Distribution Agreement

Signature:

________________________________ ________________

Date

An Application of DIKW Model in Malaria Systems Biology Study:

From NGS Data to Disease Progression Insight

By

Jung-Ting (Bob) Chien

Ph.D., Computer Science and Informatics, Emory University, Atlanta, GA, USA, 2016

Mary R. Galinski, Ph.D.

Advisor

Zhaohui "Steve" Qin, Ph.D.

Co-Advisor

Jinho Choi, Ph.D.

Committee Member

Steven E. Bosinger, Ph.D.

Committee Member

Juan B. Gutierrez Ph.D.

Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

An Application of DIKW Model in Malaria Systems Biology Study:
From NGS Data to Disease Progression Insight

By

Jung-Ting “Bob” Chien
M.S, Computer Science, Emory University, Atlanta, GA, USA, 2016

Advisor: Mary R. Galinski Ph.D
Co-Advisor: Zhaohui "Steve" Qin, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
In partial fulfillment of the requirement for the degree of
Doctor of Philosophy
In Computer Science
2017

Abstract

The data, information, knowledge and wisdom (DIKW) model has been widely used in data science fields to generate a comprehensive view of each domain. It provides a hierarchical representation of the understanding of the domain knowledge; the DIKW model can reveal insights in systems biology by integrating different types of –omics data to form a comprehensive understanding.

The foundation of systems biology is mining genomics data with machine learning. As the use of high-throughput, next-generation sequencing (NGS) applications grows, research in genomics enters the "big data" era. NGS applications can be divided into two major categories, short-read and long-read techniques, which are based on the principle differences in generating "reads". A "read" is the fundamental element of genomic information. Short-read applications have been widely applied in several fields of genomics research, while long-read applications just came to market in 2011. Long-read applications have shown the potential to handle several areas of genomic questions. However, obtaining a well-defined genome still has a number of challenges in malaria systems biology research, and these challenges block researchers' understanding the mechanism of the malaria disease progression.

To tackle these challenges, we built a novel long-read NGS pipeline with third party modules and modified them to solve *complicated Plasmodium* genome assembly questions. These techniques provided a solution where traditional, short-read technologies could not because of the *Plasmodium* genome's highly repetitive nature. We also implemented infrastructure to solve data management difficulties and developed several novel and robust pipelines to process and analyze the data. We host this pipeline along with other third party applications for data quality control, generic data visualization and data management tools. Our pipeline is also scalable and flexible to combine different technologies (long reads and short reads) to assemble the *Plasmodium* genome and conduct downstream annotations.

This dissertation describes an overview of –omics research in the big data era and reveals the possibility of applying DIKW models through mining genomics data. There will also be detailed discussion on how to apply our platform to solve questions, including multiple *Plasmodium* genome assemblies and annotations, and an initial discussion of applying machine learning approaches in a host-pathogen transcriptome analysis and its data mining applications.

An Application of DIKW Model in Malaria Systems Biology Study:
From NGS Data to Disease Progression Insight

By

Jung-Ting “Bob” Chien
M.S, Computer Science, Emory University, Atlanta, GA, USA, 2016

Advisor: Mary R. Galinski Ph.D
Co-Advisor: Zhaohui "Steve" Qin, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
In partial fulfillment of the requirement for the degree of
Doctor of Philosophy
In Computer Science
2017

Acknowledgements

My time as a Ph. D. student was marked by the opportunity to work with a number of talented scientists, both inside and outside of my field. These people took the time to share their knowledge and insights not only on science, but on life. Among these individuals is my academic advisor, Dr. Mary Galinski, whose guidance has helped me navigate my doctoral studies and whose support and care has made her a friend. I would like to recognize the many collaborators who have helped create the research described in this dissertation. Without their willingness to share samples and data, this work would not have been possible. I would like to thank Dr. Jinho Choi, Dr. Juan Gutierrez, and Dr. Steve Qin for the insights they have provided on my dissertation.

I am grateful to all the other people I met during my studies at Emory University. I would like to thank my colleagues, and the MaHPIC consortium, for support and all the discussions that helped to create an initial design of my work.

Finally, I owe my thanks to my wife, Amelia Longo for her support, patience and hard work. I must say that her presence and warm words were not only help, but also encouragement and motivation for me. Also, I would like to thank my family (including Toby and Dusk) and friends for always supporting my academic pursuits with their love, encouragement, and care packages.

Table of Contents

List of Tables

List of Figures

Chapter 1, Introduction

1.1 Motivation

The DIKW (data, information, knowledge and wisdom) model has been widely discussed and applied in data science; however, this model has not yet been widely used in biological research, especially in systems biology. Along with bioinformatics tools, the DIKW model could serve as a roadmap to tackle complicated systems biology questions. Next generation sequencing represents massively paralleled DNA sequencing technology, which has reshaped genomic research. An entire human genome can be sequenced in a single day using NGS (Liu et al. 2012). Using previous generations of sequencing technology, such as Sanger, to decipher the human genome required more than a decade to conduct the final draft. There are two categories of NGS technologies, short-read and long-read, defined by the length of their reads. Short-read sequencing technologies (<1K) have been widely used and have provided huge improvements over Sanger sequencing; however, their disadvantages, especially the read length, make some biological problems almost impossible to solve, including characterizing a complex genomic mixture, assembling a large genome with repetitive family, and others (Quail et al. 2012). Single-molecule real-time (SMRT) sequencing, developed by Pacific BioSciences (PacBio), provides a "long-read" approach (>30K) to solve the difficulties of short-read sequencing technology. However, there is currently no solid end-to-end pipeline for long-read sequencing users, and there are still many difficulties that need to be solved before the wide-spread application of long-read technology, especially the issue of the high-error rate in each sequencing read. Providing a solution to overcome the error rate issue in several applications is necessary for downstream systems biology analysis. This dissertation addresses several long-read sequencing questions with robust and computationally inexpensive solutions. This pipeline can be adapted to different biological questions, including malaria whole genome sequencing and annotation (Chien et al. 2016).

1.2 Key Questions

1. There is no fully integrated data analysis pipeline for long-read sequencing applications. Can we solve the engineering problem by integrating several third party tools into a centralized pipeline? Can it be scalable? Can this pipeline perform reproducible results?

2. Long-read NGS technology has not been applied to a *Plasmodium* full genome assembly. Can we apply this technology and design an algorithm to assembly the highly repetitive *Plasmodium* genome assembly and complete the annotation? Can this solution be reproduced? Can long-read NGS technology be combined with other NGS technologies to improve the *Plasmodium* genome assembly?

3. What kind of data mining questions can we answer with a superior, long-read genome and huge amounts of genomic information?

4. Is the DIKW model a good fit to solve systems biology questions? What are the limitations of the DIKW model in systems biology studies?

1.3 Dissertation Hypotheses

By answering the major questions stated in Section 1.2, we hypothesize that an integrated cloud-based NGS data analysis system could be applied to different fields of infectious disease research. Together, with data mining techniques, we expect this system will play a key role in discovering the hidden genomics regulatory mechanisms behind the bigger systems biology. We expect the scalable cloud-based system can not only handle variant bioinformatics data types (e.g. fasta, fastq, SAM/BAM) but also provide robust analysis results. We also expect our system can provide the most accurate long-read sequencing results while handling different tasks from high-resolution sequencing to whole genome assembling. Moreover, we expect the final genomic products generated from our system will lead to other innovative and high impact studies while revealing secrets of systems biology in infectious diseases. A study conducted by Le Roch et al. (Le Roch, Chung, and Ponts 2012) described different aspects of systems biology (Figure 1.1), and we expect our system can at least handle two major areas of systems biology: genomics and transcriptomics.

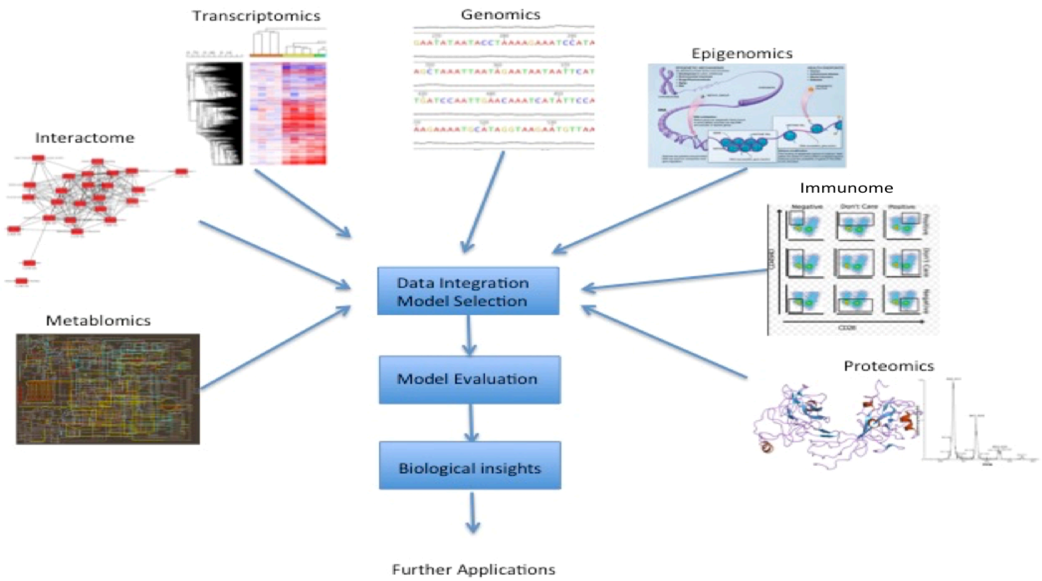


Figure 1.1, Different Components of Systems Biology (Le Roch et al. 2012).

Genomics identifies the genetic diversity and is important in detecting drug resistance, Epigenomics explores the transcriptional regulation, Transcriptome study helps us to evaluate RNA dynamics, Proteomes helps us to understand the protein levels, Immunome research characterizes protective antigens, Interactomes reveal protein-protein interactions, Metabolomic research helps us to evaluate metabolites.

1.4 Dissertation Organization

In this dissertation, we will discuss several topics. In Chapter 2, the current issues of systems biology are introduced, showing how data science has rapidly become recognized as an essential discipline in the post genomic era. Malaria scientists have been adapting to new ways of developing, analyzing and integrating large datasets using systems biology approaches. Numerous data points are generated, from different omics fields, including genomics, transcriptomics, proteomics, epigenomics, lipidomics and metabolomics. Applying machine learning based techniques to high throughput Next-Generation Sequencing (NGS) data is a promising means to construct the fundamental mechanistic understanding of such data in malaria research and identify new biological targets of intervention. We also discuss the fitness of the DIKW model along with NGS workflow strategies and the impact of NGS technologies in malaria post-genomics research, the utility of several bioinformatic approaches, and the data-driven concepts of information to knowledge and insights to tackle systems biology questions in new and powerful ways.

In Chapter 3, we demonstrate the basic workflow, from data processing to data interpretation. Our workflow has been successfully validated for the malaria whole genome assembly. We show that our workflow simply requires submission of raw data from PacBio sequencing to our pipeline; then with minimum tuning, our pipeline allows us to obtain a full genome assembly and annotation. Our methods are scalable and robust to several other *Plasmodium* genome assembly projects.

In Chapter 4, we demonstrated our full genome assembly pipeline along with other third party modules to assemble large complex genomes. We demonstrated feasibility by assembling a *Plasmodium coatneyi* genome. *Plasmodium coatneyi* is a protozoan parasite species that causes simian malaria and is an excellent model for studying disease caused by the human malaria parasite, *P. falciparum* (Moreno et al. 2013). Here we report the complete (non-telomeric) genome sequence of *P. coatneyi* Hackeri strain generated by the application of only Pacific Biosciences RS II (PacBio RS II) single-molecule real-time (SMRT) technology and assembled the genome using the Hierarchical Genome Assembly Process (HGAP) (Chin et al. 2013). This is the first

Plasmodium genome sequence reported using only PacBio technology. This approach has proven to be superior to short-read only approaches for this species. For further improvement, the Hi-C technology has been introduced along with the long-read based assembly method, and this hybrid method has resolved the contig orientation problems and improved assembly accuracy (Lapp, Geraldo, Chien *et al*. 2017).

Chapter 5 describes the use of the concatenated host and pathogen genomes to perform the dual RNA-seq analysis to understand the interaction between hosts and pathogens in disease progression. A prior assembled *Plasmodium coatneyi* genome was used to avoid sequencing bias in this dual RNA-seq analysis. By combining several data mining approaches, a comprehensive host-pathogen time-series transcriptome analysis was conducted. We will also discuss further applications of these data mining and machine learning techniques.

1.5 Contributions

This thesis describes the first long-read sequencing technology system hosted on the cloud; it produces experiment-specific, data-rich reports in industry-standard output formats. All data files are accessible directly to the user, and the system allows for easy third-party software analysis and collaboration. This system offers the ability to assemble reads into a *de novo* genome and analyze RNA-sequence data.

1.6 References

Chien, Jung-Ting, Suman B Pakala, Juliana A Geraldo, Stacey A Lapp, Jay C Humphrey, John W Barnwell, Jessica C Kissinger, and Mary R Galinski. 2016. "High-Quality Genome Assembly and Annotation for *Plasmodium Coatneyi*, Generated Using Single-Molecule Real-Time PacBio Technology." *Genome Announcements* 4 (5). 1752 N St., N.W., Washington, DC: American Society for Microbiology: e00883-16. doi:10.1128/genomeA.00883-16.

Chin, Chen-Shan, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nat Meth* 10 (6). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 563–69. http://dx.doi.org/10.1038/nmeth.2474.

Dilernia, Dario A, Jung-Ting Chien, Michael P S Brown, Zachary Ende, Martin J Deymier, Ling Yue, Ellen E Paxinos, Susan Allen, Alfredo Tirado-ramos, and Eric Hunter. 2015. "Multiplexed Highly-Accurate DNA Sequencing of Closely-Related HIV-1 Variants Using Continuous Long Reads from Single Molecule , Real-Time Sequencing," 1–13. doi:10.1093/nar/gkv630.

Ferrarini, Marco, Marco Moretto, Judson A Ward, Nada Šurbanovski, Vladimir Stevanović, Lara Giongo, Roberto Viola, et al. 2013. "An Evaluation of the PacBio RS Platform for Sequencing and de Novo Assembly of a Chloroplast Genome." *BMC Genomics* 14 (October). BioMed Central: 670. doi:10.1186/1471-2164-14-

670.

Le Roch, Karine G, Duk-Won D Chung, and Nadia Ponts. 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.

Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. 2012. "Comparison of next-Generation Sequencing Systems." *Journal of Biomedicine & Biotechnology* 2012: 251364. doi:10.1155/2012/251364.

Moreno, Alberto, Monica Cabrera-Mora, AnaPatricia Garcia, Jack Orkin, Elizabeth Strobert, John W Barnwell, and Mary R Galinski. 2013. "*Plasmodium Coatneyi* in Rhesus Macaques Replicates the Multisystemic Dysfunction of Severe Malaria in Humans." Edited by J H Adams. *Infection and Immunity* 81 (6). 1752 N St., N.W., Washington, DC: American Society for Microbiology: 1889–1904. doi:10.1128/IAI.00027-13.

Quail, Michael, Miriam E. Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas Richard Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. "A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers." *BMC Genomics* 13 (1). BMC Genomics: 1. doi:10.1186/1471-2164-13-341.

S.A. LAPP, J.A. GERALDO, J.-T. CHIEN, F. AY, S. B. PAKALA, G. BATUGEDARA, J. HUMPHREY, the MaHPIC consortium, J.D. DEBARRY, K. G. LE ROCH, M. R. GALINSKI and J. C. KISSINGER. 2017 "PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family" Parasitology.

Chapter 2, Systems Biology in Big Data Era

2.1 Introduction to the Fundamental Elements of Systems Biology

The systems biology modeling concept can be traced back to 1952, when Alan Lloyd Hodgkin and Andrew Fielding Huxley published the first cell biology numerical simulation mathematical model to explain the action potential propagation between the axon and neuronal cell (Hodgkin and Huxley 1952). This model described the interaction of two molecular components, potassium and sodium channels, at the cellular level and therefore has been considered as the first paper on computational systems biology. Alan Turing also published a paper in 1952 describing how the natural patterns of stripes, spirals, and spots, might arise naturally out of a uniform, homogeneous state; a theory which has served as classical model in theoretical biology (Turing, Aug, and Turing 2007). After several decades, systems biology has evolved naturally alongside technological progression by adapting several computational approaches. In fact, each field of systems biology is more and more computationally intensive. The main research interest in systems biology is therefore in developing and using efficient algorithms, data structures, and data visualization and communication applications for building a computational model to describe a biological system with subsystems including signal transduction pathways, gene regulation networks and networks of metabolites.

Systems biology is an inter-disciplinary field of study that focuses on interactions within a biological system by applying computational and mathematical modeling techniques (Ideker, Galitski, and Hood 2001). Recent developments in the computational sciences have resulted in the creation of tools that allow for important advanced approaches to systems biology. For example, parallel cloud computing techniques allow systems biologists to process terabytes of data on a daily basis using workflows from genomics or transcriptomics research. However, these advances require multiple data types to be assimilated into a single representation and massive quantities of data to be correctly handled. As such, the three aspects of a computational approach are referred to as the three "V"'s of big data: velocity, variety, and volume (Ott, Longnecker, and Ott 2001). Thus, the properties of systems biology questions are essentially similar to Big Data questions especially in data generating speed, data variety and data volume. In this chapter, we will discuss different elements of systems biology under the DIKW scheme, but systems biology is not limited to the elements described here.

2.2 Introduction to Current Genomics

After decades of genome sequencing progress, a critical milestone for scientific advances in medical research was the sequencing of the entire human genome at the affordable cost of less than $1000 USD (Mardis 2006). Successfully reaching this target signaled that bioinformatics had entered the post-genome big-data analysis era, and this marked the beginning of personalized medicine and systems biology strategies (Weston and Hood 2004; Gonzalez-Angulo, Hennessy, and Mills 2010). Bioinformatics researchers have since focused on big data-driven research, creating new large data collection tools, data storage, and data analyzing approaches that are being applied in malaria research. Following on the human genome success, several nonhuman (NHP) primate genomes have also been sequenced (Rogers and Gibbs 2014), greatly supporting the study of host-pathogen interactions using NHP model systems, as performed by the Malaria Host-Pathogen Interaction Center (MaHPIC) (Joyner et al. 2016; MaHPIC consortium, n.d.).

In recent years, Pacific Biosciences and Illumina technologies have become recognized as the two main next generation sequencing (NGS) approaches (Quail et al. 2012). These NGS sequencing technologies generate chain lengths of DNA data composed of either long (>500 base pairs) or short reads (25-500 bp), respectively. Illumina technologies generate short reads based on a massively parallel sequencing technique, a revolutionary and continuously improving industry standard for over 10 years. In the last five years, PacBio technologies have gained ground with the additional advantage of being capable of generating long reads (Liu et al. 2012). Once reads, whether short or long, are generated and ordered into contiguous sequences called contigs, researchers then have highly scalable and quantifiable information regarding the nature and composition of the target sequences. All NGS applications can be separated into several categories based on the different data types they assess, including not only whole genome *de novo* sequencing, but also exon sequencing, transcriptome sequencing (RNA-seq), and chromatin immunoprecipitation sequencing (ChIP-seq), to name a few.

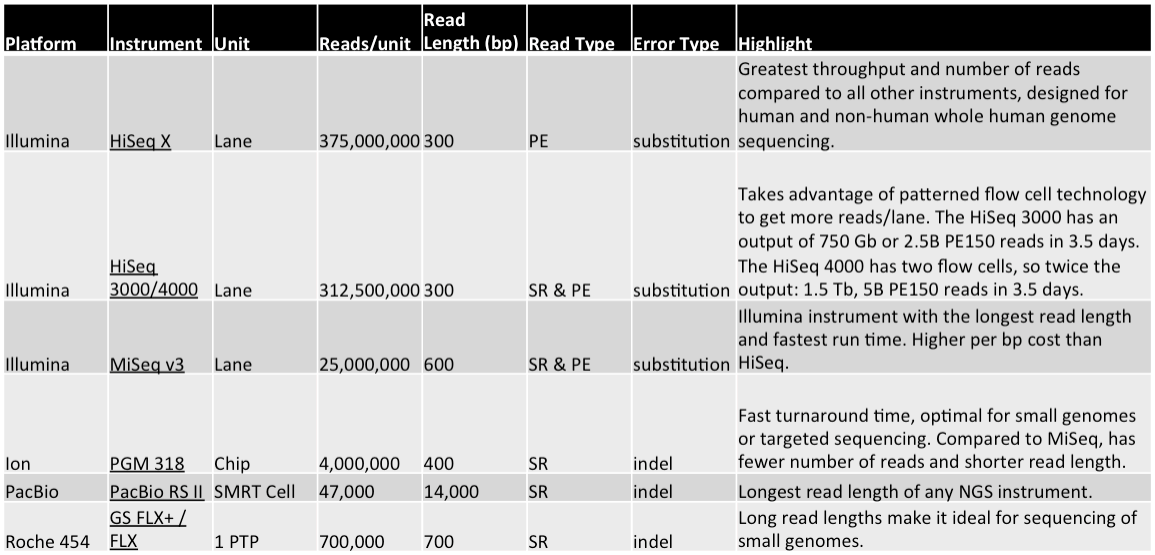
More specifically, Illumina technologies generate millions of short reads to analyze a target genome, which bioinformaticians then process and analyze with different software packages, such as the Genome Analysis ToolKit (GATK), software packages developed by the Broad Institute (DePristo et al. 2011), and SAMtools, which handles short reads datasets (H. Li et al. 2009). These two software packages are most widely used for finding single-nucleotide polymorphism (SNPs) or small insertions and deletions (INDELs), each with different basic assumptions. The basic assumption behind the GATK model is that errors are independent, but SAMtools attempts to model the error dependency. Given the frequency as 1, the SAMtools model results will be similar to those generated from the GATK model. For most practical cases, the two models are actually rather similar in the face of shallow sequencing coverage; however, these two models do differ when high sequencing coverage (>100X) is assessed (H. Li et al. 2009). For single-sample SNP calling, high phred-scaled genotype likelihoods are not needed, thus, the difference has a slightly noticeable effect on detecting SNPs (Yu and Sun 2013). Additionally, several open-sourced bioinformatics tools are available, covering a wide range of topics. For instance, these include ABySS for *de novo* assembly (Simpson et al. 2009), Bowtie for alignment (Langmead 2010), and RNA-STAR for transcriptomics research (Dobin et al. 2013). In general, technologies employing short reads are better able to locate SNPs (Hert, Fredlake, and Barron 2008). While short-read technologies possess certain advantages, these technologies require the building of large DNA scaffolds, particularly if the target sequence has numerous highly repetitive regions across the whole genome. This is especially true for inverted repeats and GGC sequences, whose presence can lead to gaps in the final genome assembly (Alkan, Sajjadian, and

Eichler 2011). As highlighted above for the *SICAvar* gene family, repetitive sequences are known to complicate the process of assembling *Plasmodium* genomes, and gaps have been commonplace until recently, with the application of PacBio technologies (Chien et al. 2016; Vembar et al. 2016; Rutledge et al. 2016).

Single Molecule, Real-Time (SMRT) DNA sequencing is a third generation sequencing technology, developed by PacBio, and this is the only NGS technology currently capable of processing long-reads datasets. The newest PacBio chemistry, P6-C4, can generate DNA sequence reads with lengths averaging 10-14 kb. Furthermore, this method has significantly higher throughput (between 500 million to 1 billion bases per SMRT Cell), from two megabases to upwards of one gigabyte. Longer DNA sequence reads have a higher chance of being appropriately mapped to the genome being sequenced, and even to highly repetitive regions, which have remained problematic for other platforms (Shin et al. 2013). However, the high sequencing error rate (Quail et al. 2012), particular to this technology, still poses a challenge. Since this technology detects the presence of individual fluorescent nucleotides from single molecules entering the DNA polymerase active site in real-time, there are two common sources of error. First, some nucleotides might be sampled by the polymerase but not immediately incorporated, yet become mistakenly detected as an additional nucleotide and leading to the insertion of an erroneous base in the resulting sequence. Furthermore, even nucleotides that are incorporated may not dwell long enough in the well to be detected, leading to the creation of an erroneous gap (Travers et al. 2010). In contrast, Illumina technologies are able to reduce the likelihood of that type of error by generally detecting ensembles of many molecules to increase signal strength. For example, Illumina cluster generation algorithms create up to 1,000 copies of the template to ensure more accurate detection and validate the existence of nucleotides in the sequence.

To compensate for these shortcomings in current long-reads technologies, “hybrid” approaches have been implemented to combine the high accuracy of Illumina technology with long reads generated from SMRT sequencing (Au et al. 2012; English et al. 2012). However, recently, the non-hybrid approach called Hierarchical Genome Assembly Process (HGAP) (Chin et al. 2013) has proven to be capable of generating accurate assemblies while relying exclusively on SMRT reads. This is accomplished by using the longest reads as ‘seed’ reads that can then be used as reference sequences for the mapping of shorter reads, leading to the formation of so called “pre-assembled reads”. These pre-assembled reads have the advantage of lower error rates since they are consensus sequences derived from the combination of multiple shorter reads. Pre-assembled reads can be further assembled together to build a final consensus sequence, producing near-complete, high-quality genomes *de novo*. However, two significant problems remain with this method. It is still necessary to have high coverage for the final consensus sequence to compensate for the high error rate that results from using long reads. Moreover, it is still not feasible to discern the nucleotide sequence of more than one genetic variant if or when multiple variants are present in the genomic DNA sample being sequenced. Complex mixtures (even diploid genomes) are problematic for long-reads-only approaches because closely related genomes have the potential to be mixed together and appear in the data as noise, leading to the mapping of a mixture of slightly different sequences, which the consensus algorithms must then resolve to create the single best consensus. To address this issue, Dilernia *et al*. developed a novel classification algorithm for overcoming this difficulty by only applying long-reads DNA sequencing to separate closely-related genomic mixtures, a technique that has been validated on HIV-1 genomes (Dilernia et al. 2015).

Table 2.1 Comparison of Major Sequencing Platform



Platform	Instrument	Unit	Reads/unit	Read Length (bp)	Read Type	Error Type	Highlight
Illumina	HiSeq X	Lane	375,000,000	300	PE	substitution	Greatest throughput and number of reads compared to all other instruments, designed for human and non-human whole human genome sequencing.
Illumina	HiSeq 3000/4000	Lane	312,500,000	300	SR & PE	substitution	Takes advantage of patterned flow cell technology to get more reads/lane. The HiSeq 3000 has an output of 750 Gb or 2.5B PE150 reads in 3.5 days. The HiSeq 4000 has two flow cells, so twice the output: 1.5 Tb, 5B PE150 reads in 3.5 days.
Illumina	MiSeq v3	Lane	25,000,000	600	SR & PE	substitution	Illumina instrument with the longest read length and fastest run time. Higher per bp cost than HiSeq.
Ion	PGM 318	Chip	4,000,000	400	SR	indel	Fast turnaround time, optimal for small genomes or targeted sequencing. Compared to MiSeq, has fewer number of reads and shorter read length.
PacBio	PacBio RS II	SMRT Cell	47,000	14,000	SR	indel	Longest read length of any NGS instrument.
Roche 454	GS FLX+ / FLX	1 PTP	700,000	700	SR	indel	Long read lengths make it ideal for sequencing of small genomes.

2.3 Introduction to Current Proteomics

In order to understand the proteomics of *Plasmodium* infected animal model or clinical cases, the main task of proteomics research is to produce detailed proteomic profiles from selected experiments quantitatively (Bautista et al. 2014). Thus, proteomics research in systems biology research serves to provide positive identifications of both the host and pathogen proteins and post-translational modifications (PTM) acquired by in-depth screening of peptide fractions that have been generated after digestion or chemical cleavage (Aebersold and Mann 2003). Normally, proteomics research provides quantitative comparisons of relevant samples by applying tandem mass spectrometry (MS/MS). This data analysis workflow can be customized, but the main idea is very similar to metabolomics research. As a result, proteomics research involves identifying the pattern spectrum peaks, validating the identifications, building up the protein inference, and then conducting the quantification of the finding (Huang et al. 2012). For the identification step, the current best approach is to utilize search engines such as Mascot (Weatherly et al. 2005) and X! Tandem (Vaudel et al. 2011; Craig and Beavis 2004), for searching the peptide MS "fingerprint" in several open-sourced online databases. Because *Plasmodium* is a higher-level eukaryote, Apicomplexan parasitic organism with complicated protein systems, one cannot gain a complete understanding of a protein's structure even after obtaining the peptide, thus applying statistics inference to predict the distribution of the peptide/proteome is essential (Swan et al. 2013). To address this issue, the software ProteinProphet can calculate the probabilities of each protein being present in the sample according to the possibilities of the association presented peptides (Nesvizhskii et al. 2003). Finally, data visualization that can be realized by applying R & Bioconductor after model construction, which can be machine-learning and statistical-analysis heavy.

2.4 Introduction to Current Metabolomics

Metabolomics research focuses on constructing the metabolic profile of experimental samples, whether plasma, serum, urine, sputum or other sample types collected from clinical cases, or *in vitro* model systems. Data acquisition is the first step for analyzing metabolomics data. Metabolomics data is collected using NMR, LC-MS, GC-MS (Jonsson et al. 2004; Young, Barrett Jr, and Beecher 2009; Ward et al. 2010), all of which yield spectral data. The spectral data then undergoes baseline correction, noise filtering, peak detection, peak alignment, normalization, and deconvolution to get an annotated data tables of features. After constructing the feature table, the analysis will move forwards in two directions, data analysis and metabolite identification (Goodacre et al. 2007; Shulaev 2006). For data analysis, machine-learning techniques that we have previously mentioned in section three will be applied (Dale, Popescu, and Karp 2010). By applying machine-learning techniques, researchers can construct the classification model or discover biomarker candidates. Machine-learning techniques are powerful when the dataset(s) is large and with numerous features, but how to select which method to apply is still dependent on the purpose and performance that one expects to acquire from the experiments (Martínez-Arranz et al. 2015). Identifying the metabolite is another important task in metabolomics research, which mainly requires relating a map of the novel metabolite back to the spectral databases based on the dataset features, including peak positions, and correlation patterns. The final step of metabolomics research is to generate a biological interpretation. Using tools like KEGG and MetaCYC (Martínez-Arranz et al. 2015; Krieger et al. 2004), the finding associated retention time and analytic conditions are first matched, then data mining is applied to find features that already exist. Applying data mining in databases can also be used for conducting pathway analysis, over-representation, and quantitative enrichment analysis. Recently, *Li et al.* presented a novel method that provides a faster solution to achieve high throughput metabolomics analysis. This method, called mummichog, determines functional activity from spectral features tables using knowledge of metabolic pathways and networks and is unique because it does not require a priori identification of metabolites (Salinas et al. 2014; S. Li et al. 2013). Once finished with these analyses, the next step usually involves constructing the model based on Gaussian graphical modeling, network topology measures and of course the model will then be presented elegantly with visualization form.

2.5 Malaria Genomics

Establishing a fundamental understanding of malaria systems biology from the genomics perspective, including a thorough examination of the complete sequence and expression characteristics of the genome of malaria-causing *Plasmodium* species is crucial for understanding the basis of *Plasmodium* biology and pathogenesis. The genome of *P. falciparum*, which has the highest mortality rate of all human malaria parasites, was published in 2002 as the first 'fully sequenced' *Plasmodium* genome (Gardner et al. 2002). Shortly afterwards, the genome sequence of *P. yoelii*, a rodent malaria parasite species that serves as a model system, was also published (Jane M Carlton et al. 2002). After decades of focused effort by the malaria research community, genome sequence information from >13 *Plasmodium* species/strains are now publicly available from PlasmoDB, an integrated multi omics database for multipurpose *Plasmodium* and malaria research studies (Aurrecoechea et al. 2009), and the genomes of most interest in this database are continually being improved upon with corrections to the original sequence or assembly data and its annotation. With the evolution of NGS technologies, the volume of novel information being generated and the diversity of different *Plasmodium* sequencing projects have increased as well. *Plasmodium* species that have been sequenced in the last 10 years include the most prevalent human malaria parasite *P. vivax* (J M Carlton et al. 2008), simian malaria parasites that serve as models for *P. falciparum* and *P. vivax*, respectively (*P. coatneyi* (Chien et al. 2016) and *P. cynomolgi* (Tachibana et al. 2012)), *P. knowlesi* (Pain et al. 2008a),which is a simian malaria parasite but also one that infects humans in South East Asia (Singh 2013), and another rodent model parasite, *P. berghei* (Janse and Waters 1995).

While each species differs, *Plasmodium* nuclear genomes each have 14 linear chromosomes encoding between 5000–5800 genes, and with the overall genome sizes ranging from 16 to 27.7 megabases; in addition, the parasite has mitochondrial and apicoplast organellar genomes (Arisue et al. 2012), (Vaidya and Mather 2009). A common feature of *Plasmodium* genomes across different species is the presence of high Adenine/Thymine-rich regions. This high A/T feature introduces sequencing difficulties due to a physical structural obstacle for constructing 'reads'. In addition, the *Plasmodium* genome has numerous repeated sequences, as well as complex multi-gene families that prove prohibitive for full genome completion with high confidence. A good example can be found in *P. knowlesi*, whose large (>100 member) and complex (multi-intron/exon) variant antigen encoding *SICAvar* multigene family (Al-Khedery, Barnwell, and Galinski 1999) is located all over the originally reported genome's 14 chromosomes, but in many instances, *SICAvar* sequences are contained in unmapped or misplaced fragments (Pain et al. 2008b; Lapp, Korir, and Galinski 2009). Several *SICAvar* ORFs have also been annotated as hypothetical proteins, a placeholder feature, due to insufficient sequence similarity when comparing to already characterized proteins in available databases. In light of these issues, long-read PacBio sequencing technology (described below) can provide a solution for the A/T bias and detect repetitive regions without linkage by taking advantage of the length of the reads being generated (Ferrarini et al. 2013). A 'PacBio-only' approach was in fact used recently to develop the first reported genome assembly for *P. coatneyi,* a species that is closely related to *P. knowlesi,* and which has a complex variant antigen gene family comparable to *SICAvar* (Chien et al. 2016). Chien *et al.* (Chien et al. 2016) brings to light the potential and necessity for long-reads technologies to support the development of new and improved, correctly assembled genomes for malaria research.

2.6 Malaria Transcriptomics

Now that several *Plasmodium* species and host genome sequencing and assembling projects have been completed, with reference genomes constructed and annotated, a solid foundation exists to investigate malaria transcriptomes, and in essence, many aspects of functional genomics. Functional genomics has been coined as a term to define research related to the functional expression and regulation of a genome, which results in the various phenotypic outcomes of a biological system or interacting systems at various levels; whether cellular, tissue, organ, organismal, or from infections involving a host and pathogen, not to mention microbiome and a multitude of environmental influence. It is one thing to study infected hepatocytes or red blood cells (RBCs) in *in vitro* cultures, and another to study them in the context of an infected host, where the host environment comes into play.

NGS technologies have advanced from genome sequencing to post-genomic utilities with RNA-seq analysis commonplace for studying transcriptome sequence data. In the past, microarrays were the standard technology for evaluating transcriptomes; however, microarrays are constrained in the amount of RNA that can be processed at a given time and this technology has limitations when it comes to quantifying expression levels with a large dynamic range, as well as the caveat of being unable to detect novel transcript regions that may not be represented on the arrays, leading to the possibility of some coding or non-coding RNAs being missed and ultimately not obtaining a full picture of gene expression activity. NGS techniques in transcriptome research have become recognized as a superior methodology for obtaining a more comprehensive view of transcriptome expression (Werner 2010).

To conduct transcriptome research, RNA samples are first isolated from biological samples, e.g. from tissue samples, or *in vitro* or *ex vivo* cultures of a model organism. Then, several preparatory steps can be applied, such as fragmentation of the extracted RNA and polymerase chain reaction (PCR) amplification. RNA fragmentation is useful for increasing the accuracy and reducing the error rate of the final sequencing step by allowing for shorter reads to be generated. Though PCR amplification can be an advantageous step for obtaining higher coverage depth or a stronger signal for reads of interest, especially when there is low input RNA, it may lead to the presence of artifacts in transcriptomic analysis where certain genes may appear to have much higher levels of expression than in actuality due to the biased presence of certain PCR amplicons. Once the extracted RNA has been processed, cDNA libraries of the RNA reads can be produced and then sequenced using an NGS platform to construct a comprehensive profile of the RNA reads sequences (Wang, Gerstein, and Snyder 2009; Martin and Wang 2011). The sequence reads can then be mapped onto host and parasite genomes, as appropriate to begin to study the changes of gene expression and how these relate to specific biological functions and infection outcomes – whether relating to immunity or disease pathogenesis and illness.

Because of the complicated life cycle of *Plasmodium*, including multiple hosts, stage-specific transcriptome studies have become an important starting point for understanding *Plasmodium* gene expression. Initial *in vitro* and *ex vivo* time course gene expression patterns have been constructed using microarray technology to understand *P. falciparum, P. vivax*, *P. knowlesi*, *P. berghei* ANKA, *P. yoelii*, *P. chabaudi* AS intraerythrocytic developmental cycles (IDC) (Bozdech et al. 2003; Llinás et al. 2006; Otto et al. 2014; Hoo et al. 2016; Zhu et al. 2016). IDC-based transcriptome profiling of these species has provided a global view of *Plasmodium* stage-specific gene expression behavior in the blood-stage parasites, however, transcriptome profiling of the liver-stage parasites remain unexplored across the species (Albuquerque et al. 2009). Of overarching interest are *Plasmodium* species that are capable of entering a hypnozoite stage in the liver, as these parasite forms can remain dormant in the host and then activate months or years later causing relapsing blood-stage infections (Douglas et al. 2012; White and Imwong 2012). This is the case for *P. vivax* and *P. ovale,* as well as several simian malaria model species including *P. cynomolgi*. While obtaining *Plasmodium* transcripts from human liver tissue is not feasible, they can be obtained from NHP models or *in vitro* cultures as a starting point for learning about the parasite's gene expression in the human liver environment (Joyner, Barnwell, and Galinski 2015).

Regarding the vector, transcriptome profiling of RNA from *P. berghei* infected *Anopheles stephensi* mosquitoes has been conducted by Xu *et al.* in 2005 (Xu et al. 2005). The advantage of profiling both *Anopheles* and *Plasmodium* simultaneously could lead to a better understanding of these host-pathogen interactions and the development of methods to block transmission by eliminating *Plasmodium* in the *Anopheles* mosquitoes without affecting the roles these mosquitoes play in the ecosystem. Likewise, rodent and NHP experimental model systems can allow for unprecedented profiling of the host and parasite, notably as infections are developing and the host is mounting a defensive response (Lovegrove et al. 2006; Hansen and Schofield 2010; Joyner et al. 2016; Reid and Berriman 2013).

Comparative studies are important for obtaining a better understanding of intra and inter species differences and to understand possible differences between model laboratory reared strains and wild strains under comparable growth conditions. Llinas *et al*. conducted a transcriptome comparison study of the IDC across three different strains of *P. falciparum* (Llinás et al. 2006). This study showed that the time series transcriptome for all three strains was highly conserved and that the main noticeable differences in gene expression between strains were for various transcripts encoding surface proteins that interact with the host immune system. More recently, Hoo *et al*. conducted an integrated transcriptome analysis of the IDC across six species of *Plasmodium*. This study revealed that the early ring stage of the lifecycle displayed the highest transcriptional variance between the different species, most significantly in genes that encode basic proteins crucial to the parasite's development and survival, such as transcription and translation. To explain this phenomenon, the authors suggested that the parasite's gene expression behavior was not in response to host-specific variations in surface receptors, but rather dependent on host-specific conditions encountered upon invading erythrocytes. This discovery implies that the earliest stages of the parasite's development may reflect genetically controlled initial adaptive behavior to adjust to the conditions inside a cellular environment, such as inside host erythrocytes (Hoo et al. 2016).

Importantly, host responses play an important role in influencing changes in the gene expression patterns. Clearly, parasites can have different up-regulated and down-regulated gene expression profiles depending on whether they are grown *in vitro* or *ex vivo*, indicating the importance of taking into consideration the possible *in vivo* influence of host effects when studying the parasite's growth and development behaviors in external experimental environments. Genomic expression control that is contingent on host conditions can be exemplified by the variable expression of the *SICAvar* multigene family (Al-Khedery, Barnwell, and Galinski 1999)(Howard, Barnwell, and Kao 1983). Barnwell *et al*. discovered that the erythrocytes of rhesus monkeys infected with *P. knowlesi* would display variant schizont-infected cell agglutination (SICA) antigens on their surface in the presence of a spleen, but would show loss of SICA protein expression (i.e., be SICA[-]) in monkeys that had their spleens removed (Howard, Barnwell, and Kao 1983). Lapp *et al.* since showed that SICA[-] parasites have downregulated - and in fact shutdown - their expression of *SICAvar* transcripts and proteins (Lapp et al. 2013). Early studies also showed differential expression of related variant antigens in *P. falciparum* whether in splenectomized or intact New World monkeys ("Surface Alterations of Erythrocytes in Plasmodium Falciparum Malaria. Antigenic Variation, Antigenic Diversity, and the Role of the Spleen" 1983).

2.7 Transcriptional Regulation: Epigenetics

Next-Generation Sequencing technologies have also benefitted malaria epigenetics research. ChIP-seq is a technology that combines chromatin immunoprecipitation (ChIP) with parallel sequencing (Landt et al. 2012). Applying ChIP-Seq can help to accurately detect the interactions between protein, DNA, and RNA as regulatory events for interpreting biological processes and disease states. Among the many advantages of ChIP-seq include the fact that it is not limited by array design, which has proven to be especially useful for species without commercially available arrays. Furthermore, it has better signal to noise ratio and dynamic ranges; as well as higher resolution than an array based predecessor ChIP-chip (Park 2009; Ho et al. 2011). Bisulfite-seq is designed to detect DNA methylation by measuring the methylation status of each nucleotide. For this sequencing technique, the DNA sample is treated with bisulfite so that un-methylated cytosine will be turned into uracil while methylated cytosine remains protected (still as cytosine), and other bases stay unaffected (Y. Li and Tollefsbol 2011).

Epigenetic mechanisms have been studied to understand *Plasmodium* gene regulation and antigenic variation (Ralph and Scherf 2005; Duffy et al. 2013; Scherf, Lopez-Rubio, and Riviere 2008), however, they have not yet been rigorously explored during malaria (Jiang et al. 2009). In recent year, Ponts *et al*. applied the Bisulfite-seq technique to reveal patterns of DNA methylation in *P. falciparum* and discovered the role of methylation in regulating transcripts (Ponts et al. 2013). Gupta *et al*. discovered 12 histone post-translational modifications during the asexual blood-stage cycle and identified eight histone modifications that were correlated with transcriptional regulation. This study shows a set of euchromatic histones that work in conjunction, generating a unique dynamic pattern of histone combinations that is related to gene expression during the development of the *Plasmodium* blood-stage life cycle (Gupta et al. 2013). A recent *P. falciparum* transcriptional start site (TSS) and transcriptional termination site (TTS) profiling study can serve as the basis for further histone modification research (Rawat, Bhosale, and Karmodiya 2016). Although much research has been focused on understanding the mechanisms in multigene families that code for variant antigens, other research has emphasized the importance of epigenetics in virulence, pathogenesis, and chromatin biology (Flueck and Baker 2014; Duraisingh and Horn 2016; Cortés et al. 2012). Being able to better characterize epigenetic regulation in *Plasmodium* may further reveal hidden regulatory mechanisms that can lead to the discovery of novel drug or vaccine targets (Ay et al. 2015).

2.8 The DIKW Model in Malaria Systems Biology

According to J. E. Cohen, "Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better" (Cohen 2004). The universe, an extremely complex entity with gravitational interaction, electromagnetic interaction, weak interaction and strong interaction, and has been investigated for centuries, yet a solution to unify all these aspects of the system continues to elude scientists. Biological systems can be viewed as similarly or more complex than the universe because of the many factors and interactions that drive the functioning and evolution of cellular and organismal components, let alone host-pathogen interactions (Alberghina and Westerhoff 2007; Voit 2016). Coincident with the rapid progression of technological advancement over the past decades, many computational tools have been developed for use in biological studies. In fact, attaining biological research results and insights by employing analytical capabilities of computing power are no longer a rarity but a daily occurrence.

The systems biology modeling concept can be traced back to 1952, when Alan Lloyd Hodgkin and Andrew Fielding Huxley published the first cell biology numerical simulation mathematical model to explain the action potential propagation between the axon and neuronal cell (Hodgkin and Huxley 1952). This model described the interaction of two molecular components, potassium and sodium channels, at the cellular level, and their work has been recognized as the first reported example of computational systems biology. Alan Turing also published a paper in 1952 describing how the natural patterns of stripes, spirals, and spots, might arise naturally out of a uniform, homogeneous state, a theory which has served as a classical model in theoretical biology (Turing, Aug, and Turing 2007). After several decades, systems biology as a scientific discipline has evolved naturally alongside technological progression by adapting several computational approaches. In fact, each field of systems biology has become more and more computationally intensive. The main research interest in systems biology is therefore focused on developing and using efficient algorithms, data structures, data visualization and communication applications to build computational models to describe a biological system with multiple, interconnected subsystems, including signal transduction pathways, gene regulation networks and networks of metabolites. The "Data, Information, Knowledge, and Wisdom" (DIKW) conceptual framework, described by Rowley et al. in 2007 (Rowley 2007), may prove useful for malaria systems biology research. In the past hundred years of malaria research, knowledge towards ending malaria has been acquired from basic research experiments, clinical observation, and clinical trials (KS et al. 2007; Galinski and Barnwell 2008; Buffet et al. 2011). During this time, humankind has suffered greatly due to the prevalence of this infectious disease. Considering the hundreds of millions of lives lost to malaria throughout human history, we currently possess relatively little knowledge about it. However, through the systems biology approach of a big picture view of malaria from the individual phenotypical perspective to the more expansive -omic level of interpretation, the speed of knowledge construction is significantly greater in today's big data era than it has ever been. According to Russell Ackoff, we can now simply see our human mind as several categories, which include data, information, knowledge (with understanding) and wisdom (Rowley 2007). Data represent the raw value of these measurements, which without significance or insights, simply exists. Furthermore, without processing, data will not become useful information. Information can provide insight on the data, when the data has been given significance by its relationship between and among each data point, or when the data has been endowed with a specific purpose. Knowledge is a sophisticated collection of information, so forming knowledge is a deterministic process, which involves selecting 'useful' information to be gathered into knowledge. With a greater understanding of knowledge, the process of synthesizing knowledge will occur naturally. Publications simply present the results of one's findings after combining several data points, interpreting them as information, and then combining previous works to form knowledge. While the previous three levels are all related to the past, wisdom is actually related to the future where we need to make a judgment or decision via sensible action to bring about beneficial consequences by applying the accumulation of knowledge through time. Moreover, wisdom is also the ability to deliver and to teach 'information' and 'knowledge'; therefore, gaining wisdom is not only the result of an individual's decisions but also the distribution of knowledge. Challenges that may seem intangible can become tangible.

Wisdom can be extracted from experience, and experience plays an important role in biological research because of the contributions of scholars who have spent years in their fields building knowledge in different subtopics. Nevertheless, experience can also be subjective because humans cannot execute certain actions as precise as machines, e.g. memorize all the contents of all articles related to malaria in PubMed. Therefore, the integration of technologies in both the biological and computational domains is one of the most important tasks for systems biologists, requiring an assembly of an objective 'wisdom' to construct systems for further research (Dale, Popescu, and Karp 2010; Ghosh et al. 2011). Based on the application of the DIKW scheme to the biological domain, Bernstam at el. describes the DIKW pyramid in biomedical information and knowledge systems from definitions to applications (Bernstam, Smith, and Johnson 2010).

In Figure 2.1, a comparable workflow of applying the DIKW model in malaria systems biology.

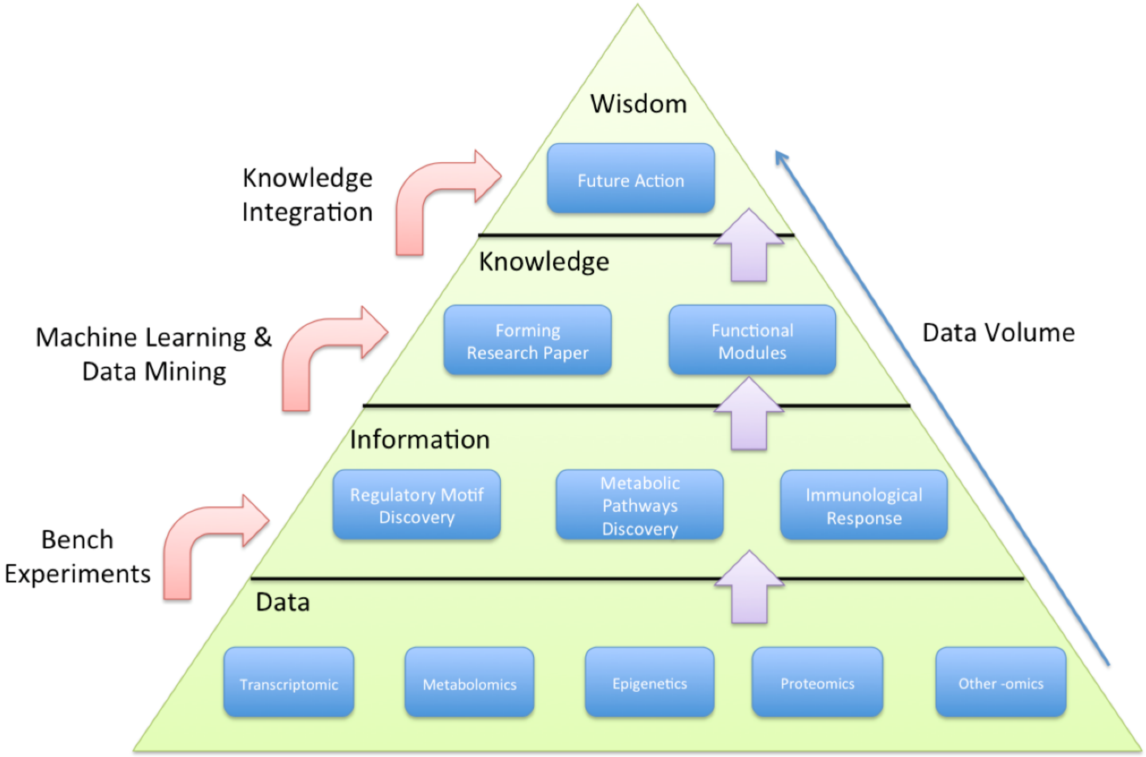


Figure 2.1 The DIKW pyramid model.

The levels of this model from bottom to top are data, information, knowledge and wisdom. In this diagram, each blue box represents the components of each level of pyramid.

2.9 Malaria Systems Biology: Data Science from Conceptualization to Action

The systems biology research consortium, known as the Malaria Host-Pathogen Interaction Center (MaHPIC), was designed in 2011 to systematically and comprehensively study malaria in nonhuman primate model systems and relate the findings to human malaria. The MaHPIC has developed since its official launch in 2012 as a working model for malaria systems biology with computational biology, mathematical modelling and iterative experimentation all being key. Scientists across the center generate multi-omic, clinical, parasitological, and immune profiling large datasets from samples obtained from nonhuman primate longitudinal experimental infections, as exemplified by Joyner *et al.* 2016 (Joyner et al. 2016), and from clinical samples obtained from malaria endemic country partners. The diverse data and associated metadata from each experiment are validated to enable analytical reproducibility, and then stored in a secure, internal data repository that serves as a file 'warehouse'. Experimental results are also loaded into the project's relational database where they can be systematically queried, including by mathematical modelers working to gain new insights by integrating diverse results (e.g. how the parasitological and clinical findings over time in a host relate to the transcriptome, metabolome and immune profiles). Finally, to foster use by the broad research community and the creation and testing of new hypotheses, raw data files, information and knowledge gained are shared with scientists worldwide in canonical public data repositories (e.g. NCBI, PlasmoDB, PRIDE, IMMport, Metabolomics Workbench, MassIVE) with links made accessible from the MaHPIC website. This method allows for the collaboration of researchers around the world from both experimental wet-labs and computational dry-labs and supports the development of confirmatory studies, validation experimentation, and open communication towards the shared goal of a better understanding of malaria. (Appendix 1)

2.10 References

Aebersold, Ruedi, and Matthias Mann. 2003. "Mass Spectrometry-Based Proteomics." *Nature* 422 (6928). Nature Publishing Group: 198–207.

Al-Khedery, Basima, John W Barnwell, and Mary R Galinski. 1999. "Antigenic Variation in Malaria: A 3′ Genomic Alteration Associated with the Expression of a *P. knowlesi* Variant Antigen." *Molecular Cell* 3 (2): 131–41.

doi:http://dx.doi.org/10.1016/S1097-2765(00)80304-4.

Albuquerque, Sónia S, Céline Carret, Ana Rita Grosso, Alice S Tarun, Xinxia Peng, Stefan H I Kappe, Miguel Prudêncio, and Maria M Mota. 2009. "Host Cell Transcriptional Profiling during Malaria Liver Stage Infection Reveals a Coordinated and Sequential Set of Biological Events." *BMC Genomics* 10 (June). BioMed Central: 270. doi:10.1186/1471-2164-10-270.

Alkan, Can, Saba Sajjadian, and Evan E Eichler. 2011. "Limitations of next-Generation Genome Sequence Assembly." *Nat Meth* 8 (1). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 61–65. http://dx.doi.org/10.1038/nmeth.1527.

Arisue, Nobuko, Tetsuo Hashimoto, Hideya Mitsui, Nirianne M Q Palacpac, Akira Kaneko, Satoru Kawai, Masami Hasegawa, Kazuyuki Tanabe, and Toshihiro Horii. 2012. "The *Plasmodium* Apicoplast Genome : Conserved Structure and Close Relationship of P . Ovale to Rodent Malaria Parasites" 29 (9): 2095–99. doi:10.1093/molbev/mss082.

Au, Kin Fai, Jason G Underwood, Lawrence Lee, and Wing Hung Wong. 2012. "Improving PacBio Long Read Accuracy by Short Read Alignment." *PLoS ONE* 7 (10). Public Library of Science: e46679. http://dx.doi.org/10.1371%252Fjournal.pone.0046679.

Aurrecoechea, Cristina, John Brestelli, Brian P Brunk, Jennifer Dommer, Steve Fischer, Bindu Gajria, Xin Gao, et al. 2009. "PlasmoDB: A Functional Genomic Database for Malaria Parasites." *Nucleic Acids Research* 37 (Database issue). Oxford University Press: D539–43. doi:10.1093/nar/gkn814.

Ay, Ferhat, Evelien M Bunnik, Nelle Varoquaux, Jean-Philippe Vert, William Stafford Noble, and Karine G Le Roch. 2015. "Multiple Dimensions of Epigenetic Gene Regulation in the Malaria Parasite *Plasmodium Falciparum*: Gene Regulation via Histone Modifications, Nucleosome Positioning and Nuclear Architecture in P. Falciparum." *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 37 (2): 182–94. doi:10.1002/bies.201400145.

Bautista, José M, Patricia Marín-García, Amalia Diez, Isabel G Azcárate, and Antonio Puyet. 2014. "Malaria Proteomics: Insights into the Parasite–host Interactions in the Pathogenic Space." *Journal of Proteomics* 97 (January): 107–25. doi:http://dx.doi.org/10.1016/j.jprot.2013.10.011.

Bozdech, Zbynek, Manuel Llinás, Brian Lee Pulliam, Edith D Wong, Jingchun Zhu, and Joseph L DeRisi. 2003. "The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium Falciparum*." *PLoS Biol* 1 (1). Public Library of Science: e5. http://dx.doi.org/10.1371%252Fjournal.pbio.0000005.

Carlton, J M, A A Escalante, D Neafsey, and S K Volkman. 2008. "Comparative

Evolutionary Genomics of Human Malaria Parasites." *Trends Parasitol* 24. doi:10.1016/j.pt.2008.09.003.

Carlton, Jane M, Samuel V Angiuoli, Bernard B Suh, Taco W Kooij, Mihaela Pertea, Joana C Silva, Maria D Ermolaeva, et al. 2002. "Genome Sequence and Comparative Analysis of the Model Rodent Malaria Parasite *Plasmodium Yoelii* ." *Nature* 419 (6906): 512–19. http://dx.doi.org/10.1038/nature01099.

Chien, Jung-Ting, Suman B Pakala, Juliana A Geraldo, Stacey A Lapp, Jay C Humphrey, John W Barnwell, Jessica C Kissinger, and Mary R Galinski. 2016. "High-Quality Genome Assembly and Annotation for *Plasmodium Coatneyi*, Generated Using Single-Molecule Real-Time PacBio Technology." *Genome Announcements* 4 (5). 1752 N St., N.W., Washington, DC: American Society for Microbiology: e00883-16. doi:10.1128/genomeA.00883-16.

Chin, Chen-Shan, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nat Meth* 10 (6). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 563–69. http://dx.doi.org/10.1038/nmeth.2474.

Cortés, Alfred, Valerie M Crowley, Alejandro Vaquero, and Till S Voss. 2012. "A View on the Role of Epigenetics in the Biology of Malaria Parasites." Edited by Glenn F Rall. *PLoS Pathogens* 8 (12). San Francisco, USA: Public Library of Science: e1002943. doi:10.1371/journal.ppat.1002943.

Craig, R., and R. C. Beavis. 2004. "TANDEM: Matching Proteins with Tandem Mass Spectra." *Bioinformatics* 20 (9): 1466–67. doi:10.1093/bioinformatics/bth092.

Dale, Joseph M, Liviu Popescu, and Peter D Karp. 2010. "Machine Learning Methods for Metabolic Pathway Prediction." *BMC Bioinformatics* 11 (1). BioMed Central Ltd: 15. doi:10.1186/1471-2105-11-15.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data." *Nat Genet* 43 (5). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 491–98. http://dx.doi.org/10.1038/ng.806.

Dilernia, Dario A, Jung-ting Chien, Michael P S Brown, Zachary Ende, Martin J Deymier, Ling Yue, Ellen E Paxinos, Susan Allen, Alfredo Tirado-ramos, and Eric Hunter. 2015. "Multiplexed Highly-Accurate DNA Sequencing of Closely-Related HIV-1 Variants Using Continuous Long Reads from Single Molecule , Real-Time Sequencing," 1–13. doi:10.1093/nar/gkv630.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. "STAR:

Ultrafast Universal RNA-Seq Aligner." *Bioinformatics (Oxford, England)* 29 (1). England: 15–21. doi:10.1093/bioinformatics/bts635.

Douglas, Nicholas M, Nicholas M Anstey, Pierre A Buffet, Jeanne R Poespoprodjo, Tsin W Yeo, Nicholas J White, and Ric N Price. 2012. "The Anaemia of *Plasmodium Vivax* Malaria." *Malaria Journal* 11 (1): 135. doi:10.1186/1475-2875-11-135.

Duffy, Michael F, Shamista A Selvarajah, Gabrielle A Josling, and Michaela Petter. 2013. "Epigenetic Regulation of the Plasmodium Falciparum Genome." *Briefings in Functional Genomics*. doi:10.1093/bfgp/elt047.

Duraisingh, Manoj T., and David Horn. 2016. "Epigenetic Regulation of Virulence Gene Expression in Parasitic Protozoa." *Cell Host & Microbe* 19 (5): 629–40. doi:http://dx.doi.org/10.1016/j.chom.2016.04.020.

English, Adam C, Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, et al. 2012. "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology." *PLoS ONE* 7 (11). Public Library of Science: e47768. http://dx.doi.org/10.1371%252Fjournal.pone.0047768.

Ferrarini, Marco, Marco Moretto, Judson A Ward, Nada Šurbanovski, Vladimir Stevanović, Lara Giongo, Roberto Viola, et al. 2013. "An Evaluation of the PacBio RS Platform for Sequencing and de Novo Assembly of a Chloroplast Genome." *BMC Genomics* 14 (October). BioMed Central: 670. doi:10.1186/1471-2164-14-670.

Flueck, Christian, and David A. Baker. 2014. "Malaria Parasite Epigenetics: When Virulence and Romance Collide." *Cell Host & Microbe* 16 (2): 148–50. doi:http://dx.doi.org/10.1016/j.chom.2014.07.012.

Gardner, Malcolm J, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, et al. 2002. "Genome Sequence of the Human Malaria Parasite *Plasmodium Falciparum*" 419.

Gonzalez-Angulo, Ana Maria, Bryan T J Hennessy, and Gordon B Mills. 2010. "Future of Personalized Medicine in Oncology: A Systems Biology Approach." *Journal of Clinical Oncology* 28 (16). American Society of Clinical Oncology: 2777–83.

Goodacre, Royston, David Broadhurst, Age K Smilde, Bruce S Kristal, J David Baker, Richard Beger, Conrad Bessant, et al. 2007. "Proposed Minimum Reporting Standards for Data Analysis in Metabolomics." *Metabolomics* 3 (3). Springer: 231–41.

Gupta, Archna P, Wai Hoe Chin, Lei Zhu, Sachel Mok, Yen-Hoon Luah, Eng-How Lim, and Zbynek Bozdech. 2013. "Dynamic Epigenetic Regulation of Gene Expression during the Life Cycle of Malaria Parasite *Plasmodium Falciparum*." *PLoS Pathog* 9 (2). Public Library of Science: e1003170. http://dx.doi.org/10.1371%252Fjournal.ppat.1003170.

Hansen, Diana S, and Louis Schofield. 2010. “Natural Regulatory T Cells in Malaria: Host or Parasite Allies?” *PLoS Pathog* 6 (4). Public Library of Science: e1000771. http://dx.doi.org/10.1371%252Fjournal.ppat.1000771.

Hert, Daniel G., Christopher P. Fredlake, and Annelise E. Barron. 2008. “Advantages and Limitations of next-Generation Sequencing Technologies: A Comparison of Electrophoresis and Non-Electrophoresis Methods.” *Electrophoresis* 29 (23): 4618–26. doi:10.1002/elps.200800456.

Ho, Joshua W K, Eric Bishop, Peter V Karchenko, Nicolas Nègre, Kevin P White, and Peter J Park. 2011. “ChIP-Chip versus ChIP-Seq: Lessons for Experimental Design and Data Analysis.” *BMC Genomics* 12 (1): 134. doi:10.1186/1471-2164-12-134.

Hodgkin, A L, and A F Huxley. 1952. “A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve.” *The Journal of Physiology* 117 (4): 500–544. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1392413/.

Hoo, Regina, Lei Zhu, Anburaj Amaladoss, Sachel Mok, Onguma Natalang, Stacey A Lapp, Guangan Hu, et al. 2016. “Integrated Analysis of the Plasmodium Species Transcriptome.” *EBioMedicine* 7 (May): 255–66. doi:http://dx.doi.org/10.1016/j.ebiom.2016.04.011.

Howard, R J, J W Barnwell, and V Kao. 1983. “Antigenic Variation of *Plasmodium Knowlesi* Malaria: Identification of the Variant Antigen on Infected Erythrocytes.” *Proceedings of the National Academy of Sciences* 80 (13): 4129–33. http://www.pnas.org/content/80/13/4129.abstract.

Huang, Honglei, Mukram M Mackeen, Matthew Cook, Eniyou Oriero, Emily Locke, Marie L Thezenas, Benedikt M Kessler, Davis Nwakanma, and Climent Casals-Pascual. 2012. “Proteomic Identification of Host and Parasite Biomarkers in Saliva from Patients with Uncomplicated *Plasmodium Falciparum* Malaria.” *Malaria Journal* 11 (1). 178. doi:10.1186/1475-2875-11-178.

Ideker, Trey, Timothy Galitski, and Leroy Hood. 2001. “A New Approach to Decoding Life: Systems Biology.” *Annual Review of Genomics and Human Genetics* 2 (1). Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA: 343–72.

Janse, C J, and A P Waters. 1995. “*Plasmodium Berghei*: The Application of Cultivation and Purification Techniques to Molecular Studies of Malaria Parasites.” *Parasitol Today* 11. doi:10.1016/0169-4758(95)80133-2.

Jiang, Lubin, María José López-barragán, Hongying Jiang, Jianbing Mu, Deepak Gaur, Keji Zhao, and Gary Felsenfeld. 2009. “Epigenetic Control of the Variable Expression of a *Plasmodium Falciparum* Receptor Protein for Erythrocyte Invasion.” doi:10.1073/pnas.0913396107.

Jonsson, Pär, Jonas Gullberg, Anders Nordström, Miyako Kusano, Mariusz Kowalczyk, Michael Sjöström, and Thomas Moritz. 2004. "A Strategy for Identifying Differences in Large Series of Metabolomic Samples Analyzed by GC/MS." *Analytical Chemistry* 76 (6). ACS Publications: 1738–45.

Joyner, Chester, John W. Barnwell, and Mary R. Galinski. 2015. "No More Monkeying around: Primate Malaria Model Systems Are Key to Understanding *Plasmodium Vivax* Liver-Stage Biology, Hypnozoites, and Relapses." *Frontiers in Microbiology* 6 (March): 145. doi:10.3389/fmicb.2015.00145.

Joyner, Chester, Alberto Moreno, Esmeralda V S Meyer, Monica Cabrera-Mora, Jessica C Kissinger, John W Barnwell, and Mary R Galinski. 2016. "*Plasmodium Cynomolgi* Infections in Rhesus Macaques Display Clinical and Parasitological Features Pertinent to Modelling Vivax Malaria Pathology and Relapse Infections." *Malaria Journal* 15 (1): 1–18. doi:10.1186/s12936-016-1480-6.

Krieger, Cynthia J, Peifen Zhang, Lukas A Mueller, Alfred Wang, Suzanne Paley, Martha Arnaud, John Pick, Seung Y Rhee, and Peter D Karp. 2004. "MetaCyc: A Multiorganism Database of Metabolic Pathways and Enzymes." *Nucleic Acids Research* 32 (Database issue). Oxford, UK: Oxford University Press: D438–42. doi:10.1093/nar/gkh100.

Landt, Stephen G, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, et al. 2012. "ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia." *Genome Research* 22 (9). Cold Spring Harbor Laboratory Press: 1813–31. doi:10.1101/gr.136184.111.

Langmead, Ben. 2010. "Aligning Short Sequencing Reads with Bowtie." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* CHAPTER (December): Unit-11.7. doi:10.1002/0471250953.bi1107s32.

Lapp, Stacey A, Cindy Korir-Morrison, Jianlin Jiang, Yaohui Bai, Vladimir Corredor, and Mary R Galinski. 2013. "Spleen-Dependent Regulation of Antigenic Variation in Malaria Parasites: *Plasmodium Knowlesi SICAvar* Expression Profiles in Splenic and Asplenic Hosts." *PLoS ONE* 8 (10). Public Library of Science: e78014. http://dx.doi.org/10.1371%252Fjournal.pone.0078014.

Lapp, Stacey A, Cindy C Korir, and Mary R Galinski. 2009. "Redefining the Expressed Prototype *SICAvar* Gene Involved in *Plasmodium Knowlesi* Antigenic Variation." *Malaria Journal* 8 (July). BioMed Central: 181. doi:10.1186/1475-2875-8-181.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, Shuzhao, Youngja Park, Sai Duraisingham, Frederick H Strobel, Nooruddin Khan, Quinlyn A Soltow, Dean P Jones, and Bali Pulendran. 2013. "Predicting Network Activity from High Throughput Metabolomics." *PLoS Comput Biol* 9 (7). Public Library of Science: e1003123. http://dx.doi.org/10.1371%252Fjournal.pcbi.1003123.

Li, Yuanyuan, and Trygve O Tollefsbol. 2011. "DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis." *Methods in Molecular Biology (Clifton, N.J.)* 791: 11–21. doi:10.1007/978-1-61779-316-5_2.

Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. 2012. "Comparison of next-Generation Sequencing Systems." *Journal of Biomedicine & Biotechnology* 2012: 251364. doi:10.1155/2012/251364.

Llinás, Manuel, Zbynek Bozdech, Edith D Wong, Alex T Adai, and Joseph L DeRisi. 2006. "Comparative Whole Genome Transcriptome Analysis of Three *Plasmodium Falciparum* Strains." *Nucleic Acids Research* 34 (4): 1166–73. http://nar.oxfordjournals.org/content/34/4/1166.abstract.

Lovegrove, Fiona E, Lourdes Peña-Castillo, Naveed Mohammad, W Conrad Liles, Timothy R Hughes, and Kevin C Kain. 2006. "Simultaneous Host and Parasite Expression Profiling Identifies Tissue-Specific Transcriptional Programs Associated with Susceptibility or Resistance to Experimental Cerebral Malaria." *BMC Genomics* 7 (1): 295. doi:10.1186/1471-2164-7-295.

MaHPIC consortium. n.d. "MaHPIC Publications." http://www.systemsbiology.emory.edu/research/Publications/index.html. http://www.systemsbiology.emory.edu/research/Publications/index.html.

Mardis, Elaine R. 2006. "Anticipating the $1,000 Genome." *Genome Biology* 7 (7). London: BioMed Central: 112. doi:10.1186/gb-2006-7-7-112.

Martin, Jeffrey A, and Zhong Wang. 2011. "Assembly." *Nature Publishing Group* 12 (10). Nature Publishing Group: 671–82. doi:10.1038/nrg3068.

Martínez-Arranz, Ibon, Rebeca Mayo, Miriam Pérez-Cormenzana, Itziar Mincholé, Lorena Salazar, Cristina Alonso, and José M. Mato. 2015. "Data in Support of Enhancing Metabolomics Research through Data Mining." *Data in Brief* 3. Elsevier B.V.: 155–64. doi:10.1016/j.dib.2015.02.008.

Nesvizhskii, Alexey I, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. 2003. "A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry Abilities That Proteins Are Present in a Sample on the Basis" 75 (17): 4646–58.

Ott, Lyman, Michael Longnecker, and R Lyman Ott. 2001. *An Introduction to Statistical Methods and Data Analysis*. Vol. 511. Duxbury Pacific Grove, CA.

Otto, Thomas D, Ulrike Böhme, Andrew P Jackson, Martin Hunt, Blandine Franke-Fayard, Wieteke A M Hoeijmakers, Agnieszka A Religa, et al. 2014. "A

Comprehensive Evaluation of Rodent Malaria Parasite Genomes and Gene Expression." *BMC Biology* 12 (1): 1–18. doi:10.1186/s12915-014-0086-0.

Pain, A., U. Bohme, A. E. Berry, K. Mungall, R. D. Finn, A. P. Jackson, T. Mourier, et al. 2008a. "The Genome of the Simian and Human Malaria Parasite *Plasmodium Knowlesi*." *Nature* 455 (7214): 799–803. doi:10.1038/nature07306.

Pain, A, U Bohme, A E Berry, K Mungall, R D Finn, A P Jackson, T Mourier, et al. 2008b. "The Genome of the Simian and Human Malaria Parasite *Plasmodium Knowlesi*." *Nature* 455 (7214). Macmillan Publishers Limited. All rights reserved: 799–803. http://dx.doi.org/10.1038/nature07306.

Park, Peter J. 2009. "ChIP-Seq: Advantages and Challenges of a Maturing Technology." *Nature Reviews. Genetics* 10 (10): 669–80. doi:10.1038/nrg2641.

Ponts, Nadia, Lijuan Fu, Elena Y Harris, Jing Zhang, Duk-Won D Chung, Michael C Cervantes, Jacques Prudhomme, et al. 2013. "Genome-Wide Mapping of DNA Methylation in the Human Malaria Parasite *Plasmodium Falciparum*." *Cell Host & Microbe* 14 (6): 696–706. doi:10.1016/j.chom.2013.11.007.

Quail, Michael, Miriam E. Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas Richard Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. "A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers." *BMC Genomics* 13 (1). BMC Genomics: 1. doi:10.1186/1471-2164-13-341.

Ralph, Stuart A, and Artur Scherf. 2005. "The Epigenetic Control of Antigenic Variation in *Plasmodium Falciparum*." *Current Opinion in Microbiology* 8 (4): 434–40. doi:http://dx.doi.org/10.1016/j.mib.2005.06.007.

Rawat, Mukul, Madhvi A Bhosale, and Krishanpal Karmodiya. 2016. "*Plasmodium Falciparum* Epigenome: A Distinct Dynamic Epigenetic Regulation of Gene Expression." *Genomics Data* 7 (March): 79–81. doi:http://dx.doi.org/10.1016/j.gdata.2015.11.026.

Reid, Adam James, and Matthew Berriman. 2013. "Genes Involved in Host–parasite Interactions Can Be Revealed by Their Correlated Expression." *Nucleic Acids Research* 41 (3). Oxford University Press: 1508–18. doi:10.1093/nar/gks1340.

Rogers, Jeffrey, and Richard A Gibbs. 2014. "Comparative Primate Genomics: Emerging Patterns of Genome Content and Dynamics." *Nature Reviews. Genetics* 15 (5): 347–59. doi:10.1038/nrg3707.

Rutledge, Gavin G, Ulrike Böhme, Mandy Sanders, Adam J Reid, Oumou Maiga-, Abdoulaye A Djimdé, Tobias O Apinjoh, et al. 2016. "Elusive *Plasmodium* Species Complete the Human Malaria Genome Set," 1–57.

Salinas, Jorge L, Jessica C Kissinger, Dean P Jones, and Mary R Galinski. 2014.

"Metabolomics in the Fight against Malaria." *Memórias Do Instituto Oswaldo Cruz* 109 (5): 589–97. doi:10.1590/0074-0276140043.

Scherf, Artur, Jose Juan Lopez-Rubio, and Loic Riviere. 2008. "Antigenic Variation in *Plasmodium Falciparum*." *Annual Review of Microbiology* 62. United States: 445–70. doi:10.1146/annurev.micro.61.080706.093134.

Shin, Seung Chul, Do Hwan Ahn, Su Jin Kim, Hyoungseok Lee, Tae-Jin Oh, Jong Eun Lee, and Hyun Park. 2013. "Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes." *PloS One* 8 (7): e68824. doi:10.1371/journal.pone.0068824.

Shulaev, V. 2006. "Metabolomics Technology and Bioinformatics." *Briefings in Bioinformatics* 7 (2): 128–39. doi:10.1093/bib/bbl012.

Simpson, Jared T, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven J M Jones, and İnanç Birol. 2009. "ABySS: A Parallel Assembler for Short Read Sequence Data." *Genome Research* 19 (6). Cold Spring Harbor Laboratory Press: 1117–23. doi:10.1101/gr.089532.108.

Singh, Balbir. 2013. "Human Infections and Detection of *Plasmodium Knowlesi*" 26 (2): 165–84. doi:10.1128/CMR.00079-12.

"Surface Alterations of Erythrocytes in *Plasmodium Falciparum* Malaria. Antigenic Variation, Antigenic Diversity, and the Role of the Spleen." 1983. *The Journal of Experimental Medicine* 157 (4). The Rockefeller University Press: 1137–48. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2186973/.

Swan, Anna Louise, Ali Mobasheri, David Allaway, and Susan Liddell. 2013. "Application of Machine Learning to Proteomics Data : Classification and Biomarker Identification in Postgenomics Biology" 17 (12). doi:10.1089/omi.2013.0017.

Tachibana, Shin-Ichiro, Steven A Sullivan, Satoru Kawai, Shota Nakamura, Hyunjae R Kim, Naohisa Goto, Nobuko Arisue, et al. 2012. "*Plasmodium Cynomolgi* Genome Sequences Provide Insight into Plasmodium Vivax and the Monkey Malaria Clade." *Nature Genetics* 44 (9): 1051–55. doi:10.1038/ng.2375.

Travers, Kevin J, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. 2010. "A Flexible and Efficient Template Format for Circular Consensus Sequencing and SNP Detection." *Nucleic Acids Research* 38 (15). Oxford University Press: e159–e159. doi:10.1093/nar/gkq543.

Turing, A M, No Aug, and B Y A M Turing. 2007. "The Chemical Basis of Morphogenesis THE CHEMICAL BASIS OF MOKPHOGENESIS" 237 (641): 37–72.

Vaidya, Akhil B, and Michael W Mather. 2009. "Mitochondrial Evolution and Functions in Malaria Parasites." *Annual Review of Microbiology* 63. Annual Reviews: 249–67.

Vaudel, Marc, Harald Barsnes, Frode S Berven, Albert Sickmann, and Lennart Martens. 2011. "SearchGUI: An Open-Source Graphical User Interface for Simultaneous OMSSA and X! Tandem Searches." *Proteomics* 11 (5). Wiley Online Library: 996–99.

Vembar, Shruthi Sridhar, Matthew Seetin, Christine Lambert, Maria Nattestad, Michael C Schatz, Primo Baybayan, Artur Scherf, et al. 2016. "Complete Telomere-to-Telomere de Novo Assembly of the *Plasmodium Falciparum* Genome through Long-Read (> 11 Kb), Single Molecule , Real-Time Sequencing" 0 (0): 1–13. doi:10.1093/dnares/dsw022.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nat Rev Genet* 10 (1). Nature Publishing Group: 57–63. http://dx.doi.org/10.1038/nrg2484.

Ward, Jane L, John M Baker, Sonia J Miller, Catherine Deborde, Mickael Maucourt, Benoit Biais, Dominique Rolin, et al. 2010. "An Inter-Laboratory Comparison Demonstrates That [1H]-NMR Metabolite Fingerprinting Is a Robust Technique for Collaborative Plant Metabolomic Data Collection." *Metabolomics* 6 (2). Springer: 263–73.

Weatherly, D Brent, James A Atwood, Todd A Minning, Cameron Cavola, Rick L Tarleton, and Ron Orlando. 2005. "A Heuristic Method for Assigning a False-Discovery Rate for Protein Identifications from Mascot Database Search Results." *Molecular & Cellular Proteomics* 4 (6). ASBMB: 762–72.

Werner, T. 2010. "Next Generation Sequencing in Functional Genomics." *Briefings in Bioinformatics* 11 (5): 499–511. doi:10.1093/bib/bbq018.

Weston, Andrea D, and Leroy Hood. 2004. "Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine." *Journal of Proteome Research* 3 (2). ACS Publications: 179–96.

White, Nicholas J, and Mallika Imwong. 2012. "Chapter Two - Relapse." In *Advances in Parasitology*, edited by Ric Price S.I. Hay and J Kevin Baird, 80:113–50. Advances in Parasitology. Academic Press. doi:http://dx.doi.org/10.1016/B978-0-12-397900-1.00002-5.

Xu, Xiaojin, Yuemei Dong, Eappen G Abraham, Anna Kocan, Prakash Srinivasan, Anil K Ghosh, Robert E Sinden, et al. 2005. "Transcriptome Analysis of Anopheles stephensi–Plasmodium Berghei Interactions." *Molecular and Biochemical Parasitology* 142 (1): 76–87. doi:http://dx.doi.org/10.1016/j.molbiopara.2005.02.013.

Young, Sidney Stanley, Thomas Henry Barrett Jr, and Christopher William Beecher. 2009. "System, Method, and Computer Program Product for Analyzing

Spectrometry Data to Identify and Quantify Individual Components in a Sample." Google Patents.

Yu, Xiaoqing, and Shuying Sun. 2013. "Comparing a Few SNP Calling Algorithms Using Low-Coverage Sequencing Data." *BMC Bioinformatics* 14 (1). BMC Bioinformatics: 274. doi:10.1186/1471-2105-14-274.

Zhu, Lei, Sachel Mok, Mallika Imwong, Anchalee Jaidee, Bruce Russell, Francois Nosten, Nicholas P Day, Nicholas J White, Peter R Preiser, and Zbynek Bozdech. 2016. "New Insights into the *Plasmodium Vivax* Transcriptome Using RNA-Seq." *Scientific Reports* 6 (February). The Author(s): 20498. http://dx.doi.org/10.1038/srep20498.

Chapter 3, Data Management Framework

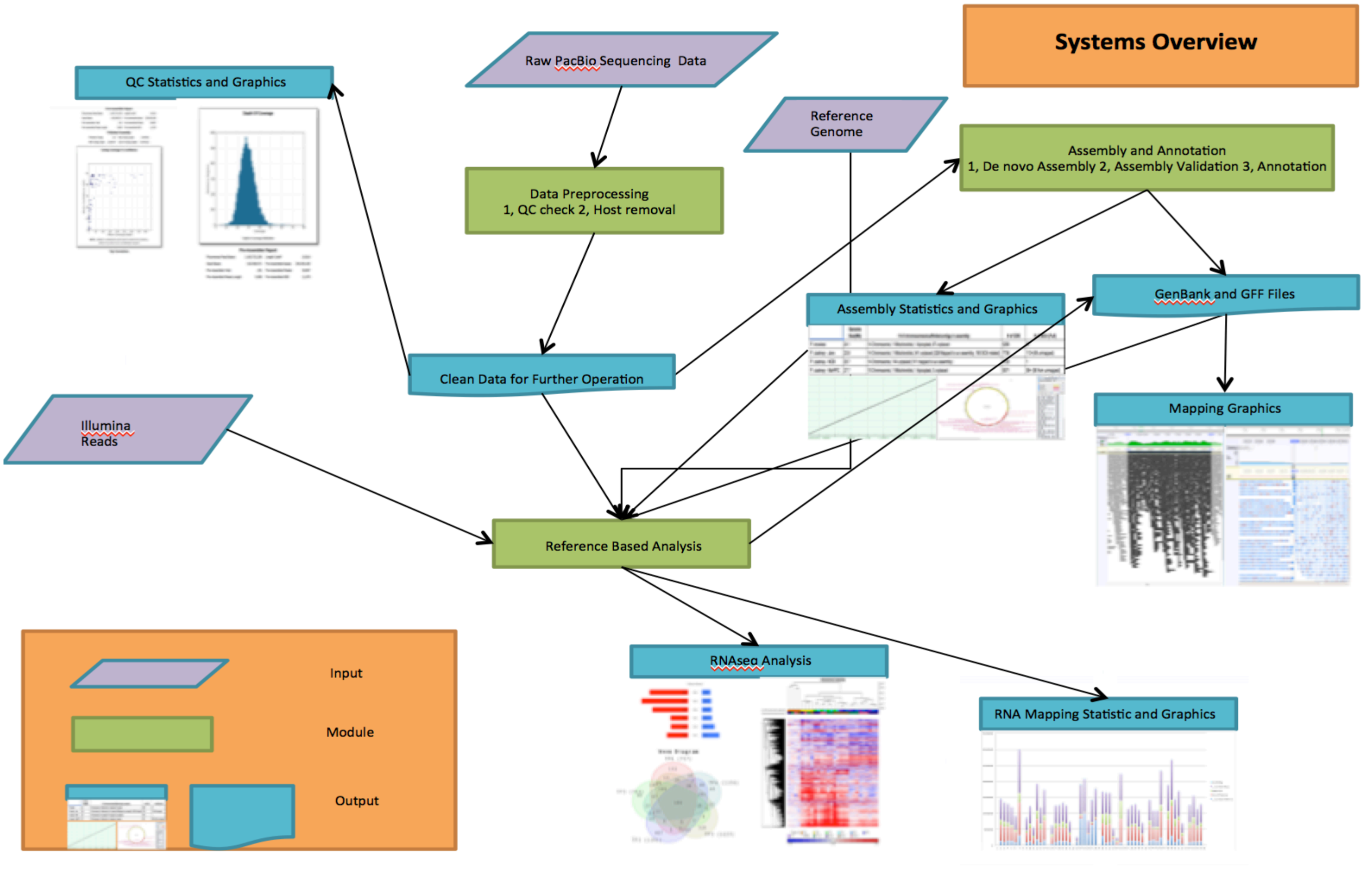
3.1 Introduction to the Infrastructure of Long-Reads Genomic Data Collection

After discussing each element of systems biology in chapter 2, an understanding the computational methods needed to obtain genomics data to build the "D" layer of the DIKW model is necessary. This chapter describes the data flow and the data management as fundamental elements of the "D" layer. In the era of big-data, biomedical investigators have increasingly focused on data-driven research (Liao, Chu, and Hsiao 2012; Bao et al. 2014). Thus, many relevant and innovative data collection tools, data storage, and data analysis approaches have been developed in the past decade. The most widely used Next Generation Sequencing (NGS) technologies, PacBio, Illumina and 454 Life Science, are shaping the modern application of sequencing techniques in genomic research (Quail et al. 2012; Ferrarini et al. 2013; Nakamura et al. 2011; Liu et al. 2012), both in scale and scope. BaseSpace, an Illumina integrative cloud-based platform, allows users to conduct bioinformatics research in several areas (https://www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps.html). Although this integrative analytical platform demonstrated outstanding performance in handling general-purpose research, a novel, flexible end-to-end solution is necessary because of the increasing complexity of different biomedical research domains and the accessibility of NGS technology.

To satisfy the need for customized long-read NGS data analysis, we developed an end-to-end workflow to address full genome assembly and other tasks.

3.2 The Information Scheme in UML

Here we demonstrate the basic workflow (Figure 3.1). In this first iteration, we integrated the Amazon cloud machine, EZ2, and cloud storage, S3, to achieve "easy I/O, easy access". We also combined several third party bioinformatics tools to our analysis


Systems Overview
Raw PacBio Sequencing Data
QC Statistics and Graphics
Reference Genome
Assembly and Annotation
1, De novo Assembly 2, Assembly Validation 3, Annotation
Data Preprocessing
1, QC check 2, Host removal
Assembly Statistics and Graphics
GenBank and GFF Files
Clean Data for Further Operation
Mapping Graphics
Illumina Reads
Reference Based Analysis
RNAseq Analysis
RNA Mapping Statistic and Graphics
Input
Module
Output

platform to let users to have multiple tools to fit their research needs.

Figure 3.1 Systematic view of the Bioinformatics pipeline

The workflow was divided into three major steps that will be discussed below:

1. Project Input and Output (I/O) and Quality Control (QC)

The first step of the framework is to building a new project. Once a project has been built, the user must input the project's purpose and parameters in main page of the user interface. Note that QC has been automatically applied inside the module of the "complexDiff". The QC module uses "PacBioEDA" (https://github.com/TomSkelly/PacBioEDA), an open-source toolset for pipeline QC.

2. Analysis Component

The main analysis modules will be described in Chapter 4, including full genome assembly and annotation, and downstream RNA-seq analysis will be addressed in Chapter 5.

3. Project Management and Storage Components

After running the pipeline, the end results are stored in the user "project" folder as end product fasta files, including the QC report and the coverage report. All results and sample data were stored on Amazon S3, encrypted and made private. The data transfer followed the Amazon S3 protocol.

3.3 Workflow Design

We applied the multiple instructions, multiple data (MIMD) model in our approach for several reasons. First, when we applied our data I/O to Amazon EC2, we expected our data source to have multiple accesses; we can withdraw different small, partial portions of read data at the same time but under different accesses. Using multiple accesses to download a data set is faster, or more time efficient. Second, when withdrawing data from Amazon EC2, our sequencing assembling pipeline will be used to process different parts of the data at the same time and generate the whole sequence with high accuracy in a short period. Presumably, the data set from each workstation will contain a large amount of data, and based on this assumption, our approach is to generate precise sequence data in a short period, thus the multiple instructions approach is reasonable and goal oriented.

For storage, we selected Amazon S3 as our platform to use after data processing, and by using Amazon S3, we can guarantee the efficiency of data access from the client's ends. Amazon S3 has several advantages. It is a stable cloud storage platform that can provide fully protected access for private datasets. It also has an immediate data recall function and a user-friendly interface. For regular users of this proposed methodology, typically researchers, hospital or healthcare facilities and pharmaceutical companies, our approach provides a reliable, standardized and certificated cloud storage platform. Amazon S3 is also relatively inexpensive because Amazon S3 can help researchers avoid hardware failure or hardware-agnostic, so it can lower the infrastructure costs.

3.4 Discussion

This pipeline shows the ability for conducting the studies of the complicated systems biology studies in malaria and the applications will be described in chapters 4 and 5. By using Amazon Web Service (AWS), our data management platform offers a specific solution and retains reproducibility and robustness by combining its use with the downstream bioinformatics pipelines. AWS allows this scheme to be scaled to more than 6,000 compute cores, multiple Petabyte of storage, and numerous core-hours of analysis. This scheme provides a graphic user interface (GUI) and a command-line tool, allowing users to upload raw sequencing data directly from any NGS instrument to our management system and to save the cost of building a computational pipeline and their own storage infrastructure.

3.5 References

Bao, Riyue, Lei Huang, Jorge Andrade, Wei Tan, Warren A Kibbe, Hongmei Jiang, and Gang Feng. 2014. "Cancer Informatics" 13: 67–82. doi:10.4137/CIN.S13779.Received.

Ferrarini, Marco, Marco Moretto, Judson A Ward, Nada Šurbanovski, Vladimir Stevanović, Lara Giongo, Roberto Viola, et al. 2013. "An Evaluation of the PacBio RS Platform for Sequencing and de Novo Assembly of a Chloroplast Genome." *BMC Genomics* 14 (October). BioMed Central: 670. doi:10.1186/1471-2164-14-670.

Giardine, Belinda, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. "Galaxy: A Platform for Interactive Large-Scale Genome Analysis." *Genome Research* 15 (10): 1451–55. doi:10.1101/gr.4086505.

Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. 2012. "Data Mining Techniques and Applications – A Decade Review from 2000 to 2011." *Expert Systems with Applications* 39 (12): 11303–11. doi:http://dx.doi.org/10.1016/j.eswa.2012.02.063.

Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. 2012. "Comparison of next-Generation Sequencing Systems." *Journal of Biomedicine & Biotechnology* 2012: 251364. doi:10.1155/2012/251364.

Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, et al. 2011. "Sequence-Specific Error Profile of Illumina Sequencers." *Nucleic Acids Research* 39 (13): e90–e90. doi:10.1093/nar/gkr344.

Quail, Michael, Miriam E. Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas Richard Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. "A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers." *BMC Genomics* 13 (1). BMC Genomics: 1. doi:10.1186/1471-2164-13-341.

Rhoads, Anthony, and Kin Fai Au. 2015. "PacBio Sequencing and Its Applications." *Genomics, Proteomics and Bioinformatics*. doi:10.1016/j.gpb.2015.08.002.

Chapter 4, Large Genome Assembly & Annotation - 3rd Generation Sequencing-Based Workflow

4.1 The Challenges of Assembling a *Plasmodium* Genome

One of the most important elements in systems biology is genomics, and all genomic knowledge comes from a well-defined genome. In this chapter, we discuss the importance of a well-defined *Plasmodium* genome as well as how to obtain a superior genome through our pipeline. Through our pipeline, we not only gather data ("D") and information ("I") but also form knowledge ("K") based on the DIKW scheme. The first *Plasmodium* genome sequence (*P. falciparum*) was published in 2002 (Gardner et al. 2002). Genome sequences for several other *Plasmodium* species have followed (Pain et al. 2008; Tachibana et al. 2012; Bright et al. 2012), but none were generated using only PacBio technology. *Plasmodium coatneyi*, which infects *Macaca mulatta* (rhesus macaques) and serves as a model of *P. falciparum* (Moreno et al. 2013), has not been assembled with PacBio technology. A preliminary draft of the *P. coatneyi* genome based on short-read (<500 bp) sequence technology is available in the NCBI database (PRJNA233970). Although the "big picture" can be gained from this genome assembly, there are over 500 sequence gaps distributed throughout the parasite's estimated 14 nuclear chromosomes. Like *P. falciparum*, the *P. coatneyi* genome has numerous repetitive sequences and complex multi-gene families, which present major difficulties that, have prohibited non-telomeric genome assembly with closure using only short-read technologies (Quail et al. 2012). Gaps prevent reliable gene content analysis, genetics and reference-based gene expression analyses, all of which are critical to understand *Plasmodium* biology and disease progression. We have implemented PacBio (RS-SMRT) sequencing technology in our computational assembly and annotation pipeline to tackle these issues.

4.2 Assembly Procedure

Given that short-read sequencing applications for the *Plasmodium* genome cannot conquer the *Plasmodium* genome's structural 'gaps', long-read sequencing technology was employed as another approach to obtain a better genome sequence. Because the read length of the long-read technology is at least 10 kb (Rhoads and Au 2015), the longer scaffolds will be more likely to cross the 'gaps' under the decent coverage, which provides better statistical confidence.

Genomic DNA (gDNA) was extracted from *ex vivo* matured schizont-stage parasites with a Qiagen DNA blood midi kit. The gDNA was further purified with a PowerClean DNA cleanup kit (Mo Bio Laboratories). Five micrograms of gDNA were subsequently used for library preparation. SMRTbell DNA libraries (Pacific Biosciences) were constructed per PacBio standard protocols with the BluePippin size-selection system (Sage Science). Sequences were generated on a PacBio RSII instrument using P6-C4 chemistry. Following cleaning, the mean assembled subread length was 5,824 bp;

the N_{50} was 7,257; the total number of bases was 1,792,197,364; and the total number of reads was 257,557. The modified "HGAP3" *de novo* assembly (Chin et al. 2013) was performed using the Amazon EC2 cloud SMRT portal; the parameters and algorithm have been provided as XML code (https://github.com/jtchien0925/PacBio_HGAP_assembly).

The error correction module was defined as a minimum sub-read length of 100 bp, a minimum read quality of 0.80, and a minimum read length of 6,000 bp. Following host (*M. mulatta*) contig removal, 15 nuclear contigs, one mitochondrial contig, and one apicoplast contig remained (51.42× average coverage). Contig identity and synteny were evaluated via BLASTn (Madden 2002) and progressive MAUVE algorithms (Darling, Mau, and Perna 2010), using the *P. knowlesi* genome from GeneDB as a reference. Two suspected inter-chromosomal rearrangements occurred within the gene family sequences located on Chr4/Chr13 and Chr12/Chr14, which could not be validated via PCR, suggesting that these sequences may in fact be correct as presented here.

4.3 Assembly Evaluation - *P. coatneyi*

Based on ontology, *P. knowlesi* and *P. coatneyi* are known to have very similar genome structures. We verified this similarity *in silico*, and we observed that the genome structure of *P. coatneyi* is very close to *P. knowlesi*, as presented in Supplemental Material 1.

First, we found the recombination event using a progressive Mauve algorithm. We observed two recombination events at two different contigs. We then reviewed the event region on contig 0 between, 1,667,610 bp to 1,780,719 bp, and we observed that this region has at least 30~40X coverage. We mapped our contig 0 with previous draft assemblies, using blastn to verify this region. A novel region was discovered that could not be mapped back to the draft assembly or the *P. knowlesi* genome. The region to left of the 3' end of this region was mapped to chromosome 14 of the *P. knowlesi* genome, and the region to the right of the 5' end of this region was mapped to chromosome 12 of the *P. knowlesi* genome. Moreover, we mapped the unplaced internal control (not shown in this dissertation) and the NCBI contigs to our contig 0 and found there are several contigs that mapped to the newly discovered region. The other recombination event was discovered in contig 8, from approximately 999,635 bp to 1,004,000bp, and the coverage of this region was also ~40X. We compared our assembly contigs with public *P. coatneyi* draft genome chromosomal contigs; in general, our contigs are longer than the draft assembly, at least 10% longer than previously existing work and free of acquisition gaps.

Based on our discovery, we might have to reconsider the order of the chromosomes. Typically, the order of the *Plasmodium* chromosomes is based on the length of the final scaffolds. In our assembly, contigs 14 and 16 are a part of chromosome 10; the remaining 13 contigs can be matched to different chromosomes. The potential 'recombination' event happens where contig 0 partially covers chromosome 14 and entirely covers chromosome 12, with a novel region discovered, and contig 8 entirely covers chromosome 4 and partially covers chromosome 13. Based on our coverage report, the ‘recombination’ region has at least 40X coverage, with continuous, non-fragmented reads, thus a possible explanation might be that the previous *Plasmodium* genome assembly projects may have misassembled their inter chromosomes due to their technology choice. For further verification, it might be necessary to apply long-read sequencing technology to any existing *P. knowlesi* assembly project.

4.4 Annotation Strategy

De novo gene prediction was performed using SNAP (Johnson et al. 2008) and Augustus (Stanke et al. 2004) for gene calls in the MAKER2 (Cantarel et al. 2008) genome annotation tool. The *P. vivax* and *P. knowlesi* predicted proteomes were included as evidence. In total, 5,516 protein-encoding genes were predicted, including up to 112 SICAvar genes. The complete annotated mitochondrial and apicoplast genomes are also included in this chapter. The annotation was validated with *P. coatneyi* RNA-Seq data, Uniprot (Apweiler et al. 2004), KEGG (Nakaya et al. 2013) OrthoMCL Orthology (Li, Stoeckert, and Roos 2003), and InterProScan5 (Jones et al. 2014). Five thousand sixty genes have strong evidence of synteny. The fourteen chromosome sequences were deposited in the NCBI database (BioProject PRJNA315987) under accession numbers CP016239 to CP016252 and provided to PlasmoDB. The workflow of the *Plasmodium* genome annotation is shown in Figure 4.1 and the source code of maker could be found at the URL below.

(https://github.com/jtchien0925/useful_bioinformatics_scripts)

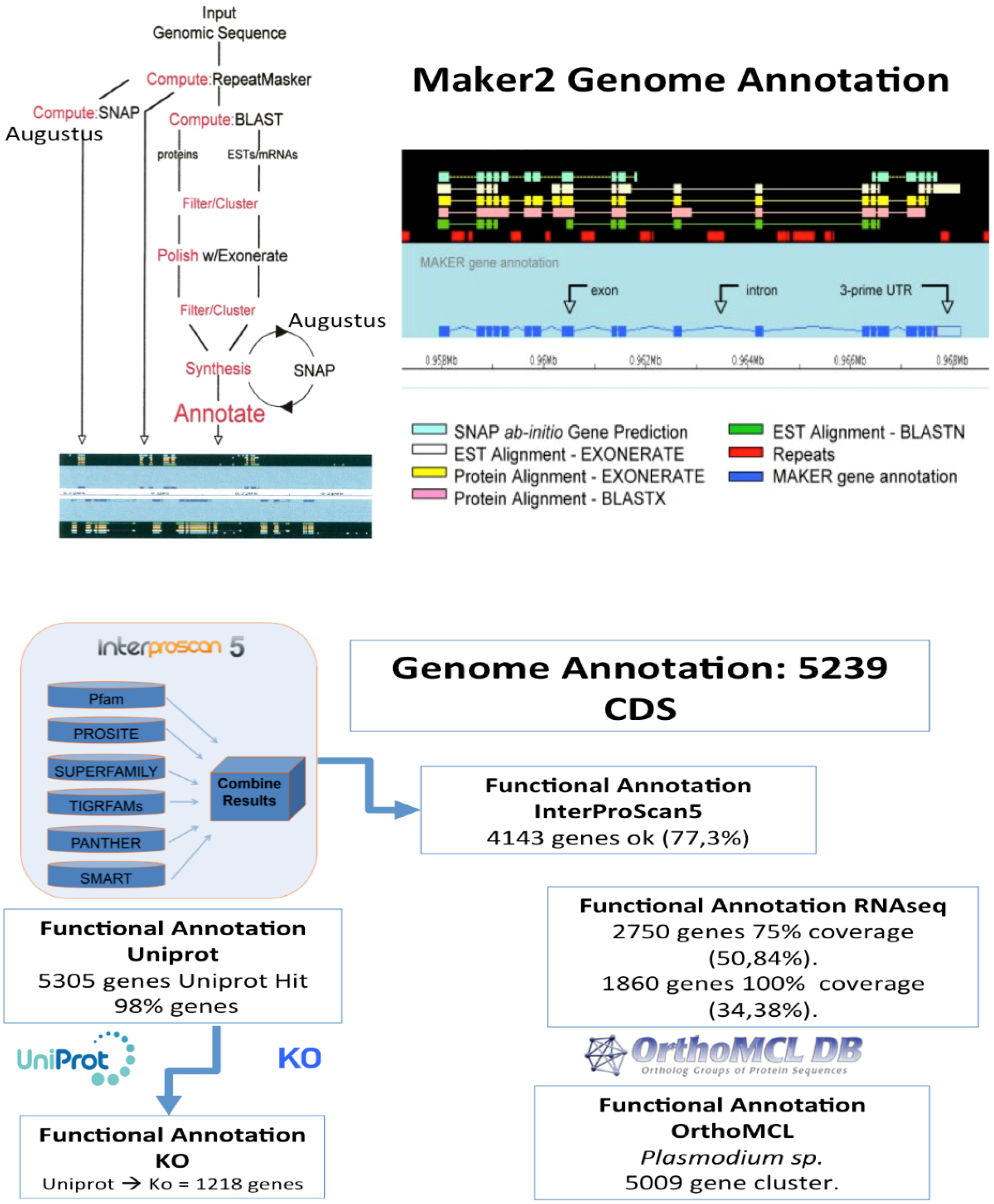


Figure 4.1 The Genome Annotation Pipeline.

The annotation pipeline can be broken into several steps. 1) RNA-seq, homologous database construction for reference purposes, 2) modification of parameters 3) training SNAP and Augustus for gene prediction and 4) functional annotation. Note that all of parameters in these steps will be provided in github link above.

4.4.1 Genome Annotation Workflow

There are several genome annotation applications available for researchers to use to achieve the genome annotation. We have selected Maker2 as our major annotation tool. We selected Maker2 because Maker2 is a flexible genome annotation pipeline. Maker2 was used to annotate smaller eukaryotic and prokaryotic genomes independently and to create genome databases. Maker2 allows the repeats identification, ESTs alignments, ab-initio gene predictions, proteins alignment and consensus all the data together to generate annotations. Maker2 has the training feature that allows outputs from previous execution to be used to retrain gene prediction algorithm and then producing gene models for subsequent runs. Maker2 generates different outputs and can be viewed and edited by using visualization tools e.g. Geneious.

The annotation workflow can be described into several steps: 1) Generating input database by upload the transcript data, expressed sequence tag (ESTs), and homologous proteins. 2) Generating and editing maker2 configuration file (provided in github). 3) Training SNAP model for *ab initio* gene prediction 4) Rerun Maker2 with SNAP output models to generate next run of annotation model. 5) Training SNAP again along with

Augustus by using the output from last run of Maker2. In this step, we introduced the Augustus *ab initio* the gene prediction algorithm to this task along with SNAP to avoid the bias of single algorithm usage. 6) Rerun Maker2 again with previous run of the SNAP and Augustus prediction models as input. 7) Obtaining the final annotation file from Maker2.

4.5 Repetitive Gene Families Analysis

The sequencing of the *SICAvar* genes have been deeply reviewed; it has a long intron and a high AT repetitive feature. Because there is currently no public gene annotation indicating the number *SICAvar* genes in *P. coatneyi*, our genome assembly project is the first to report the number of *SICAvar* genes. Furthermore, by comparing the number of *SICAvar* genes in our assembly to *P. knowlesi*, our assembly showed the number of discovered, full *P. coatneyi SICAvar* CDS (112) is slightly higher in *P. knowlesi* (97). This discovery again shows that these two species are very close from an evolutionary perspective when we look at the similarity of the number of *SICAvar* genes. Because the previous public assembly *P. coatneyi* project is not annotated, our assembly represents the first released annotation.

To verify our assembly, we mapped the rest of the NCBI WGS *P. coatneyi* bioproject 144 unplaced contigs to our assembly. One hundred thirty-eight of 144 unplaced contigs were mapped to our 15 chromosomal contigs, 1 apicoplast contig and 1 mitochondrial contig. Most of these unplaced contigs were mapped to the sub-telomerase region, the *SICAvar* and CoIR rich region. There were two unplaced contigs mapped to the newly discovered region in contig 0 (1,667,610 bp to 1,780,719 bp). This discovery indicated that the newly discovered region exists in the *P. coatneyi* naturally, and this region was not discovered in previous assemblies because of the limitations of short-read sequencing technology. In contrast, applying the long-read based sequencing technology can solve the problem of insufficient read length of short read technologies. For example, even though the *SICAvar* has >14 kb segment of genomic DNA, it is not problematic for long-read sequencing technology because the average read length is above 12000 kb.

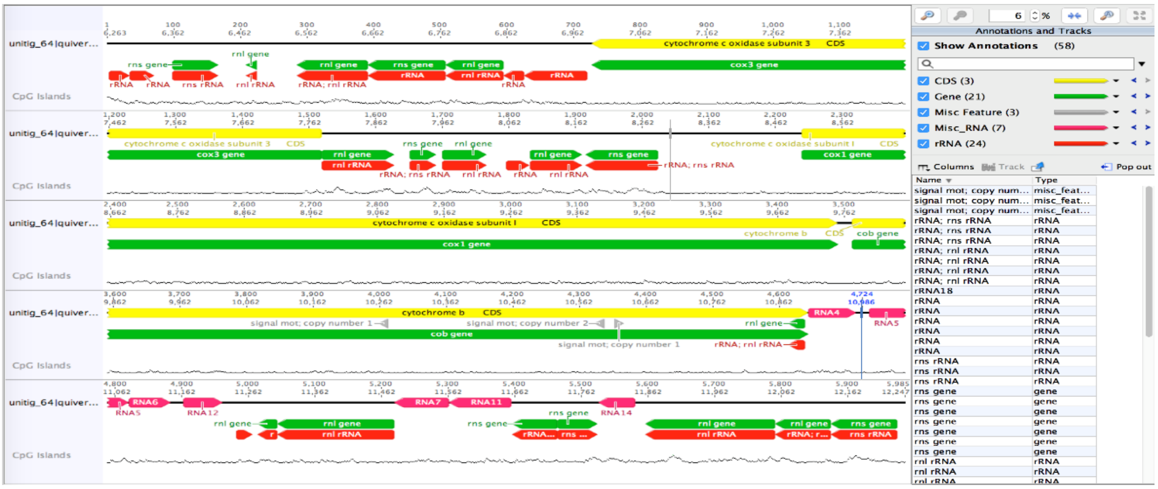
Although we did not obtain more than 100X coverage, 50X coverage allowed the assembly to be performed with confidence. The other difficulty of applying short-read technology to this region is that this region has plenty of repetitive families. After applying Repeatmasker to mark the repetitive families, we discovered that there are 101 hits of repetitive regions in contigs 0 and 33 hits in contig 8. This feature has been reported as the main difficulty for short-read sequencing and PCR because of the confusion of the sequence fragment mapping; however, the repetitive regions are not a difficulty for long-read sequencing, like in our *P. coatneyi* assembly project.

There is another newly discovered region (999,635 bp-1,004,000 bp) in contig 8. We also mapped the unmapped contigs from NCBI WGS *P. coatneyi* bioproject, but we did not acquire any hits in the newly discovered area. Based on the report from our

Maker2 annotation and RATT annotation (Otto et al. 2011), this newly discovered region partially covers the full *SICAvar* gene. This discovery indicates the difficulty of sequencing the *SICAvar* genes using short-read technology. After verified the two ‘novel’ regions, we also reviewed the rest of unmapped contigs from the NCBI WGS *P. coatneyi* bioproject by annotating the unmapped contigs in silico. The report shows that there are 26 *CoIR* related regions, 20 *SICAvar* related regions and 96 highly repetitive fragments. This finding also indicates that long-read technology can sequence ‘across’ the problematic regions that short-read technology has.

4.6 Mitochondria and Apicoplast Genome

A complete mitochondrial genome and an apicoplast genome (Figure 4.2 & 4.3) have been found. We applied the *P. knowlesi* genome from GeneDB as our annotation reference for these two genome sequences. The mitochondrial genome is around 6k bp and has 3 CDS regions, cytochrome c oxidase III (coxIII), cytochrome b (Cytb) and cytochrome c oxidase I (cox1). For Apicoplast (Figure 4.3), the genome size is about 35 kb, and it contains genes for the large subunit (LSU) and the small subunit (SSU) rRNAs, 25 tRNAs, 3 subunits of RNA polymerase, 17 ribosomal proteins, caseinolytic protease C (clpC), elongation factor Tu (tufA), sulfur mobilizing protein B (sufB), and 7 unknown open reading frames.



6 %

Annotations and Tracks

☑ Show Annotations (58)

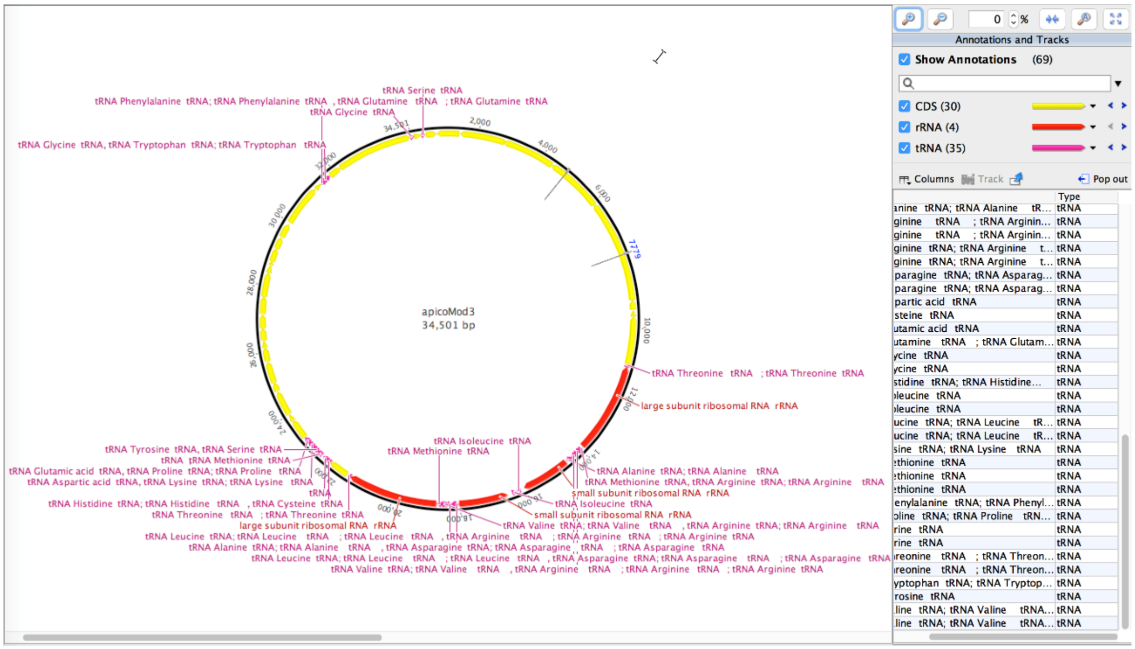
☑ CDS (3) 🟨
☑ Gene (21) 🟩
☑ Misc Feature (3)
☑ Misc_RNA (7)
☑ rRNA (24) 🟥

Columns Track Pop out

Name	Type
signal mot; copy num...	misc_feat...
signal mot; copy num...	misc_feat...
signal mot; copy num...	misc_feat...
rRNA; rns rRNA	rRNA
rRNA; rns rRNA	rRNA
rRNA; rns rRNA	rRNA
rRNA; rnl rRNA	rRNA
rRNA; rnl rRNA	rRNA
rRNA; rnl rRNA	rRNA
rRNA; rnl rRNA	rRNA
rRNA18	rRNA
rRNA	rRNA
rRNA	rRNA
rRNA	rRNA
rRNA	rRNA
rRNA	rRNA
rRNA	rRNA
rns rRNA	rRNA
rns rRNA	rRNA
rns rRNA	rRNA
rns gene	gene
rns gene	gene
rns gene	gene
rns gene	gene
rns gene	gene
rns gene	gene
rns gene	gene
rnl rRNA	rRNA
rnl rRNA	rRNA
rnl rRNA	rRNA

Figure 4.2 Mitochondrial Genome Visualization

The 3 CDS, COX1, COXIII and Cytb are present in the yellow bars, and the gene predictions are present in the green bars. There are 7 misc_RNA in pink and 22 rRNA present in red.


apicoMod3
34,501 bp
2,000
4,000
6,000
7,779
10,000
12,000
14,000
16,000
18,000
20,000
22,000
24,000
26,000
28,000
30,000
32,000
34,501
tRNA Serine tRNA
tRNA Phenylalanine tRNA; tRNA Phenylalanine tRNA
tRNA Glutamine tRNA ; tRNA Glutamine tRNA
tRNA Glycine tRNA
tRNA Glycine tRNA, tRNA Tryptophan tRNA; tRNA Tryptophan tRNA
tRNA Threonine tRNA ; tRNA Threonine tRNA
large subunit ribosomal RNA rRNA
tRNA Isoleucine tRNA
tRNA Methionine tRNA
tRNA Tyrosine tRNA, tRNA Serine tRNA
tRNA tRNA Methionine tRNA
tRNA Glutamic acid tRNA, tRNA Proline tRNA; tRNA Proline tRNA
tRNA Aspartic acid tRNA, tRNA Lysine tRNA; tRNA Lysine tRNA
tRNA
tRNA Histidine tRNA; tRNA Histidine tRNA , tRNA Cysteine tRNA
tRNA Threonine tRNA ; tRNA Threonine tRNA
large subunit ribosomal RNA rRNA
tRNA Leucine tRNA; tRNA Leucine tRNA ; tRNA Leucine tRNA , tRNA Arginine tRNA ; tRNA Arginine tRNA ; tRNA Arginine tRNA
tRNA Alanine tRNA; tRNA Alanine tRNA , tRNA Asparagine tRNA; tRNA Asparagine tRNA ; tRNA Asparagine tRNA
tRNA Leucine tRNA; tRNA Leucine tRNA ; tRNA Leucine tRNA , tRNA Asparagine tRNA; tRNA Asparagine tRNA ; tRNA Asparagine tRNA
tRNA Valine tRNA; tRNA Valine tRNA , tRNA Arginine tRNA ; tRNA Arginine tRNA ; tRNA Arginine tRNA
tRNA Alanine tRNA; tRNA Alanine tRNA
tRNA Methionine tRNA, tRNA Arginine tRNA; tRNA Arginine tRNA
small subunit ribosomal RNA rRNA
tRNA Isoleucine tRNA
small subunit ribosomal RNA rRNA
tRNA Valine tRNA; tRNA Valine tRNA , tRNA Arginine tRNA; tRNA Arginine tRNA

0 %

Annotations and Tracks

☑ Show Annotations (69)

☑ CDS (30) 🟨
☑ rRNA (4) 🟥
☑ tRNA (35)

Columns Track Pop out

	Type
anine tRNA; tRNA Alanine tR...	tRNA
ginine tRNA ; tRNA Argin...	tRNA
ginine tRNA ; tRNA Argin...	tRNA
ginine tRNA; tRNA Arginine t...	tRNA
ginine tRNA; tRNA Arginine t...	tRNA
paragine tRNA; tRNA Asparag...	tRNA
paragine tRNA; tRNA Asparag...	tRNA
partic acid tRNA	tRNA
steine tRNA	tRNA
utamic acid tRNA	tRNA
utamine tRNA ; tRNA Glutam...	tRNA
ycine tRNA	tRNA
ycine tRNA	tRNA
stidine tRNA; tRNA Histidine...	tRNA
oleucine tRNA	tRNA
oleucine tRNA	tRNA
ucine tRNA; tRNA Leucine tR...	tRNA
ucine tRNA; tRNA Leucine tR...	tRNA
sine tRNA; tRNA Lysine tRNA	tRNA
ethionine tRNA	tRNA
ethionine tRNA	tRNA
ethionine tRNA	tRNA
enylalanine tRNA; tRNA Phenyl...	tRNA
oline tRNA; tRNA Proline tRN...	tRNA
rine tRNA	tRNA
rine tRNA	tRNA
reonine tRNA ; tRNA Threon...	tRNA
reonine tRNA ; tRNA Threon...	tRNA
yptophan tRNA; tRNA Tryptop...	tRNA
rosine tRNA	tRNA
line tRNA; tRNA Valine tRNA...	tRNA
line tRNA; tRNA Valine tRNA...	tRNA

Figure 4.3 Apicoplst Genome Visualization

There are 30 CDS present in yellow, 4 rRNA present in red and 35 tRNA present in pink.

4.7 Case Studies of Robustness - Applying the Assembly & Annotation Pipeline to Other Species of *Plasmodium*

The goal of constructing an assembly and annotation pipeline is to satisfy a need in the malaria research community; reference genomes are the most crucial aspect of downstream systems biology research. We applied our pipeline to several other *Plasmodium* species, including *Plasmodium cynomolgi, Plasmodium vivax,* and *Plasmodium knowlesi.* In this study, we demonstrated the robustness of our pipeline on several *Plasmodiums* and out-performed short-read technology. In the *Plasmodium cynomolgi* genome assembly, we demonstrated that long-read based technology, along with our pipeline, could capture more information in subtelomeric regions compared to short-read technology (Figure 4.6). Moreover, the end results of the *Plasmodium knowlesi* genome were validated via HI-C technology (Ay, Bailey, and Noble 2014; Yaffe and Tanay 2011; Korbel and Lee 2013) and showed that our pipeline can generate robust results compared to other genome assembly tools (Table 5.1).

4.8 Combining Hi-C Assembly with the PacBio Assembly Pipeline

Contings scaffolding and gaps closing in assembly projects are still difficult tasks that usually require time-consuming validation by applying expensive methods. Despite novel algorithms and new technologies are constantly introduced to tackle these issues, only a few widely used *Plasmodium* published assemblies have a higher consensus of agreements on reflecting the structure of the genome. Nevertheless, repetitive regions are still a problem, for example, antigen variation gene families show a wide range of polymorphisms in chromosome structure level between different strains of the same species of *Plasmodium*. Moreover, current *P. knowlesi* assemblies do not provide the acceptable resolution for the subtelomeric regions and repetitive regions; instead, the published assemblies gather the problematic contigs into one concatenate scaffold and have been released along with the assembly projects in public domains.

Plasmodium genomes in public archives that are sequenced and assembled are only partially complete. For instance, the most current version of *Plasmodium knowlesi* -H genome, is available with a set of 14 scaffolds 14 contigs (Lapp *et al.* 2017). To conduct studies in comparative genomics and disease evolution, an incomplete genome will harm the quality of locus analyses, making genome-wide annotation and downstream expression analysis very challenging. Furthermore, the highly repetitive gene families and uncertain subtelomeric regions are related to genomic rearrangement events, making it difficult to conduct completed sequence and assemble. Chien et al in 2016 reported an approach for solving the *Plasmodium* genome complexity by applying long-reads sequencing method. However, to address these difficulties with higher accuracy and lower bias in assembly validation, applying long-reads and Hi-C sequencing approaches (Lapp, Geraldo, Chien *et al.* 2017) as a combination approach can use to guide genome assembly finalization especially in gap closing (Fig 4.5). The Hi-C methodology was originally developed to study the three-dimensional folding of the genome as well as the physical interactions that link regulatory elements with distant sequences. However, this methodology can also provide a sophisticated solution to improve chromosome- and genome-scale assembles constructed from short reads by taking advantage of the *in* *vivo* chromatin interaction frequency data to accurately position individual contigs without a requirement for sequence overlap. Compared to applying a long-reads sequencing technology only method in *Plasmodium* genome assembly, this study demonstrated the ability to apply Hi-C to guide long-reads assembly in contig orientation which can lead to a high-quality assembly and annotation in an economical way.

In this study, Hi-C can correctly scaffold the unplaced pre-assembled contigs into a reference genome correctly by identifying the interaction of each contig (Table 4.1). Our study first reported the combined application of both technologies in the *Plasmodium* genome assembly. The complete *Plasmodium knowlesi* genome structure obtained in our study enables the analysis of antigenic variation observation and could potentially lead to better understanding of the processes of *Plasmodium* parasites involved in evading the host's acquired immune system. Moreover, we expect our reference genome to increase the accuracy of future studies of *Plasmodium knowlesi* genome structure as well as increase the accuracy of quantitative expression analysis (Table 4.1).

Upon comparing our *Plasmodium* genome assembly with the most recent chromosomal reference assembly, our genome differs from the previous assembly in several important aspects (Table 4.2)(Fig 4.4). First, our approach only requires the Hi-C data and set of HGAP assembled contigs as input with minor parameter tuning. Second, unlike previous *P. knowlesi* assemblies, our approach correctly identifies and scaffolds especially well in complicated repetitive regions, which is an obvious difficulty for short-reads based genome assembly methods. Third, our approach allows researchers to detect rearrangements events in chromosomes and to correct misassemblies, with visual and computational evidence. These features have the potential to be very informative for downstream analyses, for instances, comparing different *P. knowlesi* strains genome structure to understand the antigenic variation switching events quantitatively.

Despite our approach potentially improving the *de novo* assembly of complex *Plasmodium* genomes, there are still opportunities for further enhancements. First, though the workflow can be executed smoothly, the entire workflow can be fully automated when the computing power is available. Second, the accuracy of HGAP pre-assembled contigs depends on not only the reads assembly coverage but also the (intra-chromosomal interactions). We have discovered that the HGAP algorithm misassembled contig0, the largest contig collected after running HGAP, which can be broken down into 3 smaller pieces manually based on suggestions from Hi-C interaction heat maps. These three pieces could then be mapped to three different chromosomes and to the public reference assembly (http://www.genedb.org/Homepage/Pknowlesi), so the junctions of these three contigs were associated with the complicated repetitive regions. This discovery suggests that despite the superior length of reads that the PacBio system can provide for sequencing and assembling, currently available assembly algorithms are not "one fit all". Moreover, applying any sequencing technology solely in the *Plasmodium* genome assembly project might introduce the technology selection bias to the results.

In summary, we believe our approach is robust and accurate for assembling genomes, scaffolding the complicated repetitive regions and distinguishing the differences between real chromosomal rearrangement and artifacts.

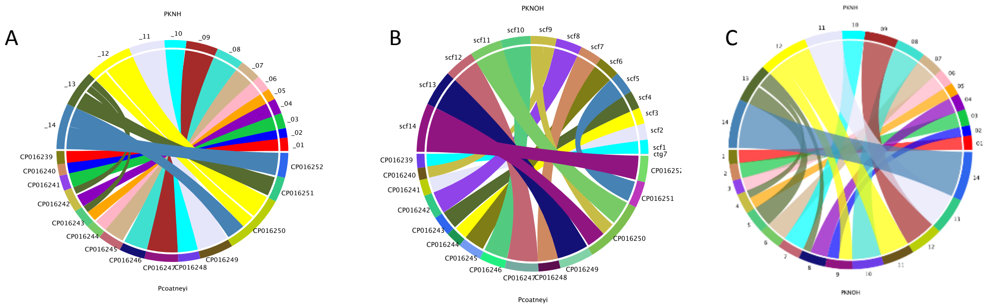


Fig 4.4 MaHPIC *PKNOH* genome comparison with the *PKNH*-V2 and *PCOAH*

A) SyMAP circular DNA comparison of the MaHPIC Pk genome sequence scaffolds to the PKNH 2015 consensus sequence. B) SyMAP circular DNA comparison of the MaHPIC Pk genome sequence scaffolds to the *P. coatneyi* HACKERI genome sequence that was assembled using PacBio technologies (Chien et al., 2016). C) SyMAP circular DNA comparison of the PKNH 2015 consensus sequence and *P. coatneyi* genome sequence.

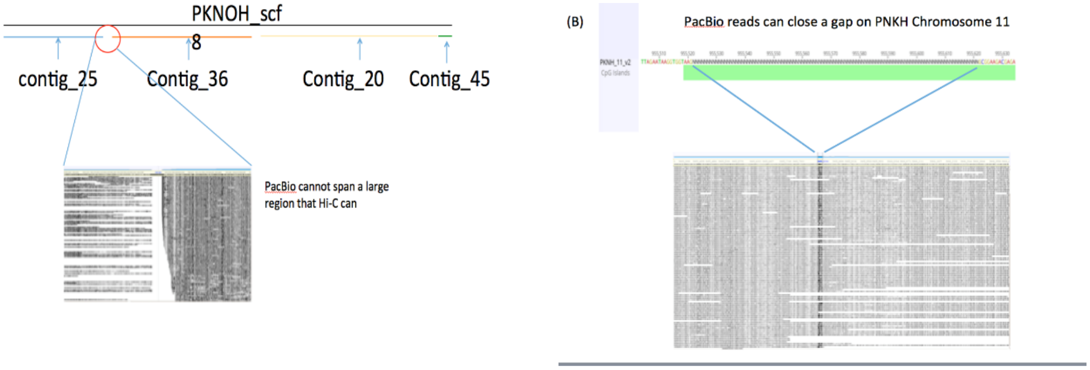

PKNOH_scf
8
contig_25
Contig_36
Contig_20
Contig_45
PacBio cannot span a large region that Hi-C can
(B)
PacBio reads can close a gap on PNKH Chromosome 11
PKNH_11_v2
CpG Islands

Fig 4.5 A Visualization of PacBio sequence closed the gap

A) PacBio sequence provides support for the novel fusions relative to PKNH, it is not able to span some gaps that Hi-C was able to span. B) PacBio was however, able to close a gap that is present in the PKNH version 2 genome sequence.

4.9 Discussion

In our *P. coatneyi* assembly, we observed the drastic difference between *P. coatneyi* and *P. knowlesi* with respect to the number of the CoIR/KIR regions. There are only 70 complete knowlesi interspersed repeats (KIR) annotated in *P. knowlesi;* however, we found 1002 complete coatneyi interspersed repeat (CoIR) regions and 283 fragmented CoIR regions. KIR was discovered and defined in 2008. This paper indicated that KIR is the second largest repetitive gene family in *P. knowlesi*; however, in our *P. coatneyi* assembly the CoIR-like region is actually the largest repetitive gene family rather than SICAvar. The discovery of the *Plasmodium* interspersed repeats (pir) super family was reported in several papers. The function of this super family has recently been discovered in *P.knowlesi* and *P.cynomolgi* in NHP model research. These two findings indicate that the 'IR' gene plays 'molecular mimicry' roles by mimicking CD99 in immunoregulatory regulation. We have not defined the function of our CoIR; however, there are clear differences in the number of groups in *P.knowlesi* and *P. coatneyi*, so conducting further investigation into CoIR will be the next step toward understanding the function of CoIR and the different evolutionary features.

In summation, the complete *Plasmodium* genome assembly demonstrates the ability of long-read DNA sequencing technology and set the standard for *de novo Plasmodium* genome assembly. We demonstrate that >40× PacBio coverage is sufficient to deliver a high-quality genome assembly (Appendix 3). However, to achieve an error-free assemble, the *Plasmodium* genome was obtained by applying the HGAP3 assembly along with other advanced technologies, such as Hi-C. Based on the combination of long-read *de novo* assembly and Hi-C, we were able to assemble a high-quality *Plasmodium* genome in a cost-effective manner (Appendix 4)). Technological advances in sequencing, as well as in sequencing chemistry, will eventually solve complex genome assembly tasks. High-quality genome assemblies will provide genomic information that allows the research community to conduct reliable analyses of *Plasmodium* genetics and gene expression, which is critical to understanding the disease and disease processes.

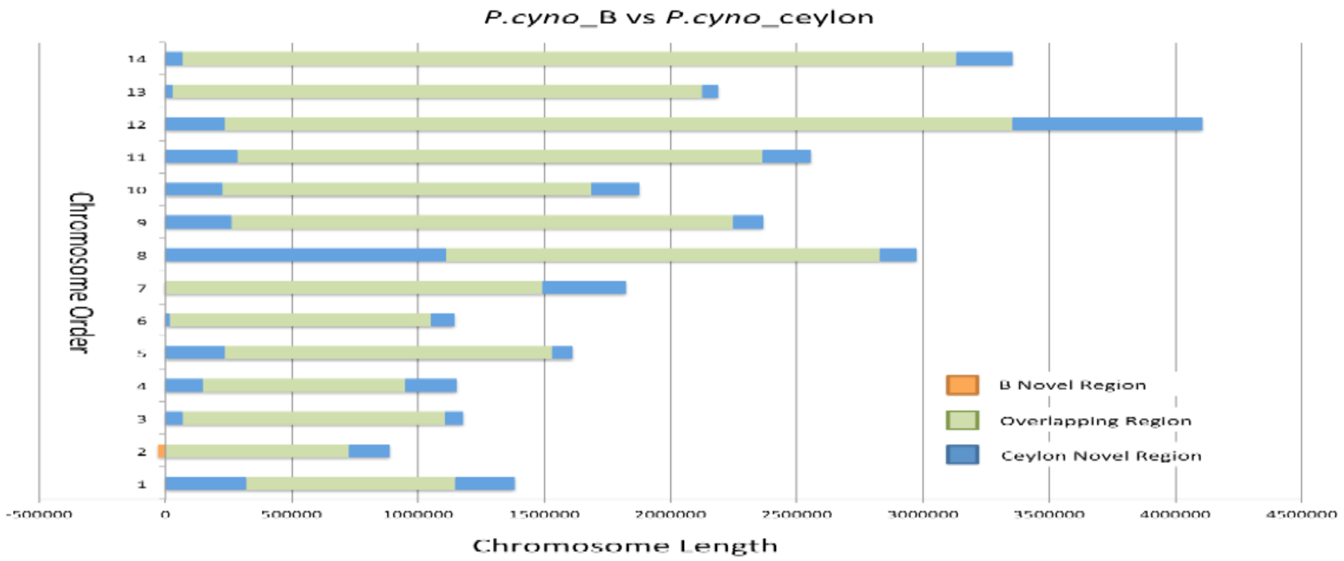


Figure 4.6 A Comparison of Different Sequencing Technology in *Plasmodium cynomolgi* Assembly

The green regions represent the genome assembled with Illumina technology; the blue regions represent the extra information acquired using PacBio technology.

Table 4.1 Mapping relationship between PacBio contigs and the HI-C assembly of the *Plasmodium knowlesi* genome

PacBio Unitig	Length	Scaffold Assignment
unitig_18_chr1	914,168	scf1
unitig56	14,772	scf2
unitig_23_chr3	719,235	
unitig_30_chr3	345,602	
unitig_0_chr6	489,988	scf3
unitig_3_chr6	606,307	
unitig_12_chr5_13	1,360,573	scf4
unitig_10_chr13	1,435,464	scf5
unitig_15_chr7	1,207,278	scf6
unitig_31_chr7	323,894	
unitig_96_chr11	13,068	scf7
unitig_7_chr10	1,543,775	
unitig_25_chr13	593,400	scf8
unitig45	2,418	
unitig_36_chr4	292,580	
unitig_20_chr4	832,288	
unitig_22_chr2	669,228	scf9
unitig_39_chr2	146,901	

unitig_16_chr12	1,016,298	
unitig_13_chr8	226,016	scf10
unitig_0_chr8	1,066,829	
unitig_35_chr8	268,932	
unitig_34_chr8	9,837	
unitig_32_chr8	410,932	
unitig_2_chr12	1,267,252	scf11
unitig_0_chr12	889,998	
unitig52	21,891	
unitig_1_chr9	2,187,873	scf12
unitig_17_chr11	980,536	scf13
unitig_9_chr11	422,355	
unitig_37_chr11	207,973	
unitig_105_chr11	5,212	
unitig_21_chr11	612,914	
unitig_38_chr11	170,838	
unitig_26_chr14	586,198	scf14
unitig_43_chr14	17,326	
unitig_4_chr14	1,880,748	
unitig_33_chr14	298,116	
unitig_28_chr14	532,183	

Table 4.2 Characteristics of genome sequences utilized in this study.

Name	Genome Size (nt)	Scaffold Number	Contig Number	Gaps	N50 Contig Length	N50 Scaffold Length	Technology
PKNOH-PacBio	24,588,173	N/A	50	N/A	1,207,278	N/A	PacBio
PKNOH-PacBio-Hi-C	24,593,696	14	14	25	16,231	1,832,627	PacBio+Hi-C
PKNH	24,359,384	14	148	77	N/A	2,162,603	Illumina
PKNA1-C.2	24,359,887	N/A	45	N/A	1,061,780	N/A	PacBio
PKNA1-H.1	23,958,038	N/A	37	N/A	1,017,166	N/A	PacBio

N/A = Not Applicable.

PKNOH data presented here have had organellar sequences removed.

4.10 References

Apweiler, Rolf, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, et al. 2004. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 32 (suppl 1): D115–19. doi:10.1093/nar/gkh131.

Ay, Ferhat, Timothy L. Bailey, and William Stafford Noble. 2014. "Statistical Confidence Estimation for Hi-C Data Reveals Regulatory Chromatin Contacts." *Genome Research* 24 (6): 999–1011. doi:10.1101/gr.160374.113.

Bright, A Taylor, Ryan Tewhey, Shira Abeles, Raul Chuquiyauri, Alejandro Llanos-cuentas, Marcelo U Ferreira, Nicholas J Schork, Joseph M Vinetz, and Elizabeth A Winzeler. 2012. "Whole Genome Sequencing Analysis of *Plasmodium Vivax* Using Whole Genome Capture." ??? 13 (1). ??? 1. doi:10.1186/1471-2164-13-262.

Cantarel, Brandi L, Ian Korf, Sofia M C Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. "MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes." *Genome Research* 18 (1). Cold Spring Harbor Laboratory Press: 188–96. doi:10.1101/gr.6743907.

Carlton, Jane M, Samuel V Angiuoli, Bernard B Suh, Taco W Kooij, Mihaela Pertea, Joana C Silva, Maria D Ermolaeva, et al. 2002. "Genome Sequence and Comparative Analysis of the Model Rodent Malaria *Parasite Plasmodium Yoelii*." *Nature* 419 (6906): 512–19. http://dx.doi.org/10.1038/nature01099.

Chin, Chen-Shan, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nat Meth* 10 (6). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 563–69. http://dx.doi.org/10.1038/nmeth.2474.

Darling, Aaron E, Bob Mau, and Nicole T Perna. 2010. "progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement." *PLoS ONE* 5 (6). Public Library of Science: e11147. http://dx.doi.org/10.1371%252Fjournal.pone.0011147.

del Portillo, Hernando A, Carmen Fernandez-Becerra, Sharen Bowman, Karen Oliver, Martin Preuss, Cecilia P Sanchez, Nick K Schneider, et al. 2001. "A Superfamily of Variant Genes Encoded in the Subtelomeric Region of *Plasmodium Vivax*." *Nature* 410 (6830). Macmillian Magazines Ltd.: 839–42. http://dx.doi.org/10.1038/35071118.

Gardner, M, N Hall, E Fung, O White, M Berriman, R Hyman, J Carlton, et al. 2002.

"Genome Sequence of the Human Malaria Parasite *Plasmodium Falciparum*." *Nature* 419: 498–511. doi:10.1038/nature01097.

Hall, Neil, Marianna Karras, J Dale Raine, Jane M Carlton, Taco W A Kooij, Matthew Berriman, Laurence Florens, et al. 2005. "A Comprehensive Survey of the Plasmodium Life Cycle by Genomic, Transcriptomic, and Proteomic Analyses." *Science* 307 (5706): 82–86. doi:10.1126/science.1103717.

Janssen, Christoph S, R Stephen Phillips, C Michael R Turner, and Michael P Barrett. 2004. "*Plasmodium* Interspersed Repeats: The Major Multigene Superfamily of Malaria Parasites." *Nucleic Acids Research* 32 (19). Oxford, UK: Oxford University Press: 5712–20. doi:10.1093/nar/gkh907.

Johnson, Andrew D., Robert E. Handsaker, Sara L. Pulit, Marcia M. Nizzari, Christopher J. O'Donnell, and Paul I W De Bakker. 2008. "SNAP: A Web-Based Tool for Identification and Annotation of Proxy SNPs Using HapMap." *Bioinformatics* 24 (24): 2938–39. doi:10.1093/bioinformatics/btn564.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9). Oxford University Press: 1236–40. doi:10.1093/bioinformatics/btu031.

Korbel, Jan O, and Charles Lee. 2013. "Genome Assembly and Haplotyping with Hi-C." *Nat Biotech* 31 (12): 1099–1101. doi:10.1038/nbt.2764.

Li, L, C J Stoeckert, and D S Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Res* 13. doi:10.1101/gr.1224503.

Madden, Tom. 2002. "Chapter 16 : The BLAST Sequence Analysis Tool." *The NCBI Handbook[internet]*, 1–15.

Moreno, Alberto, Monica Cabrera-Mora, AnaPatricia Garcia, Jack Orkin, Elizabeth Strobert, John W Barnwell, and Mary R Galinski. 2013. "Plasmodium Coatneyi in Rhesus Macaques Replicates the Multisystemic Dysfunction of Severe Malaria in Humans." Edited by J H Adams. *Infection and Immunity* 81 (6). 1752 N St., N.W., Washington, DC: American Society for Microbiology: 1889–1904. doi:10.1128/IAI.00027-13.

Nakaya, Akihiro, Toshiaki Katayama, Masumi Itoh, Kazushi Hiranuka, Shuichi Kawashima, Yuki Moriya, Shujiro Okuda, et al. 2013. "KEGG OC: A Large-Scale Automatic Construction of Taxonomy-Based Ortholog Clusters." *Nucleic Acids Research* 41 (Database issue). Oxford University Press: D353–57. doi:10.1093/nar/gks1239.

Otto, Thomas D., Gary P. Dillon, Wim S. Degrave, and Matthew Berriman. 2011. "RATT: Rapid Annotation Transfer Tool." *Nucleic Acids Research* 39 (9): 1–7. doi:10.1093/nar/gkq1268.

Pain, A, U Bohme, A E Berry, K Mungall, R D Finn, A P Jackson, T Mourier, et al. 2008. “The Genome of the Simian and Human Malaria Parasite *Plasmodium Knowlesi*.” *Nature* 455 (7214). Macmillan Publishers Limited. All rights reserved: 799–803. http://dx.doi.org/10.1038/nature07306.

Quail, Michael, Miriam E. Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas Richard Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. “A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers.” *BMC Genomics* 13 (1). BMC Genomics: 1. doi:10.1186/1471-2164-13-341.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics and Bioinformatics*. doi:10.1016/j.gpb.2015.08.002.

Stanke, M, R Steinkamp, S Waack, and B Morgenstern. 2004. “AUGUSTUS: A Web Server for Gene Finding in Eukaryotes.” *Nucleic Acids Res* 32. doi:10.1093/nar/gkh379.

Tachibana, Shin-Ichiro, Steven A Sullivan, Satoru Kawai, Shota Nakamura, Hyunjae R Kim, Naohisa Goto, Nobuko Arisue, et al. 2012. “*Plasmodium Cynomolgi* Genome Sequences Provide Insight into Plasmodium Vivax and the Monkey Malaria Clade.” *Nature Genetics* 44 (9): 1051–55. doi:10.1038/ng.2375.

Yaffe, Eitan, and Amos Tanay. 2011. “Probabilistic Modeling of Hi-C Contact Maps Eliminates Systematic Biases to Characterize Global Chromosomal Architecture.” *Nature Genetics* 43 (11): 1059–65. doi:10.1038/ng.947.

S.A. LAPP, J.A. GERALDO, J.-T. CHIEN, F. AY, S. B. PAKALA, G. BATUGEDARA, J. HUMPHREY, the MaHPIC consortium, J.D. DEBARRY, K. G. LE ROCH, M. R. GALINSKI and J. C. KISSINGER. 2017 “PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family” Parasitology.

Chapter 5, Using Data Mining Approaches to obtain the Insights from Genomics & Transcriptome Data

5.1 Machine Learning in Systems Biology

The top layer of DIKW model is "the wisdom", which involves the knowledge processing, strategy selection and decision-making. In this chapter, we describe the approaches of applying data mining techniques (including machine learning approaches) to attempt to acquire the "wisdom". Machine learning and data mining are two quickly converging fields of study that were developed from different roots. In 1950, Alan Turing published the paper "Computing Machinery and Intelligence", posting the question 'can machines think?' (Turing, Aug, and Turing 2007). Thus, the primary aim of machine learning was to build a thinking machine that can learn from and adapt to new information. The roots of machine learning can also be traced back to 1957 when Frank Rosenblatt invented the perceptron algorithm, which models the neurons in the human brain, and this led to the development of neural networks (Rosenblatt 1958). Since the 1980s, other algorithms have also been developed such as the decision-tree-like C4.5 methods (Quinlan 1993), a classic example of 80s-90s era machine-learning research. In the mid-1990s, support vector machines (SVMs) were being widely applied to a diverse number of fields, but because of their emphasis on knowledge discovery from databases, SVMs also contributed to the growth of the data-mining field (Cortes and Vapnik 1995). Furthermore, Rakesh Agrawal *at al*. proposed an elegant algorithm for finding association patterns in large databases, creating a strong influence on discovering more efficient mining algorithms and frequent patterns (Agrawal 1993). Data mining is more practical and industry oriented, since it is concerned with typically larger datasets and the speed of data processing. Although data mining is the intersection between machine learning, artificial intelligence, and statistic and database systems for identifying patterns in datasets, machine learning might play the most crucial role in data mining. A general workflow of applying machine learning in biological domains is described below (Figure 5.1).

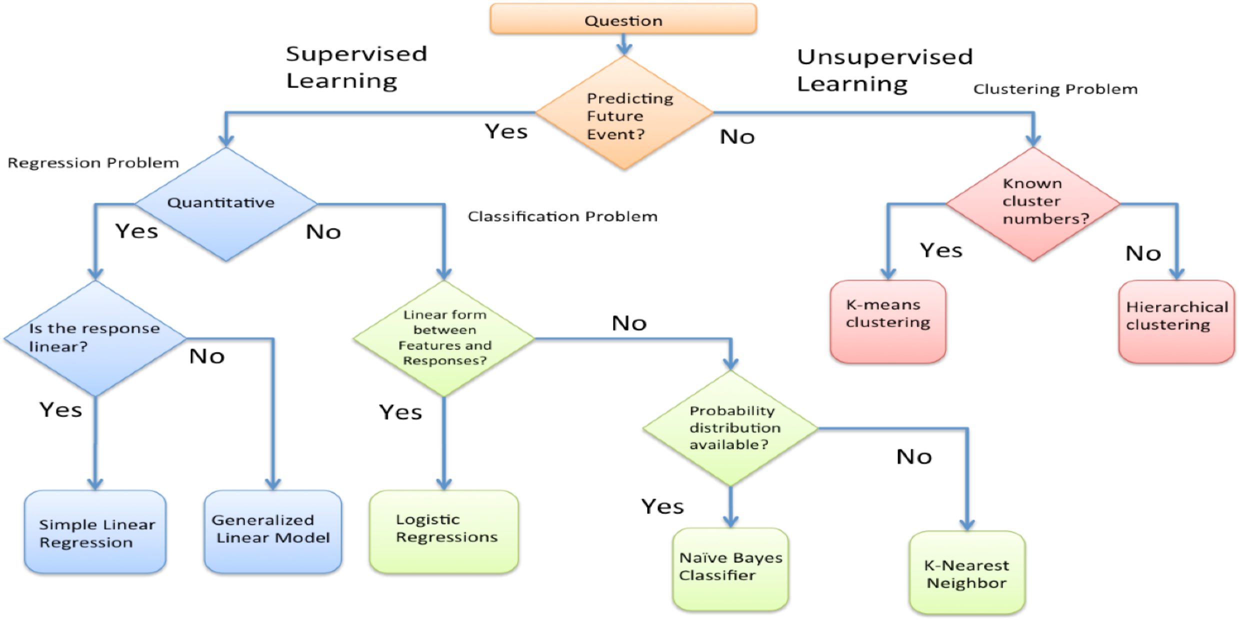


Figure 5.1, The General Workflow of Applying Machine Learning in Research Questions

The steps of applying machine learning in systems biology can be divided into data cleaning, data selection, and data modeling. Data modeling/integration is the foundation of systems biology, where computational researchers must not only process data, but also contextualize their findings. Systems biologists often receive high dimension data that has high variance because of the complex nature of biological data, making data cleaning an essential pre-processing step. Data cleaning also involves human judgment in validating and normalizing data; therefore, it must be conducted with extreme care (Rahm and Do 2000). After data cleaning, selecting the data with an eye towards knowledge discovery is an empirical process; either subjective or objective.

Following the preprocessing step, there are three major categories for understanding pre-existing relationships, utilizing algorithms such as K-means or agglomerative clustering to find partitioned or hierarchical groupings of data. Classification is a statistical method that separates data points into different categories, applying prior knowledge to the initial labeled data. Widely used classification algorithms include SVM, linear discrimination analysis, and decision forests. Finally, regression is a statistical method for modeling the relationship between a dependent variable and independent variables in order to predict the values. This works by applying supervised learning algorithms such as Bayesian linear regression and neural network regression. In this fashion, data driven knowledge discovery methods can combine model-based integration methods with different datasets to produce a model that has been constrained to data-specific properties. To combine models in a meaningful way, model-based integration must be carried out with a specific hypothesis and analysis for each data type (Kim 2015).

Using our system to obtain accurate genomic data in either reference based analysis or *de novo* genome assembling to have a better understanding of the regulatory mechanism function in infectious diseases systems biology. By applying data mining techniques, a comprehensive view of systems biology can be constructed. In this chapter, I will introduce data mining and machine learning techniques as well as introduce a potential application of this system combined with malaria host-pathogen transcriptome interaction questions, and the potential contribution of this right-left combo application will be explored.

5.1.1 Supervised Learning

Supervised learning, which has been widely used in classification problems like identifying handwritten ZIP codes, a classic example. More generally, classification is a common supervised learning approach in handling systems biology problems. In the case of neural networks, classification errors are used to adjust the network to minimize the error of the neural network. In real world applications of systems biology, neural networks were employed to demonstrate the feasibility on diagnose malaria in the Brazilian Amazon (Andrade et al. 2010). For decision tree methods, classifications are performed to determine which features provide the most information. These supervised algorithms depend on the pre-determined classification data on which they learn the statistical dependencies of the features that minimize classification error.

For classification problems, the goal is to minimize classification errors. Algorithms learn from the "training set" but the learning models created typically cannot be generalized to data outside of the training data. In other words, the ideal machine-learning algorithms will be able to learn the generalizable relationship between data points without also learning noise or unique peculiarities of a particular dataset. As a result, over-fitting to the training data is a common problem in machine learning that must be tackled using regularization techniques. Even if a model makes a few errors in the training sets, it may produce more errors when working with other datasets as well. Thus, it is prudent to split available data and allocate each part to additional datasets, such as a development set, in order to validate sample performance during training. With these points in mind, one can see the challenges in producing robust models that can be generalized to many datasets (Pintelas 2007).

5.1.2 Unsupervised Learning

The objective of unsupervised learning is to have the machine identify patterns in new information based on similarities and differences between data points without using pre-classified data. Classic unsupervised learning in systems biology is called clustering, which finds related information by putting data points in groups that minimize a given distance metric. For instance, using several attributes of iris flowers, such as length, width, etc. it is possible to separate these flowers into three species and thus three groups by clustering their attributes. Systems biology is a data-rich research field for applying clustering algorithms (Yin *et al*. 2015). Yin *et al* applied several clustering methods including K-means, hierarchical clustering, and modulated clustering before constructing a Bayesian network-based strategy for a genome scale model of malaria host-pathogen interactions over a certain period of time.

5.1.3 Semi-Supervised Learning

Semi-supervised learning uses labeled and unlabeled data together to perform tasks, which can be a highly effective approach when lacking labeled data. The process of labeling data usually requires a skilled human agent (e.g. gene annotation) or experimentation (e.g. antibody 3D structure determination), so fully labeled datasets are often not available. Although semi-supervised algorithms have been in development since the late 1960s, due to the thriving use of machine learning and data mining in other fields, semi-supervised learning is being widely applied to enhance the performance of supervised learning tasks by parsing more unlabeled data related tasks when labeled data is expensive or insufficient. This technique has been demonstrated to have great potential prediction power (Zhu 2008). In one study, Shi *et el.* (Shi and Zhang 2011) applied semi-supervised learning to overcome a difficult small sample size in cancer research, providing promising robust predicting results.

5.1.4 Feature Selection

In general, feature selection is one of the most common questions in data science, especially in systems biology. There are two reasons why feature selection is a difficult question in systems biology, the abundance and variety of the data, and the deep knowledge of the problem domain (Guyon and Elisseeff 2003). For systems biology research, the dataset usually includes genomics, transcriptomics, proteomics and other highly variant types of data, with the volume of the datasets easily reaching terabyte scale due to the high rate that biological data is currently being generated. Furthermore, due to the wide variety of data, systems biology research generally requires not only expertise in one specific field but also knowledge of the interactions between and among systems. Hence, feature selection in systems biology is important due to the interconnectivity between or among each system under a bigger system scheme (Saeys, Inza, and Larranaga 2007). A small change in one particular system might completely alter the broader outcome of the big picture. Especially in the big data era where researchers face complicated questions on a daily basis, it is almost impossible to apply prior knowledge to feature selection when datasets have thousands of features (Xing, Jordan, and Karp 2001).

Reducing the number of the features is the most intuitive solution, but another possibility is dimensionality reduction, which includes principal component analysis (PCA), singular value decomposition and Sammon's mapping (van der Maaten, Postma, and van den Herik 2009; Ridder and Duin 1997). The main difference between feature selection and dimensionality reduction is that feature selection includes a handpicked selection of the original features while dimensionality reduction creates a smaller new feature set that represents the original features. There are limitations to dimensionality reduction because the relationships between the original features are not clear. Since the dimensionality reduction method applies a feature combination solution, it might generate "hidden" feature elements that provide some variance in the data. Also, the dimensionality reduction method can sometimes remove a small but significant differentiator in the model and affect the performance (Sasan Karamizadeh, Shahidan M. Abdullah, Azizah A. Manaf, Mazdak Zamani 2013).

The purpose of applying feature selection methods in creating a model is to assist modelers in generating more accurate models while using less data, use shorter training time, and avoid over-fitting. In other words, feature selection can help modelers remove redundant features from the data that are not related to increasing accuracy and might potentially harm the accuracy of the model. There are three major approaches in feature selection: filter methods, wrapper methods and embedded methods.

Filter selection methods apply statistical methods to measure and to assign a score to the features. After measuring and assigning, the features are discarded or kept based on the rank of the score. Most filter selection methods are univariate and treat each single feature independently or related to the dependent variable. There are several examples of applying filter selection to biology, such as the chi square test in two categorical variables and the correlation coefficient scores measuring the degree of linear relationship between two variables (Talavera 2005).

The wrapper methods select a subset of features as a searching problem within different combinations of the features, then evaluate and compare with other combinations. The searching strategy for finding the best combination can be stochastic or heuristic. Stochastic methods, for instance, the hill-climbing algorithm, may not find the global maximum but it has the capability to cover the local maximum. On the other hand, heuristics methods can add or remove the features in either the forward or backward direction, and then test the score of the new combination of subsets of the features during each iteration (Kohavi and John 2011).

The final type of feature selection, the embedded method, learns the accuracy of the model based on the contribution of each feature while building the model. The most common embedded method is regularization, also known as the penalization method, which provides additional restrictions to the process of optimization while also lowering the complexity of the model. The LASSO method and Elastic Net method are commonly used methods as well (Xu et al. 2010; Zou and Hastie 2005).

5.1.5 Missing Data

In systems biology research, modelers will have to deal with missing data when obtaining datasets from bench scientists. A classic example is missing data in a microarray dataset. Of course, the issue of missing data will not only manifest itself in microarray datasets, but in other technologies and in different areas of systems biology. As a result, each sub-system is interconnected in systems biology and maintaining the accuracy of the final model in spite of the missing data effect is essential. For modelers, building an accurate model under the assumption that data will be missing is common because bench experiments are expensive and time consuming. Due to the low success rate of replicating large scale experimental results, dealing with missing data in systems biology is an essential task.

When analyzing the missing data effect, the types of missing data have to be defined case by case to find the best solution. Missing data problems can be divided into three main categories: missing complete at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR, also known as uniform non-response, is when the missing data is independent from both observed and unobserved measurements. An example of MCAR in the biomedical domain can be when an experimental sample is lost, resulting in a missing observation. A MCAR could also result from missing a random subset of parasitemia counts in a time-coursed disease progression experiment. The MAR condition can be observed when the missing data is related to a certain variable but not to the value of that variable. For example, a complete blood count (CBC) machines' accuracy might be varied in time-course animal experiments, and this will affect the data distribution. However, this phenomenon is not related to the actual CBC counts. The differences between MCAR and MAR can easily confuse researchers. The key differences are: 1, missing data in MCAR are simply missing subsets of data, and 2, that MAR is actually the data missing conditionally at random due to the missing data being dependent on the condition of another variable. The third type of missing data, MNAR, is where the data is missing for a specific reason. MNAR is a situation where even when we account for all available information, the data missing is dependent on unseen observations. For instance, time-coursed CBC data in a malaria disease progression animal model might be missing due to low blood volume or death in the late-stage of the experiment(Chiu et al. 2013; Schafer and Graham 2002).

Missing data will affect the model accuracy, so solving the missing data problem correctly is key. When the right method is used to solve the missing data problem, better accuracy in a model can be obtained. In systems biology, facing the data-missing problem while constructing sub-systems is expected; therefore understanding the workflow of bench experiments is another important task for modelers. Understanding how data were collected gives modelers the confidence to determine the types of missing data problems present. There are several strategies that modelers can apply to the missing data problem: 1, the straightforward way of only analyzing the data available, 2, imputing missing data points with replacement values as observed data points, 3, replacing missing data with uncertainty, and 4, applying statistical models to make assumptions based on the available data.

5.2 Data Mining Applications in Malaria Host-Pathogen Transcriptome Analysis

In Chapter 5 we introduced how this system assembled and annotated the first *Plasmodium coatneyi* genome and the importance of this genome to the understanding of malaria systems biology. Malaria is the most prevalent parasitic disease, with 212 million clinical reports and 429,000 deaths globally in 2015 ("WHO. World Malaria Report. Geneva: World Health Organization; 2015.," n.d.). The most prevalent regions are the tropical and subtropical areas of South Asia, Central South America, and Central and South Africa. There are five major different species of *Plasmodium* parasites that infect humans; *P. falciparum* and *P. vivax*, however, have the highest number of clinical case reports. Although the number of malaria clinical reports and mortality have decreased over the past decade, the mechanisms of asymptomatic and chronic infections are still unknown. Therefore, understanding the dynamics of chronic infections, especially transcriptomic analysis, might help malaria eradication.

Plasmodium infected non-human primates (NHP) often severe as model to study human malaria because of many shared biological features. Two simian *Plasmodium* species, *P. coatneyi* and *P. cynomolgi*, in rhesus macaques infection have been reported and serve as a model to imitate the dynamics of *P. falciparum* and *P. vivax* infections in human, respectively (Joyner, Barnwell, and Galinski 2015). These NHP-parasite models have been used to observe malaria pathogenesis, especially in clinical and hematological time-course studies. However, a comprehensive time-course studies of the host response at transcriptome level has not been address yet.

The morphological features of *P. coatneyi* are similar to *P. falciparum,* the human malaria, in rhesus macaques in a blood stage infection (Moreno et al. 2007). Here, we present the host time-course transcriptome analysis by observing the expression dynamic during the experiment. The transcriptome data was collected from *P. coatneyi* sporozoites infected rhesus macaques whole blood under a controlled time-course experiment. The time course of the infection was characterized into three phases: the acute phase in the early infection that demanded anti-malaria chemotherapy, the recrudesce phase, and a chronic phase. To characterize the dynamics of transcriptome expression in each phase, we conducted a highly robust time-coursed characterization of differential expressions in the whole blood transcriptome.

In this study, we built a robust transcriptome characterization pipeline that provides the foundation for studies of rhesus macaques (*Macaca mulatta*)- the *Plasmodium* parasites longitude infection model. The whole blood host transcriptome data has been analyzed across different time-points of parasite infection. We quantified the host transcripts across the sample batches and applied DEseq2 to normalize the RNA-seq data and conduct a differential expression analysis. Overall, a shift of the host transcriptional profile was observed throughout the phases of parasite development by applying Principle Component Analysis (PCA). The permutation of different time-point differential host expression was conducted to characterize different phases of infections, and we also applied a novel clustering method to demonstrate the dynamics of parasites expression during longitudinal observation. Moreover, we demonstrated the possibility of our pipeline to be combined with data mining techniques to generate comprehensive transcriptome analysis.

5.2.1 Experimental Design

1. Clinical follow-up of experimentally infected rhesus macaques

A detailed description of the clinical procedures used with NHP exposed to the infection with simian malaria parasites has been described previously (Ref IAI 2013, Malaria J 2016). Briefly, a cohort of five male rhesus macaques (*Macaca mulatta*, RCs13, RTi13, RUn13, RZe13, and RWr13) born and raised at the Yerkes National Primate Research Centers was assigned for the experiment. The macaques were infected with 100 freshly isolated *Plasmodium coatneyi* sporozoites, collected from salivary glands of three different species of *Anopheles* mosquitoes infected and maintained at the

Centers for Disease Control and Prevention, using intravenous inoculation. The animals were followed up for 100 days to evaluate the clinical outcome. Capillary samples collected daily in the course of the infections were used for hematological and parasitological assessment. Venous blood and bone marrow aspirates were also collected at seven time points during the follow-up period for multi-omic analyses. These time points correspond to a clinically relevant aspect of the infection. Based on the time course of the infection, the animal exhibited a spectrum of clinical phenotypes that reproduced the individual variability reported in humans. One of the macaques (RCs13) did not develop parasitemias after inoculation of sporozoites and consistent with an ineffective infection changes in hematological parameters were not recorded. The four infected macaques developed parasitemias that peaked on ~day 22 after experimental infection and received subcurative treatment with artemether to avoid clinical complications. The early chronic phase resulted in a clinical phase characterized by self-controlled parasitemias with parasite levels lower than 1,000 parasites/ml. This phase was defined as chronic. The length of the chronic phase defined the clinical phenotype as follows. 1) Severe (RTi13, characterized by a short chronic phase of 9 days); 2) mild (RWr13, characterized by a long chronic phase of 60 days) and 3) intermediate (RUn13 and RZe13, characterized by chronic phases ranging between 34 to 39 days). At the end of the follow-up period, the animals received a curative regimen of artemether. The workflow has been described in Figure 5.2.

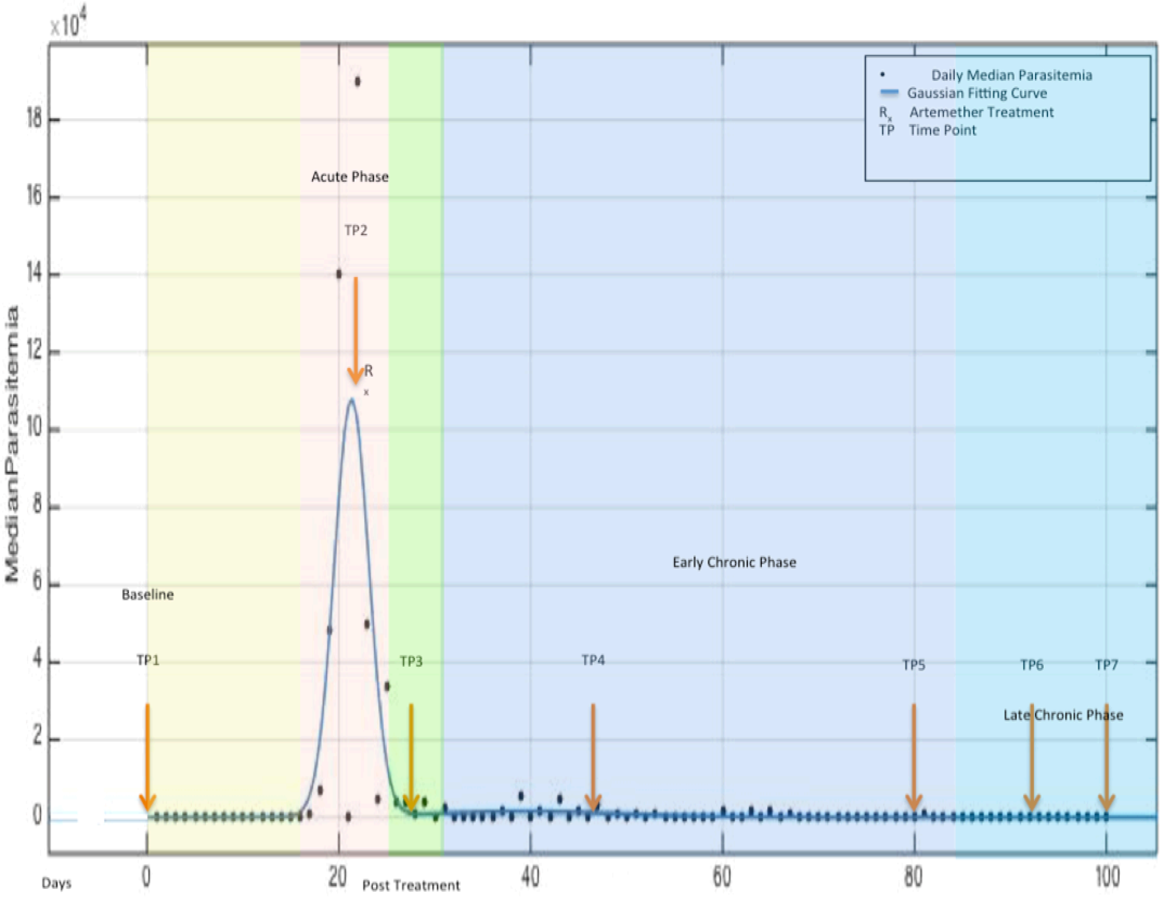


Figure 5.2. *MaHPIC* 100days NHP-*P.coatneyi* Infection Experiment Design..

The yellow area represents the baseline and after infection.; The red area in the diagram represents the acute phase; the green area represents the post treatment ;the darker blue area represents the early chronic phase; and the light blue area represents the chronic phase.

2. RNAseq Library Preparations and Sequencing

Total RNA was extracted using the Paxgene RNA isolation kit (Qiagen, CA). Globin transcripts were depleted using GLOBINclear™ Human Kit (Ambion, TX) according the manufacturer's instructions. Libraries were prepared using the Illumina (Illumina Inc., CA) TruSeq™ mRNA stranded kit as per manufacturer's instructions. Briefly, 1 ug of Globin depleted RNA was used for library preparation. ERCC (Ambion) synthetic spike-in 1 or 2 was added to each Globin depleted RNA sample. The TruSeq method (high-throughput protocol) employs two rounds of poly-A based mRNA enrichment using oligo-dT magnetic beads followed by mRNA fragmentation (120-200 bp) using cations at high temperature. First and second strand cDNA synthesis was performed followed by end repair of the blunt cDNA ends. One single "A" base was added at the 3' end of the cDNA followed by ligation of barcoded adapter unique to each sample. The adapter-ligated libraries were then enriched using PCR amplification. The amplified library was validated using a DNA tape on the Agilent 4200 TapeStation and quantified using fluorescence based method. The libraries were normalized and pooled and clustered on the HiSeq3000/4000 Paired-end (PE) flowcell on the Illumina cBot. The clustered PE flowcell was then sequenced on the Illumina HiSeq3000 system in a PE 101 cycle format. Each sample was sequenced to a target depth of 100 million pairs (50 million unique fragments) with exception of Time point 2 samples that were sequenced to 200 million pairs (100 million unique fragments).

3. Bioinformatics Pipeline

The reads were mapped to the *Macaca mulatta* reference genome V7.8.2 and concatenated with *Plasmodium coatneyi* Hackeri strain reference genome (assembled by using our pipeline as described in Chapter 4) using STAR v2.5.2b (using default aligning parameters). After aligning the reads to reference genome, the read counts for each annotated genes have been quantified by using STAR while applying HT-seq algorithm to generate the read counts. For proceeding to detect the differential expression between different time points and control, normalizing the count table has to be conducted prior the downstream differential expression comparison. In this study, we applied R package DESeq2 version 1.10.1 , implemented in R version 3.2.3, executing under OSX 10.11.4. The reason of using DESeq2 as tool for comparing differential expression is widely applied because DEseq2 adapted negative binomial generalized linear model (GLM) to evaluate differential gene expression and it has been reported as one of the appropriate differential expression detection tools (Anders, Reyes, and Huber 2012).

4. Principle Component Analysis (PCA)

We applied PCA analysis for visualize the pattern of host transcriptome dynamic in longitude observation. PCA has been widely applied on dimension reduction and it also has been widely applied on emphasizing the variation and characterizing the pattern. In our study, we treat genes (~16000) as feature (P) and host as (N) across time points. We applied Partech® genomics suite 6.16.0419 for visualization.

5. Functional Enrichment Analysis

We applied the Ingenuity Pathways Analysis (IPA, http://www.ingenuity.com) software to screen the whole gene set to determine the module that we are interested (e.g. top 100 gene sets) reached the threshold of enrichment with reported gene ontologies. We used Fisher exact test p-value as main threshold to select the significant gene sets. We also applied GSEA (http://software.broadinstitute.org/gsea/index.jsp), Metacore® (https://portal.genego.com) and DAVID (Huang et al. 2007) for gene set analysis to investigate enrichment of our gene set and compute the consensus gene sets correlate to disease status.

6. Consensus Clustering

To identify the similarity of the parasite genes expression pattern, three clustering methods (self-organized map, K-means clustering, and hierarchical clustering) were performed, and genes that share the same pattern were grouped into the same clusters. We combined those three clustering methods into a consensus clustering approach to create the consensus clusters. The consensus clustering approach can represent the consensus across different clustering methods, and it can provide a better view of cluster numbers. Most iterative descent clustering methods, like k-means and SOM clustering, overcome some of the imperfections of hierarchical clustering by providing for unambiguously defined clusters and cluster margin. In our study, we assumed that all three methods were generally feasible in identifying which genes had a strong association of expression patterns; therefore constructing a consensus cluster to eliminate the bias and assess the stability of clusters is essential. As a result, the strong correlations of parasite genes expression had been clustered together into the same consensus cluster.

5.3 The Initial Results of Malaria Host-Pathogen Transcriptome Analysis

To obtain a better understanding of the dynamic performance between the different time points during infection, gene enrichment analysis was applied to quantify the expression levels of the seven sequenced time points. We aligned the paired-end reads back to the concatenated host-pathogen genome reference-using STAR, and the alignment results were put into gene matrices. The normalized read counts were calculated and extracted by using DEseq2.

5.3.1 Pathogen Transcriptome Analysis

Given the wide range of parasitemias and resulting parasite RNAseq read counts in the different samples, we normalized for parasite read count using variance stabilizing transformation (vst) normalization. After vst normalization, we performed the consensus clustering method to identify the pattern of each cluster of genes.

Most the pathogen genes had a transcriptional peak in the acute phase. A specific cluster (#14) included the ring-stage gene that expresses a histidine-rich knob protein-like protein KPRPC, which plays a key role in the generating knob proteins. Knob proteins reportedly have a significant role in cytoadherence and cytoadherence associate with malaria coma. The antigen variant *SICAvar* gene family has significant differential expression across the different time-points with the peak in the acute phase, and declining afterward. However, in our cluster plot (Figure 5.3A), the peak was not in acute phase because the plot did not show the average of several genes in that cluster. Of the 1418 annotated genes, only a small set of genes have their peak in the chronic phase. Meiotic recombination protein DMC1-like protein presented in the chronic phase, and it belongs to cluster 3 (Figure 5.3B). This gametocyte gene is reportedly involved in gametocyte development. Gametocyte genes are likely to have higher enrichment in the later stage of the infection than the early stage, and while below the level of microscopy detection, this result may imply a switch from asexual to sexual development (Mlambo, Coppens, and Kumar 2012). We applied the consensus clustering method in finding the pattern of the pathogen gene expression; however, because of the high parasitemia in acute phase and low parasitemia in other time points, applying any normalization method will introduce random errors in each data point.

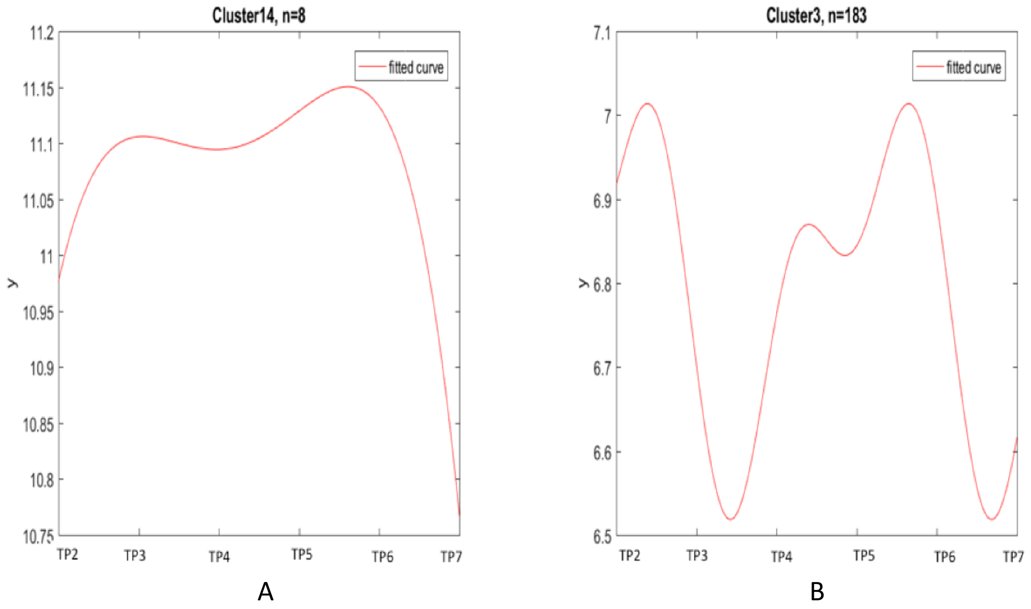


Figure Pathogen Gene Expression Pattern Clustering .

5.3A & 3B, The pattern model: 4A represents cluster 14 which contains 8 genes with similar pattern. 4B is a pattern plot representing the cluster 3 contents, 183 genes with similar pattern.

5.3.2 Host Transcriptome Analysis

To further identify genes that have significant expression differences between time points, a pairwise comparison was performed and the significance was confirmed using DEseq2. In total, we identified 13,884 differentially expressed genes through the different stages of infection and between the two time points with a P-adjust value of 0.05 and |log2 fold change| ≥ 1. We used a log2 fold change cutoff of +1 and -1 to define when a gene has been deferentially expressed. The detailed relationships between expressed genes and differentially expressed genes are shown in Figure 5.4A. 2690 genes have passed our differential expression threshold, and there were 169 genes commonly, differentially expressed, corresponding to the major diseases stages, TP2 to TP6; 1,201 (987 + 214) DEGs were specifically regulated in the acute infection stages; 1423 (1191 + 232) DEGs corresponded to TP3 (Rx), post treatment; an average of 873 DEGs were detected during the early chronic phase; and an average of 872 DEGs were detected in the chronic phase. The distributions of up and down regulated DEGs through the 6 pairwise comparisons are demonstrated in Figure 5.4B. The number of intersected DEGs detected in the pair-wise time point comparisons between the acute and chronic phase species were smaller than other pair-wised time points, indicating the significant differences between the two phases and their regulatory patterns. PGER2 and PGER3 appear in the acute phase (TP2), while the G-protein alpha-i family appeared in the chronic phase (TP6) and did not contain PGER2 and PGER3. We observed the recrudescent phase of the disease progress to attempt to understand the dynamic mechanisms occurring as the disease moves from the acute phase to the chronic phase. The number of DEGs identified in the TP2 vs. TP3 comparison was significantly higher than the number detected between TP5 to TP6 and TP6 and TP7. These observations provide resources for further investigations into the complicated regulatory mechanisms that occur during the transition from acute phase to chronic phase. The overlapping of DEGs across different time-points is profiled in Fig 5.4C.

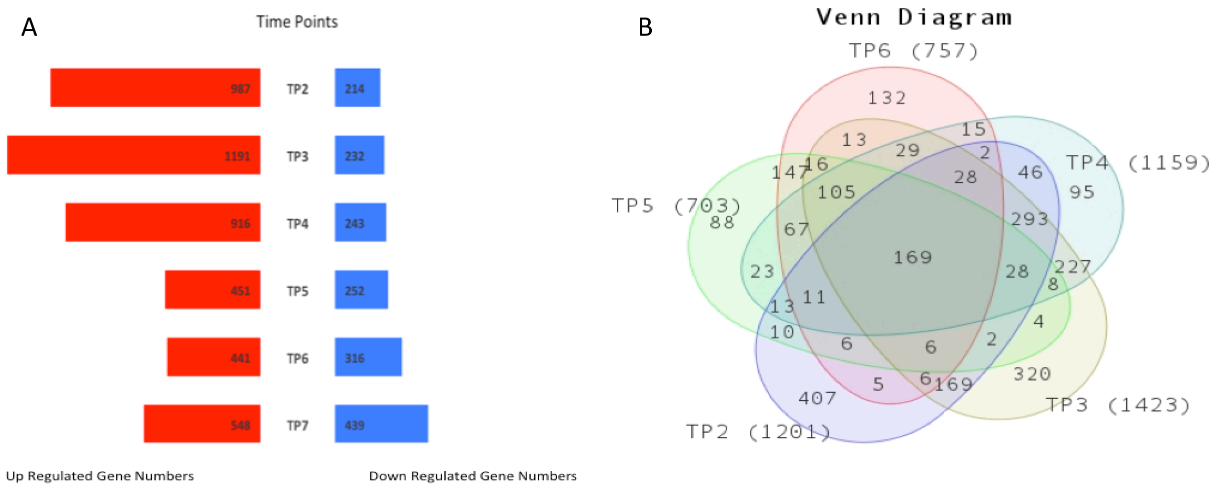
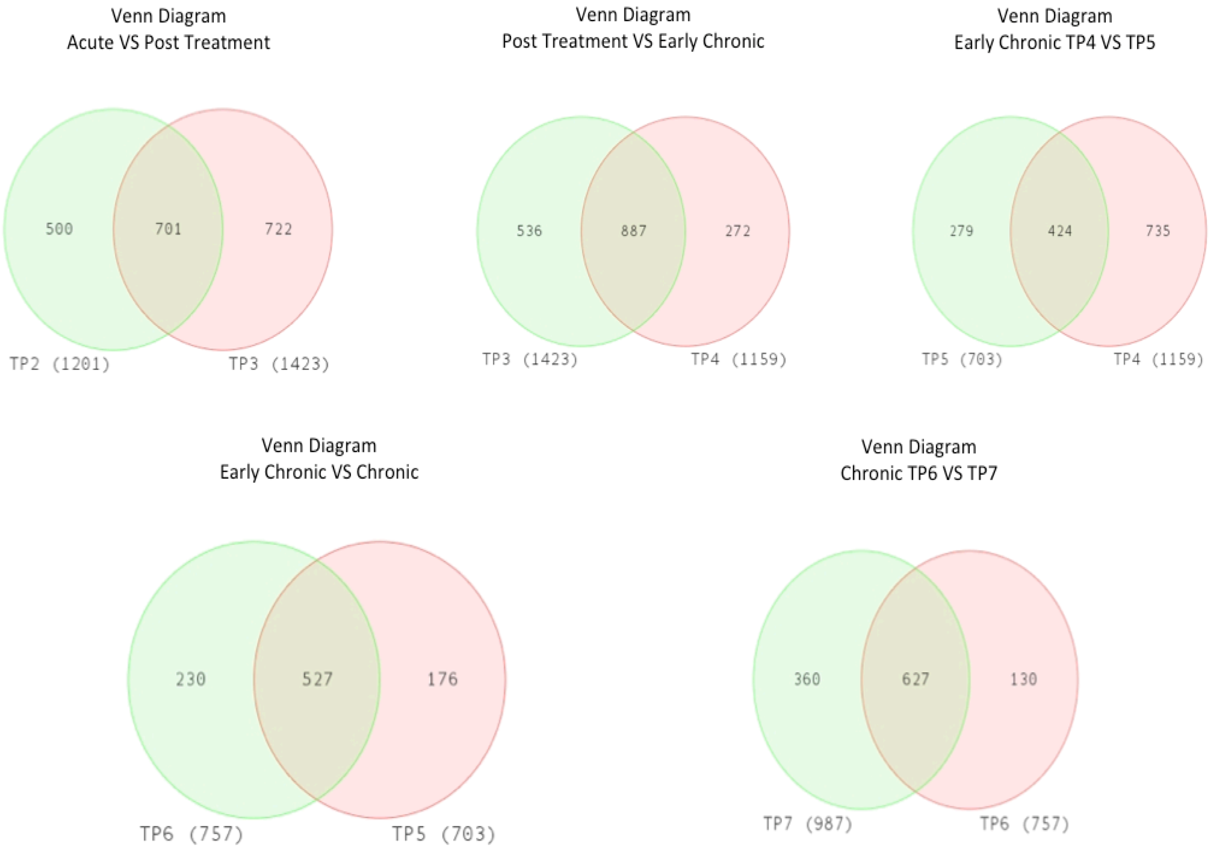


Figure 5.5 Host Differential Expression Genes in different Time Points

A) Butterfly Chart, the red bars represent the up-regulated DEGs, and the blue bars represent the down-regulated DEGs.



C) Time points Venn diagram represents the disease progression in DEGs overlapping.

After mapping the DEGs from the entire experiment with the DAVID bioinformatics database, the most significant pathways are the response to wounding, inflammation, defense response, humoral immune responses, hemostasis, cell adhesion, blood coagulation and hemeostatic process (Fig 5.5). These pathways were enriched in the data set of 2690 genes induced after *Plasmodium coatneyi* infection. By using the classification from DAVID, we focused on genes related to acute infection and chronic infection. For a better global view of the DEGs, a heat map has been produced applying hierarchical clustering (FIG heatmap), and the heat map demonstrates the clear pattern between different time points and animals and it has the similar pattern as the (PCA figure). PCA demonstrated the disease progression pattern, as the profiles of expression pattern for individual samples and profiles of gene expression in different infection stages. TP2 could serve as a mark to indicate the acute infection, and it was dissimilar from each other and from baseline after the non-human primate models acquired the *Plasmodium* infection. The animal effect has been observed, the monkey model RWr13 clearly has reached chronic phase before three other monkeys, and RTi13 was the last one to reach the chronic phase after treatment. Based on the observation from the heat map, we selected several DEGs based on the pattern of the time series expression into two groups, acute and chronic for comparison.

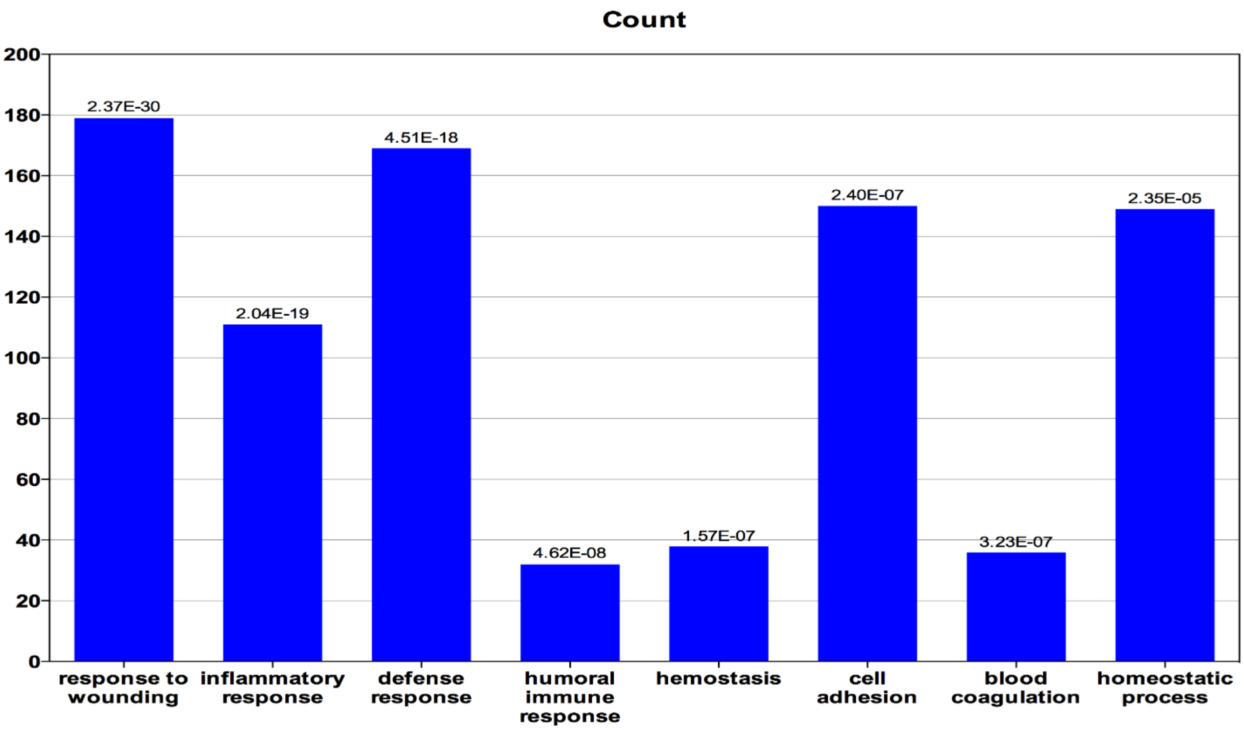


Figure 5.5 Host DEGs Pathway Ontology .

2690 DEGs Ontology profiling was based on gene ontology annotations DAVID Bioinformatics Database. Blue bars represent the number of genes in each ontology cluster and Fisher Exact (p-value) is adopted to measure the gene-enrichment in annotation terms.

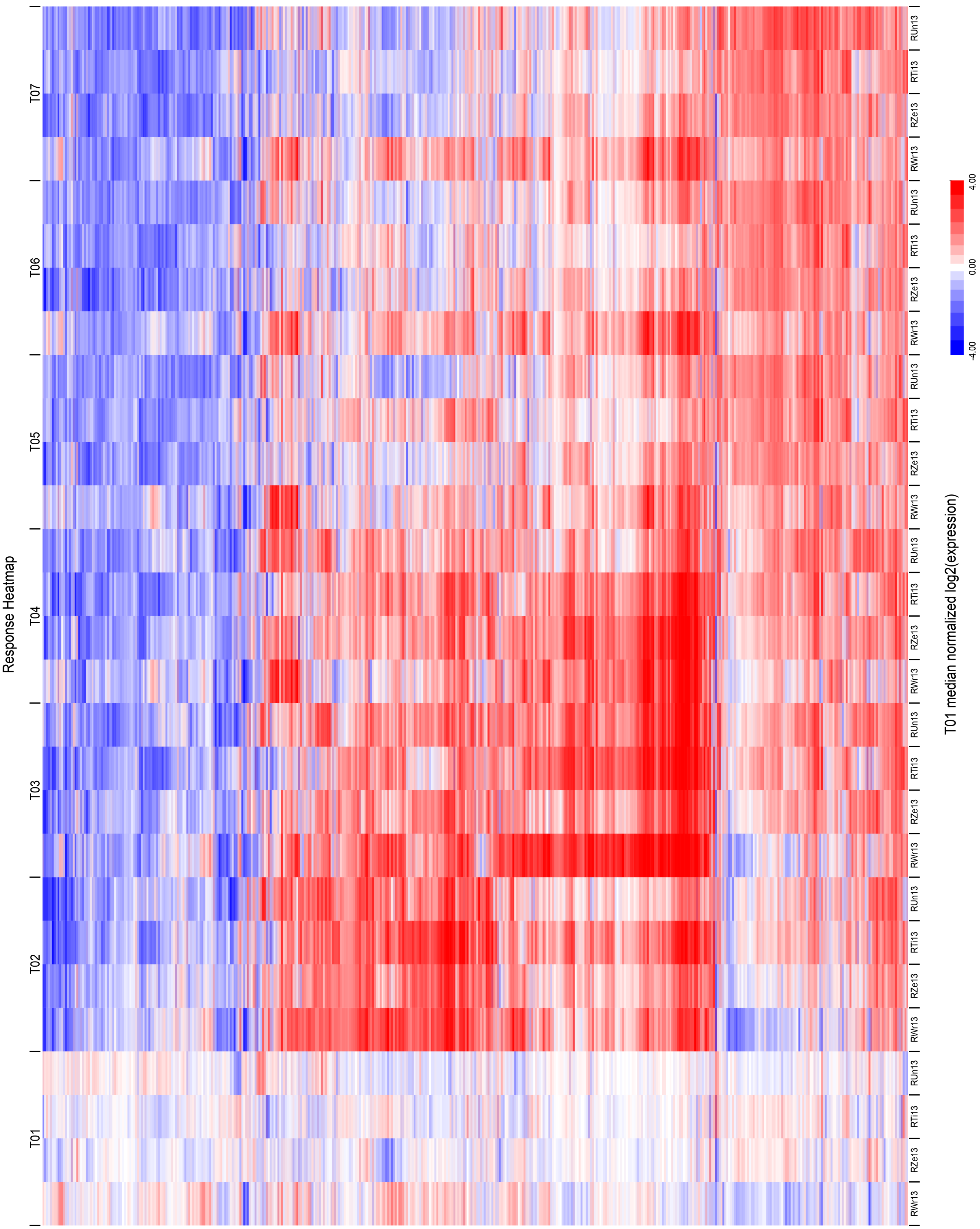


Figure 5.6 Host Global DEGs heat map. Hierarchical clustering of 2690 genes significantly regulated by *P. coatneyi* infection in different time points and individuals. Clustering was performed on the average of log10 ratios of gene expression comparing to TP1 (baseline); fold changes were calculated by subtracting the log10 intensity pre-infected measurement from after-infection measurement for individual animals prior to calculating the average.

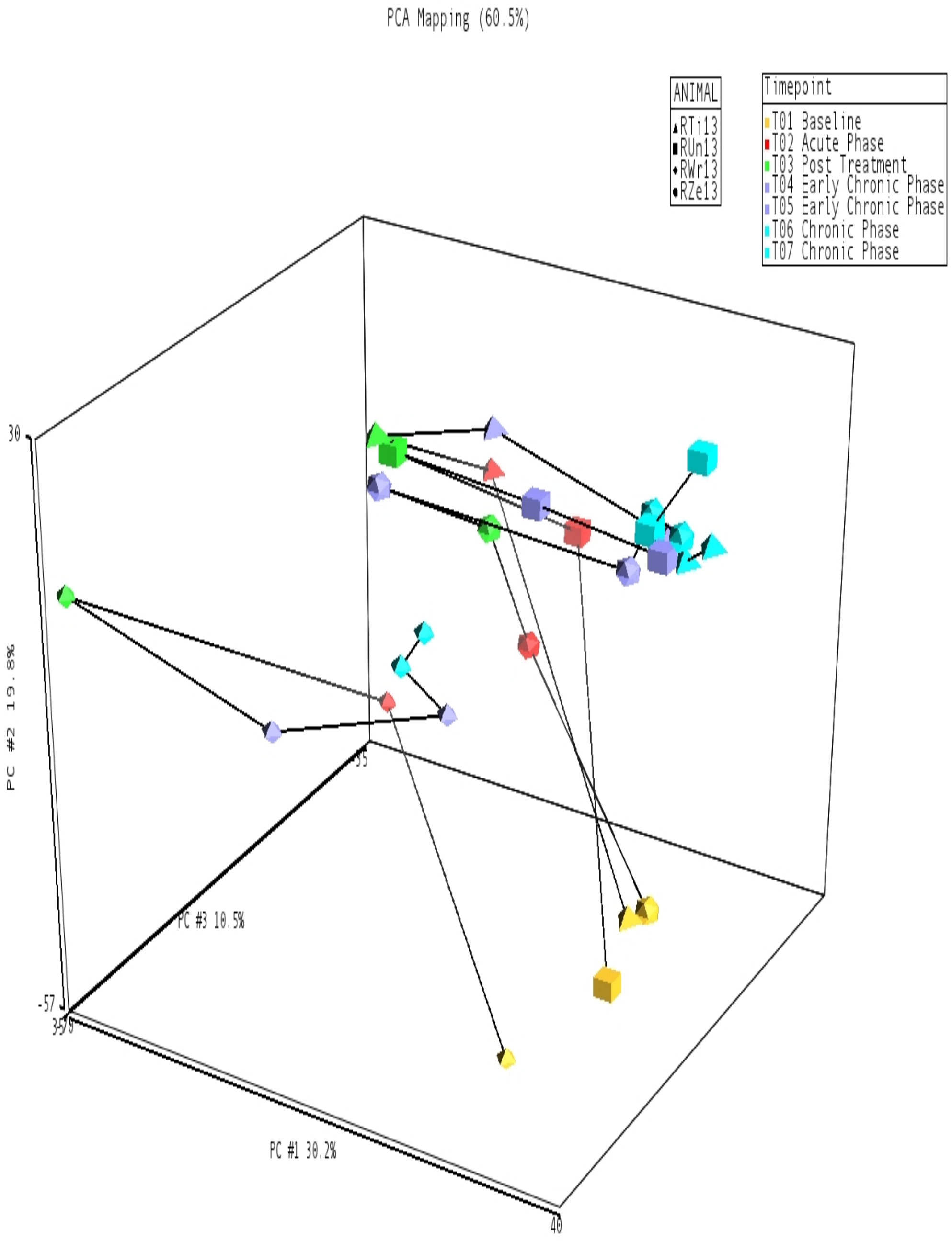


Fig 5.7 Visualization the Disease Progression by Applying PCA in Host DEGs

Log10 intensity measurements of 2690 DEGs in 4 NHP models RTi13 (triangle), RUn13 (square), RWr13 (diamond) and RZe13 (hexagonal). Colored labels represent disease stages at the time points.

5.3.3 Acute phase DEGs Analysis

The most obvious finding was that a large number of genes associate with inflammation were highly expressed in acute infection phase, including CD40, CCR2, IL10, IL27, IL2RA. Note that CD40 has significant expression in acute phase; however, it was even more up regulated in late infection. CD40 has been reported in several *Plasmodium* infection studies, but the longitudinal observation was not well defined. The acute phase also has a strong up regulation of several innate immune associated genes, including complement systems and MyD88 pathway. We also observed up regulation in type I IFN induced genes, STAT1 & STAT2. The genes clustered in the acute phase also revealed that up regulated genes are associated with a number of known cell adhesion genes, including ICAM1, TNFAIP6, ITGB5 and ITGA2. Several angiogenesis associated genes are up regulated in the acute phase, including FLT1, C5, SERPINE1 and IL1A (Fig 5.8).

These data indicate that acute *Plasmodium* infection is associated with a significant up regulation of several genes involved in triggering innate immune responses and inflammation to clear malaria infections.

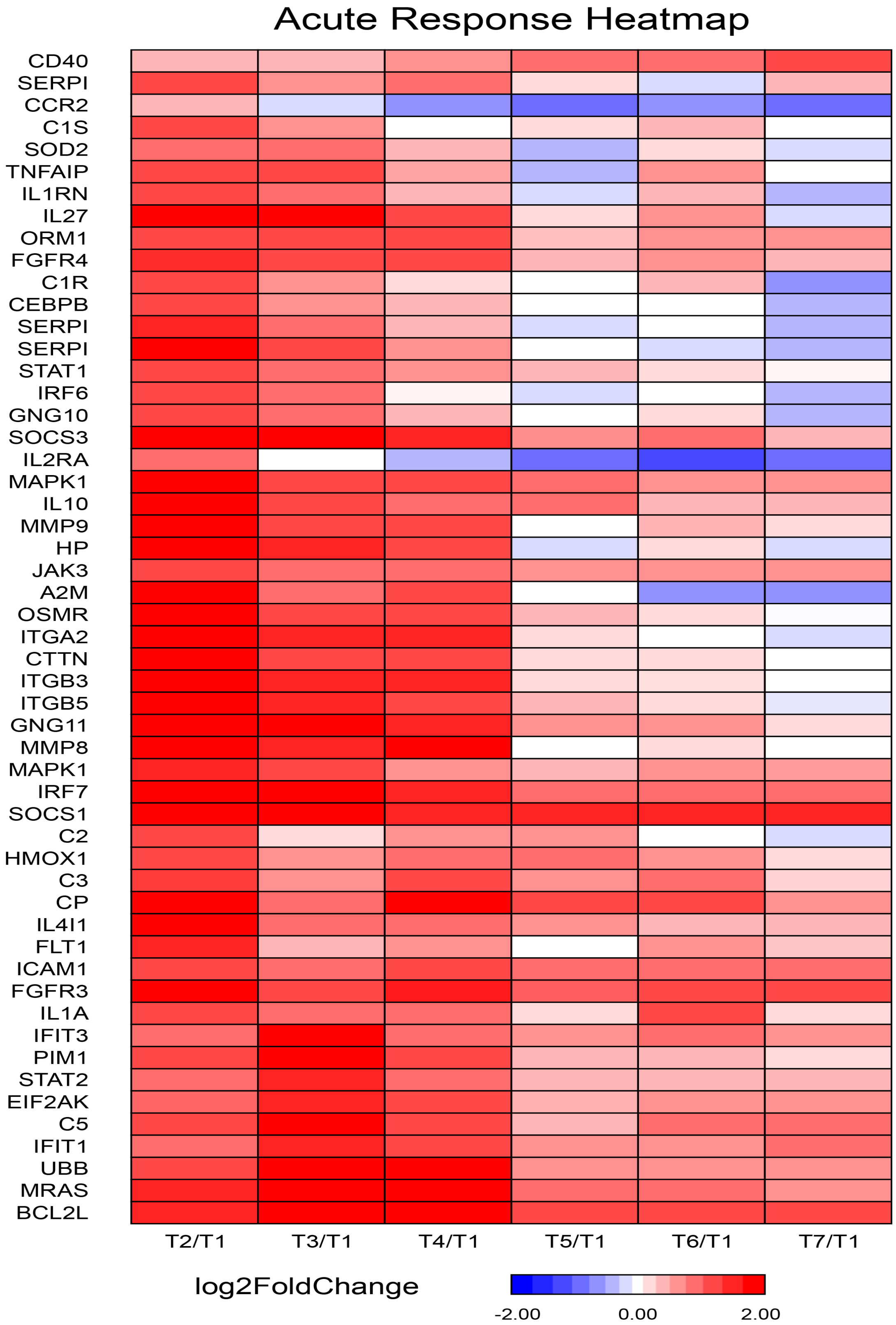


Fig 5.8 Host Acute phase DEGs heat map. Heat map of 53 selected gene sets significantly up regulated in acute phase.

5.3.4 Chronic phase DEGs Analysis

During the chronic phase, we observed that numerous genes were up regulated from TP4 – TP7. We then applied the ontology mapping and revealed up regulated genes were associated with an adaptive immune response. The innate immune response and inflammation related genes are significantly down regulated in chronic phase. The B-cell receptor pathway has highly up regulated, and the genes associated with these responses, including CD19, CD22, CD79B, CD79A, VAV2 and BLNK, are highly, differentially expressed. Genes, including TLR9, TLR10, TNF and CD80, are highly up regulated. These genes are associated with Toll-like receptor pathway, and this pathway regulated the Th1 cells and Th2 cells. We also observed genes associated with the transmembrane signals that produce immunoglobulin were upper regulated, including FGFR1, CD19, CD80, CD79B, CD79A, EBI3, CD180, and ADRB1. Collectively, the chronic phase DEGs analysis demonstrated a significant difference between acute and chronic infection host response (Fig 5.9).

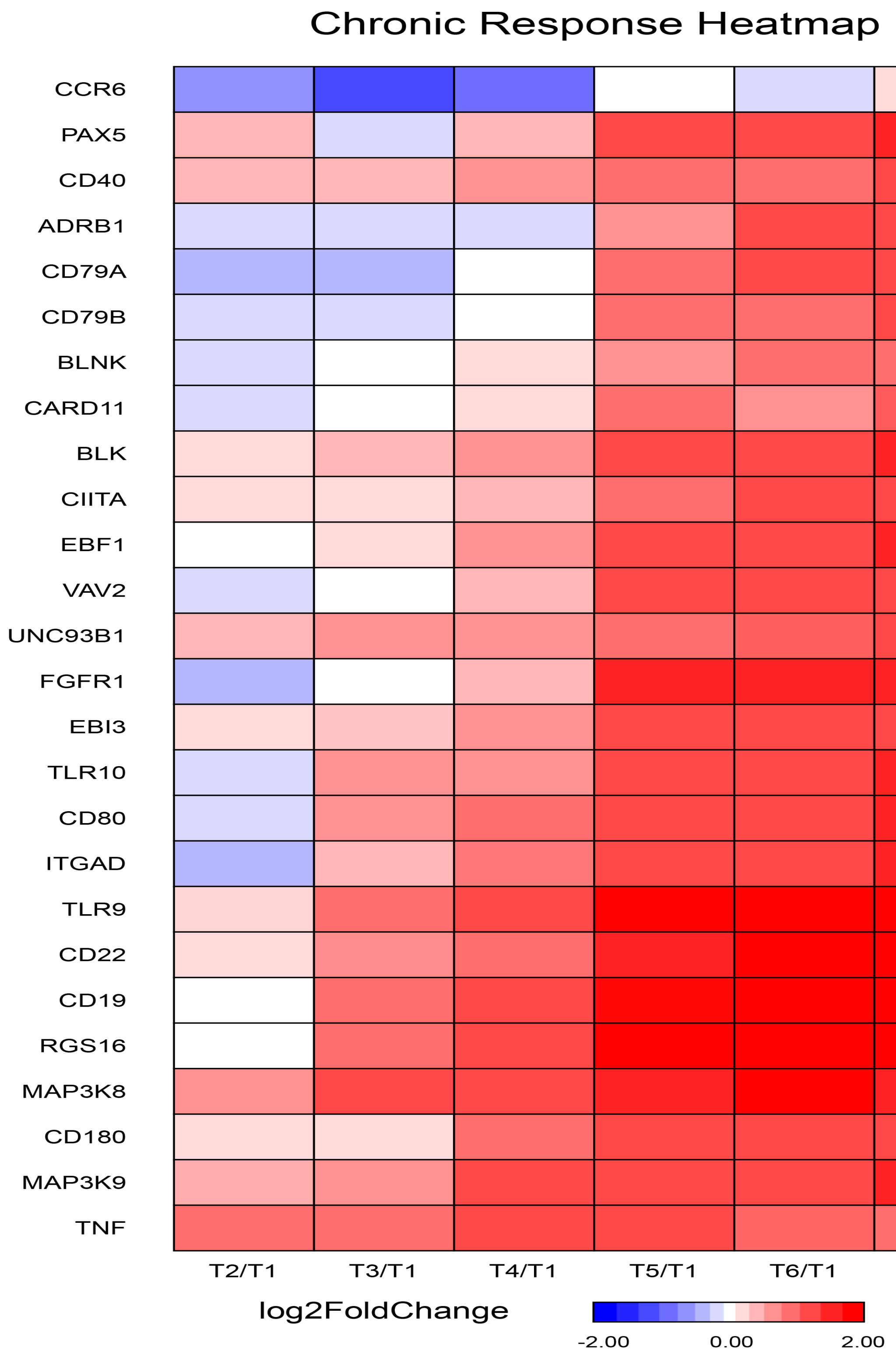


Fig 5.9. Host Chronic phase DEGs heat map. Heat map of 26 selected gene sets significantly up regulated in chronic phase.

5.4 Discussion

We can see the potential to unlock complex transcriptome dynamics by applying our system and machine learning techniques. We used malaria as our example to demonstrate the potential applications because malaria remains one of the most difficult biological paradigms to understand, and the transition from genomics to functional genomics with an emphasis on transcriptomic research has just begun. Although we only have initial results at this moment, the results show the promising indication that applying host-pathgoen dual time-series transcriptome analysis has the potential to help us understand the complicated malaria host-pathogen interaction. Due to the low sample size (n=4), the unsupervised machine learning methods are the approaches to analyze the data; however, combining prior domain knowledge to select the gene list (feature selection) for further analysis is still a better method to provide comprehensive interpretation of massive transcriptome analysis in longitudinal experiments (Appendix 5).

The pathogen gene expression analysis in this study was performed under the condition of dual RNA-sequencing, thus the low pathogen sequence coverage depth was expected other than in the acute infection phase. Low read coverage might introduce random errors while conducting the differential expression analysis; however, the scope of this study revealed the interaction between *P. coatneyi* and the NHP model rather an in-depth gene expression profiling of the disease progression. To conduct a low bias in-depth dual RNAseq analysis in malaria host-pathogen interaction study, a deeper sequencing experiment has to be performed in order to obtain a better resolution of the expression.

To conduct a comprehensive understanding of disease progression in host-pathogen response, a longitudinal transcriptome analysis is the initial step. However, this knowledge is not sufficient to reach the "wisdom" layer, thus combining other -omics data is necessary to complete the validation from different biological aspects. Combing immunological profiling could be one of the several possible approaches to investigate further because the host transcriptome data has shown strong signals in generating immunoglobulin, B-cell receptor pathways and Th1/2 regulation pathways (Supplementary Fig 1). However, without other –omics data to validate or to complete the "module" of the certain status of biological system, attempting to form the "wisdom" is still challenging.

5.5 Conclusion

The ultimate objective in malaria research is translational work with new interventions to halt all aspects of malaria's transmission and disease progression. This paper features malaria systems biology as a modern approach to this end, emphasizing current approaches and progress in genomics and functional genomics, and introduces the conceptual DIKW framework. True wisdom involves the human factor and decision-making strategies, all of which require abundant experience. Making the right decisions involves the consideration of multiple perspectives, which is important given the fact that the outcome of making various decisions during the process of ending malaria will undeniably influence the world in different realms spanning public health, the environment, economics, and eco-systems. As technologies have advanced exponentially over the past few decades, knowledge formation has similarly been just as dynamic and has progressed faster than ever. Today’s new generations of biological scientists can shorten the time frame required for the acquisition of data, information, knowledge and wisdom with the help of powerful computational tools that are now ubiquitous in the big-data era.

5.6 Supplementary

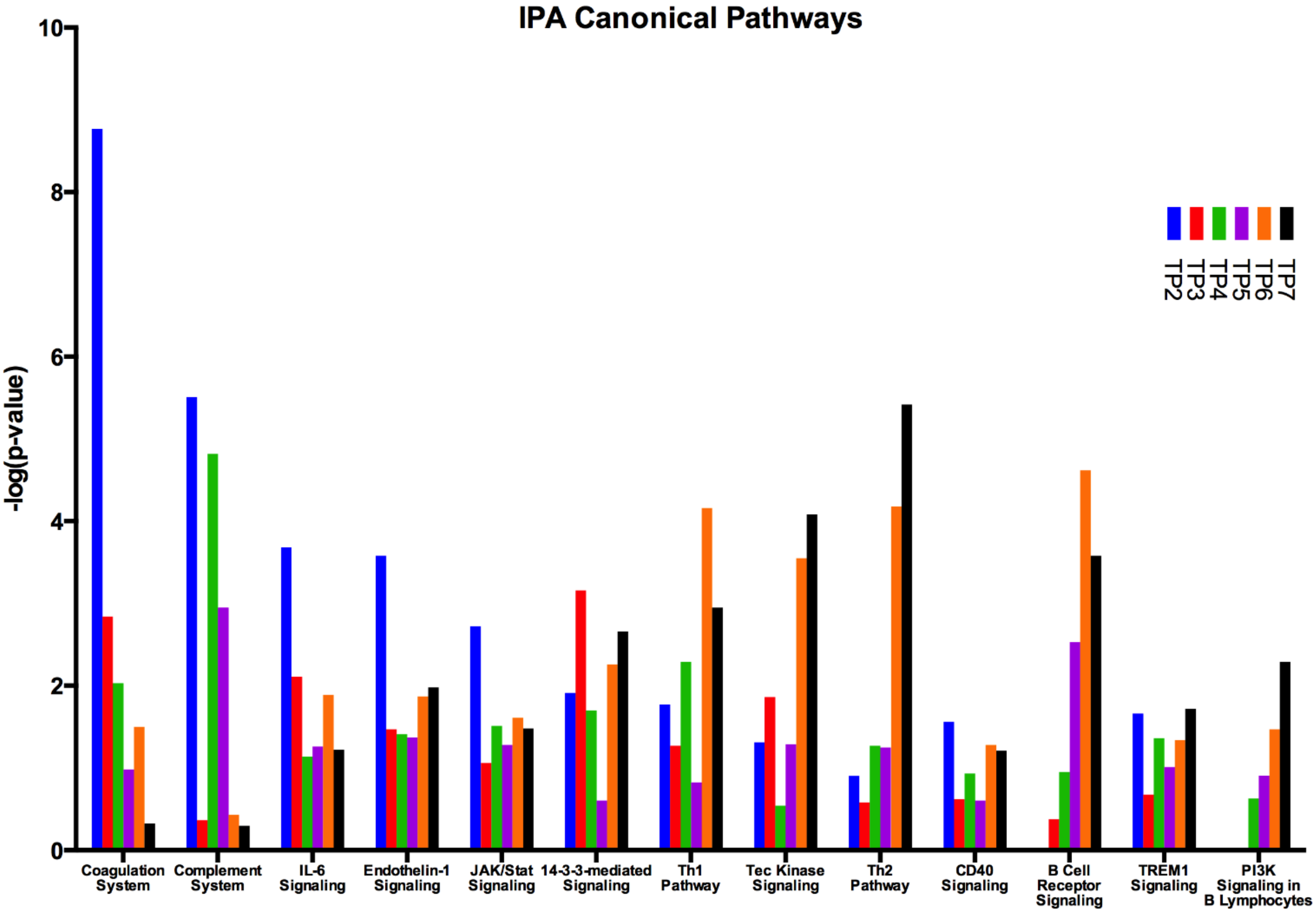


Figure 5.10 Host DEGs Canonical Pathways. Colored bar represents pathway –log(p-value) in each time points.

5.7 References

Agrawal, Rakesh. 1993. "Mining Association Rules between Sets of Items in Large Databases," no. May: 1–10.

Anders, Simon, Alejandro Reyes, and Wolfgang Huber. 2012. "Detecting Differential Usage of Exons from RNA-Seq Data." *Genome Research* 22 (10): 2008–17. doi:10.1101/gr.133744.111.

Andrade, Bruno B, Antonio Reis-Filho, Austeclino M Barros, Sebastião M Souza-Neto, Lucas L Nogueira, Kiyoshi F Fukutani, Erney P Camargo, et al. 2010. "Towards a Precise Test for Malaria Diagnosis in the Brazilian Amazon: Comparison among Field Microscopy, a Rapid Diagnostic Test, Nested PCR, and a Computational Expert System Based on Artificial Neural Networks." *Malaria Journal* 9: 117. doi:10.1186/1475-2875-9-117.

Chiu, Chia-Chun, Shih-Yao Chan, Chung-Ching Wang, and Wei-Sheng Wu. 2013. "Missing Value Imputation for Microarray Data: A Comprehensive Comparison Study and a Web Tool." *BMC Systems Biology* 7 (Suppl 6). BioMed Central Ltd: S12. doi:10.1186/1752-0509-7-S6-S12.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3). Kluwer Academic Publishers-Plenum Publishers: 273–97. doi:10.1023/A:1022627411411.

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *The Journal of Machine Learning Research* 3. JMLR. org: 1157–82.

Huang, Da Wei, Brad T. Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, et al. 2007. “DAVID Bioinformatics Resources: Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists.” *Nucleic Acids Research* 35 (SUPPL.2). doi:10.1093/nar/gkm415.

Joyner, Chester, John W. Barnwell, and Mary R. Galinski. 2015. “No More Monkeying around: Primate Malaria Model Systems Are Key to Understanding Plasmodium Vivax Liver-Stage Biology, Hypnozoites, and Relapses.” *Frontiers in Microbiology* 6 (March): 145. doi:10.3389/fmicb.2015.00145.

Kim, Dokyoon. 2015. “Methods of Integrating Data to Uncover Genotype – Phenotype Interactions.” *Nature Publishing Group* 16 (2). Nature Publishing Group: 85–97. doi:10.1038/nrg3868.

Kohavi, Ron, and H John. 2011. “Artificial Intelligence Wrappers for Feature Subset Selection” 97 (97): 273–324.

Mlambo, Godfree, Isabelle Coppens, and Nirbhay Kumar. 2012. “Aberrant Sporogonic Development of Dmc1 (a Meiotic Recombinase) Deficient Plasmodium Berghei Parasites.” *PLoS ONE* 7 (12). doi:10.1371/journal.pone.0052480.

Moreno, A, A Garcia, M Cabrera-Mora, E Strobert, and M R Galinski. 2007. “Disseminated Intravascular Coagulation Complicated by Peripheral Gangrene in a Rhesus Macaque (Macaca Mulatta) Experimentally Infected with Plasmodium Coatneyi.” *Am J Trop Med Hyg* 76.

Pintelas, S B Kotsiantis I D Zaharakis P E. 2007. “Machine Learning : A Review of Classification and Combining Techniques,” no. 2006: 159–90. doi:10.1007/s10462-

007-9052-3.

Quinlan, J Ross. 1993. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rahm, E, and Hh Do. 2000. “Data Cleaning: Problems and Current Approaches.” *IEEE Data Eng. Bull.* 23 (4): 3–13. doi:10.1145/1317331.1317341.

Ridder, Dick De, and Robert P W Duin. 1997. “Sammon ’ S Mapping Using Neural Networks : A Comparison,” 1307–16.

Rosenblatt, F. 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in ….” *Psychological Review* 65 (6): 386–408. doi:10.1037/h0042519.

Saeys, Y., I. Inza, and P. Larranaga. 2007. “A Review of Feature Selection Techniques in Bioinformatics.” *Bioinformatics* 23 (19): 2507–17. doi:10.1093/bioinformatics/btm344.

Sasan Karamizadeh, Shahidan M. Abdullah, Azizah A. Manaf, Mazdak Zamani, Alireza Hooman. 2013. “An Overview of Principal Component Analysis.” *Journal of Signal and Information Processing* 4 (August): 173–75. doi:10.4236/jsip.2013.43B031.

Schafer, Joseph L, and John W Graham. 2002. “Missing Data: Our View of the State of the Art.” *Psychological Methods* 7 (2). American Psychological Association: 147.

Shi, Mingguang, and Bing Zhang. 2011. “Semi-Supervised Learning Improves Gene Expression-Based Prediction of Cancer Recurrence.” *Bioinformatics* 27 (21). Oxford Univ Press: 3017–23.

Talavera, Luis. 2005. “An Evaluation of Filter and Wrapper Methods for Feature

Selection in Categorical Clustering." *Advances in Intelligent Data Analysis VI*, 742. doi:10.1007/11552253_40.

Turing, A M, No Aug, and B Y A M Turing. 2007. "The Chemical Basis of Morphogenesis THE CHEMICAL BASIS OF MOKPHOGENESIS" 237 (641): 37–72.

van der Maaten, Laurens, Eric Postma, and Jaap van den Herik. 2009. "Dimensionality Reduction: A Comparative Review." *Journal of Machine Learning Research* 10 (January): 1–41. doi:10.1080/13506280444000102.

"WHO. World Malaria Report. Geneva: World Health Organization; 2015." n.d.

Xing, Eric, Michael Jordan, and Richard Karp. 2001. "Feature Selection for High-Dimensional Genomic Microarray Data." *Proc. 18th International Conf. on Machine Learning*, 601–8. doi:citeulike-article-id:431564.

Xu, Zenglin, Rong Jin, Haiqin Yang, Irwin King, and Michael R Lyu. 2010. "Simple and Efficient Multiple Kernel Learning by Group Lasso." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 1175–82.

Yin, Weiwei, Jessica C Kissinger, Alberto Moreno, Mary R Galinski, and Mark P Styczynski. 2015. "From Genome-Scale Data to Models of Infectious Disease : A Bayesian Network-Based Strategy to Drive Model Development." *Mathematical Biosciences* 0. Elsevier Ltd.: 1–13. doi:10.1016/j.mbs.2015.06.006.

Zhu, Xiaojin. 2008. "Semi-Supervised Learning Literature Survey Contents."

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic

Net." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67 (2): 301–20. doi:10.1111/j.1467-9868.2005.00503.x.

Appendix

Appendix A

Applying Data Science in Malaria Research: From Infections and Disease, to Knowledge and Insights (Chien *et al*. submitted)

Appendix B

High-Quality Genome Assembly and Annotation for *Plasmodium coatneyi*, Generated Using Single-Molecule Real-Time PacBio Technology - Genome Announcement 2016

Chien's major contributions are below:

1. *Plasmodium* genome assembly and evaluation

- Algorithm evaluation
- Performing HGAP assembly, A/B test
- Performing QC
- Performing BLAST genome global alignment and evaluation
- Performing Progressive Mauve genome comparison
- Assembly gap identification
- Organelle genome assembly and evaluation
- Assembled several species including *P. coatneyi, P. knowlesi * 3, P. cynomolgi *2, P. vivax *1*

2. Building PacBio sequencing technology infrastructure on AWS and cluster computer

- SMRT analysis system installation
- Systems dependency verification

3. Performing genome annotation

- Performing RATT draft annotation

- Implementing MAKER2 genome annotation

- MAKER2 parameter tuning, annotation model training

- Involving the annotation validation and workflow designing

4. Manuscript writing, revising and submission

This work is published on Genome Announc. September/October 2016 vol. 4 no. 5 e00883-16

Appendix C

PacBio assembly of a *Plasmodium knowlesi* genome sequence with Hi-C correction and manual annotation of the *SICAvar* gene family – Lapp, Geraldo, Chien *et al*. Parasitology, Special Issue on *Plasmodium knowlesi* (*in press)*

Chien’s major contributions are below:

1. *Plasmodium* genome assembly and evaluation

- Algorithm evaluation

- Performing HGAP assembly, A/B test

- Performing QC

- Performing BLAST genome global alignment and evaluation

- Performing Progressive Mauve genome comparison

- Assembly gap identification

- Organelle genome assembly and evaluation

2. Manuscript writing and revising (in press, co-first author)

Appendix D

Multi-omic profiles differentiate acute and chronic phases of malaria in *Plasmodium coatneyi* infected rhesus macaques (Joyner et al. nanuscript in preparation)

Chien’s major contribution are below:

1. RNA-seq bioinformatics analysis pipeline construction

- RNAseq alignment

- Reads alignment counting

- Algorithm selection and evaluation

2. Differential genes expression Analysis

- Implementing several R packages for analysis includes DEseq2 and others

3. Non-supervised clustering analysis

- Performing PCA, k-means, SOM, hierarchical and Bayesian clustering methods for different tasks.

4. Differential gene sets enrichment analysis and pathway identification

- Statistical analysis

- Performing pair-wised DEGs comparisons

- RNA-seq DEGs heat-map construction

- Performing IPA, DAVID and metacore for pathway analysis and identification

- Literature Research

5. Manuscript writing and the manuscript is in preparation