.grid.minor.x = element_blank()
)

```

Figure 2 code modified for Freedman's test analysis shown in Figure 4

```

cbpalette=c("#000000", "#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
ggplot()+
geom_line(data=don3, aes(x=poscum3, y=(-Average_Counts),colour=PCA_Group))+
geom_line(data=don4,aes(x=poscum4,y=Average_Counts,colour=PCA_group))+
#add in the Virulent's frequencies and set it's colour and transparency
# custom X axis:
#to figure out what axis labels match up with what number, you need to look at the axisdf dataframe
scale_x_continuous(labels=c("10", "20", "30", "40", "50"),breaks=c(235600,572410,1015860,1262730,1418370),expand=c(0,0))
+
scale_y_continuous(labels=c("20", "15", "10", "5", "0", "5", "10", "15", "20"),breaks=c(-20,-15,-10,-5,0,5,10,15,20),expand = c(0,
1)) + # remove space between plot area and x axis and also change the axis label so that there is no negative sign
scale_colour_manual(values=cbpalette, name="PCA Group")+
#make ths ymbols in the legend a little bigger so they're easier to see
guides(colour=guide_legend(override.aes = list(size = 2)))+
xlab("Contig")+
ylab("Variants/100bp")+
ggtitle("Variant Frequency/100bp")+
# Custom the theme:
theme_bw() +
theme(panel.grid.minor.y = element_blank(),#remove some more gridlines
axis.text.x=element_text(colour="black"), #change the colour of the x axis text to black
axis.text.y=element_text(colour="black"), #change the colour of the y axis text to black
plot.title = element_text(hjust = 0.5),# centre the plot title(specified by ggtitle)
text=element_text(size=12,family="serif"),#make the front times new roman
panel.border = element_blank(),#no outline around plot
panel.grid.major.x = element_blank(), #remove some gridlines
panel.grid.minor.x = element_blank() )

```

Script and output for Friedman Analysis done in R v.3.5.1

```

Friedman_data=read.csv("F_data copy.csv")
library(agricolae)
friedman(Friedman_data$Strain,Friedman_data$Group,Friedman_data$counts.mean,console=T)

```

Study: Friedman_data\$counts.mean ~ Friedman_data\$Strain + Friedman_data\$Group

Friedman_data\$Group, Sum of the ranks

```

Friedman_data.counts.mean r
A          39 10
B          29 10
C          10 10
D          59 10
E          70 10
F          23 10
G          50 10

```

Friedman's Test

Adjusted for ties
 Critical Value: 57.25714
 P.Value Chisq: 1.620524e-10
 F Value: 187.875
 P.Value F: 0

Post Hoc Analysis

Alpha: 0.05 ; DF Error: 54
 t-Student: 2.004879
 LSD: 4.365272

Treatments with the same letter are not significantly different.

Sum of ranks groups
 E 70 a
 D 59 b
 G 50 c
 A 39 d
 B 29 e
 F 23 f
 C 10 g

First Friedman test on all of the data, not including pathogenicity, now making it is own dataset to manipulate later for P value corrections

orig_groups=friedman(Friedman_data\$Strain,Friedman_data\$Group,Friedman_data\$counts.mean,console=T,group=F)

Study: Friedman_data\$counts.mean ~ Friedman_data\$Strain + Friedman_data\$Group

Friedman_data\$Group, Sum of the ranks

Friedman_data.counts.mean r
 A 39 10
 B 29 10
 C 10 10
 D 59 10
 E 70 10
 F 23 10
 G 50 10

Friedman's Test

Adjusted for ties
 Critical Value: 57.25714
 P.Value Chisq: 1.620524e-10
 F Value: 187.875
 P.Value F: 0

Post Hoc Analysis

*Comparison between treatments
 Sum of the ranks*

	difference	pvalue	signif.	LCL	UCL
A - B	10	0e+00	***	5.63	14.37
A - C	29	0e+00	***	24.63	33.37
A - D	-20	0e+00	***	-24.37	-15.63
A - E	-31	0e+00	***	-35.37	-26.63
A - F	16	0e+00	***	11.63	20.37
A - G	-11	0e+00	***	-15.37	-6.63
B - C	19	0e+00	***	14.63	23.37
B - D	-30	0e+00	***	-34.37	-25.63
B - E	-41	0e+00	***	-45.37	-36.63
B - F	6	8e-03	**	1.63	10.37
B - G	-21	0e+00	***	-25.37	-16.63

```

C - D  -49 0e+00  *** -53.37 -44.63
C - E  -60 0e+00  *** -64.37 -55.63
C - F  -13 0e+00  *** -17.37 -8.63
C - G  -40 0e+00  *** -44.37 -35.63
D - E  -11 0e+00  *** -15.37 -6.63
D - F   36 0e+00  ***  31.63 40.37
D - G   9 1e-04  ***   4.63 13.37
E - F   47 0e+00  ***  42.63 51.37
E - G   20 0e+00  ***  15.63 24.37
F - G  -27 0e+00  *** -31.37 -22.63

```

orig_groups

\$statistics

```

Chisq Df  p.chisq  F DFErorr p.F t.value  LSD
57.25714 6 1.620524e-10 187.875  54 0.2004879 4.365272

```

\$parameters

```

test      name.t ntr alpha
Friedman Friedman_data$Group 7 0.05

```

\$means

```

Friedman_data.counts.mean rankSum  std r  Min  Max
A      0.5206515    39 0.06016552 10 0.3681623 0.5795943
B      0.4262246    29 0.06112114 10 0.3064516 0.5322581
C      0.3147643    10 0.04455880 10 0.2258065 0.3734491
D      0.6710127    59 0.10059328 10 0.4572468 0.8126137
E      0.9271045    70 0.14868601 10 0.6573134 1.1522388
F      0.3986577    23 0.06980879 10 0.3172666 0.5448444
G      0.5882955    50 0.07837898 10 0.4721591 0.7244318

```

```

Q25  Q50  Q75

```

```

A 0.5095268 0.5304241 0.5599263
B 0.4051672 0.4187575 0.4650538
C 0.2924938 0.3117245 0.3506514
D 0.6221953 0.6658581 0.7469679
E 0.8446269 0.8991045 1.0532836
F 0.3502135 0.3730933 0.4344112
G 0.5545455 0.5758523 0.6355114

```

\$comparison

```

difference pvalue signif.  LCL  UCL
A - B      10 0e+00  ***   5.63 14.37
A - C      29 0e+00  ***  24.63 33.37
A - D     -20 0e+00  *** -24.37 -15.63
A - E     -31 0e+00  *** -35.37 -26.63
A - F      16 0e+00  ***  11.63 20.37
A - G     -11 0e+00  *** -15.37 -6.63
B - C      19 0e+00  ***  14.63 23.37
B - D     -30 0e+00  *** -34.37 -25.63
B - E     -41 0e+00  *** -45.37 -36.63
B - F       6 8e-03  **   1.63 10.37
B - G     -21 0e+00  *** -25.37 -16.63
C - D     -49 0e+00  *** -53.37 -44.63
C - E     -60 0e+00  *** -64.37 -55.63
C - F     -13 0e+00  *** -17.37 -8.63
C - G     -40 0e+00  *** -44.37 -35.63
D - E     -11 0e+00  *** -15.37 -6.63
D - F      36 0e+00  ***  31.63 40.37
D - G       9 1e-04  ***   4.63 13.37
E - F      47 0e+00  ***  42.63 51.37
E - G      20 0e+00  ***  15.63 24.37
F - G     -27 0e+00  *** -31.37 -22.63

```

\$groups

NULL

attr(,"class")

```
[1] "group"
```

Making one dataset with just the P value output

```
Orig_Pvalue=orig_groups$comparison
```

```
Orig_Pvalue_red=Orig_Pvalue[,c(1,2)]
```

```
Orig_Pvalue_red
```

```
  difference pvalue
A - B      10 0e+00
A - C      29 0e+00
A - D     -20 0e+00
A - E     -31 0e+00
A - F      16 0e+00
A - G     -11 0e+00
B - C      19 0e+00
B - D     -30 0e+00
B - E     -41 0e+00
B - F       6.8e-03
B - G     -21 0e+00
C - D     -49 0e+00
C - E     -60 0e+00
C - F     -13 0e+00
C - G     -40 0e+00
D - E     -11 0e+00
D - F      36 0e+00
D - G       9.1e-04
E - F      47 0e+00
E - G      20 0e+00
F - G     -27 0e+00
```

Correcting the P values

```
Orig_Pvalue_red$Bonferroni_All=p.adjust(Orig_Pvalue_red$pvalue,method="bonferroni")
```

```
Orig_Pvalue_red
```

```
  difference pvalue Bonferroni_All
A - B      10 0e+00      0.0000
A - C      29 0e+00      0.0000
A - D     -20 0e+00      0.0000
A - E     -31 0e+00      0.0000
A - F      16 0e+00      0.0000
A - G     -11 0e+00      0.0000
B - C      19 0e+00      0.0000
B - D     -30 0e+00      0.0000
B - E     -41 0e+00      0.0000
B - F       6.8e-03      0.1680
B - G     -21 0e+00      0.0000
C - D     -49 0e+00      0.0000
C - E     -60 0e+00      0.0000
C - F     -13 0e+00      0.0000
C - G     -40 0e+00      0.0000
D - E     -11 0e+00      0.0000
D - F      36 0e+00      0.0000
D - G       9.1e-04      0.0021
E - F      47 0e+00      0.0000
E - G      20 0e+00      0.0000
F - G     -27 0e+00      0.0000
```

Making a column with the row names in order to merge later if needed in order to compare

```
Orig_Pvalue_red$comp=rownames(Orig_Pvalue_red)
```

```
Orig_Pvalue_red
```

```
  difference pvalue Bonferroni_All comp
A - B      10 0e+00      0.0000 A - B
A - C      29 0e+00      0.0000 A - C
A - D     -20 0e+00      0.0000 A - D
A - E     -31 0e+00      0.0000 A - E
A - F      16 0e+00      0.0000 A - F
A - G     -11 0e+00      0.0000 A - G
B - C      19 0e+00      0.0000 B - C
B - D     -30 0e+00      0.0000 B - D
B - E     -41 0e+00      0.0000 B - E
```

```

B - F      6.8e-03    0.1680 B - F
B - G     -21.0e+00    0.0000 B - G
C - D     -49.0e+00    0.0000 C - D
C - E     -60.0e+00    0.0000 C - E
C - F     -13.0e+00    0.0000 C - F
C - G     -40.0e+00    0.0000 C - G
D - E     -11.0e+00    0.0000 D - E
D - F      36.0e+00    0.0000 D - F
D - G       9.1e-04    0.0021 D - G
E - F      47.0e+00    0.0000 E - F
E - G      20.0e+00    0.0000 E - G
F - G     -27.0e+00    0.0000 F - G

```

Reducing again to have a smaller set of P values

```
Nodiv_Pvalues=Orig_Pvalue_red[,c(4,3)]
```

```
Nodiv_Pvalues
```

```
  comp Bonferroni_All
```

```

A - B A - B    0.0000
A - C A - C    0.0000
A - D A - D    0.0000
A - E A - E    0.0000
A - F A - F    0.0000
A - G A - G    0.0000
B - C B - C    0.0000
B - D B - D    0.0000
B - E B - E    0.0000
B - F B - F    0.1680
B - G B - G    0.0000
C - D C - D    0.0000
C - E C - E    0.0000
C - F C - F    0.0000
C - G C - G    0.0000
D - E D - E    0.0000
D - F D - F    0.0000
D - G D - G    0.0021
E - F E - F    0.0000
E - G E - G    0.0000
F - G F - G    0.0000

```

```
write.csv(Nodiv_Pvalues, file="NoDivP.csv")
```

Repeating same friedman test incorporating the pathogenicities into the factors

Going to run the split plot ANOVA in order to look at the interaction

```
splitplot.fit <-
```

```
aov(Friedman_data$counts.mean~Friedman_data$Strain*Friedman_data$Group+Error(Friedman_data$Pathogen/Friedman_data$Strain))
```

```
Warning in aov(Friedman_data$counts.mean ~ Friedman_data$Strain *
```

```
Friedman_data$Group + : Error() model is singular
```

```
summary(splitplot.fit)
```

```

Error: Friedman_data$Pathogen
      Df Sum Sq Mean Sq
Friedman_data$Strain 1 9.859e-05 9.859e-05

```

```

Error: Friedman_data$Pathogen:Friedman_data$Strain
      Df Sum Sq Mean Sq
Friedman_data$Strain 8 0.3713 0.04641

```

```

Error: Within
      Df Sum Sq Mean Sq
Friedman_data$Group      6 2.5274 0.4212
Friedman_data$Strain:Friedman_data$Group 54 0.1019 0.0019

```

Saw from the ANOVA that there was a significant interaction, now including pathogenicity in fixed factor to look at the results with the friedman test

```
groupChange=read.csv("F_data.csv")
```

```
colnames(groupChange)
```

```

[1] "X"      "Group"  "Group_P" "Strain"
[5] "Pathogen" "counts.mean" "counts.count"

```

Repeating same friedman test incorporating the pathogenicities into the factors

```
f=friedman(groupChange$Strain,groupChange$Group_P,groupChange$counts.mean,console=T,group=F)
```

Study: groupChange\$counts.mean ~ groupChange\$Strain + groupChange\$Group_P

groupChange\$Group_P, Sum of the ranks

```
groupChange.counts.mean r
A_AV      58 5
A_V       61 5
B_AV      59 5
B_V       60 5
C_AV      55 5
C_V       55 5
D_AV      84 5
D_V       85 5
E_AV      95 5
E_V       95 5
F_AV      78 5
F_V       75 5
G_AV      96 5
G_V       94 5
```

Friedman's Test

```
=====
Adjusted for ties
Critical Value: 20.33143
P.Value Chisq: 0.08724872
F Value: 1.668508
P.Value F: 0.0964919
```

Post Hoc Analysis

Comparison between treatments
Sum of the ranks

	difference	pvalue	signif.	LCL	UCL
A_AV - A_V	-3	0.8687		-38.87	32.87
A_AV - B_AV	-1	0.9561		-36.87	34.87
A_AV - B_V	-2	0.9123		-37.87	33.87
A_AV - C_AV	3	0.8687		-32.87	38.87
A_AV - C_V	3	0.8687		-32.87	38.87
A_AV - D_AV	-26	0.1538		-61.87	9.87
A_AV - D_V	-27	0.1387		-62.87	8.87
A_AV - E_AV	-37	0.0433	*	-72.87	-1.13
A_AV - E_V	-37	0.0433	*	-72.87	-1.13
A_AV - F_AV	-20	0.2718		-55.87	15.87
A_AV - F_V	-17	0.3499		-52.87	18.87
A_AV - G_AV	-38	0.0381	*	-73.87	-2.13
A_AV - G_V	-36	0.0492	*	-71.87	-0.13
A_V - B_AV	2	0.9123		-33.87	37.87
A_V - B_V	1	0.9561		-34.87	36.87
A_V - C_AV	6	0.7410		-29.87	41.87
A_V - C_V	6	0.7410		-29.87	41.87
A_V - D_AV	-23	0.2067		-58.87	12.87
A_V - D_V	-24	0.1877		-59.87	11.87
A_V - E_AV	-34	0.0630	.	-69.87	1.87
A_V - E_V	-34	0.0630	.	-69.87	1.87
A_V - F_AV	-17	0.3499		-52.87	18.87
A_V - F_V	-14	0.4411		-49.87	21.87
A_V - G_AV	-35	0.0557	.	-70.87	0.87
A_V - G_V	-33	0.0710	.	-68.87	2.87
B_AV - B_V	-1	0.9561		-36.87	34.87
B_AV - C_AV	4	0.8256		-31.87	39.87
B_AV - C_V	4	0.8256		-31.87	39.87

B_AV - D_AV	-25 0.1701	-60.87 10.87
B_AV - D_V	-26 0.1538	-61.87 9.87
B_AV - E_AV	-36 0.0492	* -71.87 -0.13
B_AV - E_V	-36 0.0492	* -71.87 -0.13
B_AV - F_AV	-19 0.2963	-54.87 16.87
B_AV - F_V	-16 0.3789	-51.87 19.87
B_AV - G_AV	-37 0.0433	* -72.87 -1.13
B_AV - G_V	-35 0.0557	. -70.87 0.87
B_V - C_AV	5 0.7830	-30.87 40.87
B_V - C_V	5 0.7830	-30.87 40.87
B_V - D_AV	-24 0.1877	-59.87 11.87
B_V - D_V	-25 0.1701	-60.87 10.87
B_V - E_AV	-35 0.0557	. -70.87 0.87
B_V - E_V	-35 0.0557	. -70.87 0.87
B_V - F_AV	-18 0.3224	-53.87 17.87
B_V - F_V	-15 0.4093	-50.87 20.87
B_V - G_AV	-36 0.0492	* -71.87 -0.13
B_V - G_V	-34 0.0630	. -69.87 1.87
C_AV - C_V	0 1.0000	-35.87 35.87
C_AV - D_AV	-29 0.1121	-64.87 6.87
C_AV - D_V	-30 0.1003	-65.87 5.87
C_AV - E_AV	-40 0.0292	* -75.87 -4.13
C_AV - E_V	-40 0.0292	* -75.87 -4.13
C_AV - F_AV	-23 0.2067	-58.87 12.87
C_AV - F_V	-20 0.2718	-55.87 15.87
C_AV - G_AV	-41 0.0254	* -76.87 -5.13
C_AV - G_V	-39 0.0334	* -74.87 -3.13
C_V - D_AV	-29 0.1121	-64.87 6.87
C_V - D_V	-30 0.1003	-65.87 5.87
C_V - E_AV	-40 0.0292	* -75.87 -4.13
C_V - E_V	-40 0.0292	* -75.87 -4.13
C_V - F_AV	-23 0.2067	-58.87 12.87
C_V - F_V	-20 0.2718	-55.87 15.87
C_V - G_AV	-41 0.0254	* -76.87 -5.13
C_V - G_V	-39 0.0334	* -74.87 -3.13
D_AV - D_V	-1 0.9561	-36.87 34.87
D_AV - E_AV	-11 0.5448	-46.87 24.87
D_AV - E_V	-11 0.5448	-46.87 24.87
D_AV - F_AV	6 0.7410	-29.87 41.87
D_AV - F_V	9 0.6202	-26.87 44.87
D_AV - G_AV	-12 0.5089	-47.87 23.87
D_AV - G_V	-10 0.5819	-45.87 25.87
D_V - E_AV	-10 0.5819	-45.87 25.87
D_V - E_V	-10 0.5819	-45.87 25.87
D_V - F_AV	7 0.6999	-28.87 42.87
D_V - F_V	10 0.5819	-25.87 45.87
D_V - G_AV	-11 0.5448	-46.87 24.87
D_V - G_V	-9 0.6202	-44.87 26.87
E_AV - E_V	0 1.0000	-35.87 35.87
E_AV - F_AV	17 0.3499	-18.87 52.87
E_AV - F_V	20 0.2718	-15.87 55.87
E_AV - G_AV	-1 0.9561	-36.87 34.87
E_AV - G_V	1 0.9561	-34.87 36.87
E_V - F_AV	17 0.3499	-18.87 52.87
E_V - F_V	20 0.2718	-15.87 55.87
E_V - G_AV	-1 0.9561	-36.87 34.87
E_V - G_V	1 0.9561	-34.87 36.87
F_AV - F_V	3 0.8687	-32.87 38.87
F_AV - G_AV	-18 0.3224	-53.87 17.87
F_AV - G_V	-16 0.3789	-51.87 19.87
F_V - G_AV	-21 0.2486	-56.87 14.87
F_V - G_V	-19 0.2963	-54.87 16.87
G_AV - G_V	2 0.9123	-33.87 37.87

Doing the same procedure as above to correct for the P values and give a column with the row names
test2=as.data.frame(f\$comparison)

```
test2$bonferroni=p.adjust(test2$pvalue,method="bonferroni")
```

```
test2$comp=rownames(test2)
```

```
test2
```

	difference	pvalue	signif.	LCL	UCL	bonferroni	comp
A_AV - A_V	-3	0.8687		-38.87	32.87	1	A_AV - A_V
A_AV - B_AV	-1	0.9561		-36.87	34.87	1	A_AV - B_AV
A_AV - B_V	-2	0.9123		-37.87	33.87	1	A_AV - B_V
A_AV - C_AV	3	0.8687		-32.87	38.87	1	A_AV - C_AV
A_AV - C_V	3	0.8687		-32.87	38.87	1	A_AV - C_V
A_AV - D_AV	-26	0.1538		-61.87	9.87	1	A_AV - D_AV
A_AV - D_V	-27	0.1387		-62.87	8.87	1	A_AV - D_V
A_AV - E_AV	-37	0.0433	*	-72.87	-1.13	1	A_AV - E_AV
A_AV - E_V	-37	0.0433	*	-72.87	-1.13	1	A_AV - E_V
A_AV - F_AV	-20	0.2718		-55.87	15.87	1	A_AV - F_AV
A_AV - F_V	-17	0.3499		-52.87	18.87	1	A_AV - F_V
A_AV - G_AV	-38	0.0381	*	-73.87	-2.13	1	A_AV - G_AV
A_AV - G_V	-36	0.0492	*	-71.87	-0.13	1	A_AV - G_V
A_V - B_AV	2	0.9123		-33.87	37.87	1	A_V - B_AV
A_V - B_V	1	0.9561		-34.87	36.87	1	A_V - B_V
A_V - C_AV	6	0.7410		-29.87	41.87	1	A_V - C_AV
A_V - C_V	6	0.7410		-29.87	41.87	1	A_V - C_V
A_V - D_AV	-23	0.2067		-58.87	12.87	1	A_V - D_AV
A_V - D_V	-24	0.1877		-59.87	11.87	1	A_V - D_V
A_V - E_AV	-34	0.0630	.	-69.87	1.87	1	A_V - E_AV
A_V - E_V	-34	0.0630	.	-69.87	1.87	1	A_V - E_V
A_V - F_AV	-17	0.3499		-52.87	18.87	1	A_V - F_AV
A_V - F_V	-14	0.4411		-49.87	21.87	1	A_V - F_V
A_V - G_AV	-35	0.0557	.	-70.87	0.87	1	A_V - G_AV
A_V - G_V	-33	0.0710	.	-68.87	2.87	1	A_V - G_V
B_AV - B_V	-1	0.9561		-36.87	34.87	1	B_AV - B_V
B_AV - C_AV	4	0.8256		-31.87	39.87	1	B_AV - C_AV
B_AV - C_V	4	0.8256		-31.87	39.87	1	B_AV - C_V
B_AV - D_AV	-25	0.1701		-60.87	10.87	1	B_AV - D_AV
B_AV - D_V	-26	0.1538		-61.87	9.87	1	B_AV - D_V
B_AV - E_AV	-36	0.0492	*	-71.87	-0.13	1	B_AV - E_AV
B_AV - E_V	-36	0.0492	*	-71.87	-0.13	1	B_AV - E_V
B_AV - F_AV	-19	0.2963		-54.87	16.87	1	B_AV - F_AV
B_AV - F_V	-16	0.3789		-51.87	19.87	1	B_AV - F_V
B_AV - G_AV	-37	0.0433	*	-72.87	-1.13	1	B_AV - G_AV
B_AV - G_V	-35	0.0557	.	-70.87	0.87	1	B_AV - G_V
B_V - C_AV	5	0.7830		-30.87	40.87	1	B_V - C_AV
B_V - C_V	5	0.7830		-30.87	40.87	1	B_V - C_V
B_V - D_AV	-24	0.1877		-59.87	11.87	1	B_V - D_AV
B_V - D_V	-25	0.1701		-60.87	10.87	1	B_V - D_V
B_V - E_AV	-35	0.0557	.	-70.87	0.87	1	B_V - E_AV
B_V - E_V	-35	0.0557	.	-70.87	0.87	1	B_V - E_V
B_V - F_AV	-18	0.3224		-53.87	17.87	1	B_V - F_AV
B_V - F_V	-15	0.4093		-50.87	20.87	1	B_V - F_V
B_V - G_AV	-36	0.0492	*	-71.87	-0.13	1	B_V - G_AV
B_V - G_V	-34	0.0630	.	-69.87	1.87	1	B_V - G_V
C_AV - C_V	0	1.0000		-35.87	35.87	1	C_AV - C_V
C_AV - D_AV	-29	0.1121	.	-64.87	6.87	1	C_AV - D_AV
C_AV - D_V	-30	0.1003		-65.87	5.87	1	C_AV - D_V
C_AV - E_AV	-40	0.0292	*	-75.87	-4.13	1	C_AV - E_AV
C_AV - E_V	-40	0.0292	*	-75.87	-4.13	1	C_AV - E_V
C_AV - F_AV	-23	0.2067		-58.87	12.87	1	C_AV - F_AV
C_AV - F_V	-20	0.2718		-55.87	15.87	1	C_AV - F_V
C_AV - G_AV	-41	0.0254	*	-76.87	-5.13	1	C_AV - G_AV
C_AV - G_V	-39	0.0334	*	-74.87	-3.13	1	C_AV - G_V
C_V - D_AV	-29	0.1121	.	-64.87	6.87	1	C_V - D_AV
C_V - D_V	-30	0.1003		-65.87	5.87	1	C_V - D_V
C_V - E_AV	-40	0.0292	*	-75.87	-4.13	1	C_V - E_AV
C_V - E_V	-40	0.0292	*	-75.87	-4.13	1	C_V - E_V
C_V - F_AV	-23	0.2067		-58.87	12.87	1	C_V - F_AV
C_V - F_V	-20	0.2718		-55.87	15.87	1	C_V - F_V

C_V - G_AV	-41 0.0254	* -76.87 -5.13	1 C_V - G_AV
C_V - G_V	-39 0.0334	* -74.87 -3.13	1 C_V - G_V
D_AV - D_V	-1 0.9561	-36.87 34.87	1 D_AV - D_V
D_AV - E_AV	-11 0.5448	-46.87 24.87	1 D_AV - E_AV
D_AV - E_V	-11 0.5448	-46.87 24.87	1 D_AV - E_V
D_AV - F_AV	6 0.7410	-29.87 41.87	1 D_AV - F_AV
D_AV - F_V	9 0.6202	-26.87 44.87	1 D_AV - F_V
D_AV - G_AV	-12 0.5089	-47.87 23.87	1 D_AV - G_AV
D_AV - G_V	-10 0.5819	-45.87 25.87	1 D_AV - G_V
D_V - E_AV	-10 0.5819	-45.87 25.87	1 D_V - E_AV
D_V - E_V	-10 0.5819	-45.87 25.87	1 D_V - E_V
D_V - F_AV	7 0.6999	-28.87 42.87	1 D_V - F_AV
D_V - F_V	10 0.5819	-25.87 45.87	1 D_V - F_V
D_V - G_AV	-11 0.5448	-46.87 24.87	1 D_V - G_AV
D_V - G_V	-9 0.6202	-44.87 26.87	1 D_V - G_V
E_AV - E_V	0 1.0000	-35.87 35.87	1 E_AV - E_V
E_AV - F_AV	17 0.3499	-18.87 52.87	1 E_AV - F_AV
E_AV - F_V	20 0.2718	-15.87 55.87	1 E_AV - F_V
E_AV - G_AV	-1 0.9561	-36.87 34.87	1 E_AV - G_AV
E_AV - G_V	1 0.9561	-34.87 36.87	1 E_AV - G_V
E_V - F_AV	17 0.3499	-18.87 52.87	1 E_V - F_AV
E_V - F_V	20 0.2718	-15.87 55.87	1 E_V - F_V
E_V - G_AV	-1 0.9561	-36.87 34.87	1 E_V - G_AV
E_V - G_V	1 0.9561	-34.87 36.87	1 E_V - G_V
F_AV - F_V	3 0.8687	-32.87 38.87	1 F_AV - F_V
F_AV - G_AV	-18 0.3224	-53.87 17.87	1 F_AV - G_AV
F_AV - G_V	-16 0.3789	-51.87 19.87	1 F_AV - G_V
F_V - G_AV	-21 0.2486	-56.87 14.87	1 F_V - G_AV
F_V - G_V	-19 0.2963	-54.87 16.87	1 F_V - G_V
G_AV - G_V	2 0.9123	-33.87 37.87	1 G_AV - G_V

R code used to indentify genes that had specific classes of variants

```
ggplot(Merged_37R)+  
geom_bar(aes(x=reorder(X.GeneName,variants_impact_MODIFIER),y=variants_effect_frameshift_variant),position="dodge",st  
at="identity")+  
  theme(axis.text.x = element_text(angle = 45,hjust=1,size=8),panel.grid.major.x = element_blank(),plot.title = element_text(hjust  
= 0.5))+  
  ggtitle("37R")+  
  xlab("SnpEff Name")
```

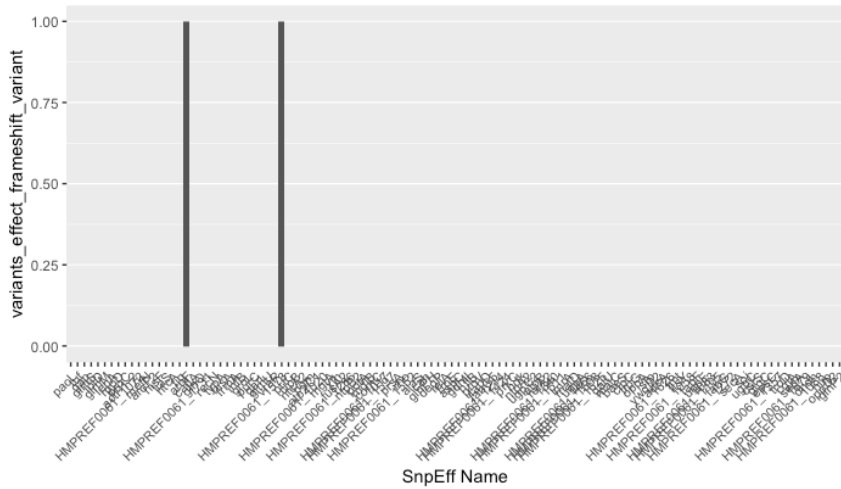


Figure 6. Frameshift Variants in 37R. Genes of impact level modifier present in avirulent strain 37R. Counts only shown for frameshift variant mutations. Plot generated in R v 3.5.1 using data from SnpEff.