Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Christopher C. Tseng April 15, 2018

Multi-omic Analysis to Define Mechanisms of Antigenic Variation in Malaria

by

Christopher C. Tseng

Mary R. Galinski, Ph. D.
Adviser

Department of Biology

Mary R. Galinski, Ph. D.

Adviser

Rustom Antia, Ph. D.

Committee Member

Jinho Choi, Ph. D.

Committee Member

2018

Multi-omic Analysis to Define Mechanisms of Antigenic Variation in Malaria

By

Christopher C. Tseng

Mary R. Galinski, Ph. D.

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Biology

2018

Abstract

Multi-omic Analysis to Define Mechanisms of Antigenic Variation in Malaria
By Christopher C. Tseng

During *Plasmodium knowlesi* malarial parasitic infections, distinctive variations of antigens are made by the parasite and are presented at the surface of infected erythrocytes. These antigens are significant to malaria pathogenesis, since they can allow infected erythrocytes to avoid detection by the host immune system, making them a potential target for vaccine development. Due to the important role SICA (schizont-infected cell agglutination) antigens play in the virulence of the malaria parasite *P. knowlesi*, studying the conditions for the expression of the *SICAvar* genes can lead to a clearer picture of how variant antigens contribute to malaria pathogenesis. We are especially curious about how *P. knowlesi* establishes and expresses five different cloned phenotypes depending on host conditions, like a missing spleen in SICA(-) parasites, or having been cured of a past infection by *P. knowlesi* A or B clones, then reinfected (B and C cloned phenotypes, respectively). To investigate this, we first used long read results from PacBio next generation sequencing to confirm high coverage of the recently generated *P. knowlesi* B and C genome assemblies. With completed *P. knowlesi* A, B, and C genomes, RNA-Seq data of *P. knowlesi* A+, B+, and C+ clones, as well as between A+ and A- clones, are compared to determine if there were any significant, large-scale differences in gene expression, perhaps due to switching at the genomic levels, which could result in the expression of these different protein repertoires. Finally, we apply different machine learning techniques to analyze the RNA-Seq data and further explore how they can be utilized to detect additional patterns of association between parasitic genes from the data. Building upon our recent efforts

in developing the first PacBio-based *Plasmodium* genome sequence and studying *P. knowlesi* gene expression through transcriptomic analysis, we use the established *P. knowlesi* model system to gain novel insights into the underlying causes behind antigen variability and virulence. By better understanding how the parasite adapts to specific host environments, we can contribute to the development of more effective control measures and the eventual eradication of malaria.

Multi-omic Analysis to Define Mechanisms of Antigenic Variation in Malaria

By

Christopher C. Tseng

Mary R. Galinski, Ph. D.

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Biology

2018

Acknowledgements

I would like to wholeheartedly thank my adviser Dr. Mary Galinski for mentoring me over the past several years, for continuing to enthusiastically encourage my research aspirations in systems biology, and for her incredible support throughout every step of this thesis. I also would like to thank Dr. Rustom Antia and Dr. Jinho Choi for taking the time to be on my honors thesis committee. Special thanks to Dr. Jung-Ting Chien, Mr. Stacey Lapp, and the Malaria Host-Pathogen Interaction Center (MaHPIC) Consortium for their technical advice and guidance that helped this thesis come to fruition.

This study builds upon collaborations with Dr. Julian Rayner, Dr. Lia Chappell, and colleagues at the Sanger Institute, Cambridge, UK. Moreover, this research was supported in part by Federal funds from the U.S. National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract # HHSN272201200031C, which has supported MaHPIC.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Malaria: The disease

Malaria is one of the deadliest diseases in human history. While significant progress has been made against malaria since its eradication was declared a United Nations Millennium Development Goal in 2000, there remains numerous regions of the world where human malaria continues to be transmitted and is endemic, namely Africa, Southeast Asia, the Eastern Mediterranean, Western Pacific, and Central and South America (WHO World Malaria Report 2017). Though incidences of malaria have been on the decline, there was still an estimated 216 million cases worldwide in 2016, with 445,000 deaths (WHO World Malaria Report 2017). Moreover, poorer, tropical areas have had to withstand the worst of the disease, with the majority of cases and deaths being concentrated in Africa and Southeast Asia (Sachs and Malaney 2002, WHO World Malaria Report 2017). Out of the five known malaria parasite species known to cause malaria in humans, *P. falciparum* remains the deadliest, accounting for 99% of deaths (WHO World Malaria Report 2016). In response to this continued threat, much funding has been put towards fighting malaria. Worldwide expenditures to control and eliminate malaria in 2016 total about $2.7 billion for preventative measures like mosquito nets and insecticides, treatment, and logistics costs, with $572 million spent in 2015 on funding malaria research, mostly focused on the discovery of new antimalarial drugs and vaccines (WHO World Malaria Report 2017). Moreover, there are socioeconomic costs associated with malaria that are not as easily measured. According to Sachs and Malaney (2002), there are

social costs like changes in family behavior and local demographics, and macroeconomic costs like loss of trade opportunities and lower rates of economic development that can be associated with the prevalence of malaria in a society. As a result, malaria remains an important global health issue facing our world today.

1.2 Malaria biology

The disease is caused by parasites of the genus *Plasmodium*, a protist defined by several distinct characteristics. It has an "apical complex" structure at one end of the organism, used for host cell invasion, and a special plastid organelle called the apicoplast (Escalante and Ayala 1995, Waller et al 1998). Moreover, the parasite demonstrates merogony, or asexual reproduction in host red blood cells, as well as produces the pigment hemozoin from metabolizing hemoglobin (Rich and Xu 2011). Though *Plasmodium* parasites are known to infect many different types of organisms, there are specifically five species of malaria parasites that have been identified to infect humans: *P. falciparum, P. vivax, P. ovale, P. malariae,* and *P. knowlesi,* with *P. falciparum* and *P. vivax* being the most common causes of malaria worldwide (WHO World Malaria Report 2016). The parasite is bloodborne, and it is transmitted via bites from a female *Anopheles* mosquito vector (Francis et al 1997). In this manner, its life cycle encompasses two hosts, the human and the female *Anopheles* mosquito (Francis et al 1997). In the mosquito gut, a male microgamete fertilizes a female macrogamete to produce a zygote, which eventually matures into an ookinete (Josling and Llinas 2015). Ookinetes are motile and migrate to the stomach lining of the mosquito to form an oocyst, which eventually bursts to

release haploid sporozoites that enter the mosquito salivary glands ready to be transmitted to the next human host (Josling and Llinas 2015). Thus, after sexual reproduction occurs inside the mosquito, sporozoites can then be transmitted to the next human the host mosquito feeds on by being deposited along with the mosquito's saliva into the site of the bite in the human skin (Frevert 2004). From there, the sporozoites can then migrate to the liver and infect the new human host's liver cells (Miller et al 2002). Once successful replication has occurred in the liver cells, the parasite will cause lysis of infected liver cells, releasing 20,000 merozoites per liver cell, which will then move on to start infecting red blood cells (Fonseca et al 2017). This leads to the human blood stage, or intraerythrocytic stage, which will cycle multiple times as the parasite replicates itself, producing upwards of 32 merozoites per red blood cell (Francis et al 1997, Cowman and Crabb 2006). This intraerythrocytic stage is the most well documented of all the stages of the malaria life cycle, deservedly so due to it being the symptomatic phase of malaria (Miller et al 2002; WHO, "Malaria vaccine: WHO position paper").

1.3 Intraerythrocytic cycle

The intraerythrocytic cycle itself is composed of multiple stages. During the merozoite stage, red blood cell invasion occurs, with the parasite displaying surface proteins for recognizing red blood cells, an actin-myosin motor for guiding the merozoite into the correct orientation for invasion, and apical organelle proteins for actually penetrating the red blood cell membrane (Florens et al 2002). During the trophozoite stage, the parasite undergoes a resource intensive phase where it is focused on nutrient acquisition and macromolecule

synthesis in order to replicate quickly inside the host cell, expressing transport proteins to mobilize nutrients in the host's cytoplasm for parasitic metabolism (Florens et al 2002). At this time, the parasite may develop a food vacuole with an acidic inner environment to digest intracellular contents, especially hemoglobin (Francis et al 1997). The parasite also produces new organelles for synthesizing antigens to be displayed on the host cell's surface membrane, like PfEMP1 from *P. falciparum*, or proteins that assist in their transport (Maier et al 2009). In the subsequent schizont phase, parasites inside infected red blood cells undergo asexual replication to produce more individual parasites in the merozoite form, which will lyse the host cell and spread through the bloodstream to infect new red blood cells (Josling and Llinas 2015). In the case of *P. falciparum*, less than 10% of merozoites produced from a replicating parasite inside an infected red blood cell will commit to sexual development instead of continuing to develop into mature asexual merozoites (Josling and Llinas 2015). These sexually differentiated parasites will develop into male or female gametocytes before they can be transmitted to the mosquito vector (Josling and Llinas 2015).

1.4 Symptoms, Treatment, Vaccine

Throughout the human blood stage of the disease, and as the parasite passes through multiple iterations of its life cycle, the infected patient will typically experience different symptoms indicative of classical malaria, including fever, chills, nausea, vomiting, and aching pain (CDC, "Malaria – Disease"). In more serious cases, typically caused by *P. falciparum* infections, severe malaria can result in impaired consciousness, severe anemia, renal

impairment, and possibly organ failure (WHO, “Severe Malaria”). As such, the blood stage has been the focal point of treatments for malaria. One common approach has been the drug chloroquine, which is still recommended by the CDC and the WHO for treatment of *P. vivax* infections (CDC, “Treatment of Malaria: Guidelines For Clinicians (United States)”; WHO, “Malaria – Overview of malaria treatment”). Chloroquine functions as an antimalarial drug by inhibiting the parasitic crystallization of the toxic molecule heme into the nontoxic hemozoin when inside an infected red blood cell, causing an increase in the amount of toxic heme present in the infected cell and premature destruction of it with the parasite (Hempelmann 2007). By the 1990s, though, chloroquine had become less effective as chloroquine resistant strains of *Plasmodium*, and especially *P. falciparum*, have arisen and become more prevalent in many regions of the world endemic for malaria, especially where *P. falciparum* predominates (Kim et al 2013). As a result, another common approach for treating malaria now includes the drug artemisinin and its derivatives like artesunate and artemether, which form the basis for artemisinin-based combination therapies (ACTs), currently the most effective treatment for non-severe malaria recommended by the WHO (WHO, “Malaria – Overview of malaria treatment”). These combination therapies work by including different treatments that specifically target different parasitic biological processes, making it less likely that a malaria strain resistant to multiple drugs will arise (Antony and Parija 2016). In terms of artemisinin, the mechanisms by which it works is still not clearly known, but research indicates that it may inhibit various parasitic metabolic pathways, as well as induce oxidative stress and therefore damage the parasite (Cui and Su 2014). Even so, the issue of drug resistant species and strains

of *Plasmodium* due to several factors including costly, insufficient, or incomplete treatment, continues to be a growing cause of concern in the ongoing fight against malaria (Kim et al 2013).

While the use of these treatments, as well as better diagnosis and prevention, has contributed to malaria's decline in worldwide prevalence over the past several years, a fully efficacious malaria vaccine has yet to be developed. There currently is only one approved malaria vaccine: RTS,S, which after a phase 3 trial with children aged 5-17 months has a vaccine efficacy of 39% for cases of malaria and 32% for cases of severe malaria. With RTS,S still undergoing a pilot program in three African countries, the WHO is awaiting further results before recommending large-scale distribution of the vaccine (WHO, "Malaria vaccine: WHO position paper"). As such, there continues to remain a critical need to identify potential targets for drug and vaccine treatments in the intraerythrocytic stage of the malaria life cycle.

1.5 Parasitic antigens

One intriguing aspect of the blood stage, as mentioned above, is the presentation of parasitic antigens on the surface of infected red blood cells. Towards this end, one of the key aspects of the malarial life cycle is parasitic modification of the surface of infected erythrocytes with specific antigens. These antigens can serve multiple functions, such as enabling infected erythrocytes to evade the host immune system while the parasite reproduces, therefore playing a considerable role in malaria's pathogenesis and many of its symptoms (Lapp et. al. 2015; Galinski et. al. 2017).

In the case of *P. falciparum*, infected red blood cells produce antigens, specifically *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) on its surface that allow infected red blood cells to adhere to and sequester itself to the endothelial lining of blood vessels (Berendt et al 1990, Hisaeda et al 2005). Since these antigens are produced at the later trophozoite stage, the more mature trophozoite and schizont stage infected red blood cells are usually not found in circulating blood and thus more likely to avoid destruction by host immune system and successfully release its merozoites (Hisaeda et al 2005). The parasite can further escape immune recognition by hiding intracellularly inside red blood cells, having a high degree of surface antigenic variability which make them more difficult to identify by antibodies, and producing toxins that suppress the maturation of antigen presenting cells in the immune system (Hisaeda et al 2005). In this study, we will be focusing on one type of these proteins, specifically the SICA antigens presented on the surfaces of RBCs infected by the human malaria parasite species, *P. knowlesi*.

1.6 *Plasmodium knowlesi*

Originally known solely as a simian malaria species, awareness of *P. knowlesi* has grown in significance over the past fifteen years since it was fully recognized as a zoonotic transmitted malaria parasite species that contributes to a significant number of cases of human malaria in Malaysia (Singh et al 2004). Since then, cases of *P. knowlesi* have been identified throughout Southeast Asia, though it continues to remain especially prevalent in Malaysia, where it accounted for 81% of reported malaria cases in 2014 (Cox-Singh et. al. 2008, World Malaria

Report 2015). Moreover, the *P. knowlesi* intraerythrocytic cycle takes only 24 hours, allowing the parasite to reproduce faster than the other known human malaria species, which reproduce in about 48 hours or 72 hours in the case of *P. malariae* (Antinori et al 2013). Thus, *P. knowlesi* has the potential to be extremely virulent and proliferate rapidly in the body, especially as it can be difficult to differentiate from *P. malariae* (Singh et al. 2004). As a result, the misdiagnosis of a *P. knowlesi* infection as an infection by less severe parasites like *P. malariae* can lead to the onset of a severe case of malaria and possibly death if not correctly and quickly detected and aggressively treated (Galinski and Barnwell 2009; Daneshvar et al 2009).

In addition, as *P. knowlesi* can be studied in a wider range of settings, including *in vivo*, *ex vivo* and *in vitro* studies, it makes for a useful model for examining characteristics of other human malaria parasites that may not be as easily cultured (Lapp et. al. 2013; Lapp et. al. 2015; Galinski et. al. 2017). As an example, *P. knowlesi* is genetically more similar to *P. vivax*, but whereas a fully successful *in vitro* blood culture of *P. vivax* has yet to be developed, *P. knowlesi* can serve as such a model system for furthering basic research on the infected RBC biology (Lapp et. al. 2015). Furthermore, the *SICAvar* gene family of *P. knowlesi* is genetically and functionally similar to the *var* gene family of the more predominant human malaria parasite *P. falciparum*, both coding for variant surface antigens. Thus, *P. knowlesi* research can lead to developments addressing the malaria blight worldwide.

1.7 SICA variant proteins

As mentioned in the previous section, *P. knowlesi* can serve as a useful model organism is in the study of antigen variability. In addition, the types of antigens being presented are dependent on multiple factors including the malaria species and the host. *Plasmodium knowlesi* has two known gene families that code for variant antigens, namely *SICAvar* and *kir*. Of particular interest for this project are the Schizont Infected Cell Agglutination (SICA) variant antigens (Howard et al 1983) coded for by the large *SICAvar* gene family of *P. knowlesi* (Al-Khedery et al 1999), significant because they have been associated with *P. knowlesi*'s virulence (Barnwell et al 1983). Monkeys with infected red blood cells displaying SICA antigens would experience a daily 10-fold increase in red blood cells infected by the parasite, leading to a more severe form of malaria which would eventually lead to death if not treated. By contrast, monkeys with infected red blood cells that did not display the SICA antigens would experience peak parasitemia of 6% after a few days, but after which the percentage of infected red blood cells declined naturally, indicating that the infection had been controlled (Barnwell et al 1983). As SICA antigens are biologically similar to *P. falciparum*'s PfEMP1 variant antigens, they likely play a similar role in allowing infected red blood cells to avoid detection by the host immune system and via adhesion to the inner lining of blood vessels, although to a much lesser extent (Korir and Galinski 2005; Galinski et al. 2017).

As of the most up to date *P. knowlesi* genome sequence (Lapp et al 2017), there are 136 full-length *SICAvar* genes recognized, along with 59 fragments. Moreover, the *SICAvar* genes can be further classified into type I and type II. Type I *SICAvar* genes are in general longer with

7-14 exons, while type II *SICAvar* genes have 3-4 exons (Pain et al 2008). In addition, there are only 19 type II *SICAvar* genes, with the majority being type I (Lapp et al 2017). The *SICAvar* genes are randomly distributed across all 14 chromosomes of the *P. knowlesi* genome (Pain et al 2008, Lapp et al 2017), which itself spans 24.7 megabases of DNA and encodes 5384 genes based on the most recent published sequence (Lapp et al 2017). The *SICAvar* gene family codes for SICA antigens, which also have a well-defined basic protein structure. SICA antigens are characterized by a series of cysteine-rich domains positioned throughout the majority of the surface-exposed protein, followed by a transmembrane domain and a cytoplasmic domain at the C-terminus (Al-Khedery et al, 1999; Lapp et al 2009; Pain et al 2008; Lapp et al 2017). The number of cysteine rich domains can vary, which are each coded for by a number of exons (Pain et al 2008; Lapp et al 2017). Furthermore, SICA proteins were originally defined as ranging in size from 180 kDa to 225 kDa (Barnwell et al 1983), determined since to correspond to the expression of different *SICAvar* genes from amongst the many options in the large multigene family (Al-Khedery et al. 1999; Lapp et al 2013; Lapp et al 2017).

For *P. knowlesi* infected erythrocytes, variant expression of the *SICAvar* genes has been demonstrated to be dependent on environmental factors such as the presence of a spleen and whether the host had been previously exposed to *P. knowlesi* infection (Barnwell et al 1982; Barnwell et al 1983; Howard et. al. 1983; Lapp et. al. 2013). Thus, due to the important role SICA antigens play in the virulence of the malaria parasite *P. knowlesi*, studying the conditions for the expression of the *SICAvar* genes can lead to a clearer picture of how variant antigens contribute to malaria pathogenesis. We are especially curious about how *P. knowlesi*

establishes and expresses five different clones depending on host conditions, like a missing spleen in SICA[-] parasites, or having been cured of a past infection by *P. knowlesi* A+ or B+ SICA[+] clones, then reinfected with the homologous clone (resulting in the *in vivo* switched B+ and C+ SICA[+] populations and cloned phenotype, respectively). The related SICA[+] clones have been characterized by the expression of different SICA proteins while SICA[-] parasites lack SICA protein expression (Howard et. al. 1983; Barnwell et al 1982; Barnwell et al 1983; Lapp et. al. 2013; and unpublished data). The process for how these and other different *P. knowlesi* clones were created is illustrated in the figure below. The C+ clone was created at Yerkes using a similar approach, essentially by infecting a rhesus that had a B+ immune response with the B+ parasites and then expanding one micromanipulated iRBC of the resulting C population in a naïve animal to generate the C+ clone.

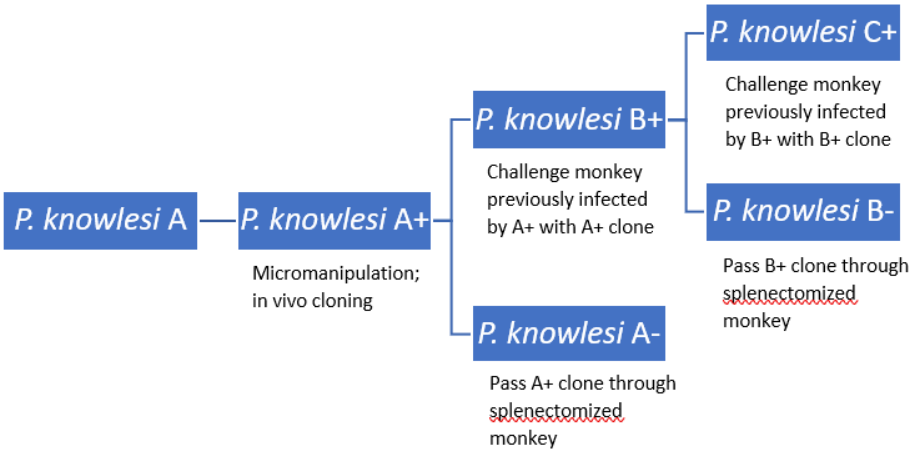


Figure 1: Description of the process utilized to create the different P. knowlesi clones, adapted from Al-Khedery et al 1999.

In terms of the spleen, monkeys with a spleen have infected red blood cells that display SICA antigens (SICA[+]), whereas infected red blood cells of splenectomized monkeys do not display a detectable level of SICA antigens (SICA[-]) (Barnwell et al 1982; Barnwell et al 1983). Of significance is the fact that when SICA[-] *P. knowlesi* parasites were used to infect a monkey with a spleen, infected red blood cells reverted back to presenting SICA antigens, indicating the significance of the spleen (Barnwell et al 1983). In addition, the SICA protein repertoire being displayed also depends on the host's immune response to malarial infection (Al-Khedery et al 1999). As the host immune system produced antibodies specific to the SICA antigens being displayed on infected red blood cells, successive parasitized red blood cells display a different set of SICA variant antigens (Brown and Brown 1965; Barnwell et al 1983). Thus, changing the variant SICA antigens being presented depending on the anti-SICA antigen antibodies generated by the host allows *P. knowlesi* infected red blood cells to evade the host's immune system response (reviewed in Galinski et al 2017). Though there have been past efforts to study *P. knowlesi* gene expression using microarray analysis (Lapp et. al. 2015), new RNA sequencing techniques and bioinformatics tools that have been developed since then can now be applied to introduce more accurate and multifaceted insights into the underlying molecular mechanisms behind antigen variability (Garber et. al. 2011; Chappell, Lapp et. al., in preparation).

1.8 Systems Biology/Omics

In the modern day, with the advent of big data, biological research is producing larger and more complex datasets at a staggering pace (Stephens at al 2015). Sequencing, initially confined to the analysis of DNA, has been further expanded to the realms of RNA and proteins, leading to the advances in the fields of transcriptomics and proteomics, respectively (Soon et al 2013, Chou 2009). Beyond sequencing, advances in other technologies have allowed for greater exploration in studying other types of biological data. For an example, recent developments in mass spectroscopy and nuclear magnetic resonance imaging have made it easier to rapidly quantify the thousands of different small molecules produced during human metabolism, or metabolites, revealing novel insights into our body's internal biochemistry (Patti et al 2012). In this manner, our capability to generate whole organism profiles of a variety of biological molecules has led to remarkable progress in the study of these datasets, or -omics, with significant practical impacts on other fields of research, like personalized medicine and disease diagnosis (Soon et al 2013).

This explosion of such technological knowledge and capabilities has been accompanied by an equally dramatic increase in computational tools to process and analyze these types of data, such as the application of clustering algorithms to classify genes that may share similar functionality (Lockhart and Winzeler 2000). In addition, all this knowledge has brought up the nontrivial challenge of where and how to store all this newly generated data. According to one study, it will take an estimated 2 – 40 exabytes of memory just to store sequencing data for the

entire human population, over 10,000 times the 3.6 petabytes of genomic data currently being stored by the National Institutes of Health National Center for Biotechnology Information (NIH/NCBI)'s Sequence Read Archive, which contains sequences from most research publications (Stephens et al 2015). Moreover, new databases have been developed for a variety of organisms, including *E. coli* (EcoCyc) and *Plasmodium* (PlasmoDB), giving scientists the opportunity to not only obtain raw datasets, but view different analyses and visualizations of that data, allowing for a comprehensive perspective of an organism's inner workings all stored in a readily accessible, convenient location (Lockhart and Winzeler 2000, Aurrecoechea et al 2009, Chin et al 2013). As a result, now more than ever before, researchers need to be able to determine how to make use of all this data that has been generated. As John Naisbitt wrote in his book *Megatrends*, "We are drowning in information but starved of knowledge." How then do we not only identify which pieces of data are actually relevant, but also piece them all together into a useful result that can solve a practical problem facing the world today?

In order to address the difficult question of how to effectively combat a disease as multifaceted as malaria, the pathogen needs to be addressed from many different angles, with research that is able to identify specific aspects of the parasite that can be targeted and lead to the development of more effective approaches to treatment (Le Roch et al 2012). This all plays into the emerging field of systems biology and its far-reaching goal to be able to synthesize many distinct types of biological data into coherent results that reveal insights into how those many different systems interact (Alberghina and Westerhoff 2007). By using a diverse range of raw datasets that have been and are being generated by different wet labs, systems biologists

can gain a big picture understanding of the inner workings of malaria and the precise means by which this disease affects the infected host. Thus, to investigate those underlying mechanisms of the malaria disease, my project will apply multi-omic, specifically genomic and transcriptomic, analyses. By building upon the Malaria Host-Pathogen Interaction Center's (MaHPIC's) recent efforts developing the first PacBio-based *Plasmodium* genome sequences (Chien et. al. 2016; Lapp et. al. 2017) and *P. knowlesi* gene expression studies through transcriptomic analysis using RNA-Seq methods (Chappell, Lapp et. al., in preparation), I set out to make use of the established *P. knowlesi* – monkey model system to gain novel insights into the underlying causes and mechanisms behind antigen variability and malaria's virulence.

1.9 Genomics/Next Generation Sequencing

DNA sequencing, a field long dominated by Sanger sequencing (Sanger et al 1977), has been recently succeeded by next generation sequencing (NGS) (Soon et al 2013). The main advantage of NGS approaches are that they are high throughput, incorporating parallelization to sequence multiple DNA strands simultaneously, which gives NGS techniques the capability to sequence genomes at an increasingly faster rate and lower cost (Soon et al 2013, Heather and Chain 2016). As such, the field of genomics has expanded greatly in scope, with current technologies creating an estimated 35 petabytes of data each year and growing (Stephens et al 2015). At the current rate, the genomes of 1.2 million animal and plant species will likely be sequenced within a decade (Stephens et al 2015). This wave of genome sequencing driven by NGS has largely been accomplished by short-read sequencing, so named because it generates

many small DNA fragments that are several hundred base pairs long (Heather and Chain 2016). By identifying overlaps between different fragments, reads can be aligned together and assembled to recreate the full DNA strand that was being sequenced, a process called *de novo* assembly (Chin et al 2013, Heather and Chain 2016). One ongoing issue with short-read sequencing, though, is the need for DNA amplification in order to generate a high enough read depth for assembly, a process that can lead to additional sequence errors (Chin et al 2013, Heather and Chain 2016). In addition, highly repetitive regions in the DNA being sequenced can be difficult to account for using short reads, creating gaps in the genome assembly (Chaisson et al 2014). To address these issues, we turn to what has been termed third generation sequencing, or long-read sequencing (Heather and Chain 2016).

For this project, we make use of recently assembled *P. knowlesi* genomes produced using Pacific Biotechnology (PacBio)'s long-read sequencing technique, which can generate reads of upwards of 10,000 bases in length (Heather and Chain 2016). Though current long reads generally tend to have lower accuracy compared to short reads, the lack of amplification means less reads that need to be generated and higher overall coverage of the genome being sequenced (Stephens et al 2015, Chaisson et al 2014, Chin et al 2013). Since each read produced by PacBio techniques is significantly longer than the average short read, PacBio long reads can be used to assemble genomes with fewer gaps, and moreover can provide coverage for current existing gaps in genomes assembled with short-read sequences, leading to resolution of those missing sequences (Chin et al 2013, Chaisson et al 2014, Heather and Chain 2016). As PacBio long reads can function as "seed" reads that shorter reads can be mapped to

in order to create consensus reads of higher quality (Chin et al 2013), they can furthermore be used to validate recently assembled *P. knowlesi* genomes and assess their overall accuracy and completeness.

For this study, I first used the recently developed and annotated *P. knowlesi* A+ genome sequence, based on long-read results from PacBio NGS and High-throughput Chromosome Conformation Capture (Hi-C) technology (Lapp et. al. 2017) to validate *P. knowlesi* B+ and C+ genomes that have been recently assembled using similar PacBio long-read *de novo* assembly procedures. Since a genomic alteration was previously identified with pre-NGS methods between *P. knowlesi* A+ and B+ clone genomes that coincided in *SICAvar* switching (Al-Khedery et. al. 1999, Corredor et. al. 2004), comparing the *P. knowlesi* A+ clone to B+ clone genomes and B+ clone to C+ clone genomes can indicate if and what type of differences in total may exist between these clones.

1.10 Transcriptomics/RNA-Seq

Similar to the field of genomics, transcriptomics has also been radically transformed through the NGS. Transcriptomics seeks to study the RNA produced by a cell, or population of cells or tissues, and thus determine the types of genes being expressed and understand their functionality and how they contribute to that specific state of the specimen under study (Wang et al 2009). If a higher than normal quantity of a specific transcript is measured at a certain time, it can be interpreted as the gene corresponding to that transcript being expressed at a higher level, indicating that the more active gene may play a more significant role with regards

to the specimen's current condition. In order to create such a profile of a sample's transcripts, past approaches have utilized microarrays, where probes corresponding to specific genes are placed onto an array and transcript levels determined by how much they bind to certain probes on the microarray (Wang et al 2009). While the microarray technique is high throughput and has proven to be an effective tool in the study of transcriptomics, it has several drawbacks, including difficulty in filtering out background noise and only being able to detect transcripts that correspond to predetermined gene sequences (Wang et al 2009).

As such, the application of sequencing techniques to transcriptomics has led to the development of RNA sequencing, or RNA-Seq. Using this method, RNA strands are directly sequenced, then mapped to a reference genome to see which genes those transcripts are matched to (Wang et al 2009). As such, a higher frequency of one particular RNA sequence can indicate that the gene it corresponds to is more active at that particular point, and the protein that it corresponds to also is more important to the organism and is potentially produced at a higher rate. This all leads back to my project's interest in the factors that play a role in the variant expression of the *SICAvar* genes and by association the different phenotypes of SICA antigens they code for. By being able to quantify the levels of gene expression by the malaria parasite during the intraerythrocytic development cycle while it is producing SICA antigens, we can better understand what other genes and biological pathways play a role in the upregulation or downregulation of the *SICAvar* genes.

Thus, along with completed *P. knowlesi* A+, B+, and C+ reference genomes, newly generated and already available RNA-Seq data of *P. knowlesi* A+, A-, B+, B- and C+ clones can be compared to determine if there were any significant, large-scale differences in gene expression, perhaps due to switching at the genomic or epigenomic levels, which could result in the expression of these different phenotypes. Moreover, if such differences are found, the transcriptomes of *P. knowlesi* A+, B+, and C+ clone genomes can be further compared to determine additional factors which would relate to SICA antigen variation.

To accomplish this, I used the aforementioned stage-specific transcriptomic data that have been collected from the infected red blood cells of five variant antigen phenotypes of *P. knowlesi* to construct a time course of gene expression for each during *P. knowlesi's* 24-hour life cycle by utilizing a pipeline of RNA sequencing and bioinformatics approaches. I then applied machine learning techniques and gene ontology analysis to predict which *P. knowlesi* genes and corresponding biological pathways are associated with regulating the expression of SICA antigens, which will lead to a better understanding of the factors that contribute to *P. knowlesi* antigen variability and the parasite's virulence. By learning how the parasite undergoes switches in gene expression as it adapts to specific host environments, my project can contribute to the development of novel interventions in line with today's global goal of malaria eradication. This research represents a significant part of the Malaria Host-Pathogen Interaction Center's (MaHPIC)'s mission as it works to identify host-parasite interactions and biological mechanisms of pathogenesis using systems biology approaches.

Chapter 2: Methods

2.1 Genome Sequence Analysis

Potential sequence differences between *P. knowlesi* A+, B+, and C+ genomes can be determined by multiple sequence alignment, using tools like the functionality included in Geneious (Kearse et al 2012) like the Mauve multiple sequence aligner. By performing such alignments, we can detect matching regions of the different genomes that exhibit significant variation and thus identify if differences in gene expression are due in part to switching at the genomic level (Al-Khedery et. al. 1999; Corredor et al 2004). Moreover, I can make use of the Artemis Comparison Tool (ACT) (Carver et al 2005) as well as the Mauve alignment viewer built into Geneious to perform a comparison between the sequences of the genomes of the different *P. knowlesi* phenotypes and visualize the results of the analysis.

As long reads generated from PacBio NGS have been used for *de novo* assembly of the *P. knowlesi* A+ genome (Lapp et al. 2017), they can serve as a viable reference for recently assembled *P. knowlesi* B and C genomes. Thus, using the bioinformatics software tool Geneious, specifically its Mauve multiple sequence aligner, PacBio generated B and C long reads can be aligned with the recently published *P. knowlesi* A genomes to ensure the quality and completeness of the *P. knowlesi* B and C genome assemblies. This was accomplished by first taking all the quality controlled reads from the *P. knowlesi* B and C sequencing runs and attempting to align them to the *P. knowlesi* A genome. As the A genome is published and has been annotated, it should serve as a reference for ensuring that the PacBio reads used to create

their respective B and C assemblies are suitable for assembling the B and C scaffolds. This mapping was completed using the "Map to Reference" mapping tool implemented in Geneious.

Once good coverage is ascertained, the next step will be to assemble the 14 scaffolds for both the B and C genomes. This was done using the polished *P. knowlesi* B and C unitigs that were put together by a graduate student that had been working on that specific project. Again using "Map to Reference" tool from Geneious, the B and C unitigs were assembled to the reference A genome to generate B and C scaffolds for their respective genomes. Any remaining reads that were unmapped from the previous step were then attempted to be individually mapped again using the Geneious mapping tool to all the scaffolds, particularly ones with large gaps in their assemblies to see if they can be matched to a specific scaffold. If that fails, remaining reads were aligned with the *P. knowlesi* A scaffolds again, this time using the more precise progressive Mauve algorithm as implemented in Geneious. If the Mauve algorithm was unable to identify which scaffold the unmapped unitig corresponds to, that unitig was excluded from the final *P. knowlesi* B and C assemblies. Finally, the complete scaffolds from A, B and C can be compared to each other. This comparison was done using the Mauve aligner, with a one to one comparison done between scaffolds (i.e. A scaffold 1 to B scaffold 1), and the A genome compared to the B genome, and the B genome compared to the C genome. These results can be visualized using the Mauve alignment viewer. The progressive Mauve algorithm tries to align homologous regions between different sequences together into "locally colinear blocks" (Darling et al 2004). Thus, by comparing the genomes of two different *P. knowlesi* clones together using Mauve, we can visually identify which regions have been grouped together into

locally colinear blocks and are thus conserved between the different clones, and which regions fall outside of those locally colinear blocks and are sections of those sequences that may have experienced recombination when transitioning to a different protein repertoire.

2.2 RNA-Seq

The RNA-Seq work developing most of the RNA sequencing data was completed prior to the start of my project in collaboration with the Sanger Institute in the United Kingdom. As noted below, some data was also generated at Yerkes. Our collaborators at the Sanger Institute had been provided RNA samples from *ex vivo* cultures of *P. knowlesi* infected erythrocytes that were acquired from infected rhesus monkeys. The RNA reads were produced by *P. knowlesi* expressing five different cloned phenotypes, A+, B+, A-, B-, and C. These RNA reads were collected over a 24-hour time period in 4-hour intervals. To isolate the mRNA from the other types of RNA present in the sample, the poly-A tails of mRNA were targeted using poly-T strands attached to magnetic beads. Poly-A tails of mRNA bind to the poly-T strands attached to the magnetic beads and those beads isolated were then washed to remove the mRNA strands. The RNA was then fragmented to produce shorter reads. Reverse transcriptase was then used to create cDNA libraries from the RNA reads, specifically using the dUTP protocol described in Zhong et al 2011. RNA transcripts were then sequenced using Illumina technology and the transcriptome for each time point assembled using RNA-Seq techniques (Figure 1) (Wang et al 2009, Otto et al 2010).

2.3 Quality Assessment

FastQC is used to perform a data quality assessment of the RNA sequences by applying various metrics for measuring read quality. Quality checking is necessary to detect multiple issues that may have occurred during sequencing and indicate if additional RNA data modifications, such as trimming extraneous adapter sequences that may remain from the creation of the RNA-Seq libraries will be necessary to attempt to improve quality. Specifically, after consulting with the Sanger Institute lab that generated this RNA-Seq data, adapter trimming at this stage would be unnecessary as there would only be a few reads that may be unmappable due to extraneous adapters, and therefore this additional computational step would lead to a very minor improvement in percentage of uniquely mapped reads. In general though, if any modifications are made, those modified reads will then need to be reassessed to ensure that data quality had improved.

2.4 Splice Transcript Aligned to a Reference (STAR)

The Spliced Transcripts Alignment to a Reference (STAR) software is a high performing, RNA-Seq aligner that is designed to map RNA transcripts to a reference genome (Dobin et al 2013). For this project, STAR was used to align the RNA reads from RNA-Seq to the *P. knowlesi* genome (Anders and Huber 2010). Mapping the data allows for the determination of which transcripts correspond to which section of the *P. knowlesi* genome at each time point.

Ran on STAR version 2.4.1c

Parameters:

runThreadN 2

sjdbGTFtagExonParentTranscript Parent

maxIntronLength 15000

2.5 High-Throughput Sequencing (HT-Seq)

After determining what each read's position is on the reference genome, it can be compared to an annotated features file for *P. knowlesi* to identify which exon the read corresponded to using High-Throughput Sequencing (HT-Seq) (Anders et al 2015). Since each exon corresponds to a gene, HT-Seq will count how many reads were mapped to each exon (the feature) and therefore mapped to a gene. It will then return the number of reads that were mapped to each gene in the annotation file. Therefore, the more reads that are mapped to a gene in a time point, the more that will be counted and the higher the level of expression for that gene. Results were then cleaned up where all genes with no expression (i.e., fragments that were not mapped to any chromosomes in the annotations) across all time points were removed. Finally, aligned reads will be concatenated together by gene and time point into a unified matrix for each phenotype of *P. knowlesi* using the R programming language. Concatenation of the data will allow for improved visualization and comparison of all results across the entire 24-hour window of observation.

Ran on HTseq version 0.6.1p1

Parameters:

Mode=union

Stranded=reverse

Order=name

Type=exon

Idattr=Parent

2.6 Normalization

Using R, this project undertook the nontrivial task of exploring and developing different approaches for normalizing levels of RNA expression across multiple time points among all genes. With all RNA expression data standardized to a specific scale, the magnitude of changes in gene expression could be more accurately compared so that the subsequent clustering produced more meaningful results. Normalization was accomplished using the same methodology as that implemented in the counts function in DE-Seq (Anders and Huber 2010). Each time point was first transformed by the natural log. All zeros were ignored in order to avoid returning infinity. The geometric mean of each time point was then calculated and each value in the time point is multiplied by that geometric mean and that product put into a separate column. The median of all gene expression levels multiplied by the geometric mean is then determined, and the actual expression levels in that time point are multiplied by the median of the geometric means. After each time point is adjusted in such a manner, the entire data set has been normalized.

2.7 Preliminary Classification

In order to determine if clustering would produce relevant and significant results, I first needed to determine if it is possible for a machine to learn the similar patterns and associations in activity that exist between genes and be able to identify genes that are part of the same biochemical or biological pathways in the organism. Thus, using supervised machine learning, I attempted to train a logistic regression classifier using our experimental RNA-Seq data and by labelling each gene according to their most significant biological pathway from the KEGG profile, and see if the machine is capable of classifying genes into the correct biological pathway category. This was implemented using the Scikit-Learn toolbox implemented in Python (Pedregosa et al 2011).

2.8 Clustering

I then applied different clustering algorithms to group together genes whose levels of expression are closely correlated. These algorithms identify patterns of expression between different genes that are temporally similar and cluster them together. A detailed description of the three algorithms used: hierarchical, k-means, and self-organizing map (SOM) follows in the next three sections.

2.9 Hierarchical Clustering

Hierarchical clustering involves first assigning each data point to its own cluster. The algorithm tries to merge the most similar clusters together based on a certain threshold. That

threshold is increased until eventually all data points have been classified into one single cluster. As a result, hierarchical clusters are frequently illustrated as dendrograms, with data points that are more similar to each other being connected by nodes closer to the bottom of the tree. There are different considerations to take when running the algorithm. For our method, I applied the parameter for calculating the distance between data points through Euclidean distance, which is merely the straight-line distance from one point to another. Ward's clustering is then applied to combining the clusters, which seeks to minimize the within-cluster variance. This is accomplished by combining clusters with the shortest distance between their respective cluster centers. Note that the Euclidean distance is not squared as indicated in true Ward's criterion. Once hierarchical clustering is completed, each "cluster" is determined by finding the cutoff height of the tree that will lead to k number of subtrees below the cutoff threshold. Implemented using the stats package in R (R Core Team 2016).

2.10 K-means Clustering

The elbow method was applied to determine the optimal number of clusters for the data set. This was done by applying the k-means algorithm iteratively with successively increasing number of clusters (k) and calculating the within groups sum of squares for each clustering result. Though the within groups sum of squares will decrease with increasing k, viewing the graph plotting the within group sum of squares against k should display one point where the within group sums of square for a certain value of k does not significantly decrease

for higher values of k. This characteristic will appear as an “elbow” that can be visually appraised from a plot. How significant of a bend the elbow needs to have is subjective.

After specifying k number of clusters, all data points are assigned to a cluster so that there are roughly equal numbers of data points per cluster. The centroid of each cluster is then calculated. Each data point is then visited and if it is closer to the centroid of another cluster then to the centroid of its own cluster, it is reassigned to that cluster and the centroid of that cluster is recalculated. This indicates that the algorithm seeks to minimize the dissimilarity between points within their own cluster while maximizing the dissimilarity between different clusters. This is repeated until a convergence point is reached where none of the clusters overlap and each data point in a cluster is closer to the centroid of its own cluster than to the centroid of any other cluster. Note that k-means clusters are globular and occupy a multidimensional space. In our experiment, the Hartigan and Wong version of the k-means algorithm was employed using the stats package in R (R Core Team 2016).

2.11 Self-Organizing Map (SOM)

SOM is a type of artificial neural network, an unsupervised learning approach. It is used to map a 2D grid onto a multidimensional dataset. The points of the SOM are first initialized to random points. One point in the dataset is then randomly selected. The closest node of the SOM is then determined and it is moved a certain distance towards that data point. Furthermore, neighboring nodes are also moved to a less extent along with that closest node towards the data point. Afterwards, another random point in the dataset is selected and the

process is repeated all over again. After enough iterations, the shape of the SOM should have taken on a form encompassing the topography of the dataset. We used 100 iterations for creating the SOM with a Kohonen network. There is a risk of overfitting, but we believed that drawback of neural networks is countered by the advantages of a more organic approach to finding the underlying organization of the dataset. This was implemented using the kohonen package in R (Wehrens et al 2007).

2.12 Plotting Clusters

We plotted the clustering results of each method using the plotcluster method from the fpc package in R [17]. It creates a discriminant projection plot that converts a multidimensional space like clustering results into a 2-dimensional space. Though this approach allows us to illustrate the clusters, it can lead to some misconceptions such as why the largest clusters seem small and why the clusters seem to overlap. This is all due to the projection, since a perspective on the multidimensional space had to be selected to slice it and make it 2-dimensional.

2.13 Consensus Clustering

To determine which genes were grouped into the same clusters by each method, a consensus clustering approach was applied (Bansal et al 2014). A simple heuristic was utilized to create the consensus clusters. We assumed that all three methods were generally effective in correctly identifying which genes had expression dynamics correlated together, as the proportion of genes split between each cluster was about the same. Thus, the biggest clusters

identified by each method were assumed to be the same. Therefore, to create the consensus cluster, I had to first assign the same clustering number to each corresponding cluster, hence "1" for the largest clusters from each method, "2" for the second largest, and so on and so forth to create a consensus cluster numbering system. Therefore, each gene will have been assigned a consensus cluster number from each method. A Venn diagram can then be used to illustrate the results of the consensus clustering. As a result, the genes that have been clustered into the same consensus cluster will be the most strongly correlated with each other, as each method had identified those genes with the same consensus clustering number. Optimally, we seek to minimize the number of genes that fall in the overlap between different consensus clusters, since that would indicate those genes lie on the boundaries between different clusters and are not as strongly related to the genes in either cluster.

2.14 Second stage of clustering

We isolated the specific *SICAvar* genes and the clusters that each *SICAvar* gene was assigned to. Using the consensus clustering results, we could identify which cluster most of the *SICAvar* genes fell into. After the initial clustering and creation of the consensus cluster, I discovered that a very large majority of the *SICAvar* fell into the same cluster. To make the results more precise in regard to the genes that correlate with *SICAvar* expression, we repeated the same clustering process again on the largest cluster that contained the majority of *SICAvar* genes and analyzed those results specifically.

2.15 Gene Pathway/PlasmoDB Analysis

We then used hierarchical clustering to identify which genes are the most correlated with *SICAvar* by reclustering into 20 clusters and again checking where most *SICAvar* genes were located. A list of the genes IDs was collected for genes that were clustered into the largest cluster, then uploaded to the *Plasmodium* database PlasmoDB for analysis with Gene Ontology (GO) Enrichment, a functionality that has been integrated into PlasmoDB (Aurrecoechea et al 2009). GO analysis can be used to identify biological function and relationships between clustered genes. In this manner, we can quantify the distribution of parasitic gene expression levels during the different stages of the *P. knowlesi* intraerythrocytic cycle and determine which sets of genes are upregulated and/or downregulated with *SICAvar*, indicating the biological pathways which may play a potential regulatory role in the variation of SICA antigen expression.

PlasmoDB instructions

1. On PlasmoDB homepage, on top toolbar go to New Search -> Genes -> Annotation, curation, and identifiers -> Gene IDs
2. Click on Upload a text file and upload a file with all gene IDs, then click Get Answer
3. Click on blue Analyze Results button
4. Click on Gene Ontology Enrichment, then click submit
5. Results generated and displayed in table form

2.16 Machine learning approach

For machine learning, we used *P. knowlesi* A+, B+, and C+ single timepoint RNA-Seq data that had been recently sequenced at Yerkes. For each *P. knowlesi* clone there were three samples, each with one timepoint taken at the ring stage of the malaria life cycle where the expression levels of each malaria parasite gene were measured. These values were then normalized using the DE-Seq2 normalization approach, or dividing by the geometric mean (Love et al 2014), to get the final dataset used for linear regression analysis. Moreover, the *P. knowlesi* A+ 7 timepoint RNA-Seq data was used for the training and testing dataset for the artificial neural network. In general, for both datasets, each column is a timepoint and each row is a gene. All machine learning techniques used the implementation given by the Python library Scikit-learn.

2.17 Linear Regression

Preprocessing the data for linear regression involved first concatenating the columns (each column corresponding to a sample) from two of the *P. knowlesi* clones together to create an X dataset. Next, the columns of the remaining *P. knowlesi* clone had to be merged together into one, using some representative value for the expression level of that gene across all three samples to create a Y dataset. Different measurements were tested, including minimum, maximum, mean, and median. This process was repeated for three different combinations (columns of two *P. knowlesi* clones combined to make X dataset in order to predict merged columns of remaining *P. knowlesi* clone which is the Y dataset): A and B to predict C, A and C to predict B, and B and C to predict A.

For this study, the LinearRegression class from Scikit-learn was used. Simple linear regression attempts to fit a linear model using the dependent input values (X dataset) to independent output values (Y dataset) (Pedregosa et al 2011). In this manner, we will be able to learn a linear function in the form y = f(x) that can be used to predict output values from new input values. This linear function is found by drawing a line that minimizes the distance between the given data points and the line, so that the residual sum of squares between the given Y output values and the approximated Y output values given by the line is as small as possible (Pedregosa et al 2011). Thus, the coefficients of this best fit line describe the amount of variation each feature/column of the X dataset has on the Y dataset, which can be observed by the magnitude of its coefficient in the resulting linear model after fitting.

2.18 Neural Network

Preprocessing the RNA-Seq data involved first associating each *P. knowlesi* gene with a specific biological pathway as assigned from gene ontology (GO) terms in PlasmoDB. Genes with no biological pathway listed in PlasmoDB were excluded. The dataset was then split into a training and testing dataset using StratifiedKFold validation, which split the data into k subsets, with each subset further split into a training set and a testing set. A stratified split means that each k subset would have a similar proportion of genes from each pathway as the proportions of genes from those pathways in the original dataset (Pedregosa et al 2011). Thus the classifier can be trained k times with each training dataset in that k subset, then tested with its corresponding testing dataset to measure the accuracy of the neural net in predicting the

biological pathway of the genes in the testing dataset. In this study 10-fold validation was performed, so k = 10.

For this study, the MLPClassifier class from Scikit-learn was used (Pedregosa et al 2011). The Multi-layer Perceptron (MLP) is a type of neural network, a classifier that attempts to learn a non-linear regression model from a training dataset that can be used to predict an output value from a set of input values. The MLP is made up of an input layer, multiple hidden layers, and an output layer. An entry from the training dataset is fed into the input layer, which is then entered into the first hidden layer. Each hidden layer is composed of several artificial neurons (perceptrons), where each neuron transforms values from the last layer into new values by multiplying the input values by a weighted vector then mapping those weighted values to new output values by an activation function which are then fed into the next layer. The output layer takes values from the last layer and output a predicted class for the original input entry. The MLP classifier is trained by adjusting the weights of the perceptrons in the hidden layers depending on how accurately the current neural network is able to predict the known class of a training set of input values. Thus, for each new dataset used to train and further refine the weights of the neural network, the MLP classifier should become more accurate in categorizing a set of values from the testing dataset into a specific class. Some notable aspects of MLPClassifier is that it is trained via backpropagation, which involves backpropagating any mistakes in classification back through the hidden layers by adjusting the weights in the perceptrons. Moreover, in order to perform multiclass classification, the SoftMax function is

utilized to calculate from the output layer the probabilities of each possible class being the correct class for the given input set of values and returns the class with the highest probability.

Chapter 3: Results

3.1 Genomics results

P. knowlesi clone	PacBio reads assembled to A genome	PacBio reads not assembled to A genome	Total PacBio reads	Percentage PacBio reads assembled
B	28,361	231	28,592	99.2%
C	36,861	129	36,990	99.7%

Table 1: Coverage percentages for Pk B and C from Geneious

A to B scaffold alignment comparison

Note: In Mauve alignment visualizations, gaps between colinear blocks are indicative of nonhomologous regions that didn't match between the two sequences, which may be areas of recombination (Darling et al 2004). 14 B scaffolds were assembled from aligning 64 B unitigs to 14 A scaffolds.

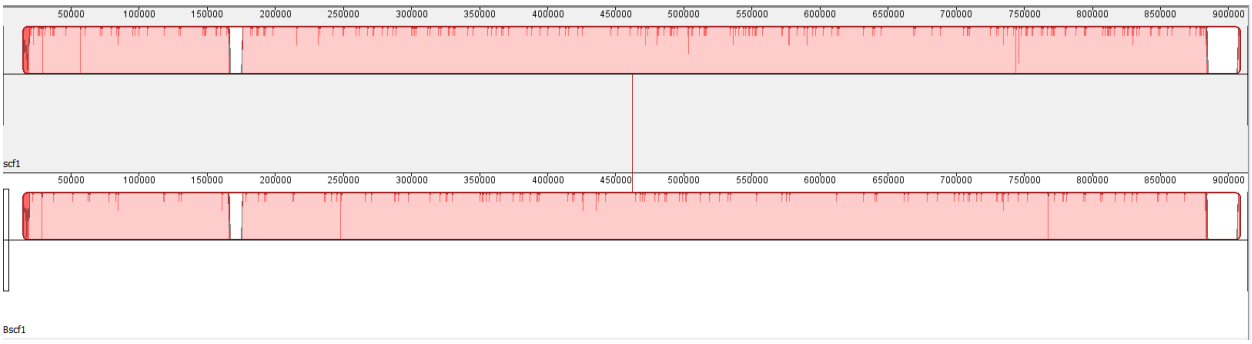


Figure 2.1: B scaffold 1 aligned to A scaffold 1. 98.3% coverage of A reference scaffold 1 by B unitigs to assemble B scaffold 1.

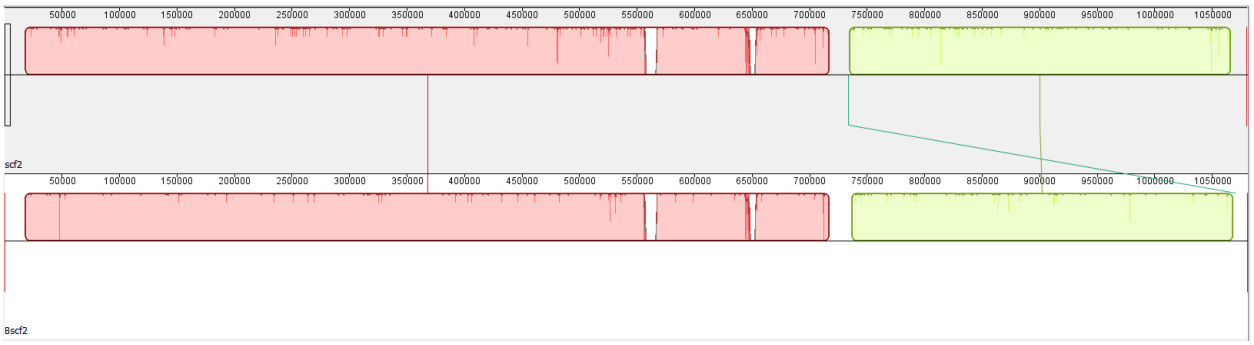


Figure 2.2: B scaffold 2 aligned to A scaffold 2. 99.4% coverage of A reference scaffold 2 by B unitigs to assemble B scaffold 2.

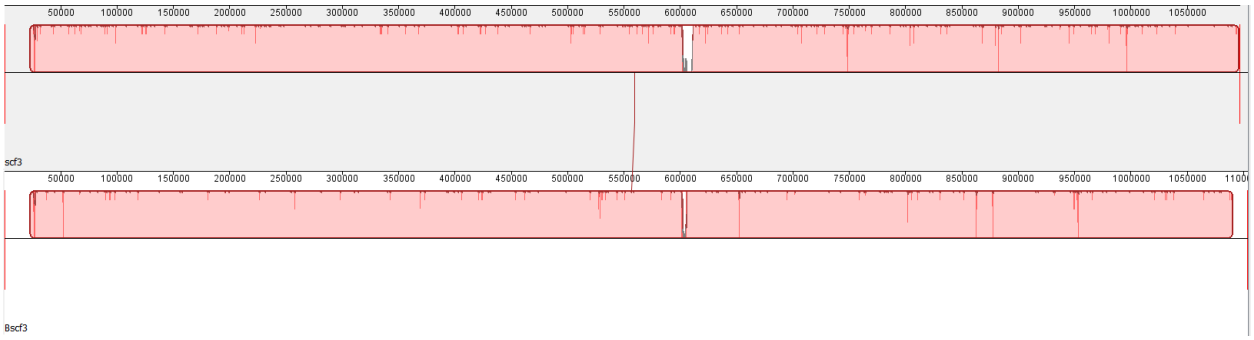


Figure 2.3: B scaffold 3 aligned to A scaffold 3. 99.7% coverage of A reference scaffold 3 by B unitigs to assemble B scaffold 3.

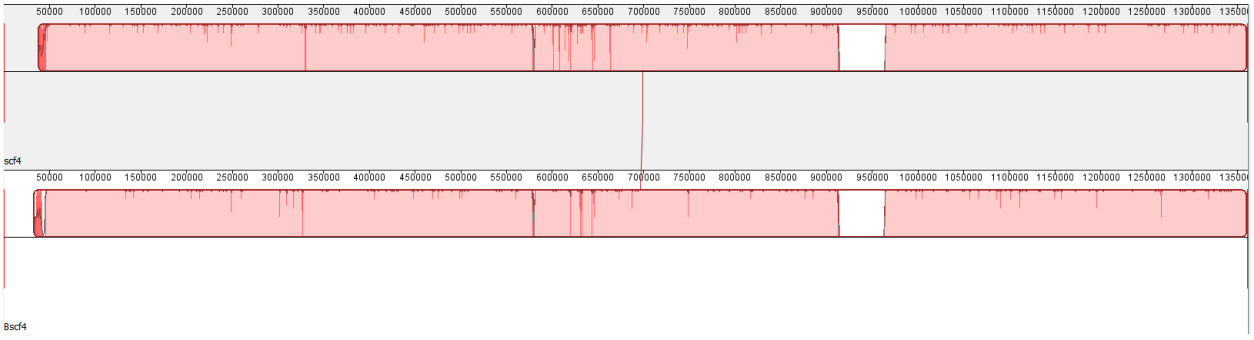


Figure 2.4: B assembled scaffold 4 aligned to A scaffold 4. 99.1% coverage of A reference scaffold 4 by B unitigs to assemble B scaffold 4.

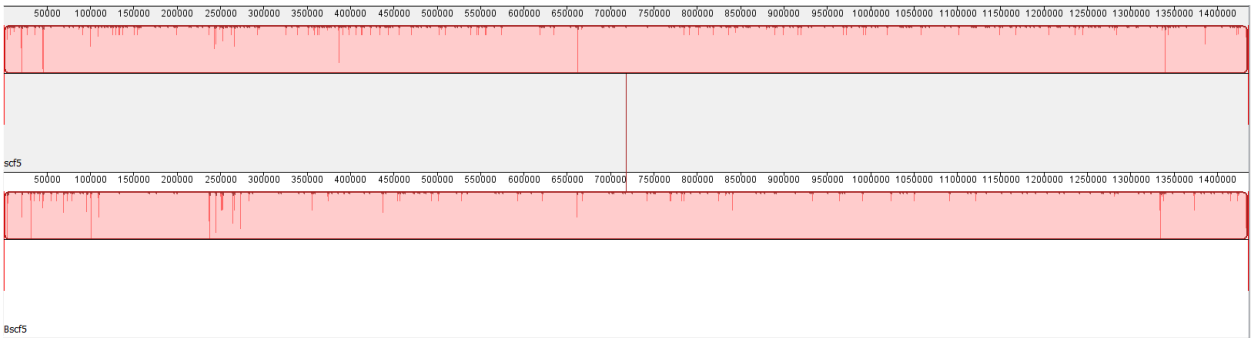


Figure 2.5: B scaffold 5 aligned to A scaffold 5. 100% coverage of A reference scaffold 5 by B unitigs to assemble B scaffold 5.

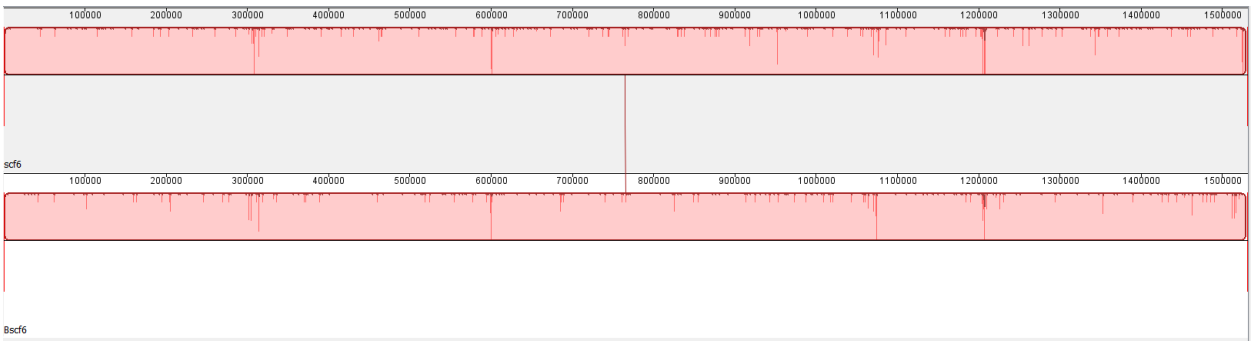


Figure 2.6: B scaffold 6 aligned to A scaffold 6. 99.8% coverage of A reference scaffold 6 by B unitigs to assemble B scaffold 6.

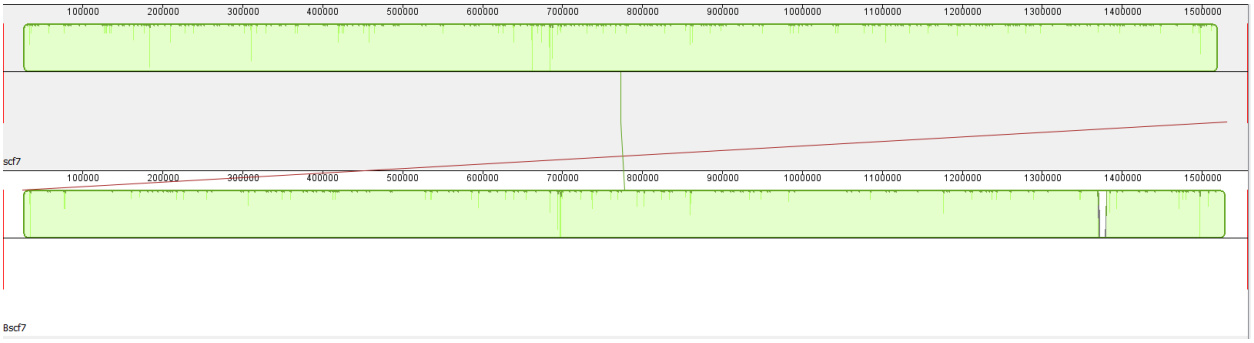


Figure 2.7: B scaffold 7 aligned to A scaffold 7. 99.99% coverage of A reference scaffold 7 by B unitigs to assemble B scaffold 7.

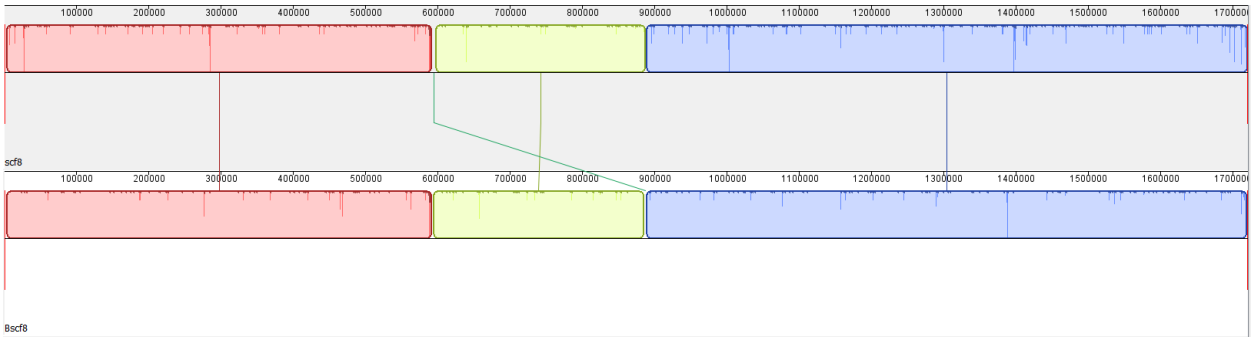


Figure 2.8: B scaffold 8 aligned to A scaffold 8. 99.8% coverage of A reference scaffold 8 by B unitigs to assemble B scaffold 8.

Figure 2.9: B scaffold 9 aligned to A scaffold 9. 99.7% coverage of A reference scaffold 9 by B unitigs to assemble B scaffold 9.

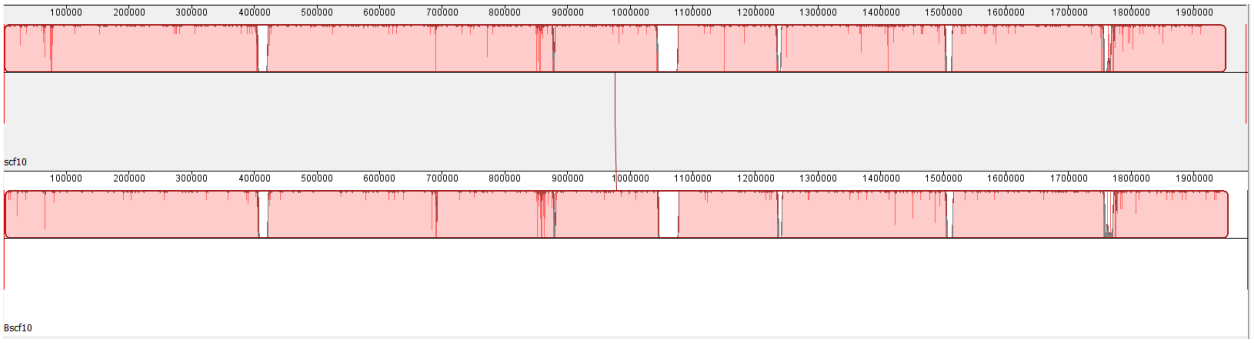


Figure 2.10: B scaffold 10 aligned to A scaffold 10. 99.3% coverage of A reference scaffold 10 by B unitigs to assemble B scaffold 10.

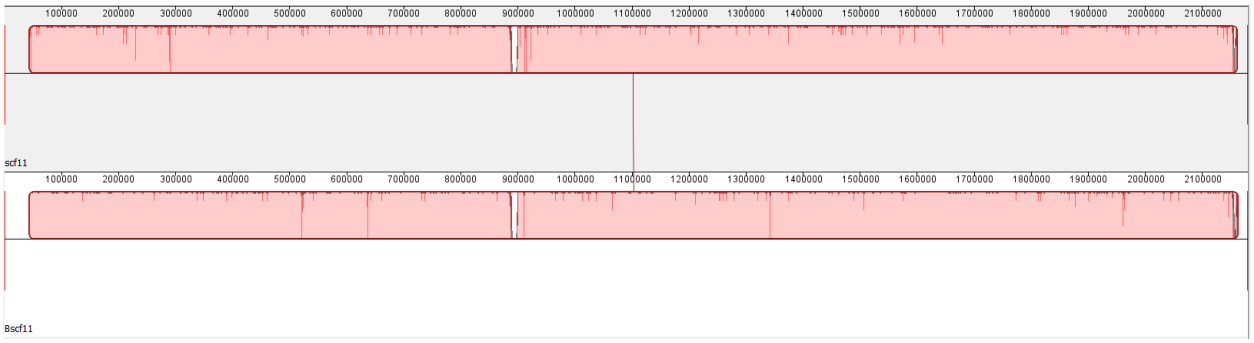


Figure 2.11: B scaffold 11 aligned to A scaffold 11. 99.99% coverage of A reference scaffold 11 by B unitigs to assemble B scaffold 11.

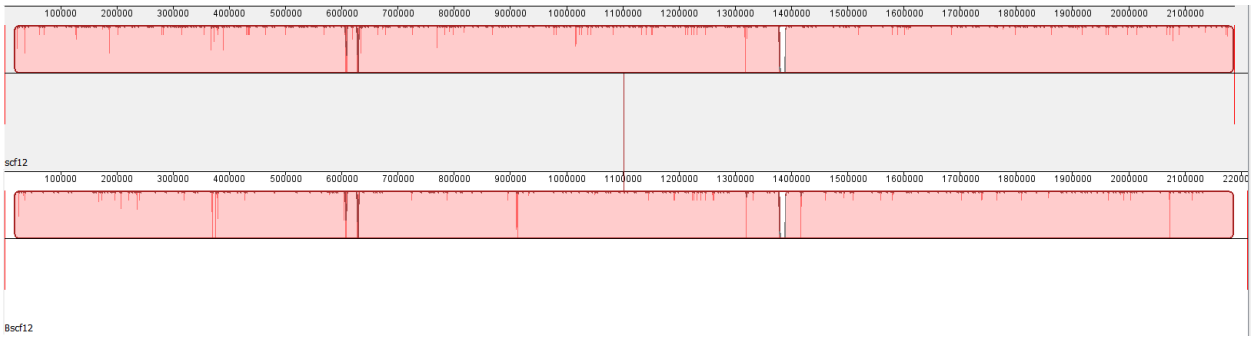


Figure 2.12: B scaffold 12 aligned to A scaffold 12. 99.6% coverage of A reference scaffold 12 by B unitigs to assemble B scaffold 12.

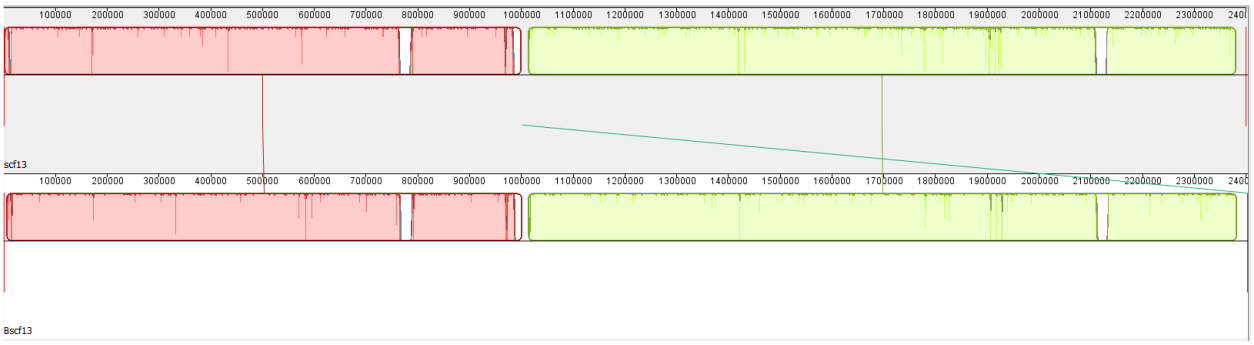


Figure 2.13: B scaffold 13 aligned to A scaffold 13. 99.2% coverage of A reference scaffold 13 by B unitigs to assemble B scaffold 13.

Figure 2.14: B scaffold 14 aligned to A scaffold 14. 99.7% coverage of A reference scaffold 14 by B unitigs to assemble B scaffold 14.

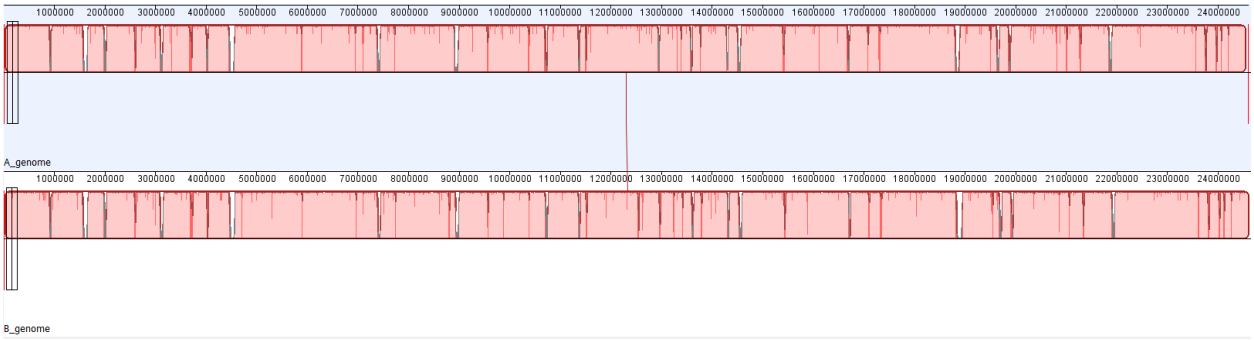


Figure 2.15: B whole genome aligned to A whole genome

B to C scaffold alignment comparison

Note: 14 B scaffolds were assembled from aligning 63 B unitigs to 14 A scaffolds.

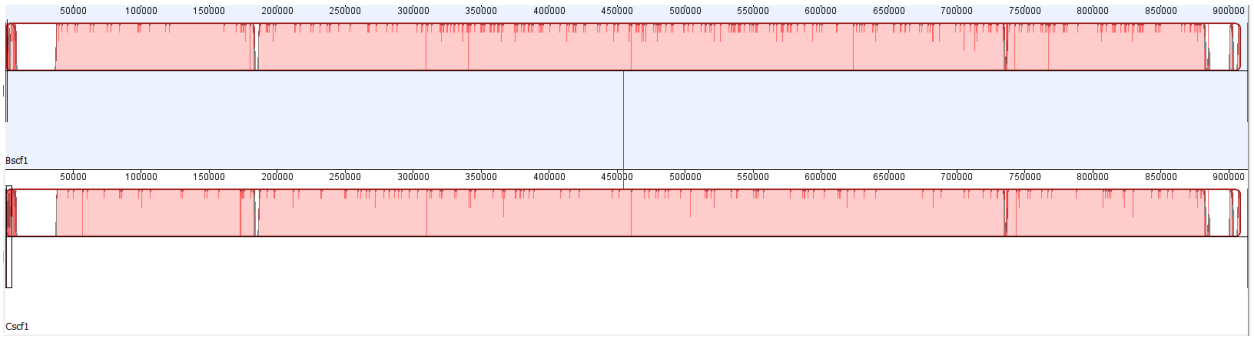


Figure 3.1: C scaffold 1 aligned to B scaffold 1. 99.6% coverage of A reference scaffold 1 by C unitigs to assemble C scaffold 1.

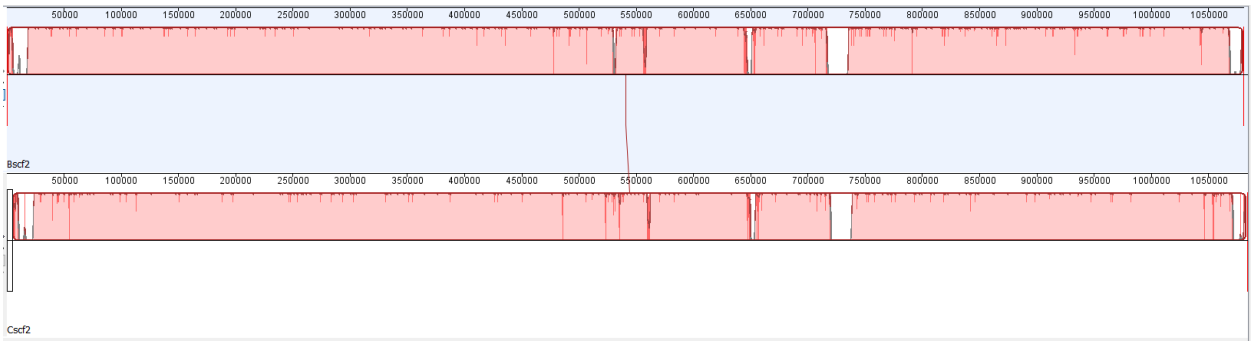


Figure 3.2: C scaffold 2 aligned to B scaffold 2. 99.8% coverage of A reference scaffold 2 by C unitigs to assemble C scaffold 2.

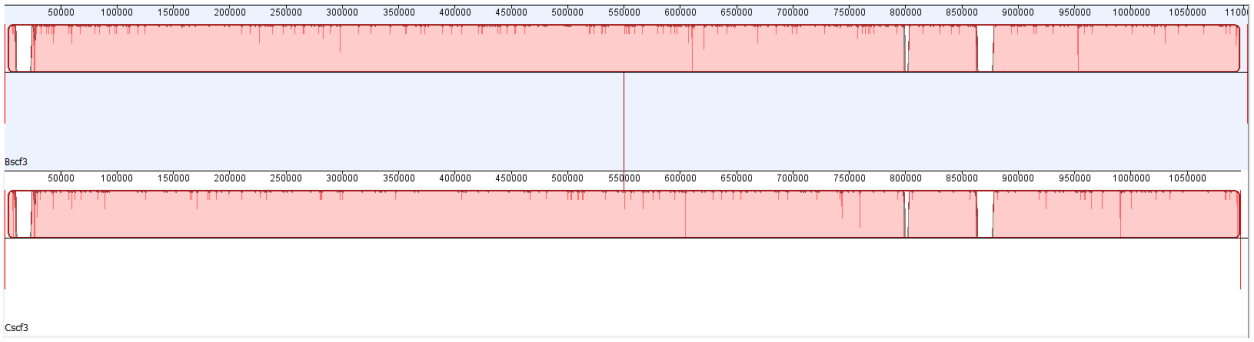


Figure 3.3: C scaffold 3 aligned to B scaffold 3. 99.7% coverage of A reference scaffold 3 by C unitigs to assemble C scaffold 3.

Figure 3.4: C scaffold 4 aligned to B scaffold 4. 99.7% coverage of A reference scaffold 4 by C unitigs to assemble C scaffold 4.

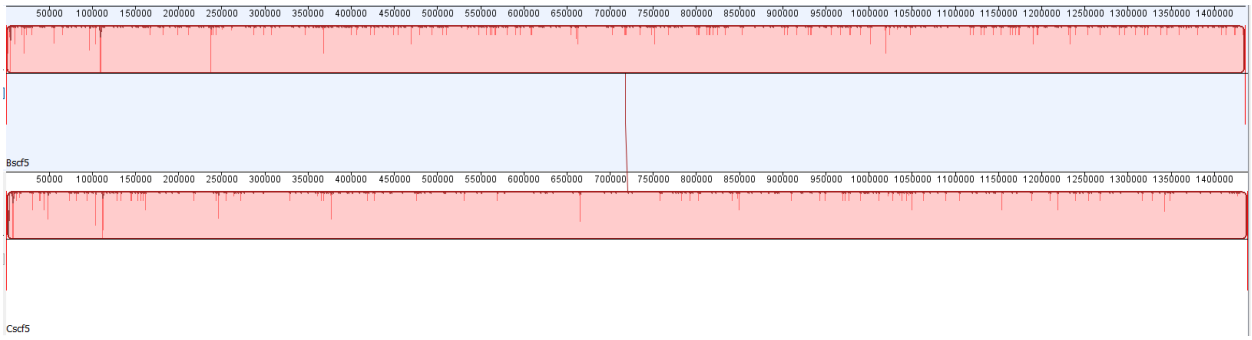


Figure 3.5: C scaffold 5 aligned to B scaffold 5. 100% coverage of A reference scaffold 5 by C unitigs to assemble C scaffold 5.

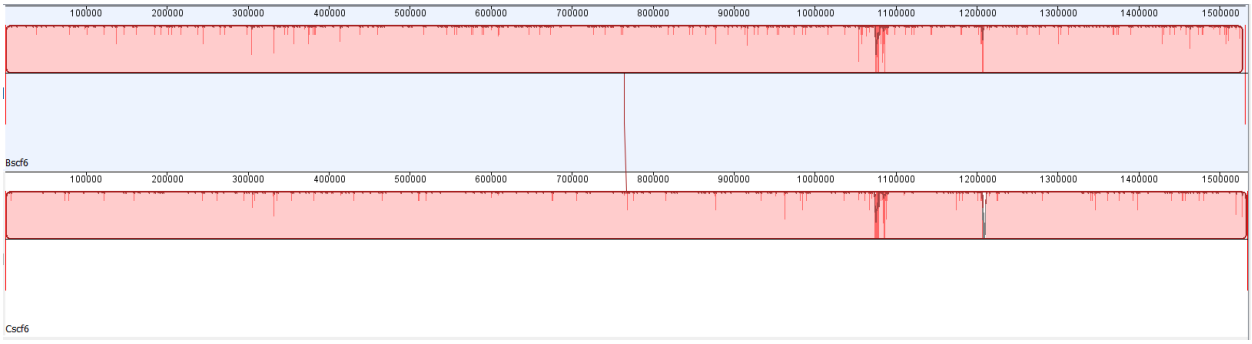


Figure 3.6: C scaffold 6 aligned to B scaffold 6. 99.9% coverage of A reference scaffold 6 by C unitigs to assemble C scaffold 6.

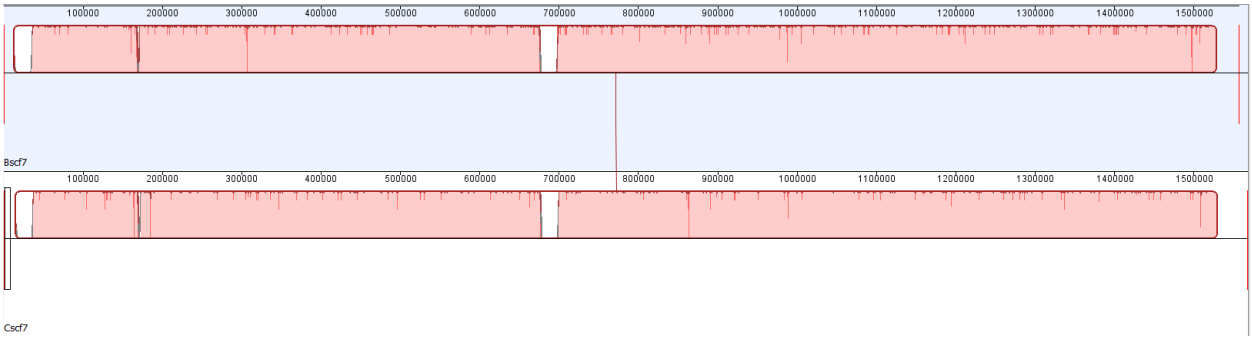


Figure 3.7: C scaffold 7 aligned to B scaffold 7. 99.9% coverage of A reference scaffold 7 by C unitigs to assemble C scaffold 7.

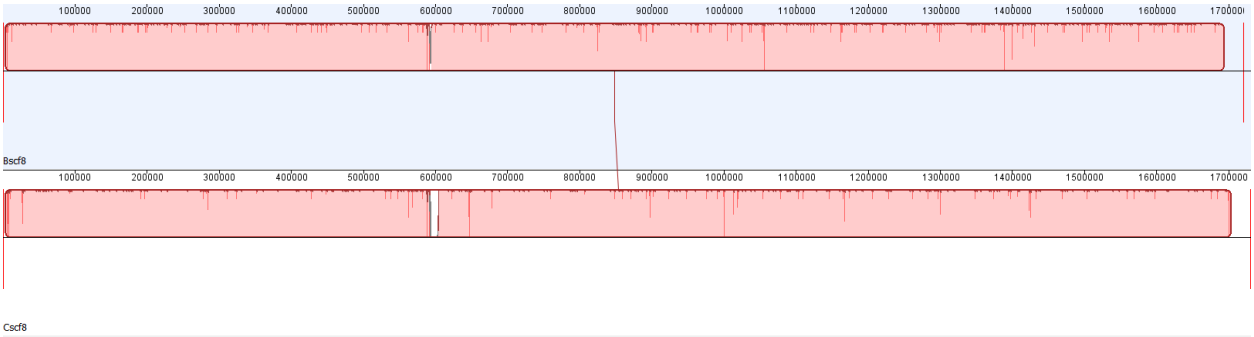


Figure 3.8: C scaffold 8 aligned to B scaffold 8. 98.3% coverage of A reference scaffold 8 by C unitigs to assemble C scaffold 8.

Figure 3.9: C scaffold 9 aligned to B scaffold 9. 99.7% coverage of A reference scaffold 9 by C unitigs to assemble C scaffold 9.

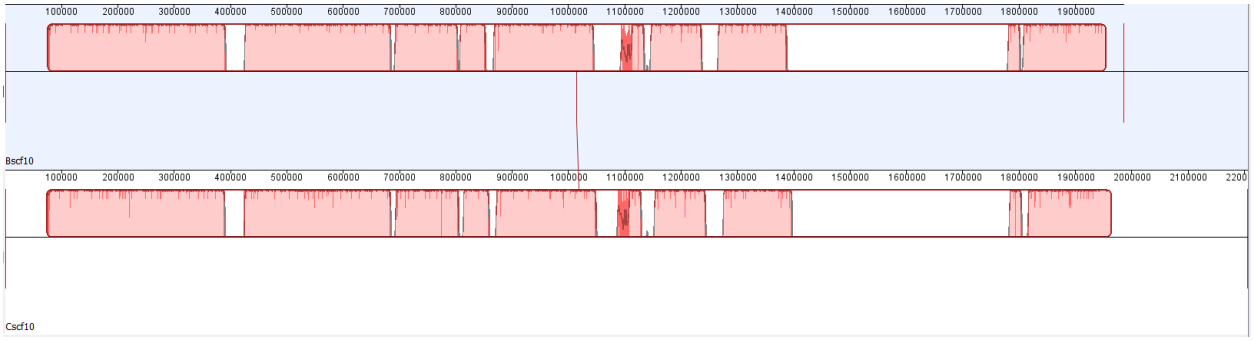


Figure 3.10: C scaffold 10 aligned to B scaffold 10. 38.8% coverage of A reference scaffold 10 by C unitigs to assemble C scaffold 10.

Figure 3.11: C scaffold 11 aligned to B scaffold 11. 42.7% coverage of A reference scaffold 11 by C unitigs to assemble C scaffold 11.

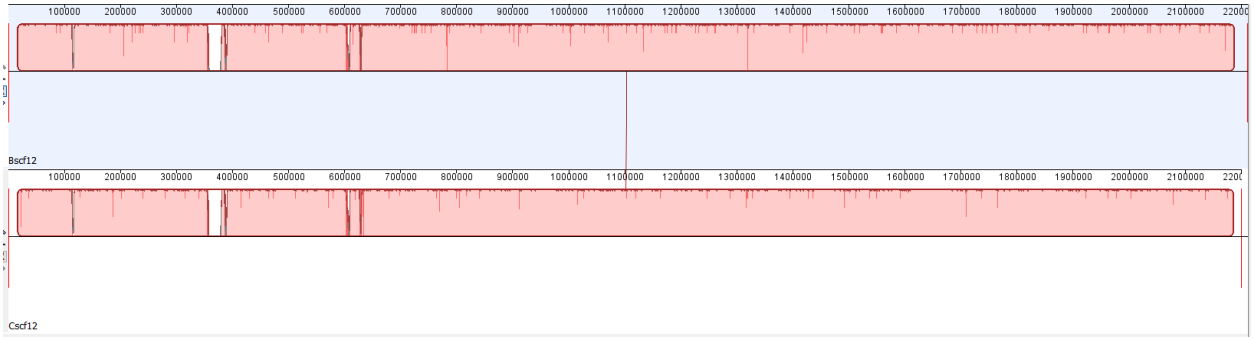


Figure 3.12: C scaffold 12 aligned to B scaffold 12. 99.5% coverage of A reference scaffold 12 by C unitigs to assemble C scaffold 12.

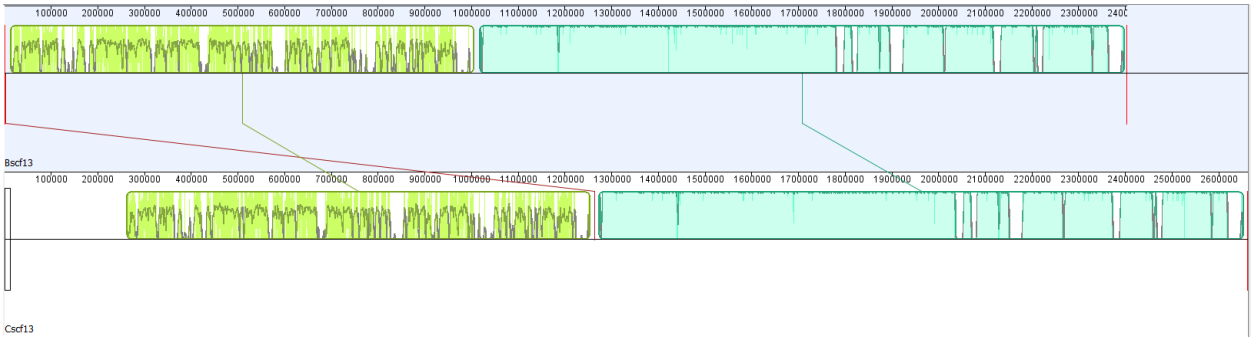


Figure 3.13: C scaffold 13 aligned to B scaffold 13. 99.99% coverage of A reference scaffold 11 by C unitigs to assemble C scaffold 11.

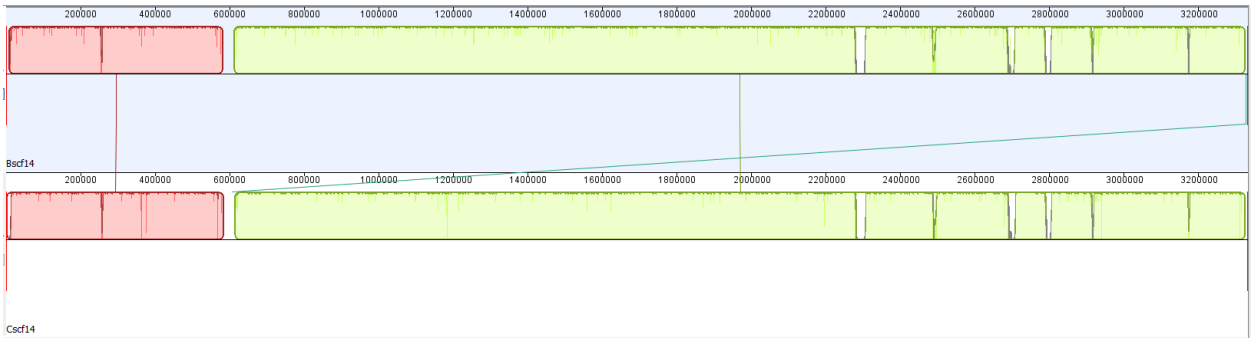


Figure 3.14: C scaffold 14 aligned to B scaffold 14. 99.9% coverage of A reference scaffold 14 by C unitigs to assemble C scaffold 14.

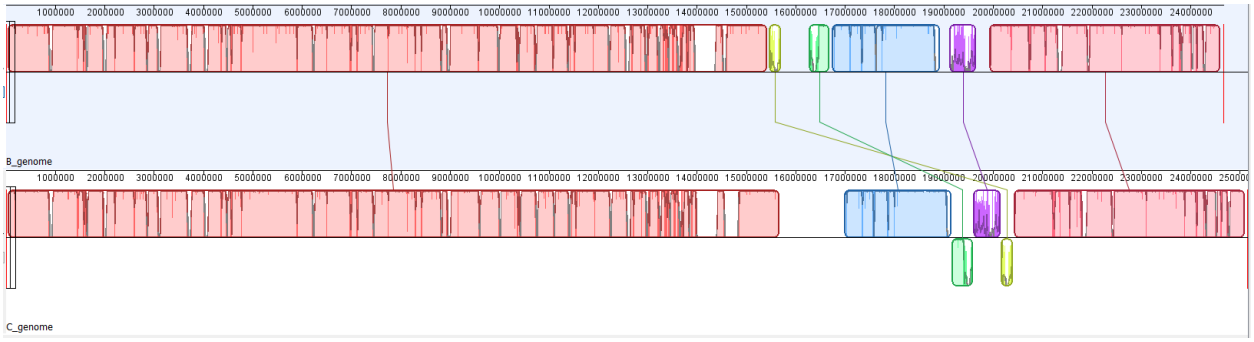


Figure 3.15: C whole genome comparison to B whole genome

Overall, after assembly, most assembled scaffolds from both the *P. knowlesi* B and C unitig assemblies matched fairly closely to the scaffold they were compared with. Scaffolds with large regions that do not match with the corresponding position in the scaffold it is being compared to are described further in the discussion section.

3.2 RNA-Seq analysis results

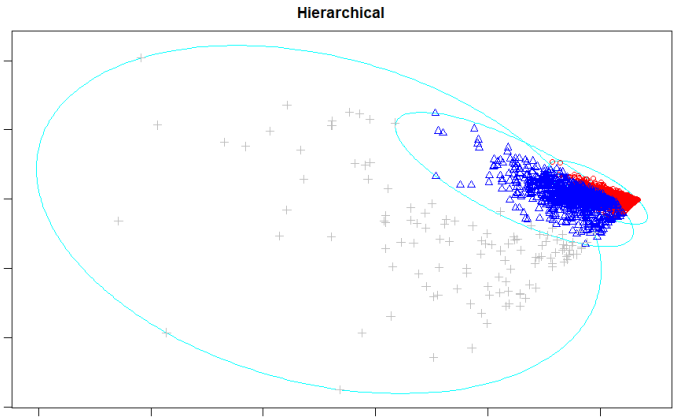


Figure 4.1: SICA[+] 3-cluster results from hierarchical clustering

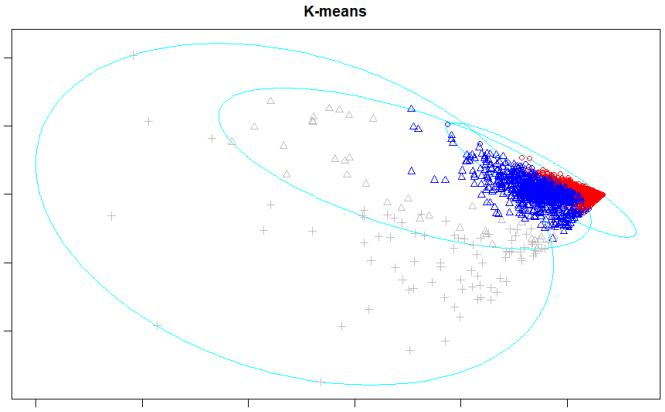


Figure 4.2: SICA[+] 3- cluster results from k-means clustering

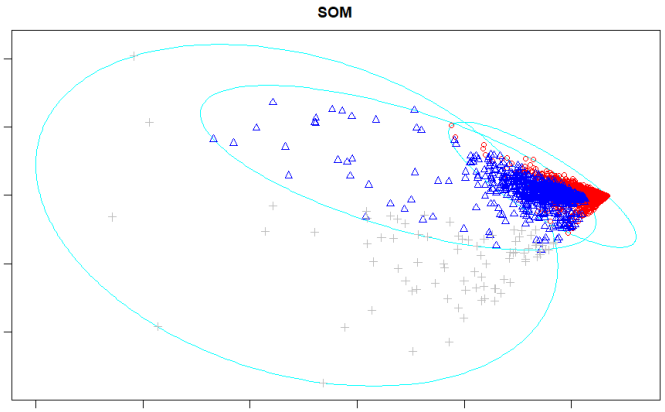


Figure 4.3: SICA[+] 3-cluster results from SOM clustering

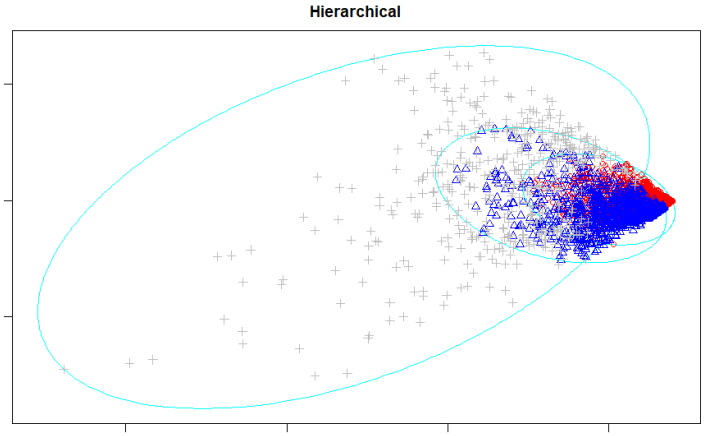


Figure 4.4: SICA[-] 3-cluster results from hierarchical clustering

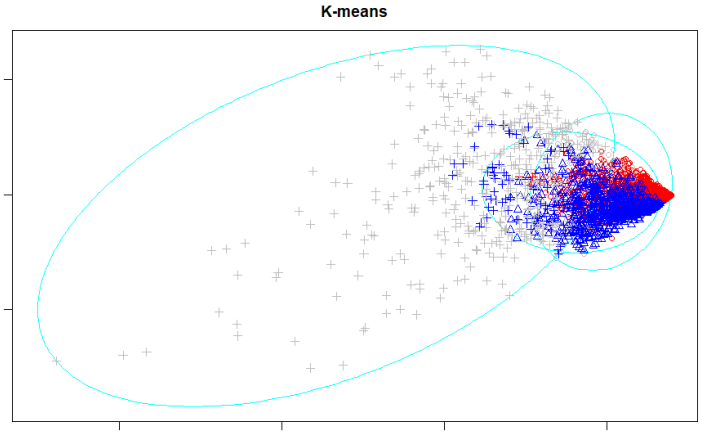


Figure 4.5: SICA[-] 3-cluster results from k-means clustering

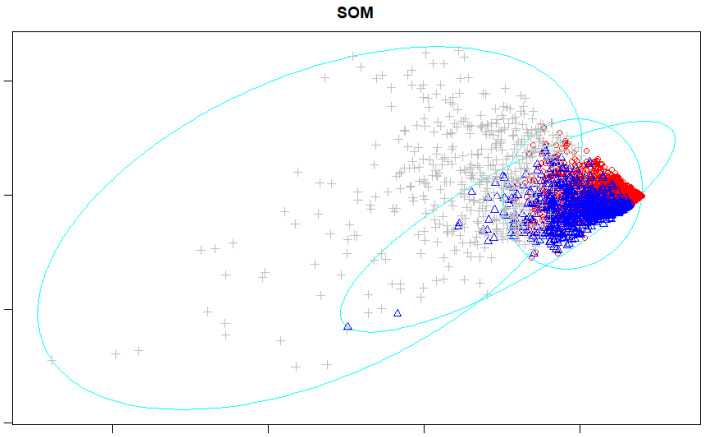


Figure 4.6: SICA[-] 3-cluster results from SOM clustering

The clustering results from both the SICA[+] and SICA[-] RNA-Seq clustering show that all three clustering algorithms return fairly similar clusters for their respective clones. Furthermore, the returned clusters look noticeably different when comparing results from the same clustering algorithm between SICA[+] and SICA[-], indicating that overall expression patterns change between the two *P. knowlesi* clones.

Gene Ontology (GO) results

Rank	Biological Process	Number of Genes	P-Value
1	cellular lipid metabolic process	26	4.29e-7
2	lipid biosynthetic process	22	6.68e-7
3	lipid metabolic process	32	1.07e-6
4	glycerolipid biosynthetic process	10	1.30e-5
5	glycerophospholipid biosynthetic process	10	1.30e-5
6	cellular response to stimulus	40	1.52e-5
7	GPI anchor biosynthetic process	9	4.02e-5
8	glycolipid biosynthetic process	9	4.02e-5
9	phosphatidylinositol biosynthetic process	9	4.02e-5
10	membrane lipid biosynthetic process	9	4.02e-5

Table 2.1: SICA[+] gene ontology (GO) enrichment analysis from PlasmoDB

Rank	Biological Process	Number of Genes	P-Value
1	nucleic acid metabolic process	77	2.32e-3
2	cellular component assembly	14	2.96e-3
3	nucleobase-containing compound metabolic process	95	4.75e-3
4	biological process	324	5.66e-3
5	cellular aromatic compound metabolic process	99	5.68e-3
6	metallo-sulfur cluster assembly	6	5.83e-3
7	iron-sulfur cluster assembly	6	5.83e-3
8	RNA processing	28	7.02e-3
9	cellular component organization	24	7.18e-3
10	heterocycle metabolic process	99	8.19e-3

Table 2.2: SICA[-] gene ontology (GO) enrichment analysis from PlasmoDB

Initial exploratory analysis shows that the largest cluster with the most number of *SICAvar* genes had many gene pathways associated with lipid biosynthesis. Furthermore, a high percentage, almost 50%, of all *SICAvar* genes were clustered together with genes associated with the lipid biosynthesis pathway, indicating that many *SICAvar* genes share similar expression patterns with that of genes in lipid biosynthesis pathways, pointing to possible co-expression. This is corroborated with the high proportion of *SICAvar* genes, almost 75%, in that largest initial cluster that were then placed in the cluster that was identified as mainly composed of genes from the lipid biosynthesis pathway.

3.3 Machine learning results

Linear Regression results

Note: A two clone combination of A+, B+ or C+ are utilized to make the X dataset and different measurements are used to merge the values of the remaining third clone into the Y values. As such, the X dataset should have 6 columns and the Y dataset should have 1 column.

A	Coefficient 1	Coefficient 2	Coefficient 3	Coefficient 4	Coefficient 5	Coefficient 6
min	0.38073176	-0.68174925	0.5083056	0.88185737	-0.46040636	0.08262304
max	1.05553482	-1.38071666	0.72804362	1.27239154	-0.97436005	0.23049192
mean	0.6580248	-0.85928201	0.5451133	0.98794954	-0.67948443	0.16953683
median	0.53780783	-0.51538012	0.39899069	0.8095997	-0.60368687	0.19549553

Table 3.1: Coefficients of the linear model after fitting with input X composed of B+ and C+ to predict Y composed of A+

B	Coefficient 1	Coefficient 2	Coefficient 3	Coefficient 4	Coefficient 5	Coefficient 6
min	0.2187913	0.4656963	-0.39002695	-0.19072088	0.68493413	0.07128083
max	0.69517188	0.05569992	-0.39325701	-0.36805606	0.9970095	0.06602753
mean	0.3381604	0.39227459	-0.40387973	-0.28924499	0.83891659	0.07064362
median	0.100518	0.65542754	-0.42835522	-0.30895802	0.83480615	0.0746225

Table 3.2: Coefficients of the linear model after fitting with input X composed of A+ and C+ to predict Y composed of B+

C	Coefficient 1	Coefficient 2	Coefficient 3	Coefficient 4	Coefficient 5	Coefficient 6
min	-0.11734976	0.89581503	0.23399677	1.06284919	-1.84272797	0.81660261
max	-3.37297332	4.29313856	0.61803816	2.19335273	-1.43975418	-0.89746027
mean	-0.60426603	1.02481031	0.42287111	0.23711845	0.15810789	-0.03367449
median	-1.33531309	1.29426721	0.81182321	0.86733697	-0.45549382	-0.15264293

Table 3.3: Coefficients of the linear model after fitting with input X composed of A+ and B+ to predict Y composed of C+

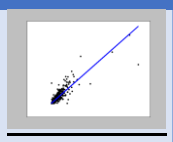
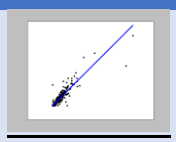
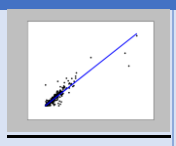
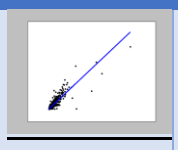
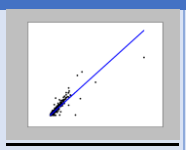
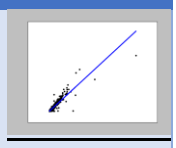
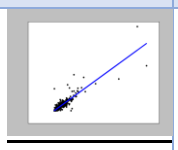
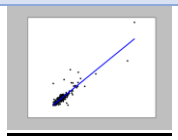
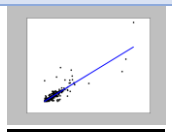
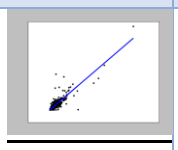
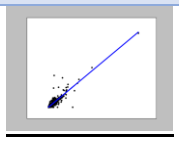
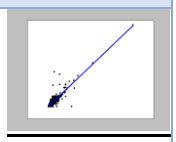
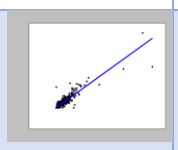
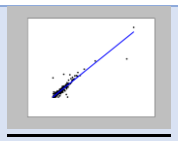
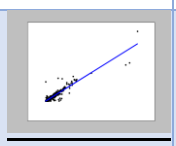
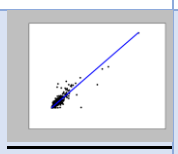
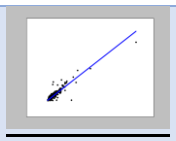
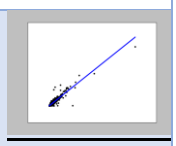
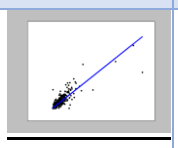
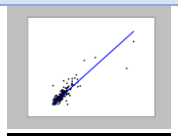
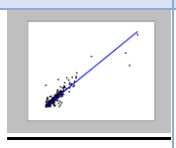
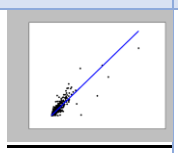
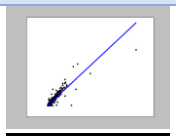
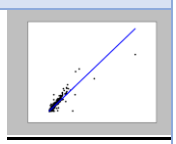
C	A1	A2	A3	B1	B2	B3
min						
max						
mean						
med						

Table 3.4: Scatter plots of the linear model after fitting each individual sample from A+ and B+ to predict C+

Multi-layer Perceptron (MLP) results

K	1	2	3	4	5	6	7	8	9	10
Accuracy	3.0%	6.5%	8.7%	11.2%	13.1%	15.4%	15.5%	17.1%	19.1%	21.0%

Table 4: Accuracy of MLP classifier in identifying the correct class of the testing dataset after training at each k fold iteration

After 10 fold validation, average accuracy was 13.1%.

Chapter 4: Discussion

4.1 Genomics Analysis Discussion

Coverage with the several thousand corrected reads from *P. knowlesi* B and C assemblies was good, indicative that the vast majority of the PacBio reads correspond to the *P. knowlesi* A reference genome. As we would expect the sequences of both the *P. knowlesi* B and C genomes to be mostly similar to the *P. knowlesi* A genome, the fact that over 99% of reads from both assemblies were successfully aligned to the A genome

It is possible that misalignments observed in the assembled B and C scaffolds are due in part to recombination, as a known rearrangement of a *SICAvar* gene can be seen as a region with missing coverage at around the same position of that gene in the corresponding *P. knowlesi* B scaffold 4 (Al-Khedery et al 1999, Corredor et al 2004; Lapp et al 2017). Due to this observation and the multiple regions with missing coverage that can be seen in Mauve alignment comparisons between the A genome and the B genome, and the B genome and the C genome, there is evidence pointing towards the presence of other regions that have experienced genome-level rearrangements.

One of the central difficulties that had to be confronted in this approach was the lack of additional data to validate the position and direction of unitigs from the given *P. knowlesi* B and C assemblies in the actual B and C genomes. While both the B and C genomes are likely to be very similar in sequence to the A genome, using the A genome as a reference for mapping the assembly unitigs to the correct scaffold makes it more difficult to ascertain if regions with no

coverage in the assemblies are due to gaps from PacBio sequencing or due to actual genomic rearrangements. We would need validation from other sources, such as short reads from Illumina sequencing, or Hi-C results to confirm that the position and orientation of a unitig is accurate as had been determined from assembly using Geneious. Such validation was applied in part by MaHPIC to create the current most up to date *P. knowlesi* genome (Lapp et al 2017), and thus would be necessary to verify that the *P. knowlesi* B and C genome assemblies are correct as well.

Moreover, while the *P. knowlesi* B genome assembly went fairly smoothly, the C genome assembly ran into some additional difficulties. Significant regions of missing coverage were present in C scaffold 10, 11, and 13, after the first assembly by the Geneious read mapper. We attempted to resolve these misalignments using unmapped unitigs with progressive Mauve. This result was achieved with some success with the C scaffold 10 assembly. As can be seen from the figure below, after the first mapping run there was no coverage of almost the first 1 million bases of scaffold 10.

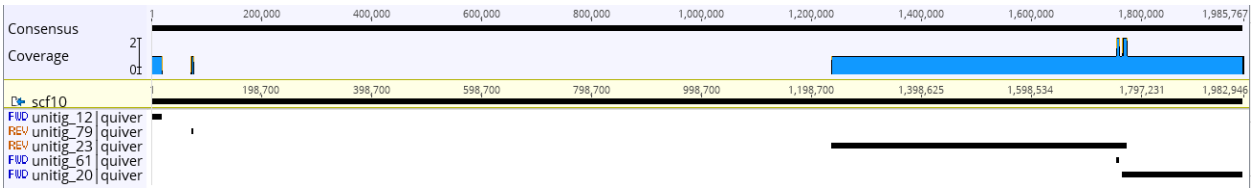


Figure 5: C scaffold 10 assembly after first mapping run with C assembly unitigs

That region of no coverage, though, was proved by progressive Mauve to be due to the mapper being unable to align unitig 10 from the C assembly, which had been labelled as an

unused read. This can be seen in the figure below, as untig 10 fills in the misaligned region observed in the previous figure.

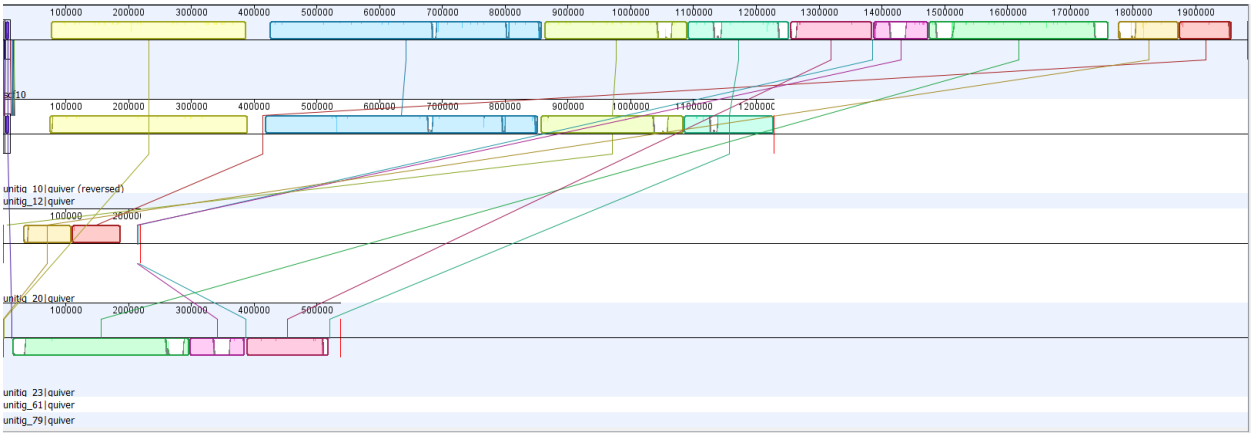


Figure 6: Alignment of longest C unitigs to A scaffold 10 using progressive Mauve algorithm

To resolve that region I then extracted the localized collinear blocks that had been identified to match with A scaffold 10 (the topmost sequence in the above figure), and realigned them to A scaffold 10, resulting in a consensus sequence with higher coverage than after the first mapping run and which was then used as the assembled C scaffold 10, though full coverage of the A scaffold by C assembly unitigs continues to remain elusive.

Another notable issue was the inability of the Geneious read mapping tool to find any C unitigs that matched with scaffold 8. A second alignment attempt using Mauve determined that unitig 4 was just a reverse complement of *P. knowlesi* A scaffold 8. This issue was resolved by extracting the reverse complements of those unitigs and using those reverse complements to map to the reference sequence, as the direction and complementariness of the unitigs are arbitrarily assigned when put into the data file.

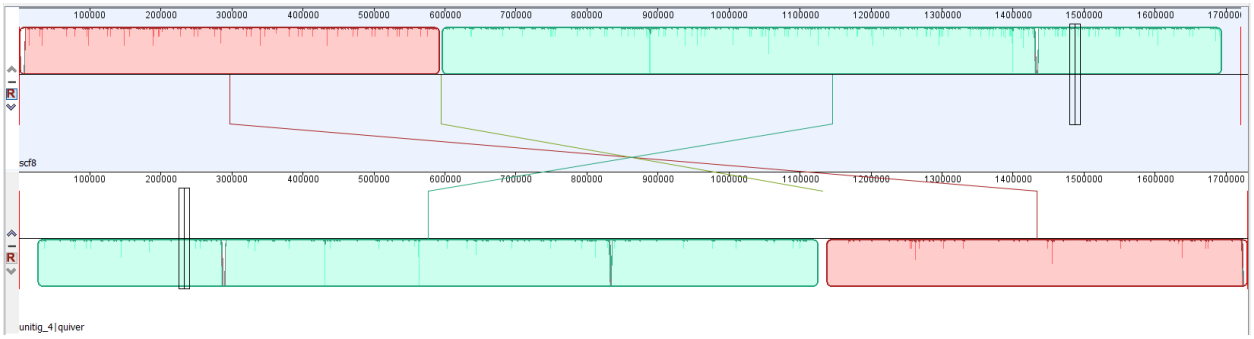


Figure 7: Mauve alignment of A scaffold 8 to unitig 4

Most interesting to this discussion is the presence of a long unitig 0 from the *P. knowlesi* C assembly that appears to be made up of sequences from *P. knowlesi* A scaffold 11 and scaffold 13. This issue turned out to be nontrivial as unitig 0 could not just be split where it diverges between aligning to scaffold 11 and scaffold 13 since the exact position of that unitig in the C genome has not been validated. As such the large regions with no coverage in the assembled C scaffold 11 and 13 had to be left in place. This unitig 0 may potentially correspond to a completely rearranged chromosome, but such a conclusion would require additional data to validate. Furthermore, the region where the two different sections corresponding to different scaffolds "meet" on untig 0 may be intriguing to investigate to see if it's a region with many repeats, or other interesting features.

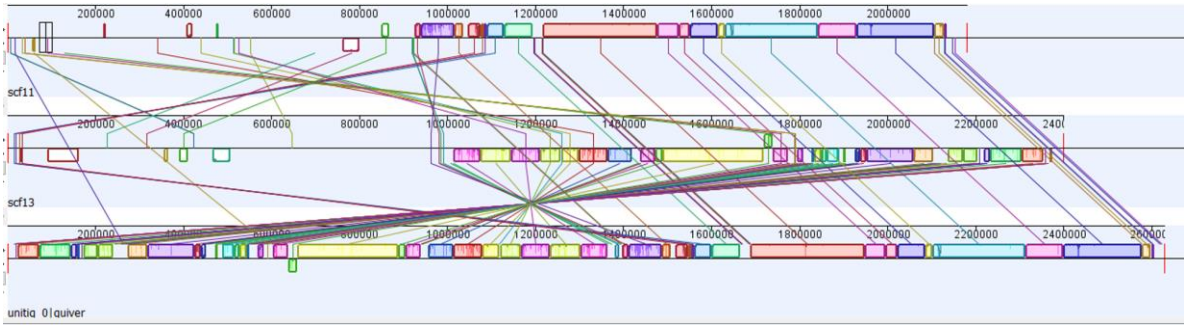


Figure 8: Matching alignments between C unitig 0 and A scaffold 11 and A scaffold 13

Thus, significant gaps remain in both the *P. knowlesi* B and C genome assemblies, with several unitigs from both assemblies failing to be placed into any of the current scaffolds for both assemblies. These gaps are attributable to multiple reasons, including potential errors from the PacBio sequencing, a lack of robustness of the Geneious read mapper to correctly place short length unitigs into a scaffold, and potential large-scale recombination events in those genomes that would make using the *P. knowlesi* A genome as a reference not very useful. All these explanations are ones that can be addressed with more careful validation of the current B and C genomes, with the hope that those gaps can be resolved.

4.2 RNA-Seq Analysis Discussion

Using a state-of-the-art quantification of the distribution of parasitic gene expression levels during the different stages of the *P. knowlesi* infected RBC cycle, we could determine parasitic cellular processes that may be associated with *SICAvar* expression. Of interesting note is the clear difference between the SICA[+] and SICA[-] RNA-Seq data analysis results. First of all, as would be expected due to the given clone names, there are much more *SICAvar* genes and fragments present in the SICA[+] RNA-Seq data than the SICA[-] RNA-Seq data, 163 compared to 30, respectively. Furthermore, there is also a marked difference in the gene ontology (GO) enrichment results from the genes in the largest cluster after the second clustering. In the SICA[+] GO enrichment results the majority of the top 10 processes identified by PlasmoDB are related to lipid biosynthesis, as mentioned above. In the SICA[-] GO enrichment results, though, the associated GO processes are much more diverse and general. With this in mind, it is evident

that in SICA[+] clones the parasite is much more focused on what is likely to be surface protein production with extensive upregulation of genes that code for lipid biosynthesis, whereas in SICA[-] the parasite expresses genes for more general cellular processes. This decrease in lipid biosynthesis activity may also be indicative of the downregulation of the production of other parasitic surface proteins that help infected red blood cells escape detection by the host immune system, further explaining how infection by SICA[-] clones are less virulent. While lipid biosynthesis would be an expected correlated biological function as SICA antigens are produced and displayed on the infected red blood cell membrane surface, further investigation of the pathways and cellular mechanisms associated with clusters whose genes do not appear to have closely related functionality can point towards a clearer picture of how changing levels of gene expression due to different host conditions can lead to variations in parasite-derived antigens. In addition, it would be intriguing to also determine the specific mechanisms that lead to the observed differences in gene expression profiles between SICA[+] and SICA[-] as transcription level modifications and potential translational repression are known to play a role in how the production of SICA antigens is regulated during the malaria lifecycle (Lapp et al 2013).

4.3 Consensus Clustering Discussion

Consensus clustering is a useful method to identify a more accurate set of clusters from those generated by the hierarchical, k-means, and SOM approaches to compensate for potential biases that may arise from the inherent drawbacks of each individual method. The consensus clustering approach (consensus by order of cluster size) I utilized to determine which

clusters from all three clustering algorithms matched and could be merged together to produce consensus clusters is a relatively naïve approach. In order to assess this aspect, I produced a Venn diagram showing how much overlap there was between clusters of relatively similar size from the results of the three clustering algorithms.

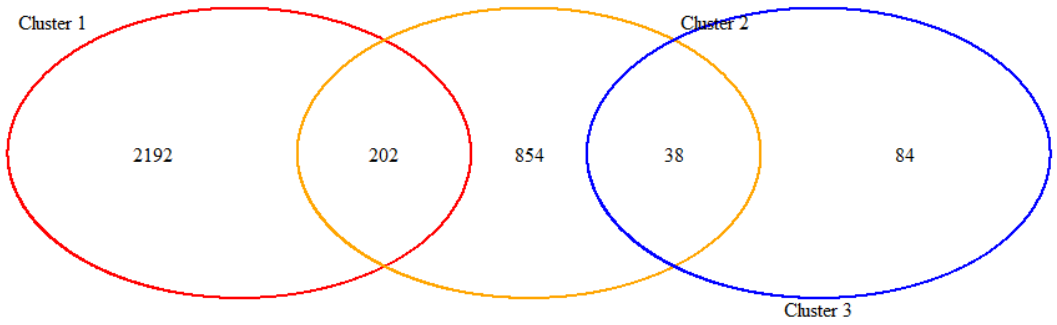


Figure 9.1: SICA[+] Venn diagram comparing similarity between the results of the three clustering algorithms when clusters are matched by size

Figure 9.2: SICA[-] Venn diagram comparing similarity between the results of the three clustering algorithms when clusters are matched by size

By using this Venn diagram, one should be able to visualize the level of similarity between the different clusters, as if the approach is valid we should see the majority of genes in only one circle of the Venn diagram but not in the overlap between the other circles. Genes that do fall in those overlaps can be determined to be unclear in terms of which consensus cluster they should be placed in and are confounds for the validity of this approach. Despite the Venn diagram, a more thorough approach to consensus clustering outside of sorting by largest sized cluster and placing the gene in its highest frequency cluster would ultimately be necessary to point to a more valid result. Such a comprehensive approach would require much more parameter tuning in terms of running more iterations of each clustering algorithm and finding

consensus between results from the same algorithm, the parameters of the clustering algorithms themselves, the selection of the value of k, comparison between the consensus clusters from different consensus approaches, just to name a few.

Furthermore, this approach also assumes that each clustering algorithm is capable of detecting significantly different patterns of expression levels between the different *P. knowlesi* genes from our RNA-Seq time series data. This validity is explored in the machine learning discussion section. Even so, applying clustering algorithms to RNA-Seq data has been previously attempted in other studies and has been determined to be useful for identifying potential regulatory interactions between different genes that have been clustered together (Lockhart and Winzeler 2000). As such, clustering should have a useful contribution towards exploring our own RNA-Seq data and looking for novel interactions between parasitic genes.

In addition, it may be necessary to undertake an alternate approach to consensus clustering. To determine relevant results from clustering, I need to be able to obtain a useful comparison between the results of the three different clustering algorithms. This can be accomplished by classifying each cluster that is produced by each algorithm by the biological pathway/cellular function that has the most corresponding genes in that cluster. After all the clusters from each result are classified, the clusters will be ranked according to percentage of *SICAvar* genes contained within that cluster. Thus, by ordering the clusters by the aforementioned rank, we can also order the biological pathways by that rank, and determine which pathways/functions are most highly associated with the expression of *SICAvar* genes. In

this manner, we can identify the biological pathways that potentially play a significant part in the regulation of the production of SICA antigens and the variation in the specific phenotypes being expressed.

The final potential issue is how useful consensus clustering even is in the first place. While there is research that indicates that consensus clustering may not be necessary (Nguyen and Caruana 2007), there is also significant evidence that points to the "wisdom of the crowds" and how consensus approaches can lead to results that are more accurate than individual algorithmic approaches when assessing using gold standards (Marbach et al 2012). With this in mind, exploring how consensus clustering can be applied to producing better results with RNA-Seq analysis warrants future study.

4.4 Machine learning discussion

A variety of machine learning approaches were applied to the *P. knowlesi* A+, B+, and C+ single timepoint RNA-Seq data. These approaches include linear regression and multilayer perceptron (MLP). The main issue with these results was a general lack of features per gene entry. For linear regression, there was only one timepoint measured for each sample, with nine samples in total, with three for each *P. knowlesi* clone (specifically A+, B+ and C+). Though each timepoint is from a different individual monkey, the monkeys that correspond to each *P. knowlesi* clone share similar host conditions, as *P. knowlesi* parasites isolated from those monkeys will be the same respective *P. knowlesi* clone expressing their specific protein repertoire. Still, the fact that each sample is technically a different individual presents a

potential issue of how the gene expression levels between the samples of different *P. knowlesi* clones can accurately be compared, since the purpose of the linear regression analysis is to describe how different features in the data contribute the most variability to the observed output results. As such, more timepoints for each sample would be useful for providing more data so that different samples can be compared with each other.

Moreover, after preprocessing the RNA-Seq dataset for the MLP classifier, 3953 genes out of the total 5384 (73.4%) *P. knowlesi* genes had to be excluded as they did not have a biological pathway associated with them on PlasmoDB. This reduced the usable data down to 1431 genes, making the dataset smaller than it already was, a complication further compounded by the large number of biological pathways, specifically 425 classes that the neural network would need to be trained to categorize different genes into, making the classification task difficult. In addition, using raw accuracy as a measurement for assessing the performance of the MLP classifier is a very strict definition as it involves multiclass classification with many more potential wrong classifications than the one correct classification.

Finally, the MLP neural network used in this study was prebuilt and run with default parameters. The architecture of a neural network should be designed specifically with the type of data that it will be trained with in mind. This level of specificity tailored to the dataset would require building a neural network from scratch, a complex task outside of the amount of time given to complete this study. Also, the performance of a neural network is known to be heavily dependent on parameter tuning to accommodate for the unique quirks of the dataset being

used to train the classifier, again a nontrivial task that would require multiple performance tests at different parameter settings and difficult to decide fully given the time frame of this project. As such, a higher level of optimization would be necessary to produce better results from the MLP neural network (Pedregosa et al 2011).

Chapter 5: Conclusion

Analyses were completed comparing the differences between the *P. knowlesi* A+, B+, and C+ clones, in addition to the differences between the SICA[+] and SICA[-] (specifically A+ and A-) clones. There is promising evidence of genomic recombination events between the *P. knowlesi* A+ and B+ clones and between the B+ and C+ clones due to multiple regions of sequence misalignments when comparing the different genomes to each other. Though these apparent "gaps" in the resulting assembled B+ and C+ scaffolds could point towards potential rearrangements, they are also attributable to multiple other factors like inherent PacBio sequencing errors and would thus require more data to validate the actual position of the *P. knowlesi* B+ and C+ assembly unitigs in their respective genomes. This necessary step of validation would allow for the assembly of more accurate scaffolds and thus improved genome comparisons.

Moreover, the RNA-Seq data analysis indicated a distinct difference in the genes being significantly expressed between the SICA[+] and SICA[-] clones. In the cluster with the most *SICAvar* genes, parasitic genes coding for the biological process associated with lipid biosynthesis predominated as the highest ranked GO processes in the SICA[+] clones whereas only more general biological processes were identified as the highest ranked GO processes for the SICA[-] clone. This decrease in lipid biosynthesis activity in SICA[-] clones may also be indicative of a decrease in the production of other parasite produced surface antigens, making it more difficult for infected red blood cells to evade the host immune system and therefore

allowing the infection to be controlled. As such, the biological processes associated genes assigned to clusters outside the ones with the most *SICAvar* genes would warrant future investigation, as well as utilizing other techniques like the R software package TCseq (Wu and Gu 2017) with our RNA-Seq data.

Finally, though preliminary results from applying machine learning approaches to the RNA-Seq A+, B+, and C+ data were produced, those tools turned out to require much more work in data preprocessing and parameter tuning than we had time to complete in the course of this study. As such, the given results using linear regression and neural networks are promising explorations towards applying those techniques to analyze malaria RNA-Seq data and reveal underlying patterns of parasitic gene expression. Overall, multi-omic analysis of a wide variety of *P. knowlesi* data using several different computational approaches can confirm already known biological characteristics of *P. knowlesi* as well as provide fascinating insights into the numerous genomic and transcriptomic level differences between the clones of *P. knowlesi*, and thus the different factors that play a role in the SICA antigen variation.

References

Alberghina, L., & Westerhoff, H. V. (2007). Systems biology: definitions and perspectives. Springer Science & Business Media, 13.

Al-Khedery, B., Barnwell, J. W., & Galinski, M. R. (1999). Antigenic Variation in Malaria: a 3′ Genomic Alteration Associated with the Expression of a P. knowlesi Variant Antigen. Molecular Cell, 3(2), 131-141. doi:https://doi.org/10.1016/S1097-2765(00)80304-4

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology, 11(10). doi:10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166-169. doi:10.1093/bioinformatics/btu638

Antinori, S., Galimberti, L., Milazzo, L., & Corbellino, M. (2013). Plasmodium knowlesi: The emerging zoonotic malaria parasite. Acta Tropica, 125(2), 191-201. doi:https://doi.org/10.1016/j.actatropica.2012.10.008

Antony, H. A., & Parija, S. C. (2016). Antimalarial drug resistance: An overview. Tropical Parasitology, 6(1), 30-41. doi:10.4103/2229-5070.175081

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., . . . Wang, H. M. (2009). PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Research, 37, D539-D543. doi:10.1093/nar/gkn814

Bansal, M., Yang, J. C., Karan, C., Menden, M. P., Costello, J. C., Tang, H., . . . Community, N.-D. (2014). A community computational challenge to predict the activity of pairs of compounds. Nature Biotechnology, 32(12), 1213-U1269. doi:10.1038/nbt.3052

Barnwell, J. W., Howard, R. J., Coon, H. G., & Miller, L. H. (1983). SPLENIC REQUIREMENT FOR ANTIGENIC VARIATION AND EXPRESSION OF THE VARIANT ANTIGEN ON THE ERYTHROCYTE-MEMBRANE IN CLONED PLASMODIUM-KNOWLESI MALARIA. Infection and Immunity, 40(3), 985-994.

Berendt, A. R., Ferguson, D. J. P., & Newbold, C. I. (1990). SEQUESTRATION IN PLASMODIUM-FALCIPARUM MALARIA - STICKY CELLS AND STICKY PROBLEMS. Parasitology Today, 6(8), 247-254. doi:10.1016/0169-4758(90)90184-6

Brown, K. N., & Brown, I. N. (1965). IMMUNITY TO MALARIA - ANTIGENIC VARIATION IN CHRONIC INFECTIONS OF PLASMODIUM KNOWLESI. Nature, 208(5017), 1286-&. doi:10.1038/2081286a0

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M.-A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis comparison tool. Bioinformatics, 21(16), 3422-3423. doi:10.1093/bioinformatics/bti553

Centers for Disease Control and Prevention (CDC). (2015). Malaria - Disease. Retrieved from https://www.cdc.gov/malaria/about/disease.html

Centers for Disease Control and Prevention (CDC). (2015). Treatment of Malaria: Guidelines For Clinicians (United States). Retrieved from https://www.cdc.gov/malaria/diagnosis_treatment/clinicians2.html

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., . . . Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature, 517(7536), 608-611. doi:10.1038/nature13907

Chien, J.-T., Pakala, S. B., Geraldo, J. A., Lapp, S. A., Humphrey, J. C., Barnwell, J. W., . . . Galinski, M. R. (2016). High-Quality Genome Assembly and Annotation for Plasmodium coatneyi, Generated Using Single-Molecule Real-Time PacBio Technology. Genome Announcements, 4(5), e00883-00816. doi:10.1128/genomeA.00883-16

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., . . . Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10, 563. doi:10.1038/nmeth.2474 https://www.nature.com/articles/nmeth.2474#supplementary-information

Chou, K. C. (2009). Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. Current Proteomics, 6(4), 262-274. doi:10.2174/157016409789973707

Corredor, V., Meyer, E. V. S., Lapp, S., Corredor-Medina, C., Huber, C. S., Evans, A. G., . . . Galinski, M. R. (2004). A SICAvar switching event in Plasmodium knowlesi is associated with the DNA rearrangement of conserved 3 ' non-coding sequences. Molecular and Biochemical Parasitology, 138(1), 37-49. doi:10.1016/j.molbiopara.2004.05.017

Cowman, A. F., & Crabb, B. S. (2006). Invasion of Red Blood Cells by Malaria Parasites. Cell, 124(4), 755-766. doi:https://doi.org/10.1016/j.cell.2006.02.006

Cox-Singh, J., Davis, T. M. E., Lee, K. S., Shamsul, S. S. G., Matusop, A., Ratnam, S., . . . Singh, B. (2008). Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. Clinical Infectious Diseases, 46(2), 165-171. doi:10.1086/524888

Cui, L., & Su, X. (2009). Discovery, mechanisms of action and combination therapy of artemisinin. Expert Rev Anti Infect Ther, 7(8), 999-1013. doi:10.1586/eri.09.68

Daneshvar, C., Davis, T. M. E., Cox-Singh, J., Rafa'ee, M. Z., Zakaria, S. K., Divis, P. C. S., & Singh, B. (2009). Clinical and Laboratory Features of Human Plasmodium knowlesi Infection. Clinical Infectious Diseases, 49(6), 852-860. doi:10.1086/605439

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 14(7), 1394-1403. doi:10.1101/gr.2289704

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics. doi:10.1093/bioinformatics/bts635

Escalante, A. A., Freeland, D. E., Collins, W. E., & Lal, A. A. (1998). The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. Proceedings of the National Academy of Sciences of the United States of America, 95(14), 8124-8129. doi:10.1073/pnas.95.14.8124

Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., . . . Carucci, D. J. (2002). A proteomic view of the Plasmodium falciparum life cycle. Nature, 419(6906), 520-526. doi:10.1038/nature01107

Fonseca, L. L., Joyner, C. J., Galinski, M. R., & Voit, E. O. (2017). A model of Plasmodium vivax concealment based on Plasmodium cynomolgi infections in Macaca mulatta. Malaria Journal, 16(1), 375. doi:10.1186/s12936-017-2008-4

Francis, S. E., Sullivan, D. J., & Goldberg, D. E. (1997). Hemoglobin metabolism in the malaria parasite Plasmodium falciparum. Annual Review of Microbiology, 51, 97-123. doi:10.1146/annurev.micro.51.1.97

Frevert, U. (2004). Sneaking in through the back entrance: the biology of malaria liver stages. Trends in Parasitology, 20(9), 417-424. doi:10.1016/j.pt.2004.07.007

Galinski, M. R., & Barnwell, J. W. (2009). Monkey malaria kills four humans. Trends in Parasitology, 25(5), 200-204. doi:10.1016/j.pt.2009.02.002

Galinski, M. R., & Corredor, V. (2004). Variant antigen expression in malaria infections: posttranscriptional gene silencing, virulence and severe pathology. Molecular and Biochemical Parasitology, 134(1), 17-25. doi:10.1016/j.molbiopara.2003.09.013

Galinski, M. R., Lapp, S. A., Peterson, M. S., Ay, F., Joyner, C. J., Le Roch, K. G., . . . Voit, E. O. (2017). Plasmodium knowlesi: a superb in vivo nonhuman primate model of antigenic variation in malaria. Parasitology, 1-16. doi:10.1017/S0031182017001135

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods, 8(6), 469-477. doi:10.1038/nmeth.1613

Haas, B. J., & Zody, M. C. (2010). Advancing RNA-Seq analysis. Nature Biotechnology, 28(5), 421-423. doi:10.1038/nbt0510-421

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1-8. doi:https://doi.org/10.1016/j.ygeno.2015.11.003

Hempelmann, E. (2007). Hemozoin biocrystallization in Plasmodium falciparum and the antimalarial activity of crystallization inhibitors. Parasitology Research, 100(4), 671-676. doi:10.1007/s00436-006-0313-x

Hisaeda, H., Yasutomo, K., & Himeno, K. (2005). Malaria: immune evasion by parasites. International Journal of Biochemistry & Cell Biology, 37(4), 700-706. doi:10.1016/j.biocel.2004.10.009

Howard, R. J., Barnwell, J. W., & Kao, V. (1983). ANTIGENIC VARIATION IN PLASMODIUM-KNOWLESI MALARIA - IDENTIFICATION OF THE VARIANT ANTIGEN ON INFECTED ERYTHROCYTES. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 80(13), 4129-4133. doi:10.1073/pnas.80.13.4129

Josling, G. A., & Llinas, M. (2015). Sexual development in Plasmodium parasites: knowing when it's time to commit. Nature Reviews Microbiology, 13(9), 573-587. doi:10.1038/nrmicro3519

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., . . . Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12), 1647-1649. doi:10.1093/bioinformatics/bts199

Kim, Y., & Schneider, K. (2013). Evolution of Drug Resistance in Malaria Parasite Populations. Nature Education Knowledge, 4(8), 6.

Lapp, S. A., Geraldo, J. A., Chien, J. T., Ay, F., Pakala, S. B., Batugedara, G., . . . Kissinger, J. C. (2017). PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family. Parasitology, 145(1), 71-84. doi:10.1017/S0031182017001329

Lapp, S. A., Korir-Morrison, C., Jiang, J., Bai, Y., Corredor, V., & Galinski, M. R. (2013). Spleen-Dependent Regulation of Antigenic Variation in Malaria Parasites: Plasmodium knowlesi SICAvar Expression Profiles in Splenic and Asplenic Hosts. PLoS ONE, 8(10), e78014. doi:10.1371/journal.pone.0078014

Lapp, S. A., Mok, S., Zhu, L., Wu, H., Preiser, P. R., Bozdech, Z., & Galinski, M. R. (2015). Plasmodium knowlesi gene expression differs in ex vivo compared to in vitro blood-stage cultures. Malaria Journal, 14. doi:10.1186/s12936-015-0612-8

Le Roch, K. G., Chung, D. W. D., & Ponts, N. (2012). Genomics and integrated systems biology in Plasmodium falciparum: a path to malaria control and eradication. Parasite Immunology, 34(2-3), 50-60. doi:10.1111/j.1365-3024.2011.01340.x

Le Roch, K. G., Zhou, Y. Y., Blair, P. L., Grainger, M., Moch, J. K., Haynes, J. D., . . . Winzeler, E. A. (2003). Discovery of gene function by expression profiling of the malaria parasite life cycle. Science, 301(5639), 1503-1508. doi:10.1126/science.1087025

Lockhart, D. J., & Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. Nature, 405, 827. doi:10.1038/35015701

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. doi:10.1186/s13059-014-0550-8

Maier, A. G., Cooke, B. M., Cowman, A. F., & Tilley, L. (2009). Malaria parasite proteins that remodel the host erythrocyte. Nature Reviews Microbiology, 7(5), 341-354. doi:10.1038/nrmicro2110

Marbach, D., Costello, J. C., Kueffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., . . . Consortium, D. (2012). Wisdom of crowds for robust gene network inference. Nature Methods, 9(8), 796. doi:10.1038/nmeth.2016

Miller, L. H., Baruch, D. I., Marsh, K., & Doumbo, O. K. (2002). The pathogenic basis of malaria. Nature, 415(6872), 673-679. doi:10.1038/415673a

Nguyen, N., & Caruana, R. (2007). Consensus clusterings. In N. Ramakrishnan, O. R. Zaiane, Y. Shi, C. W. Clifton, & X. D. Wu (Eds.), Icdm 2007: Proceedings of the Seventh Ieee International Conference on Data Mining (pp. 607-612).

Otto, T. D., Wilinski, D., Assefa, S., Keane, T. M., Sarry, L. R., Böhme, U., . . . Llinás, M. (2010). New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Molecular Microbiology, 76(1), 12-24. doi:10.1111/j.1365-2958.2009.07026.x

Pain, A., Bohme, U., Berry, A. E., Mungall, K., Finn, R. D., Jackson, A. P., . . . Berriman, M. (2008). The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature, 455(7214), 799-U797. doi:10.1038/nature07306

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Metabolomics: the apogee of the omics trilogy. Nature Reviews Molecular Cell Biology, 13(4), 263-269. doi:10.1038/nrm3314

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Rich, S. M., & Xu, G. (2011). Resolving the phylogeny of malaria parasites. Proceedings of the National Academy of Sciences of the United States of America, 108(32), 12973-12974. doi:10.1073/pnas.1110141108

Sachs, J., & Malaney, P. (2002). The economic and social burden of malaria. Nature, 415(6872), 680-685. doi:10.1038/415680a

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463-5467.

Singh, B., Sung, L. K., Matusop, A., Radhakrishnan, A., Shamsul, S. S. G., Cox-Singh, J., . . . Conway, D. J. (2004). A large focus of naturally acquired Plasmodium knowlesi infections in human beings. The Lancet, 363(9414), 1017-1024. doi:10.1016/S0140-6736(04)15836-4

Soon, W. W., Hariharan, M., & Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. Molecular Systems Biology, 9(1). doi:10.1038/msb.2012.61

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., . . . Robinson, G. E. (2015). Big Data: Astronomical or Genomical? PLOS Biology, 13(7), e1002195. doi:10.1371/journal.pbio.1002195

Team, R. C. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Waller, R. F., Keeling, P. J., Donald, R. G. K., Striepen, B., Handman, E., Lang-Unnasch, N., . . . McFadden, G. I. (1998). Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum. Proceedings of the National Academy of Sciences, 95(21), 12352-12357. doi:10.1073/pnas.95.21.12352

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63. doi:10.1038/nrg2484

Wehrens, R., & Buydens, L. M. C. (2007). Self- and super-organizing maps in R: The kohonen package. Journal of Statistical Software, 21(5), 1-19.

World Health Organization (WHO). (2014). Severe Malaria. Tropical Medicine and International Health, 19, 7-131.

World Health Organization (WHO). (2015). World Malaria Report 2015. Retrieved from http://www.who.int/malaria/publications/world-malaria-report-2015/report/en/

World Health Organization (WHO). (2016). Malaria vaccine: WHO position paper – January 2016. Retrieved from http://www.who.int/wer/2016/wer9104.pdf?ua=1

World Health Organization (WHO). (2016). World Malaria Report 2016. Retrieved from http://www.who.int/malaria/media/world-malaria-report-2016/en/

World Health Organization (WHO). (2018). Malaria - Overview of malaria treatment. Retrieved from http://www.who.int/malaria/areas/treatment/overview/en/

Wu, M., & Gu, L. (2017). TCseq: time course sequencing data analysis. Retrieved from https://www.bioconductor.org/packages/release/bioc/vignettes/TCseq/inst/doc/TCseq.pdf

Zhong, S., Joung, J.-G., Zheng, Y., Chen, Y.-r., Liu, B., Shao, Y., . . . Giovannoni, J. J. (2011). High-Throughput Illumina Strand-Specific RNA Sequencing Library Preparation. Cold Spring Harbor Protocols, 2011(8), pdb.prot5652. doi:10.1101/pdb.prot5652