The Dissertation Committee for Fanping Kong Certifies that this is the approved version of the following dissertation:

Translating Biomedical Research Data to Knowledge Through Bioinformatics

Committee:

Bruce A. Luxon, Ph.D., Supervisor

Werner Braun, Ph.D., Chair

Peter C. Melby, MD

Allan R. Brasier, MD

Heidi M. Spratt, Ph.D.

David G. Gorentein, Ph.D.

_____

Dean, Graduate School

TRANSLATING BIOMEDICAL RESEARCH DATA TO KNOWLEDGE THROUGH BIOINFORMATICS

by

Fanping Kong, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas Medical Branch

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas Medical Branch

July 2015

Dedication

To Kaijian

Acknowledgements

This work would not have been possible without my mentors, Drs. Bruce A. Luxon and Heidi M. Spratt to whom I am eternally grateful. I also like to thank Drs. Melby and Braiser for leading me into their amazing biological research. Thank you, Drs. Braun and Gorenstein, for setting excellent role models for me. I have learned that there's a lot more to science than just performing experiments.

Additionally, I would like to thank Drs. Peter Melby, Thomas Green and Naomi Forester for believing in me and giving me more than one chance at success. It was very amazing experience to work closely with their groups. I also would like to thank Dr. Omar Saldarriaga, from whom I learned how to interpret many biological results and as a friend. I would also like to thank Dr. Wei Wang for his support and assistance. I wish to collectively thank all of the laboratory and research staff who have helped me throughout theses years.

I wish to thank the Jeane B. Kempner committee for granting me a one-year pre-doctoral fellowship in pursuing my Ph.D.

Finally, I would like to thank my spouse Kaijian Liu, and my parents, Xiurong Liang and Qingbao Kong, who are my inspiration.

Translating Biomedical Research Data to Knowledge Through Bioinformatics

Publication No._____

Fanping Kong, Ph.D

The University of Texas Medical Branch, 2015

Supervisor: Bruce A. Luxon

In this dissertation, we present the application of biomedical informatics, including data mining and knowledge discovery, to extract knowledge from genomics, transcriptomics, and clinical data. To improve our understanding of biomedical problems such as infectious disease and addiction research, we have successfully applied next generation sequencing (NGS) techniques to generate both transcriptomics and genomics data. First, through the investigation of transcriptomics data on rats in environmentally enriched and isolated conditions, we concluded that the pathways of *retinoic acid receptor activation*, *eukaryotic initiation factor 2 signaling*, and *protein ubiquitination* play significant roles in addictive behavior and thus direct our focus on individual differences in susceptibility to addiction. Second, through mining Syrian golden hamster transcriptomics data during visceral leishmaniasis, we discovered that splenic macrophages experienced mixed classic and alternative polarization/activation and whole spleen tissue experienced massive inflammatory response during visceral leishmaniasis. We proposed several mechanisms to understand the pathogenesis of L. *donovani*.

Additionally, we investigated genomics data of several Venezuelan equine encephalitis virus mutants and showed that the replication fidelity of Tc-83 can be increased by incorporating point mutations at the RNA-dependent RNA polymerase region. These findings should accelerate the development of a new live attenuated vaccine for Venezuelan equine encephalitis virus.

In addition to the NGS data mining and knowledge discovery, we developed a novel ensemble data analysis method to improve the predictive ability of classic bagging and AdaBoost methods. By evaluating forty-one online datasets, we demonstrate the ability of our ensemble method in increasing predictive accuracy, which could be particularly useful for identifying novel diagnostic biomarker panels. Furthermore, we assessed two different intervention strategies for schistosomiasis using the meta-analysis method and demonstrated that implementation of the new integrated strategy reduces the infection risk by ~3–4 times compared to the conventional strategy. This approach is applicable to evaluate any new prevention, diagnosis, or treatment strategy.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Illustrations

List of Abbreviations

AUC	Area Under ROC Curve
cDNA	Complementary DNA
CHO	Chinese Hamster Ovary
CPM	Count Per Million Mapped Reads
CRT	Cyclic Reversible Termination
DEG	Differentially Expressed Gene
DIKW	Data, Information, Knowledge, Wisdom
DPM	Dirichlet Process Mixture
EC	Enriched Condition
FDR	False Discover Rate
FET	Fisher's Exact Test
GSBS	Graduate School of Biomedical Science
GSEA	Gene Set Enrichment Analysis
hg38	GRCh38 Version Release 77
IC	Isolated Condition
IGV	Integrative Genomics Viewer
IPA	QIAGEN's Ingenuity® Pathway Analysis
LRT	Likelihood Ratio Test
MDS	Multidimensional Scaling
NEC	Non-Endemic Control
NGS	Next Generation Sequencing

PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
QA/QC	Quality Assurance and Quality Control
RNA-Seq	RNA Sequencing
ROC	Receiver Operating Characteristic
RR	Relative Risk
SBL	Sequencing by Ligation
SMRT	Single Molecule Real-Time
SNA	Single-Nucleotide Addition
TACC	Texas Advanced Computing Center
TDC	Thesis and Dissertation Coordinator
TMM	Trimmed Mean of M-values
UTMB	University of Texas Medical Branch
VEEV	Venezuelan Equine Encephalitis Virus
VL	Visceral Leishmaniasis

Chapter 1. Biomedical Informatics and Knowledge Discovery

Biomedical informatics is an interdisciplinary field that focuses on how to effectively use biomedical data to improve human health. It integrates strategies from computer science and the quantitative disciplines (e.g. statistics, data science, decision science, etc.) to solve challenging problems in the biological and medical sciences.

The data mining and knowledge discovery process in biomedical informatics follows the data, information, knowledge, wisdom (DIKW) hierarchy. This hierarchy represents the relationships among data, information, knowledge and wisdom (Figure 1).

Data are quantitative and qualitative descriptions of events. For example, the data can be 40 million 50 base pair nucleotide reads from RNA sequencing (RNA-Seq) or they can be the gravities of an illness. These data are just symbols with no significant meaning by themselves.

To obtain information, we endow the data with meanings by including relational connections through correction, categorization, calculation, and so on. In our case, the 40 million small sequences can be aligned to an appropriate reference genome to obtain the expression level of each gene. We can also classify samples into experimental or control groups. These data processes turn the data into information, offering us better understanding of the current state.

Knowledge is meaningful information that can be used for guiding and inferring. Typically, information is massive. For example, we can simultaneously evaluate the expression of thousands of genes in one RNA-Seq experiment. Not all genes are equally important. Some of them may closely relate to the pathogen stimuli, while the others may not. Diseases are more likely to be relevant to genes differentially represented between

two conditions. Knowledge discovery is the procedure to identify and integrate of key and useful information.

Wisdom (i.e. understanding) is a collective application of knowledge. It is the hardest to achieve but the best way to direct further research and make predictions. Once we have enough knowledge about a disease, we can help the patients with better prevention, earlier diagnosis and timely treatment. The ultimate objective for biomedical informatics and translational research is to enhance the patient care.

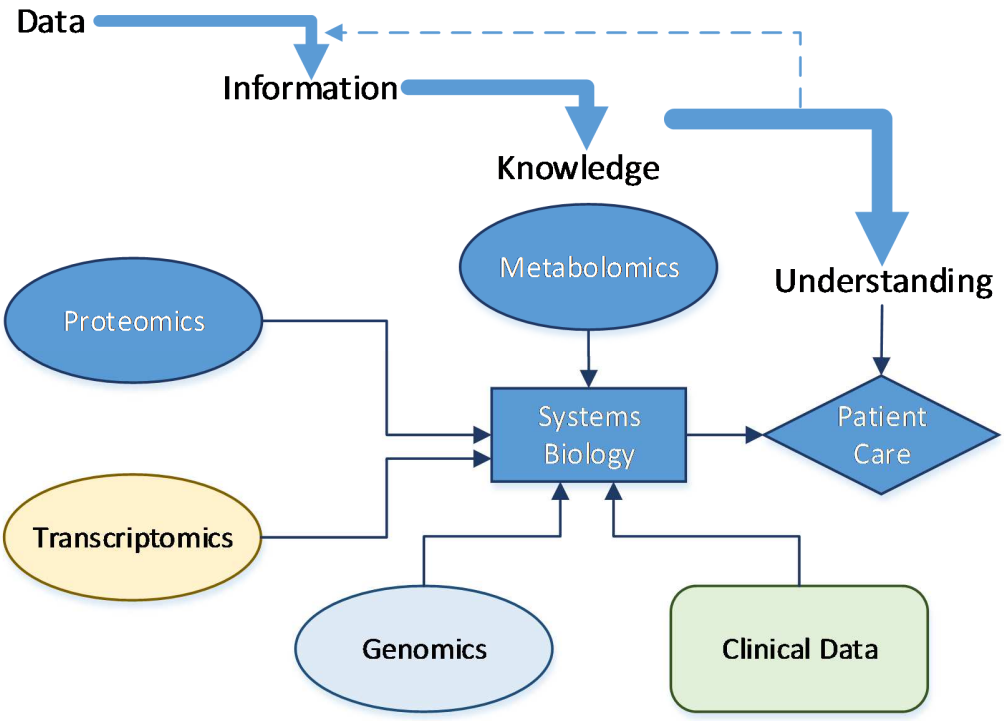

Figure 1.1 DIKW and systems biology in biomedical informatics (Adapted from Starkey JM, et al., J Clin Trans Res, 2012)

Data analysis in biomedical informatics follows the DIKW hierarchy. Its central study object is systems biology. Systems biology is a scientific approach that focuses on

complex interactions and/or processes within a biological system. The goal of systems biology is to identify emergent properties. For any given system, the systems biology combines analyses from a variety of sources including clinical response and biomolecules on a variety of levels including DNA, RNA, protein, metabolites, etc. (Figure 1). A study in systems biology fashion links all of these different levels of molecules and sub-systems into networks, describes the over-all function of the study model, enables novel understanding, and, finally, benefits patient care.

In this entire dissertation, we will address three different perspectives of the systems biology: genomics at the DNA/RNA level, transcriptomics at the mRNA level, and clinical data.

1.1 GENOMICS DATA IN BIOMEDICAL RESEARCH

Genomics describes research associated with the genome of any specific model. The investigated biological system is highly complex, and its mechanism cannot be interpreted as the effect of one individual gene [1, 2]. The term, genome, was first coined by the German botanist Dr. Hans Winkler in 1920 [3]. The genome of an organism is its genetic material encoded in DNA or RNA, which encodes all mRNA/proteins, tRNA, rRNA, siRNA, miRNA, etc. inside the specific creature. Both “genome” and “genomics” are derived from the Greek word “genesis”, with the meaning of "birth" or “origin”. As the names suggest, genomics data are typically related to the discovery-based genome-wide researches on complex biological systems at the fundamental DNA/RNA level.

The first DNA sequencing techniques were developed in the early 1970s [4-6], approximately two decades after the establishment of the DNA double helix structure in 1953 by James Watson and Francis Crick [7]. These sequencing strategies went through a

huge revolution in 1977 when Frederick Sanger and his colleagues developed both the Sanger sequencing and the shotgun sequencing methods [6]. In Sanger sequencing, DNA fragments with different lengths were radioactively or fluorescently labeled and then separated using gel electrophoresis. This method is appropriate for fairly short sequences (100 ~ 1000 base pairs). The shotgun sequencing technique is applied to the longer sequences, which are first randomly subdivided into smaller fragments, sequenced, and then re-assembled to obtain the overall sequence. Sanger Sequencing had been dominant and most widely used for approximately 25 years until 2005, when next-generation sequencing (NGS) technologies became commercially available [8]. NGS technologies conceptually adopt the Sanger method, but automate it to offer cost-effective sequencing with dramatic increases in speed. Presently, NGS technologies are the most powerful and efficient methods to obtain genomics data and have been gradually replacing the traditional Sanger methods in most genomic research studies.

1.1 TRANSCRIPTOMICS DATA IN BIOMEDICAL RESEARCH

Transcriptomics investigates the large scale and comprehensive transcriptome in an organism or a cellular system. As initial products of genetic material, the transcriptome include mRNA, rRNA, tRNA, and other non-coding RNA. mRNAs are the transcripts of protein-coding genes that can be translated into proteins. The other RNAs are non-coding RNAs, which cannot be translated into protein but regulate the synthesis of proteins.

The transcriptome for a given cell or cell population varies with external environmental conditions. This differs from genomics data, which is fixed for a given organism. The expression levels of mRNA transcripts approximately reflect the proteins

that are being actively expressed in the system at a given time. The exceptions include mRNA degradations and posttranslational modifications. The objective of the majority of transcriptome profiling is to examine the variation of mRNA expression under different experimental conditions and to further identify the altered mechanisms.

To obtain a profiling of the transcriptome, we usually convert RNA into complementary DNA (cDNA) using the enzyme reverse transcriptase. cDNA is a double-strand DNA. All of the sequencing techniques we discussed above for the genomics data are equally applicable to the sequencing of cDNA. Therefore, NGS technology is widely used to study the transcriptomics.

1.3 DATA PATTERN RECOGNITION IN BIOMEDICAL RESEARCH

Data pattern recognition and interpretation is a promising field in biomedical research when the data include a significantly large number of events. The previously unknown but interesting patterns represent novel information in a dataset. Our understanding of this information may shed light on future patient care. For instance, a biomarker panel can be obtained from the data pattern that distinguished diseased subjects from health controls. The biomarker panel can direct further biomedical research or the disease diagnosis in the clinic.

In this report, we will present a novel classification tool. We mainly focus on the classification algorithms in machine learning, particularly in pursuit of ensemble methods. Classification refers to algorithmic procedures that assign objects into groups [9, 10]. The product of a classification is a classifier, also known as a model, which can map input data to classes. The diagnostic problem in clinical research is a classification event that determines which patients will be assigned to the disease group given a set of

symptoms as the input data. We emphasize the use of ensemble methods for model construction. As a result, we developed a Dirichlet Process Mixture (DPM) integrated ensemble method to enhance the accuracy of classification.

1.4 CLINICAL DATA IN BIOMEDICAL RESEARCH

The study of clinical data using systems biology methods can also benefit patient care. The genomics and transcriptomics research discussed previously, together with other proteomics and metabolomics research, accelerates the translation from biological bench research to clinical usage. When a novel approach of diagnostics, treatment, or prevention emerges, it becomes imperative to compare it with the existing methods. We will exploit the meta-analysis to assess two different intervention strategies based upon the clinical data.

Meta-analysis summarizes findings from independent but similar studies to seek patterns of agreement or disagreement among them [11-13]. As a popular approach for systematic review, meta-analysis provides a quantitative assessment on the effects of interest by combining many available and pertinent data. This result is more precise than any individual constitutive study. This pooled analysis method has become increasingly more popular for evaluating the clinical effectiveness of health care interventions [14-19].

This dissertation covers five different projects, spanning four different fields of the biomedical research in transcriptomics, genomics, data pattern recognition, and clinical data. Since our transcriptomics and genomics data were generated by the Illumina HiSeq 1000 sequencer, we will first present in Chapter 2 a brief introduction of this state-of-the-art NGS technology, especially the Illumina HiSeq platform. After that, two

transcriptomics projects will be discussed in Chapters 3 and 4. In the first project, we investigated an enriched environment introduced protective addiction phenotype against cocaine in rats. We will describe the detailed data validation, analysis, and interpretation in this project to illustrate the procedure of data mining and knowledge discovery. In the second transcriptomics project, we worked to understand the pathogenesis mechanisms of visceral leishmaniasis. This project used Syrian golden hamster as our animal model. The third project, described in Chapter 5, is a genomics project. It aims to develop vaccine candidates against the Venezuelan equine encephalitis virus (VEEV) infection through studying the intra-host variation of VEEV in African Green monkey kidney cells and mouse model. The data analysis in this project shares some common algorithms as in the novel ensemble method discussed in Chapter 6, which is the forth project. In chapter 6, we developed and evaluated a novel Dirichlet Process Mixture (DPM) integrated ensemble method for classification. In our last project, discussed in Chapter 7, we assessed two intervention strategies for schistosomiasis using meta-analysis. In Chapter 8, we close this dissertation with a short conclusion, which readdresses the role of biomedical informatics in knowledge discovery.

Chapter 2. Next Generation Sequencing

Next Generation Sequencing (NGS), different from capillary electrophoresis-based Sanger sequencing, are a family of automated, fast, and parallel DNA sequencing technologies. The fundamental principles of NGS sequencing were developed since middle 1990s [20-23]. NGS technologies experienced many evolutions until 2005 when GS20 from Roche 454 Life Sciences became the first commercially available NGS sequencer [8]. Since then, more than 30 NGS platforms have been developed. These state-of-the-art technologies have been generating an unprecedented wealth of data with high resolution, which has revolutionized the genomics and transcriptomics research [24-27]. The size of data produced from NGS sequencers has increased by at least two-fold every year since it was invented, which outpaces Moore's law. Meanwhile, the cost of sequencing has been dropping faster than the Moore's law (Figure 2.1).

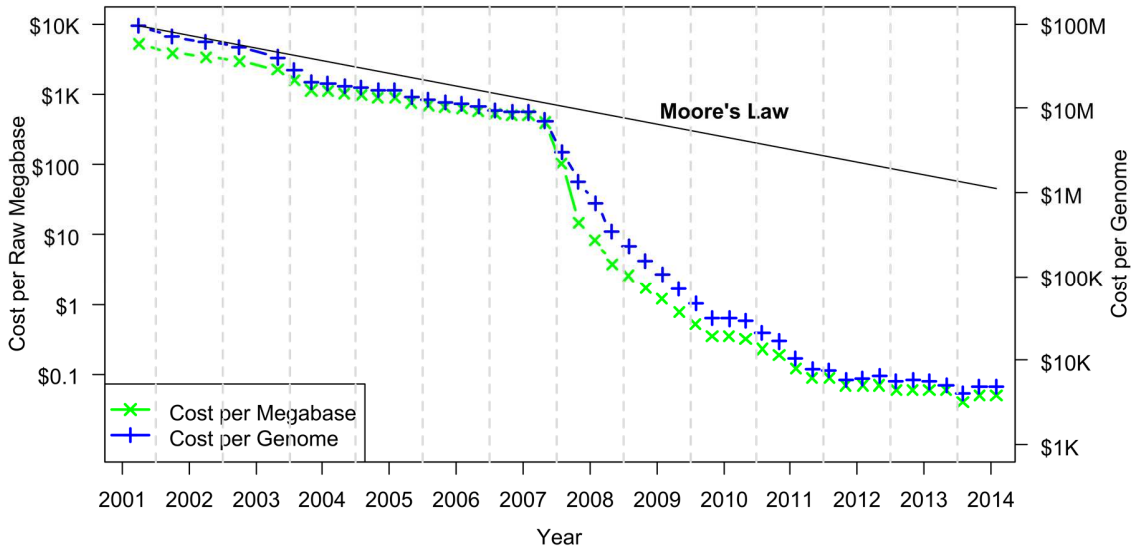

Figure 2.1 Sequencing cost 2001-2014 (Adapted from National Human Genome Research Institute website [28])

In this chapter, we will first discuss the general principles and mechanisms of most in-market NGS sequencers. After that, we will focus on the Illumina HiSeq platform, from which we generated our transcriptomics and genomics data for the following three chapters. At the end of this chapter, we will summarize the applications of NGS in biomedical research.

2.1 NGS MECHANISMS

All NGS protocols from mRNA/DNA samples to NGS data have two essential steps: template preparation and sequencing. Each step has several different methods. The unique combination of methods from each step distinguishes one NGS platform from the other and determines its associated sequencing error modes. We will summarize these methods as they apply to various NGS platforms in appendix A.

2.1.1 NGS template preparation

Template preparation is the first step of NGS protocols. The goal of this step is to create non-biased and representative nucleic acid material from the investigated genome or transcriptome. There are primarily two existing templates for NGS reaction, the clonal amplification template and the single molecule template.

Both template preparation strategies begin with randomly shearing the large genomic DNA or cDNA into small fragments. Different platforms prefer different fragment sizes: less than 600 bp for the Illumina, SOLiD and Ion Torrent platforms, less than 1kb for Roche 454 and less than 20kb for PacBio [29]. The fragmentation strategy also varies among platforms but mainly falls into three categories: physical (e.g. acoustic,

sonication, hydrodynamic shear), enzymatic (e.g. endonuclease, transposase) and chemical (e.g. heat and divalent metal action) fragmentation.

The next step in template preparation is to ligate the fragments with adapters – common or universal nucleic acids – to form the sequencing library. The clonally amplified templates require both fragment amplification and attachment / immobilization, while the single molecule templates only need to attach/immobilize the fragments to a solid surface or support. Most NGS platforms (e.g. Illumina, Roche 454, SOLiD and Ion Torrent) with an imaging system favor the clonally amplified templates to enhance the fluorescence/light signal. The single molecule sequencers are utilized in the PacBio platform and the Oxford nanopore platform. These various approaches between the platforms have determined that the clonally amplified templates need about 3-20μg of starting genomic DNA material, while 1μg is enough for the single molecule templates. Moreover, the platforms using clonal templates are more likely to have substitution errors due to the mutations in amplification.

2.1.2 NGS sequencing mechanisms

All NGS instruments follow the sequencing-by-synthesis principle [30] to sequence a template, which enables massive simultaneous sequencing. Several different mechanisms have been deployed in the widely used NGS platforms including cyclic reversible termination (CRT), single-nucleotide addition (SNA), sequencing by ligation (SBL), single molecule real-time (SMRT) sequencing, semiconductor sequencing and nanopore strand sequencing. We will briefly describe each mechanism here.

CRT utilizes reversible 3'-blocked or 3'-unblocked terminators that terminate DNA synthesis after incorporating one nucleotide. In each cycle, the sequencing

comprises three steps: nucleotide incorporation, fluorescence imaging and cleavage [31]. This method is used by the Illumina platforms, which have been recently improved from the 4-dye sequencing chemistry to a 2-dye system. This improvement halves the number of images and thus increases the sequencing efficiency [32].

The SNA method, also known as the pyrosequencing method, detects the pyrophosphate (PPi) released when a deoxynucleoside triphosphate incorporates into the sequencing template. The PPi is converted to ATP, which then converts luciferin to oxyluciferin and emits light. A high-resolution charge-coupled device (CCD) detects the light in order to monitor the sequencing process. This method has been implemented in the Roche 454 platforms [33].

SBL, another cyclic method, incorporates additional nucleotides using DNA ligases rather than polymerases as used in the CRT and SNA methods. Each cycle is comprised of three steps including ligation of a fluorescent-labeled probe, imaging the fluorescence, then cleavage in preparation for the next run. This method is implemented in the SOLiD platforms, which utilizes four fluorescently labeled di-base probes with five nucleotides added in each cycle, such that each sequencing cycle demands five rounds of primer reset.

Semiconductor sequencing is used in the Ion Torrent platforms, which utilize a semiconductor sensor, as opposed to the optics-based technology discussed above. A hydrogen ion (H^+) is released each time a nucleotide is incorporated, resulting in a local pH change. Ion Torrent monitors the sequencing process with a semiconductor sensor to detect and convert the pH chemical information to digital information [34]. The four nucleotides are added and washed off sequentially. A peak indicates that the

corresponding nucleotide matches the template and its amplitude indicates the number of matched nucleotides. Since this approach has no modified nucleotides or optical imaging, it is a relatively faster and lower cost method compared to the previous methods.

SMRT sequencing, implemented by PacBio NGS sequencers, utilizes the zero-mode waveguide (ZMW) nanophotonic structure [35]. In this method, the sequencing template and polymerase complexes are immobilized at the bottom of the ZMW. As discussed in the CRT and SNA methods, a fluorescent-labeled nucleotide is incorporated into the template thereby releasing fluorescence. The observation volume provided by the ZMW is small enough to reduce the diffusing background fluorescence and the fluctuation of the signal fluorescence so that we can sufficiently distinguish the fluorescent signal from the background noise to determine the incorporated nucleotide [36]. This method requires less genomic starting material and returns extremely long and accurate reads so that it is ideal for *de novo* sequencing.

Nanopore strand sequencing as implemented by Oxford Nanopore Technologies is another single molecule sequencing method. A protein nanopore set is an electrically resistant membrane bilayer with a DNA enzyme complex attached. The enzyme can guide and ratchet DNA into an intact DNA polymer to pass through the nanopore, which is so small that only one nucleotide could enter at a time. Each nucleotide obstructs the nanopore with a unique modulation, leading to a characteristic change in current. An electrically sensitive sensor then records the disruption to distinguish between the four nucleotides and even some modified bases. As a result, the individual DNA bases are identified as the DNA molecule passes through the system. This method requires no

modification of the nucleotides, thus has no deterioration in accuracy when long DNA strands are sequenced [37].

2.2 ILLUMINA HISEQ SYSTEM

The Illumina HiSeq sequencing system utilizes clonal amplification templates and CRT sequencing mechanisms. Specifically, this platform adopts the reversible terminator-based sequencing-by-synthesis (SBS) chemistry. The resulting sequences can be either single-end or paired-end. The sequencers are able to generate a large number of reads in one run so they are suitable to analyze large animal or plant genomes. The Illumina HiSeq 1000 platform has one flow cell for sequencing but it can be upgraded to the HiSeq 1500 by incorporating an additional rapid run mode or to the HiSeq 2000 by adding another flow cell.

We illustrate the sequencing procedures of Illumina HiSeq 1000 and other HiSeq platforms in Figure 2.2. The genomic DNA (or cDNA) of interest are first randomly sheared into segments and then ligated with adapters at both ends to build the library (Figure 2.2A). In addition to the adapters, we can link the library with an index, a unique identifier sequence that distinguishes it from other libraries pooled in the same lane of a flow cell. The reads are further immobilized on the surface of a flow cell (Figure 2.2B). Each read is then isothermally amplified using solid-phase bridge amplification to form clusters. Each cluster can contain up to 1,000 identical copies of the original nucleic acid chain (Figure 2.2C-F). There are tens of millions such clusters at each square centimeter of the surface of a flow cells. All of these clusters will then be processed in parallel. The four deoxynucleoside triphosphates are fluorescently-labeled with different dies. During

each sequencing cycle, we add all four different deoxynucleoside triphosphates, but only one deoxynucleoside triphosphates can be incorporated into the nucleic acid chain and then terminate the polymerization (Figure 2.2G). We identify the nucleotide base for each cluster by imaging the fluorescent dye in situ. To enable the detection of the next nucleotide, we enzymatically cleave the dye and then repeat the whole process again (Figure 2.2H). Eventually, we can determine the order of bases in a read to generate the sequence (Figure 2.2I). Billions of sequencing reads can be generated in one run using the Illumina HiSeq 1000 sequencer; therefore, we utilized this sequencer for the projects discussed in Chapters 3, 4 and 5.

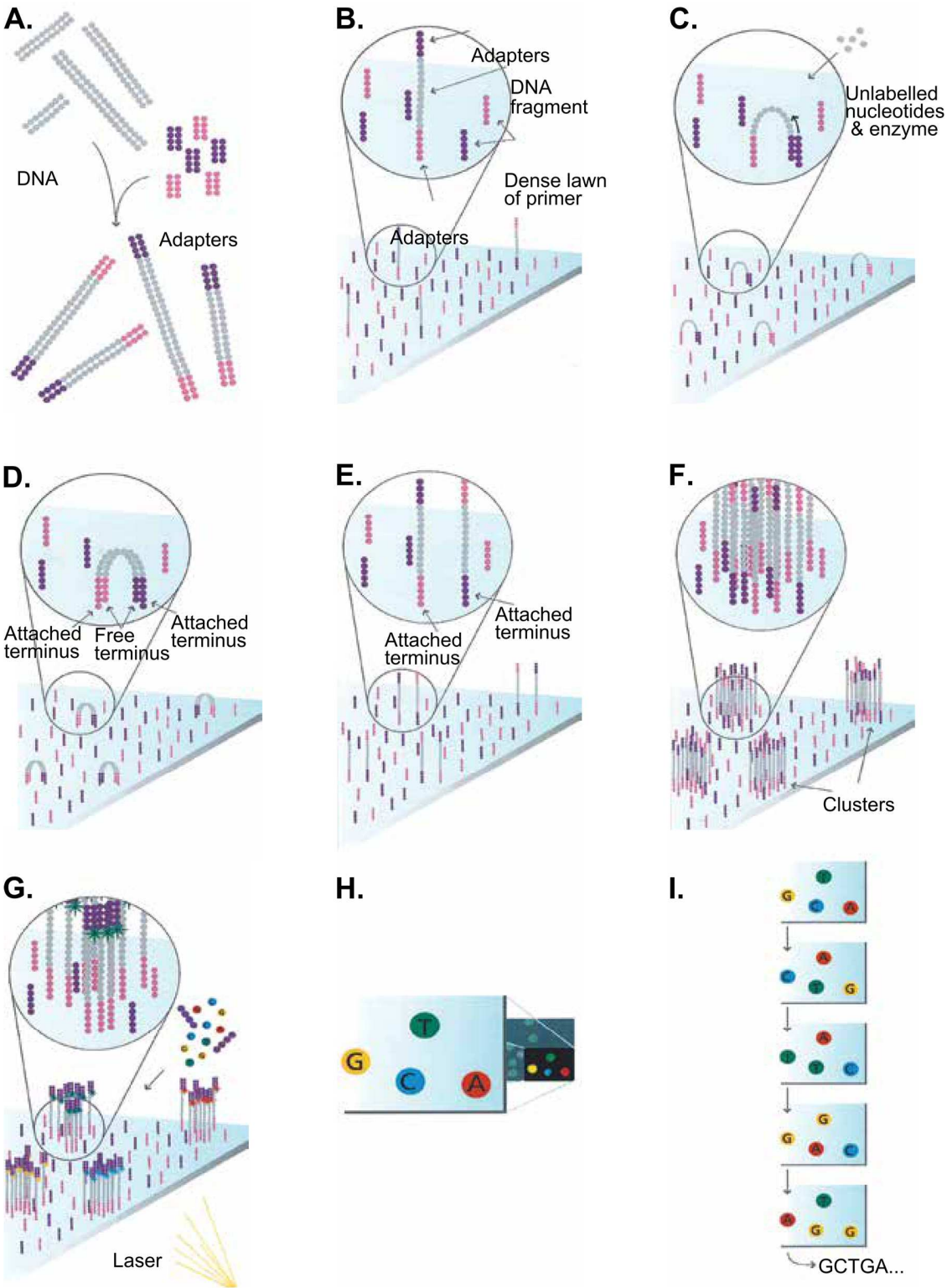

Figure 2.2 Illumina sequencing overview (Adapted from Illumina Webpage [38])

2.3 APPLICATIONS OF NGS TECHNOLOGY

NGS platforms can generate a large amount of low-cost but high quality reads within a short time, making them useful in many applications. These include but are not limited to 1) transcriptome profiling to analyze gene expression levels in cells, tissues and organisms (generally known as RNA-Seq) [39]; 2) creating *de novo* transcriptome assemblies without a reference genome; 3) discovering variants by sequencing the target region of interest or the whole genome; 4) genome-wide profiling of proteins, especially the binding sites of transcription factors based upon chromatin immunoprecipitation sequencing (e.g. ChIP-Seq); 5) studying genome-wide DNA methylation (methyl-Seq) and DNase I hypersensitive sites (DNase-Seq); and 6) inventorying co-exiting organisms (Metagenomics).

With so many applications, NGS technologies are experiencing rapid development. Moreover, competitive pricing has revolutionized the applied markets such as biomarker discovery, disease diagnostics, drug discovery, agriculture and animal research, and personalized medicine. Biomedical informatics plays, and will continue to play, a key role in all of these processes in order to glean valuable information from the raw short sequencing. However, the sheer scale of the data remains a significant challenge for data analysis.

Chapter 3. Enriched Environment Induced Protective Phenotype [1]

3.1 INTRODUCTION

Humans display vast individual differences in susceptibility to drug addiction. Some individuals become addicted after only a single exposure while others remain resistant to addiction even after many exposures to high doses of a drug. Understanding the mechanisms of the resistance to addiction will provide new targets for the treatment and even the prevention of addiction. As in previous research [40-42], rats provided the animal model to study the environmental enrichment paradigm for addiction resistance.

Currently there are no Food and Drug Administration (FDA) approved pharmacotherapeutics for cocaine addiction in the United States, in spite of decades of targeted studies of known pharmaceuticals. For this reason, finding and identifying completely novel targets for pharmacotherapeutic and genetic intervention is paramount for developing successful treatments for addiction.

3.2 OBJECTIVES AND EXPERIMENTAL DESIGN

The goal of this project is to determine intrinsic differences between environmentally enriched conditions (EC) and isolated conditions (IC) on the brain's response to cocaine administration. Our ultimate objective is to identify completely novel targets for addiction prevention or treatment. We focus on the biological significance of environmental enrichment and explore the molecular mechanism of this protective phenotype.

[1] In collaboration with Dr. Thomas A. Green's group at UTMB.

In the environmental enrichment paradigm, rats were randomly assigned to either an EC or IC group. The EC rats had daily exposure to novelty (children's plastic toys), exercise, and social contact with conspecifics (i.e. group housing). Meanwhile, the IC rats were housed singly with no exposure to novelty. In each condition, we provided 8 rats with access to self-administrated cocaine and another 7 rats with access to saline only as a control. Thus, the environmental enrichment paradigm experiment consisted of four groups: 7 rats for IC saline, 8 rats for IC cocaine, 7 rats for EC saline, and 8 rats for EC cocaine. All rats were first trained to use the self-administration system and then harvested after 14 days of cocaine/saline self-administration when their response to cocaine was stabilized [42]. For each rat, we sequenced the left nucleus accumbens mRNA samples using an Illumina HiSeq 1000 sequencer to generate 50bp paired-end reads. The protein extraction from the right nucleus accumbens was investigated using liquid chromatography mass spectrometry. Please refer to Litchi C.F. [42] and the manuscript (Y.F. Zhang, et. al., not shown) for detailed information about the experimental design.

3.3 RNA-SEQ DIFFERENTIAL EXPRESSION ANALYSIS

RNA-Seq, short for RNA-Sequencing, is the technology applied to quantitatively characterize genes at the transcription level by measuring the differential expression of messenger RNA (mRNA) [39]. The work flow for RNA-Seq differential analysis typically begins with quality assurance and quality control (QA/QC) to assess the quality of the raw sequence reads. Sequencing reads with acceptable quality are then mapped to a reference sequence. The numbers of reads that map to specific features – genes, transcripts, or exons – are counted to measure their expression levels. The work flow

ends with statistical analysis of the read counts to identify differentially expressed features. The whole process is consistent with the DIKW hierarchy as discussed in Chapter 1.

3.3.1 RNA-Seq NGS data

The mRNA of the left nucleus accumbens from each rat was first converted to cDNA, sheared into segments with length of about 270 base pairs, then sequenced using an Illumina HiSeq 1000 sequencer according to the manufacturer's directions. In each lane, we pooled four samples, which are distinguished by different index sequences. The outputs were paired-end sequences. Thus, each segment had been sequenced from both ends to produce two sequencing reads. The sequencer records all nucleotide information using massive image files. A program CASAVA converts the image files into text-enriched file in fastq format. Please refer to session 2.2 for details about the sequencing procedure.

A fastq file includes millions of sequencing reads. Each read has four lines: identifier, read sequence, connector, and coded Phred score (Box 1). As an example, the identifier "@UT344:39:C0WBUACXX:2:1101:2495:2088 1:N:0:TGACCA" indicates that the read comes from a sequencer named UT344 in run number 39. Identifiers are usually shared by all the reads within the same fastq file. Additionally, this specific read comes from the position (2495, 2088) of the 1101st tile of the 2nd lane of a flow cell named C0WBUACXX. "1:N:0:TGACCA" means that it is the forward read (1) in a paired-end sequencing effort, without filter (N) or control bits (0) on and with index sequence TGACCA. The length of the reads varies depending on the sequencing parameter as well as the sequencing platforms. In the current project, all reads have 50

base pairs. The Illumina HiSeq 1000 is a 4-dye system with four different fluorescent colors for the four standard nucleotide bases. For each base, four images will be recorded to show the intensity of each color. The Phred score directly relates the ratio of the dominant color intensity to the other three, more noisy colors. More specifically, the quality Phred score is logarithmically related to the base-calling accuracy or error probabilities. The Phred scores are then further encoded as ASCII characters by adding 33. This gives the relationship: $Phred\ Score = -10 * \log_{10}(error\ rate) = ASCII\ (Coded\ Phred\ Score) - 33$. For example, the first base of the 4-line read in Box 1 below has the quality value character "C", which corresponds to the Coded Phred Score ASCII value 67, so its Phred score is $67 - 33 = 34$ thus its error rate is $10^{-34/10} \approx 0.0004 < 0.001$. One RNA-Seq fastq file contains millions of such 4-line reads to profile the investigated transcriptome.

@UT344:39:C0WBUACXX:2:1101:2495:2088 1:N:0:TGACCA	⇨ Identifier
CCNGGAGCGGAACCACAGTCCTGTCCAGGTGGAGGCAGATGAGCACCTAT	⇨ Read Sequence
+	⇨ Connector
CC#4ADDBFHDHGJIJIJHIGHFHGIJJJBGH?GEGFGCFHGIIJJIJIJ	⇨ Coded Phred Score

Box 1. Example of one 50bp read from Illumina HiSeq 1000

3.3.2 RNA-Seq pre-analysis

All NGS data analyses including RNA-Seq begin with data quality assurance and quality control (QA/QC). This essential step aims to detect any systemic errors before or during sequencing. Some platforms support internal quality and calibration control with spiked-in RNA from a small, diverse and well-defined genome, such as PhiX virus. Meanwhile, the raw sequencing reads from all platforms can be evaluated based upon

their inherent features. Currently, several NGS data QA/QC statistical analysis tools are publicly available including FastQC, HTQC, FaQCs, FASTX-Toolkit, NGS QC Toolkit, QC-Chain and so on.

QA/QC programs generally assess raw NGS data quality from three perspectives: the representativeness of the input samples, the performance of the sequencer and the reliability of the sequenced reads. If the input sample is randomly selected from one organism, the GC content per sequence, theoretically, follows a normal distribution. A significant contamination or amplification bias, however, may introduce secondary peaks into the main distribution and alter its shape. The duplication level (i.e. the frequency of duplicated reads) can be used to evaluate the diversity of the input sample. The sequencing qualities in tile level, the output sequence length, N content percentage and so on are assessed. Read sequences with a higher Phred score are more reliable. We may also want to assess other properties, such as overrepresented sequences, adapter content, and k-mer content. Adapters, low quality reads and nucleotides can be filtered out as needed based upon customized criteria.

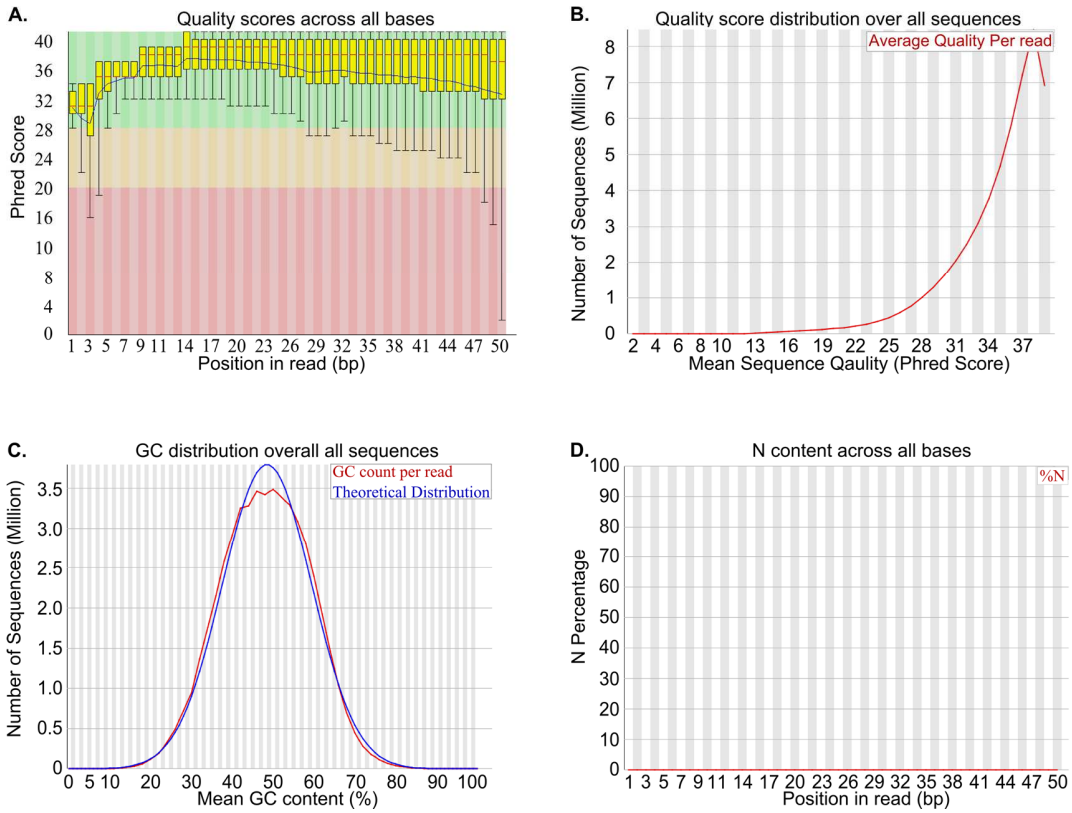

Figure 3.1 FastQC output results. Representative output when we checked quality of the reads for samples in the current projects using fastQC. Here we only show the plots of quality scores across all bases (A), quality score distribution over all sequences (B), GC distribution overall all sequence (C) and N content across all bases (D).

In this project, no internal control had been incorporated so we assessed the sequencing output quality using the software FastQC (version 0.9.1). All the raw NGS data in this project had high Phred scores (Figure 3.1A, B), low N percentages (Figure 3.1D) and an overall normal GC sequence distribution (Figure 3.1C). Thus, the overall sequencing quality was high. We then aligned the reads against the Rattus norvegicus reference genome (version 3.4) using the Tophat (version 2.0.6) and Bowtie (version 2.0.2) software without any filtering or trimming. For all samples, 93.24±2.63% of the raw reads successfully mapped to the rat genome using the default parameters. We

further examined the alignment depth using the Integrative Genomics Viewer (IGV) visualization software. An alternative approach would be to visualize the results using genomic browser or to directly calculate the alignment depth along the regions of interest. As an example, we visualized the expression of the early growth response 4 gene (Egr4) in four different conditions (Figure 3.2A). The results clearly indicate that Egr4 was over expressed in the IC brain then further increased with cocaine stimulus.

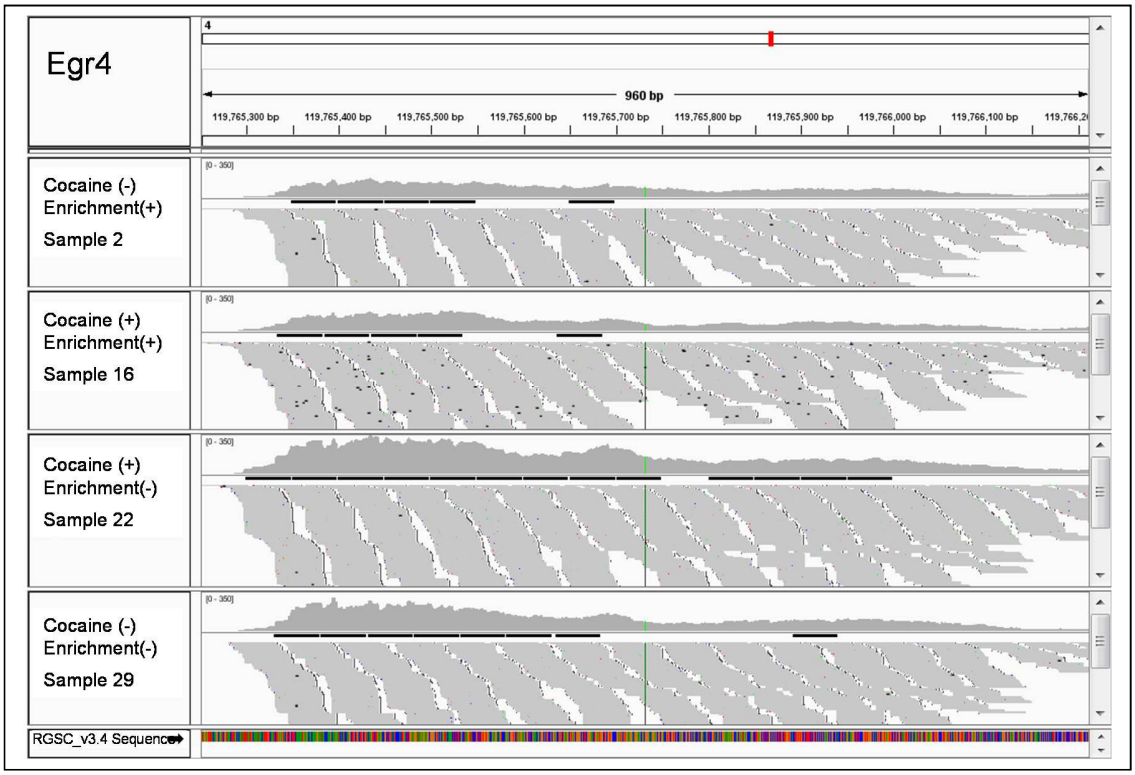

Figure 3.2. Alignment results visualization. Here we showed the visualization of the alignment results at the region of early growth response gene 4 (Egr4) using the IGV software. For each group, one representative sample was presented.

Not all regions in a genome are equally interesting. We have to integrate the alignment output with the genome annotation information and then narrow down to the investigated features, such as genes, exons, and transcripts. For an RNA-Seq gene differential expression analysis, the goal is to identify genes that are differentially

expressed in different situations. We can quantify the expression level of a gene with the number of reads that successfully map to its exon region. A higher expression level is more likely to be associated with a larger number reads. Ambiguity in region assignment occurs when one read maps to the overlap regions of two genes and / or only part of the read is mapped. The majority of currently available read count measurement software, such as *htseq-count*, *eXpress*, and *GenomeFeature*, have different options for dealing with the assignment of ambiguous reads. We measured the expression of each read using the *htseq-count* python program (version 0.5.3p9) in the union mode. In this mode, ambiguous reads are ignored. The results described 22,518 rat genes with the number of reads mapped to them.

In summary, we started the pre-analysis of RNA-Seq with data quality control and assurance. The raw data were enriched in short sequences with no assignable genetic meaning. We then mapped them against an annotated reference genome. The alignment revealed connections between the sequencing reads to the annotated genes. This relational connection provides information on the expression levels of each gene. All above approaches follows the DIKW hierarchy, as discussed in Chapter 1. In the following RNA-Seq differential expression analysis, we include more relationships such as correction, categorization, and calculation, to identify the essential information which is the precursors of knowledge.

3.3.3 RNA-Seq gene differential expression analysis

Although the amount of input genomic material for NGS sequencing is comparable at very beginning, the library size – the number of sequenced fragments – varies from sample to sample, thus it is inappropriate to directly compare the number of

reads between samples. To address this, two normalization methods are mainly used: the rescue method [43] and the scaling factor method.

The rescue method reports the abundance of gene expression in units of fragments or reads per kilobase of exon per million mapped reads (FPKM or RPKM). FPKM is used for paired-end sequencing where one segment produces two reads. FPKM reduces to RPKM in single-end sequencing when one fragment produces only one read. This method integrates the between- and within-sample normalization by rescaling the read counts over both library sizes and gene lengths. It is effective for comparing genes within one sample and suitable for between sample comparison, but may introduce biases in the per-gene variances [44]. This method has been popularly used in the Cufflinks approach.

The scaling factor normalization method employs many different strategies including total counts, mean, median, ratio, quantile, and others. The most popular approach is the Trimmed Mean of M-values (TMM) method [45]. TMM assumes the majority of genes are not differentially expressed between samples. To obtain the TMM scaling factors, we first calculate the gene-wise log-fold-changes of a reference sample (i.e. the M-values) as well as the gene-wise absolute (average) expression with the reference sample (A-values). We then exclude the most highly expressed genes having the largest fold changes based upon the M and A values. After trimming, the weighted M value of all other genes is directly related to the TMM factors. The TMM factors are typically close to 1, and serve as scaling factors to normalize library sizes and read counts. TMM normalization has been implemented in many RNA-Seq data analysis packages in BioConductor R including *edgeR*, *DESeq*, *DEXSeq* and others.

In addition to the two above approaches, strategies using housekeeping genes [46], GC-content biases [47], and technical effects [48] have also been proposed. All these normalization methods aim to reduce the technical bias inherent in the sample preparation steps. With normalization, quantitative approximations of target gene abundance become comparable.

We employed TMM factors to correct the library size. The raw library sizes showed large variation (Figure 3.3A): the sample with the largest library size had 2 times more reads than the smallest one. After TMM normalization, the effective library sizes for all samples were almost the same (Figure 3.3B).

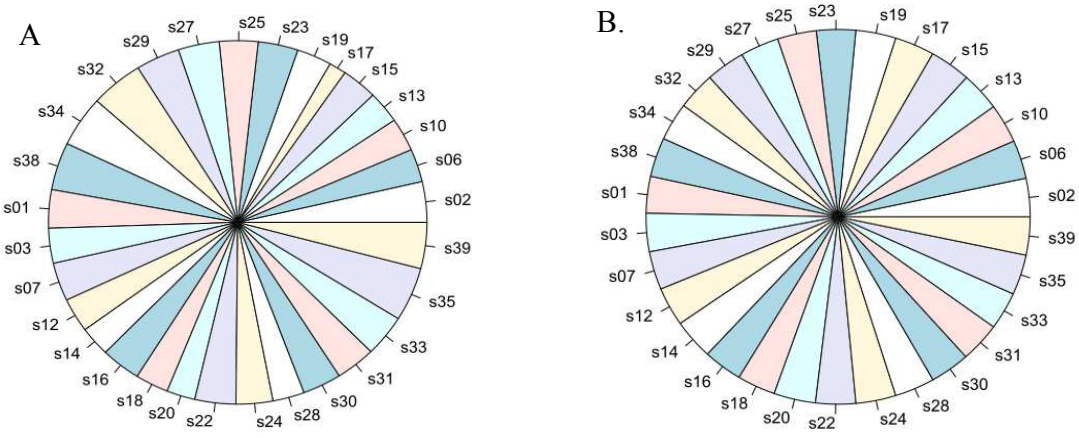

Figure 3.3 Effects of TMM normalization on library size. Effective library size of each sample before (A) and after (B) TMM normalization.

To perform statistical comparisons, we can model the normalized reads with various distributions including the normal, negative binomial [49, 50], and non-parametric distributions [51]. The Cufflinks NGS data analysis toolkit assumes that the expression measurements in FPKM / RPKM units approximately follow a normal distribution. Cufflinks performs its differential expression tests using a Bayesian

inference procedure to maximize the a posteriori count estimate [52]. The negative binomial distribution method, though more complex, is by far the most technically accurate and appropriate method and is widely used by expert analysts. It has been implemented in many BioConductor packages including *edgeR*, *DESeq*, *NBPSeq*, *baySeq*, *BBSeq*, etc. These approaches take the raw NGS count numbers and their normalization factors as input. The negative binomial distribution requires two parameters: the mean and the dispersion. To calculate the dispersion, several methods are used. They include: Cox-Reid adjusted profile likelihood, weighted quantile-adjusted conditional maximum likelihood, quasi-likelihood method, and dispersion shrinkage for sequencing method, etc. These methods have been implemented by several different analysis packages, none of which significantly outperforms the others. The resulting dispersions have three forms: common, trended/splined, and tag-wise. Common dispersion is shared by all genes in the sample. Tag-wise dispersions are different for each gene (i.e. tag). Meanwhile, genes with similar features (e.g. expression levels) are grouped together and share the same trended / splined dispersion. We can fit the read counts with a generalized linear model because we assume the counts follow negative binomial distribution, a distributions in the exponential family form. Once a model is built, we can perform statistical tests, such as the likelihood ratio test, with any dispersion to assign *p-values* to genes. However, the tagwise dispersion with a moderate degree of shrinkage is more likely to maximize performance [53]. The returned *p-values* are related to the contribution of the interested factor to the total dispersions. The non-parametric methods require no prior assumption of data distribution. The corresponding models are data-adaptive models. RNA-Seq data analysis packages such as *LFCseq*, *NOISeq*,

SAMSeq, and *MRFSeq*, utilize the non-parametric methods. They share a similar strategy, which entails creating a noise population and then comparing the signal to noise. The final *p-values* are associated with the probability that the signal is distinguished from noise.

In this project, we utilized the BioConductor package *edgeR* (version 3.0.4) for gene differential expression analysis. 14,309 (63.54%) out of the total 22,518 rat genes had more than one count in per million mapped reads (CPM) in at least five out of the 30 samples. The other genes were characterized by very low expression levels across the entire sample population. We filtered out these genes because the low expression level degrades the reliability of further statistical analyses. We performed TMM normalization only on the 14,309 genes that passed the filter. Using functions in the *edgeR* package, we estimated the common, trended, and tagwise dispersions in sequence because the calculation of the latter dispersion utilizes the former one as bootstrap input. Specifically, we first estimated the common dispersion across the whole sample, which was then utilized to calculate the trended dispersions. Finally, we obtained the tagwise dispersions, which took into account the trended dispersions. With the tagwise dispersion, we were able to fit the differential gene expression to a generalized linear model then perform a likelihood ratio test to generate the *p-values*. Genes with a *p-value* < 0.05 were treated as significant. The results indicated that the environmentally enriched condition led to 3393 genes being differentially expressed (2186 down-regulated, and 1207 up-regulated) and that cocaine administration effected the expression of 1274 genes (768 down-regulated and 506 up-regulated) (Figure 3.4A). Additionally, 1121 genes (288 down-regulated and

832 up-regulated) were associated with the interaction effects of environmental enrichment and the cocaine administration.

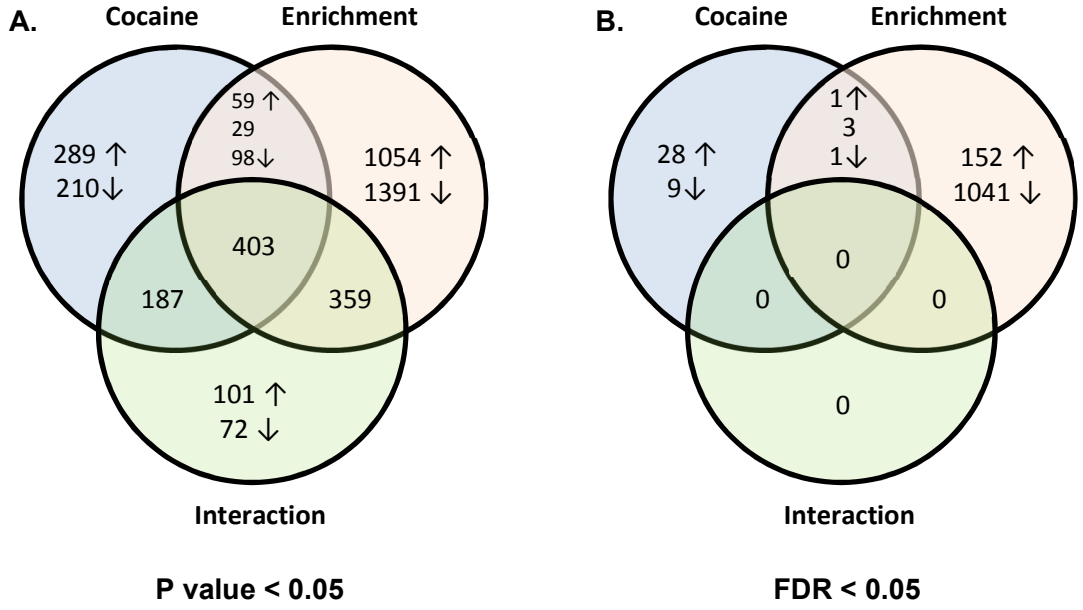

Figure 3.4 RNA-Seq expression analysis results. We presented the Venn diagrams of RNA-Seq gene expression analysis results for main cocaine effect (cocaine), main enrichment effect (enrichment) and their interaction (cocaine × enrichment) at two different cutoff *p-value* < 0.05 (A) and FDR < 0.05 (B). Number of up-regulated and down-regulated genes was indicated with up arrow (↑) and down arrow (↓), respectively. Numbers with no arrow means the regulation patterns of the gene were not consistent in two or more conditions.

To correct for multiple comparison tests, we further performed the Benjamini–Hochberg (BH) procedure to obtain the false discover rate (FDR). Genes with FDR < 0.05 were treated as significant. 1198 differentially expressed genes (DEGs) were associated with the enrichment main effect and 42 genes with the cocaine main effect (Figure 3.4B). No genes associated with the interaction effect of the enrichment and cocaine factors were identified in this level of significance.

In the following analysis, an FDR < 0.05 was utilized when we investigated relationships among molecules (session 3.4.1 and 3.4.2). The cut off was loosened to *p-value* < 0.05 (not FDR, which is much more stringent) to include more genes when we investigate the canonical pathways of interest using QIAGEN's Ingenuity® Pathway Analysis suite (IPA®, QIAGEN Redwood City,www.qiagen.com/ingenuity) [54] (session 3.4.3 and 3.4.4). Although the corresponding FDR for *p-value* < 0.05 for the main enrichment effect was 21.06%, the distribution of the false positive errors should be randomly distributed across the entire sample. For the canonical pathways in IPA, a group of genes with functions in concert are studied simultaneously, which would compensate for the relatively large false discovery rate.

3.4 RESULTS AND DISCUSSION

This project was discovery-driven to identify essential pathways related to the EC-induced addiction protective phenotype. In the following, we will first compare the RNA-Seq findings with known knowledge about cocaine addiction and EC-related protective phenotype to validate the significance of our RNA-Seq results. Then, we will introduce several novel target pathways, which are or will be investigated by our collaborative research group. Finally, we will compare the RNA-Seq transcriptomics results with the proteomics study results.

3.4.1 Effects of cocaine addiction

Cocaine is a powerful and addictive drug, whose effects have already been well studied. When people smoke or snort cocaine, the cocaine quickly travels to the brain and affects the function of the reward pathway. Cocaine acts as monoamine reuptake

transporter inhibitor to block transportation of monoamine from the synaptic cleft back into the terminal synapses in the nucleus accumbens, which is the central component of the mesolimbic reward pathway. This process leads to a huge increase and accumulation of monoamine in the nucleus accumbens.

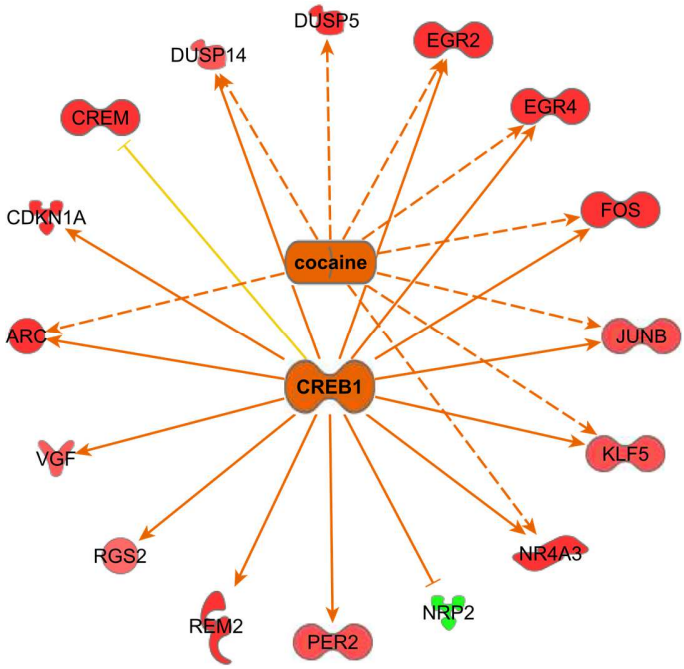

Figure 3.5 Effect of the cocaine administration. We performed up-stream regulator analysis for cocaine administration associated differentially expressed genes using IPA and presented a network of several activated regulators of interest and their target genes. (Legend for all networks and pathways from IPA. **Molecule color**: red (up-regulated) and green (down-regulated). **Regulator color**: orange (activated / increased) and blue (inhibited / decreased). **Connection line**: solid (direct connection) and dash (indirect connection). **Line arrow:** ⟶ (up-regulate / activate), ⊣ (down-regulate / inhibit) and ─ (affect). **Line color:** red (consistent finding and activated prediction), orange (consistent finding and inhibited prediction), grey (affect connection) and yellow (inconsistent finding and prediction).)

We performed IPA up-stream regulator analysis with differentially represented genes associated with cocaine administration (Figure 3.5). Many early genes that were shown to be immediately induced by cocaine administration in previous studies were up-regulated. These genes include FOS, FOSB, JUNB, EGR2, EGR4, etc. Based upon the regulation of these genes, the up-stream regulator analysis in IPA predicted that the cocaine stimulus was activated (*p-value* = 2.03e-12, z-score = 2.924). At the molecular level, cAMP response element binding protein 1 (CREB1) is known to be highly enhanced after chronic cocaine taking, thereby further enhancing reinforcement in cocaine self-administration by rats [55]. Our analysis indicates that CREB1 was predicted to be activated (*p-value* = 5.01e-16, z-score = 3.343).

The above results thus confirmed previous findings about cocaine addiction and gave us confidence in the high quality of our data and data analysis. These findings therefore assured us that our RNA-Seq technique has the power to distinguish between the samples of the two experimental conditions at both the treatment and molecular level.

3.4.2 Effect of EC induced protective phenotype

Environmentally enriched rats show a protective addiction phenotype in rat drug self-administration paradigms [40-42]. Specifically, EC rats, with daily access to novel toys, exercises and social contact, exhibit less bar pressing under acquisition, maintenance, extinction and reinstatement of cocaine taking and seeking compared to the control IC rats. Very interestingly, many genes associated with cocaine addiction were down-regulated in the EC condition. Up-stream regulator analysis of the differentially expressed genes (DEGs) associated with EC main effect indicated that the response of cocaine administration was inhibited (*p-value*=4.8e-4, z-score = -1.793) under EC. Thus,

an enriched environment leads to opposite effects as the cocaine stimulus, which can be interpreted as a protective phenotype. Additionally, we identified an inhibitory activity of potassium chloride in EC rats compared to IC rats (*p-value* 1.63e-4, z-score = -2.276). Cocaine greatly increases the response to potassium chloride [56], however, the EC rats showed decreased gene expression related to the activity of potassium chloride. This response behavior confirms that the enriched condition leads to a protective phenotype for cocaine addiction.

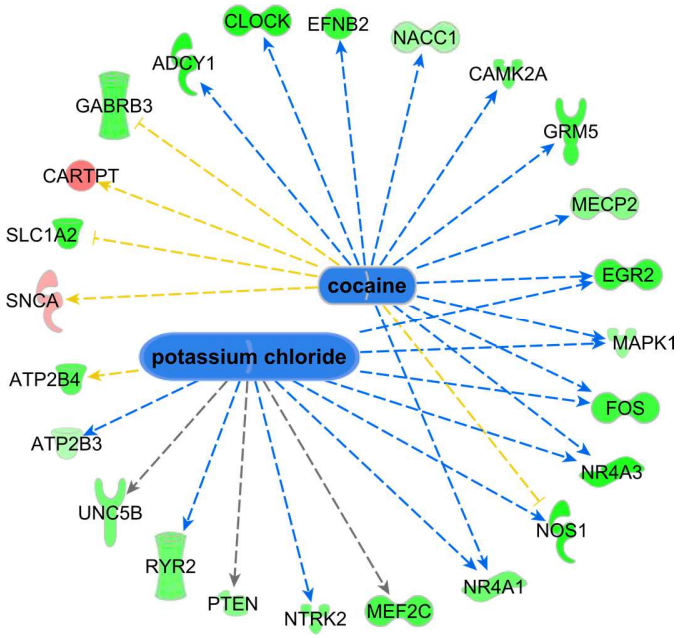

Figure 3.6 Effect of enriched condition. We performed up-stream regulator analysis for enriched environment associated differentially expressed genes using IPA and presented a network of several activated regulators of interest and their target genes.

It is known that blocking the activity of the transcription factor CAMP Responsive Element Binding Protein 1 (CREB1) in the nucleus accumbens can reproduce

the EC induced protective phenotype [56]. However, the CREB1 target genes underlying this protective phenotype have yet to be identified. Although our RNA-Seq results indicated no differential expression of CREB1, 50 CREB1 target genes were shown to be differentially expressed due to the EC environment, with the majority genes down-regulated (Figure 3.7). The altered genes included EGR2, FOS, IL6ST, NR4A1, NR4A3, SLC38A1, SLC6A11, SMOC1, TSPYL4, etc. CREB1 is activated by phosphorylation by several protein kinases as results of various cellular stimuli. The findings suggest the phosphorylation signaling for CREB1 is altered by environmental enrichment and confirm our previous discovery that CREB1 plays an essential role in the resistance of the protective phenotype against cocaine addiction.

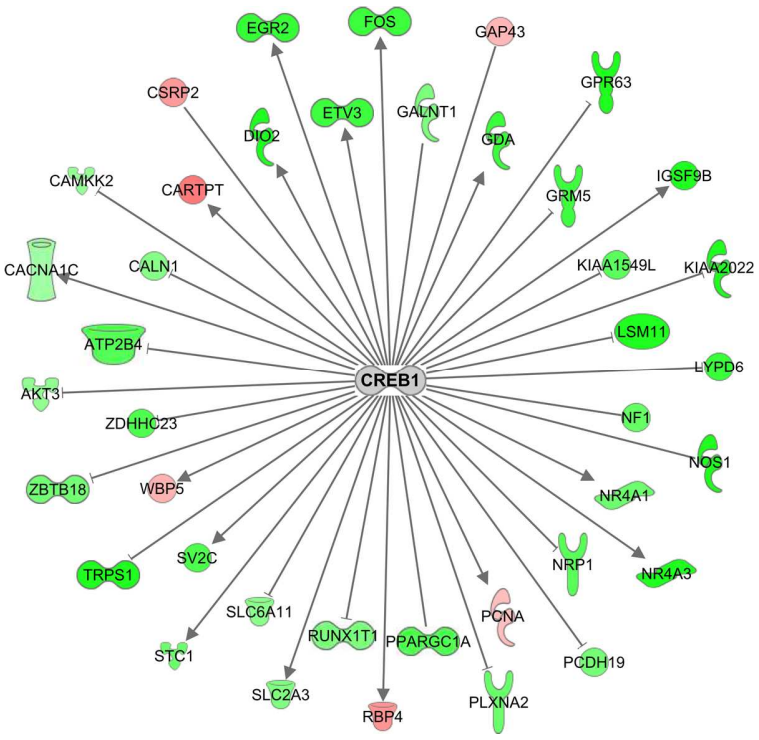

Figure 3.7 EC effect on the CREB1. We presented a network of CREB and its target genes that were differentially expressed in the enriched condition.

3.4.3 Novel EC induced protective phenotype related pathways

The objective of the current project is to identify novel pathways related to the EC-induced protective phenotype. With the DEGs associated with the environmental enriched condition main effect, Dr. Green's lab preformed an IPA canonical pathway analysis and selected three representative pathways for further research.

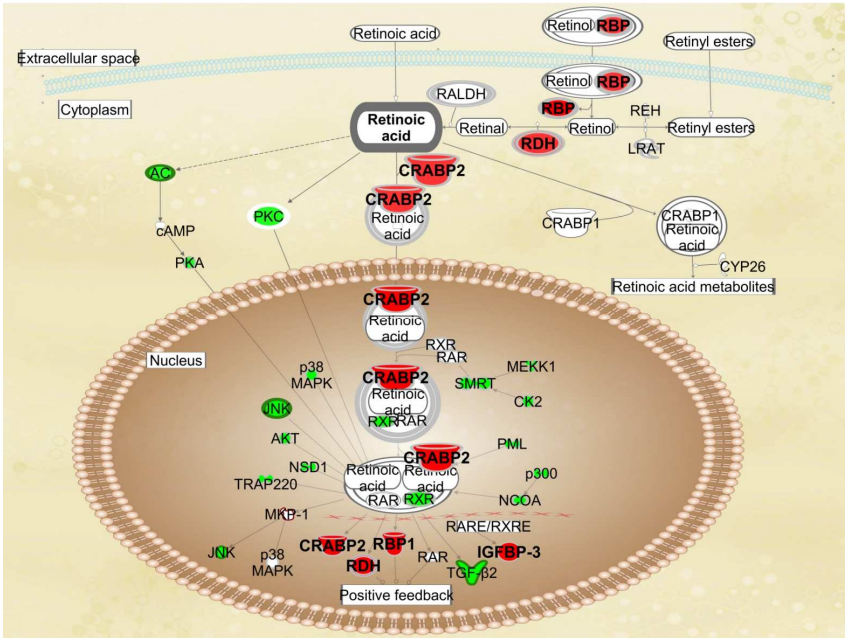

Figure 3.8 Retinoic acid receptor (RAR) activation pathway. A selected top canonical pathway enriched with genes that were differentially expressed in the enriched condition.

The first selected pathway was the retinoic acid receptor (RAR) activation pathway (Figure 3.8). In this pathway, we found up-regulation of the mRNA in the majority of retinoic acid target genes, as well as proteins involved in retinoic acid synthesis and binding. Retinol dehydrogenase (RDH) consists of a group of enzymes that

function to dehydrogenate retinol and covert it to retinal, a precursor of retinoic acid (RA). Retinol binding protein (RBP) and cellular retinoic acid binding protein (CRABP2) are retinoid-binding proteins which are necessary in RA synthesis and protect RA in the cytoplasmic environment. In addition, many RAR target genes are also up-regulated by environmental enrichment, such as RDH and CRABP2. Increased expression of these genes further enhances RA synthesis and RA signaling, which in turn forms a positive feedback loop. In contrast, most down-regulated genes are RAR pathway inhibitors, such as PCK, MAPK, JNK, AKT, etc. At the mRNA level, RA pathways transcripts showed coordinated regulation by environmental enrichment, therefore we hypothesize that the high expression of retinoic acid may relate to the EC-induced protective phenotype.

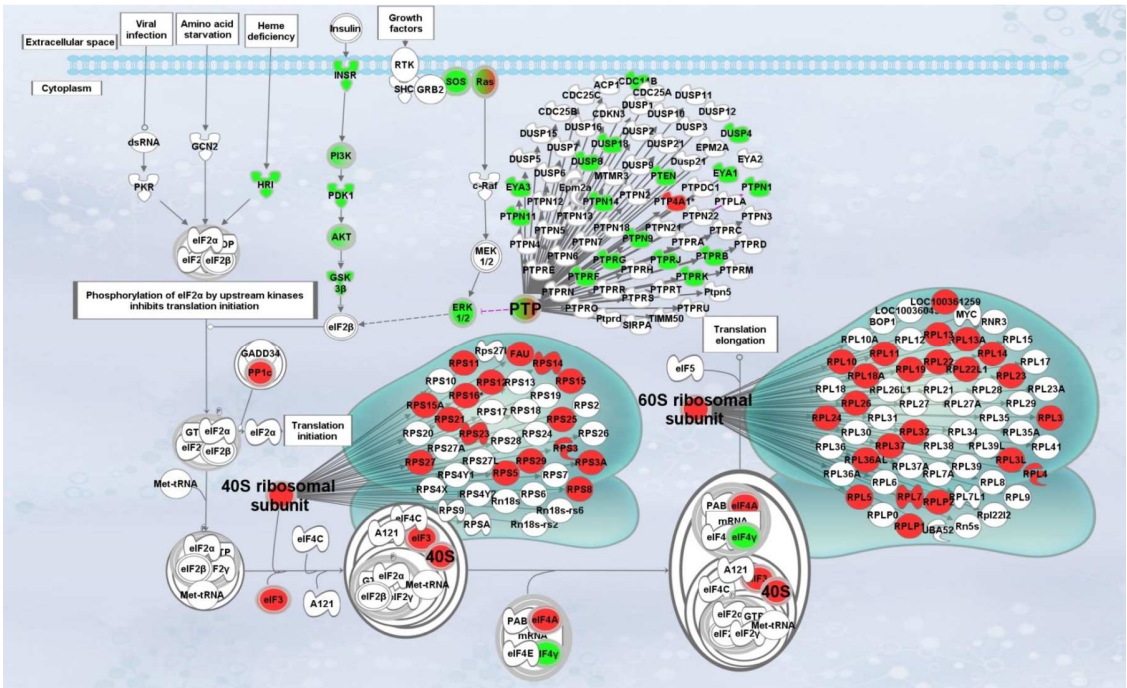

Figure 3.9 EIF2 signaling pathway. A selected top canonical pathway enriched with genes that were differentially expressed in the enriched condition.

The second selected pathway is the eukaryotic initiation factor 2 (EIF2) signaling pathway (Figure 3.9). As the name suggests, EIF2 is required in the initiation of translation. It delivers charged initiator methionyl-tRNA to the ribosome and also functions to identify the translational start site. The activity of EIF2 and EIF2β, the guanine nucleotide exchange factor of EIF2, are inhibited when they are phosphorylated. The enriched environment caused down-regulation of the up-stream kinases of both EIF2 and EIF2β. As a result, EIF2 and EIF2β were activated to initiate translation. We identified the up-regulated expression of many members belonging to the 40S and 60S ribosomal subunit-mRNA complex, indicating increased activity of protein translation. We hypothesize that the effectiveness of protein translation regulated by EIF2 may contribute to the EC-induced protective phenotype.

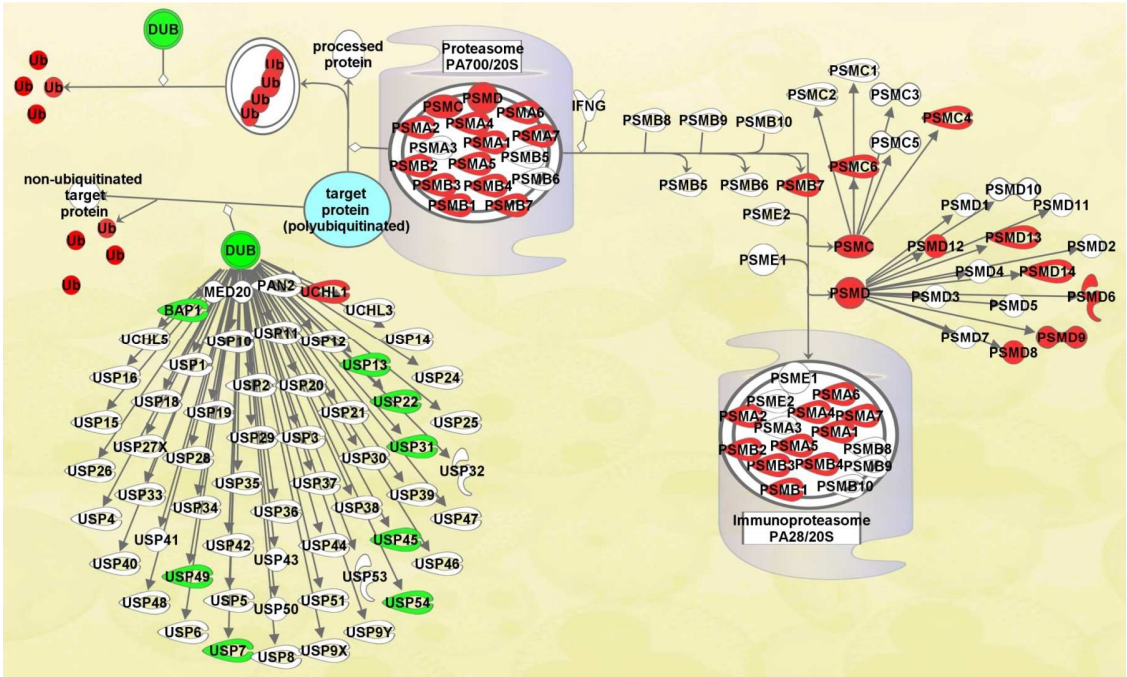

Figure 3.10 Protein ubiquitination pathway. A selected top canonical pathway enriched with genes that were differentially expressed in the enriched condition.

The third selected pathway is the protein ubiquitination pathway. This pathway regulates the degradation of short-lived or regulatory proteins through ubiquitination. A target protein is first conjugated with multiple ubiquitin moieties to form poly-ubiquitinated protein. The deubiquitinating enzymes (DUB) can remove unbiquitin chains from the tagged proteins for further recycling. Otherwise, the poly-ubiquitinated proteins will be proteolyzed by the proteasome complex PA700 / 20S. The enriched condition up-regulated many members in the proteasome complex but down-regulated many DUB enzymes. This led to more protein degradation activity in EC rats than in IC rats. We hypothesize that the effectiveness of protein degradation through ubiquitination may contribute to the EC-induced protective phenotype.

3.4.4 Comparison between transcriptomics and proteomics results

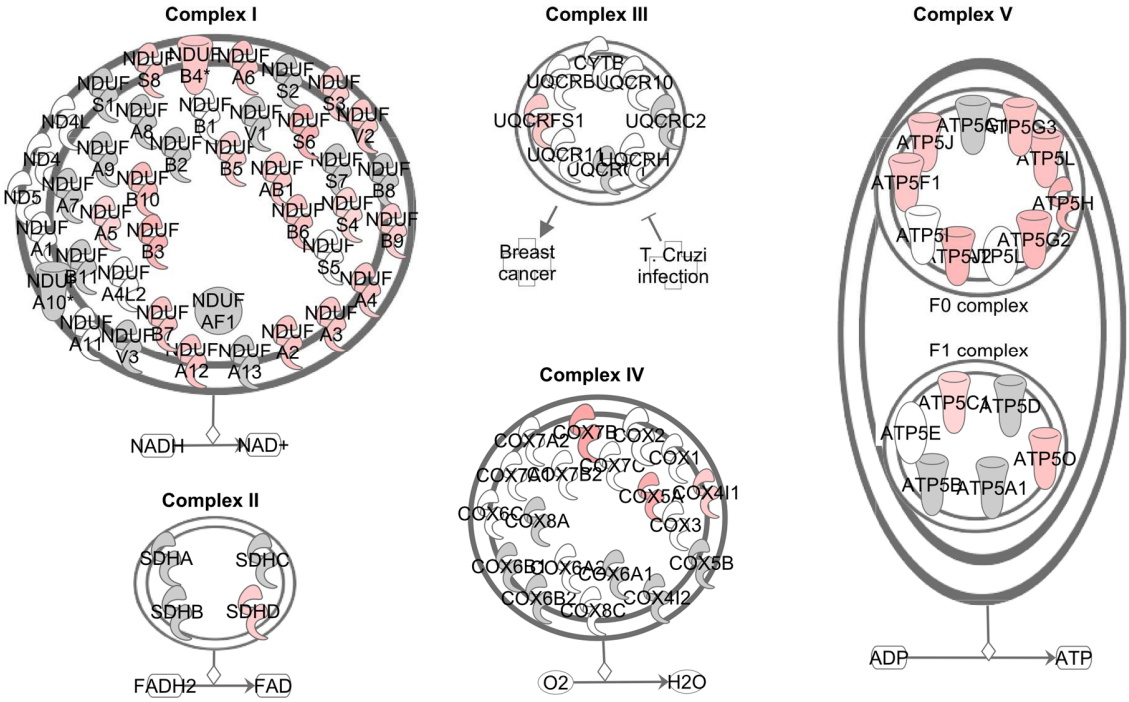

Figure 3.11 Mitochondrial redox carriers. We examined the effects of enriched condition in the component proteins of five inner-membrane-bound complexes in mitochondria.

We harvested the left nucleus accumbens for RNA-Seq transcriptomics research and the right nucleus accumbens for proteomic investigation using liquid chromatography mass spectrometry (LC/MS). The most striking result from the proteomics study is that proteins involved in energy production are highly expressed in EC rats [42]. Both the proteomics data and our RNA-Seq transcriptomics data had "Mitochondrial dysfunction" as one of the top-regulated canonical pathways. We examined the inner-membrane-bound complexes in mitochondria. All five complexes showed members that were up-regulated under the EC environment (Figure 3.11). Lipid-soluble or water-soluble electron carriers electrically connect the first four complexes. They function interactively to transfer electrons from the mitochondrial matrix into the inter-membrane space then reduce molecular oxygen to water. More specifically, complex I can pass electrons from NADH to coenzyme Q and complex II passes electrons from FADH2 to coenzyme Q. The electrons from coenzyme Q are further passed to cytochrome c through complex III. Cytochrome c then passes electrons to complex IV and leads to the reduction reaction. This process results in a proton gradient across the mitochondrial inner membrane, which is then used by complex V, also known as ATP synthase, to make ATP. The up-regulation of these complex members indicated that the EC rats had more active energy production than the IC rats. Thus, the findings at the proteomics and transcriptomics level mutually support each other.

3.5 CONCLUSION AND LIMITATIONS

The current project utilized an unbiased discovery-based transcriptomic method to search for differential expression of genes in the nucleus accumbens of EC and IC rats that self-administer cocaine or saline. Our results suggest that environmental enrichment plays a significant role in addictive behavior and provide future direction in the study of individual differences in susceptibility to addiction. With further research, we expect to unravel the biochemical and physiological mechanisms of the protective phenotype at the functional level.

The NGS approach has several significant advantages. First, RNA-Seq is a genome wide study with high throughput and sensitivity. We could detect differential expression of 10–20 thousand genes in one run. Second, RNA-Seq requires no a priori knowledge of the genomic features or cross-hybridization between similar sequences as in microarray analysis. Third, nanograms of DNA material are sufficient for RNA-Seq. The low amount of sample requirement enables us to sequence each nucleus accumbens rather than pool samples together as was necessary in the proteomics investigation.

In addition to its relatively high cost, the NGS strategy has several intrinsic limitations. First, RNA-Seq investigates the system at the transcriptomics level without considering mRNA degradations and post-translational modifications. Second, the sequencing reads are typically of short length which limits its ability to sequence highly repetitive regions. Third, the Illumina HiSeq platform utilizes a clonally amplified template technique which may cause progressive replication errors during library preparation.

Another limitation pertinent to our experimental design is that samples were harvested three hours after cocaine/saline self-administration thus only that time point

was examined. Since mRNA regulation is a dynamic process, we may fail to catch earlier or later regulations that could be essential to the environmental enriched condition induced protective phenotype.

Chapter 4. Investigation of Pathogenies of Visceral Leishmaniasis Through Transcriptional Profiling [2]

4.1 INTRODUCTION

Visceral leishmaniasis (VL), also known as kala-azar, is an endemic disease in tropical and subtropical areas. The disease is mainly caused by parasitic protozoa Leishmania *donovani*. There are four different forms of the disease, including: cutaneous, diffuse cutaneous, mucocutaneous and VL. VL is the most severe form. VL patients have a high fatality rate (up to 100%) within two years, if untreated. Over 12 million people in 88 countries are known to have leishmaniasis but many cases (>90%) are asymptomatic. Each year, 1 to 2 million more people become sick with leishmaniasis, of which 0.2 to 0.4 million new infections are deadly VL. The disease threatens about 350 million people mainly in Indian subcontinent (India, Bangladesh, and Nepal), East Africa (Sudan, South Sudan, and Ethiopia), and Brazil. These areas typically are associated with malnutrition, population displacement, poor housing, weak immune system and lack of resources [57]. All these disadvantages increase the risk of infection, and make the diagnosis and treatment difficult.

The life cycle of Leishmania occurs in the vertebrate host and the vector. The parasites live and replicate in the mid-gut of a sandfly vector in the form of flagellated mobile promastigote. During a natural infection, the sandfly injects promastigotes into the skin of susceptible hosts, such as humans, dogs, cats, etc. The promastigotes get phagocytized by macrophages where they transform into amastigotes. Amastigotes replicate and eventually increase the parasite burden [58, 59]. Clinical findings of VL

[2] In collaboration with Dr. Peter C. Melby's group at UTMB.

patients are usually characterized by fever, splenomegaly, pancytopenia, and cachexia. However, the pathogenesis of VL is not clearly understood. Our objective in the current project is to understand the key molecular mechanisms by which the parasite causes pathology. Our ultimate goal is to develop new therapeutic strategies to prevent and control this devastating infectious disease.

4.2 EXPERIMENTAL DESIGN

Two experiments were implemented to investigate VL. In the first one, we employed an animal model – Syrian golden hamsters – because they closely mimic the chronic VL pathology found in humans.

In the hamster study, six- to eight-week old outbred female Syrian hamsters (*Mesocricetus auratus*) were randomly assigned to either uninfected or infected group. Each group had four animals. We infected the hamsters with $1x10^6$ Leishmania *donovani* (MHOM/SD/001S-2D) metacyclic promastigotes, intracardially. All hamsters were sacrificed at 28 days post infection to harvest the whole spleen cells and the adherent spleen cells (after 2–3 hours of adherence). We further isolated mRNA, and verified its quality by Agilent Bioanalyzer. The mRNA samples were submitted to Illumina HiSeq 1000 using the same approach as discussed in Chapter 3, to generate 50 base paired-end reads. Please refer to the manuscript (F. Kong et. al.) for detailed information about the experimental design.

In summary, our hamster experiment resulted in four groups: uninfected whole spleen cells, infected whole spleen cells, uninfected adherent spleen cells and infected adherent spleen cells. Each group had four samples. We analyzed these RNA-Seq data using different biomedical informatics tools to investigate VL in hamster.

4.3 HAMSTER VL TRANSCRIPTOME ANALYSIS.

A transcriptome analysis aims to identify genes, transcripts, or exons that are likely to be differentially represented among different experimental conditions. A very essential step is to map the sequencing reads to the transcriptome or genome of the study model. Neither the hamster genome nor its transcriptome has been fully sequenced or annotated. As a result, we first had to assemble a draft hamster transcriptome from the raw RNA-Seq data before any differential expression analysis.

4.3.1 *De Novo* assembly

We assembled a *de novo* transcriptome for the Syrian hamster because no completed reference genome or transcriptome were available. We had access to the sequences derived from Chinese Hamster Ovary cells (from its near relative *Crisetulus griseus*) [60], and a draft genome of *Mesocricetus auratus* (NCBI BioProject PRJNA210213) [61]. Both were incompletely sequenced and/or annotated. To obtain a transcriptome for the hamster spleen, we applied the *de novo* assembly with all sequencing reads from hamster samples including both whole spleen cells and adherent spleen cells.

To prepare the data for *de novo* assembly, we extracted high quality and clean reads from all raw sequencing reads. We first performed a quality control to assess the raw sequencing data using FastQC (v0.10.1) [62]. Phred score medians at all 50 bases were ≥ 30 (i.e. error rate ≤ 0.001) and the majority of the reads had average Phred score > 37 (i.e. error rate ≤ 0.0002) (Figure 4.1A, B). The CG distribution overall all sequences content and its theoretical distribution had a similar behavior (Figure 4.1C). Moreover, the N contents across all bases were <5% (Figure 4.1D). N indicates unknown or

undetermined when a sequencer fails to determine the nucleotide in situ. Therefore, the overall sequencing quality is high. To avoid the contamination of pathogen sequences, we filtered out reads that aligned to the Leishmania *donovani* BPK282A1 genome (NCBI BioProject PRJEA61817) [63] using Bowtie2 (v2.0.0-beta5) [64] with default parameters. We further filtered out artifacts and the reads with less than 28 Phred score in more than 10% of nucleotides using FASTX-Toolkit software (v0.0.13) [65] to reduce the effect of low quality reads. Both forward and reverse reads were removed if any of them failed to pass the filters. Typically, artifacts and low quality reads were less than 2% of the whole sequencing. These results indicated that control and infected samples generated high quality sequencing reads. We pooled all the clean reads after the above filtering for further transcriptome assembly.

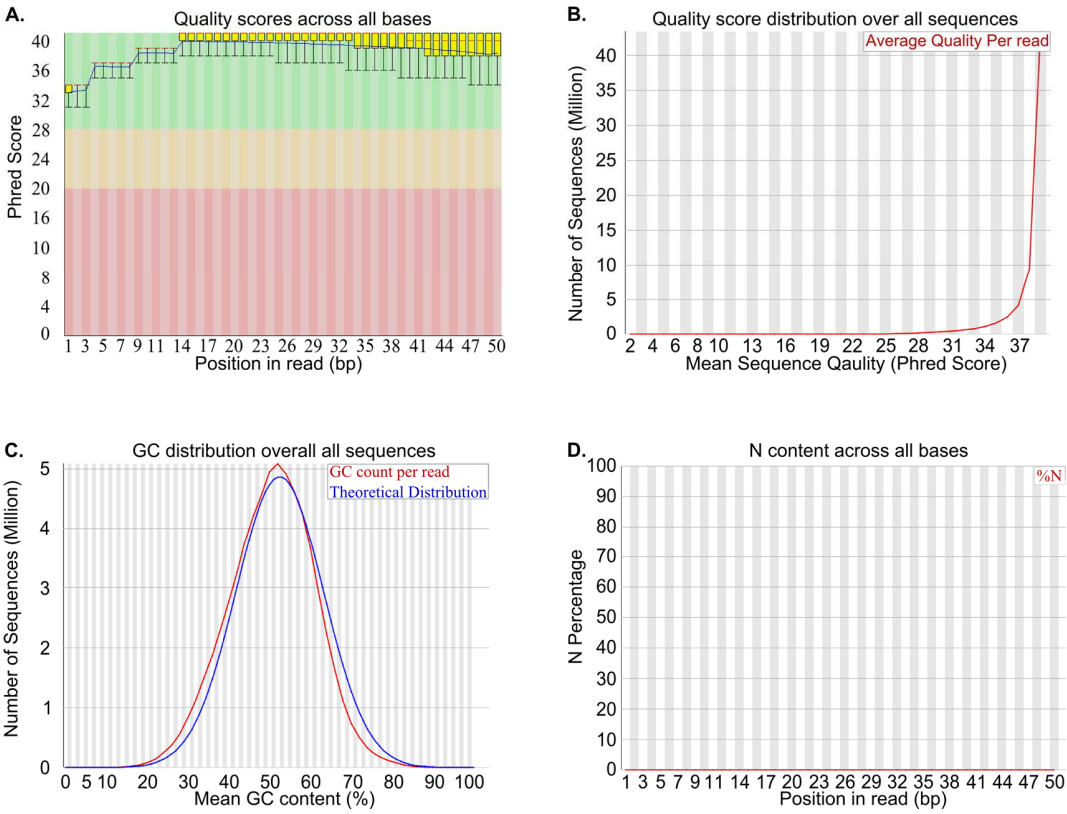

Figure 4.1 FastQC output results. Representative output when we checked quality of the reads for samples in the current projects using FastQC. Here we only showed the plots of quality scores across all bases (A), quality score distribution over all sequences (B), GC distribution overall all sequence (C) and N content across all bases (D).

To obtain a complete hamster spleen transcriptome, we used two steps. First, the cleaned sequencing reads from different spleen samples were pooled together and *de novo* assembled using Trinity software [66]. Second, the resulting transcriptome, all cleaned reads from hamster spleen and adherent cells, and CHO-K1 RefSeq genome [60] were further used to perform our second *de novo* assembly using the BRANCH software [67]. Both assembling steps were run at the Texas Advanced Computing Center (TACC) at the University of Texas at Austin. A summary of the workflow is shown in Figure 4.2A.

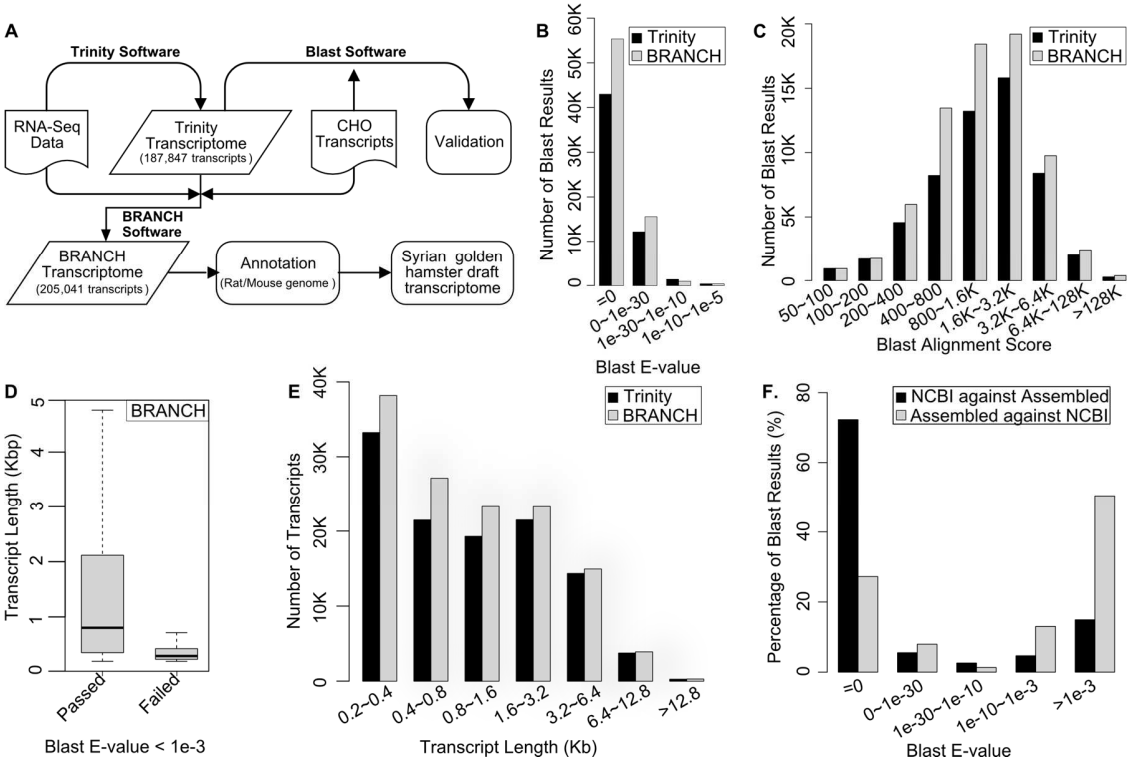

Figure 4.2 Hamster splenic transcriptome *de novo* assembly. We assembled a transcriptome draft for the hamster spleen with the work flow shown in (A). The intermediate (trinity) and final (BRANCH) transcriptomes were compared to transcripts of CHO cells (B, C). We then checked the E-value cutoff (D), and compared the final (BRANCH) transcriptomes with intermediate (trinity) transcriptome (E) and NEBI *Mesocricetus auratus* transcriptome draft (F).

The Trinity package is a software platform for de novo transcriptome assembly of RNA-Seq data from non-model organisms [66]. Trinity produced 187,847 transcripts with lengths ranging from 201 to 23,840 nucleotides. To validate the assembled results, we compared each transcript against the CHO Ref-Seq transcripts using BLAST (version 2.2.27+) [68]. 35.42% of the transcripts from Trinity returned a hit under the default BLAST parameters. We examined the lowest E-value and the largest alignment score associated with each Trinity transcript. The results showed that the largest reported E-value among all the hits was 1e-5 and 78% of the hits had an E-value equal to 0 (Figure 4.2B). 85.83% of the successful hits returned alignment scores > 500 (Figure 4.2C). These data indicated that the Syrian hamster *de novo* assembled transcriptome was highly homologous to sequences in the CHO-K1 genome. Meanwhile, the pool of Trinity transcripts contained more transcript sequences than what is represented or annotated in the CHO-K1 genome.

We further used the BRANCH software to expand the Trinity transcriptome into a more complete transcriptome [67]. In total, 205,041 transcripts with length ranging from 201 to 23,840 nucleotides were obtained. We then created our own blast library including both *Rattus norvegicus* (Rnor_5.0.73) and *Mus musculus* (GRCm38.73). All transcripts from BRANCH were compared against our customized library using BLAST (v 2.2.28+) [68]. 64% (131,021) of the BRANCH transcripts had a BLAST E-value <1e-3 when

compared to the rat and mouse genomes. We assigned each transcript with the targeted gene names, which had the lowest E-value. The transcripts that passed the E-value cutoff were typically longer and thus more informative than those failed (Figure 4.2D). After application of the <1e-3 cutoff to the assembled transcripts from BRANCH and Trinity, BRANCH produced more long transcripts compared to Trinity (Figure 4.2E), confirming that BRANCH improved the Trinity assembly.

We finally pooled 131,021 BRANCH transcripts together as a draft reference transcriptome. All of them had a blast hit with E-value <1e-3 against the mouse and rat library. Using the Bowtie2 software package (v2.1.0), we aligned all RNA-Seq reads that failed to map to the Leishmania genome against our draft transcriptome and got a 92.76 ± 0.68% alignment rate. This was considerably higher than the 58.46 ± 1.38% and 37.77 ± 1.25% obtained when aligned against the NCBI *Mesocricetus auratus* transcriptome (NCBI BioProject PRJNA210213) and CHO ref seq transcripts. Addtionally, we compared our assembled transcriptome with the NCBI *Mesocricetus auratus* draft transcriptome and found >70% NCBI transcripts could be found in our *de novo* assembled transcriptome, while <30% assembled transcripts were identified in the NCBI transcriptome (Figure 4.2E). We also tested to incoporate the NCBI *Mesocricetus auratus* genome into our *de novo* assembly by a third running of BRANCH. Furher analysis failed to identify differential expression of some genes (e.g. CCL6/7/8/9/25/28, CXCL10), known to be differentially expressed by PCR. Collectively, these data indicated that with the initial BRANCH analysis, we had assembled the most complete hamster spleen transcriptome available. We used it as a reference transcriptome for RNA-Seq differential expression analysis.

4.3.2 RNA-Seq gene differential expression analysis

All the non-leishmania-like raw sequencing reads were first mapped to our assembled reference transcriptome using Bowtie2 (v2.1.0) with default options but allowing one read to map to as many as 500 different transcripts. We then measured the expression abundance, i.e. the number of reads mapped to each transcript, using the software eXpress. The effective counts were used for RNA-Seq differential expression analysis because they corrected biases caused by multiple alignments and mismatches during alignment.

We analyzed the spleen samples and the adherent spleen samples separately because they contained different cell populations and were sequenced in two independent runs. We first examined the sample clustering patterns using multidimensional scaling (MDS, Principal Coordinate Analysis) plots (Figure 4.3A, B) with the top 500 most differentially expressed genes. The root-mean-square deviation was used to calculate the distance between each pair of samples. The infected and uninfected samples were appropriately clustered in both spleen and splenic macrophage samples. However, the first dimensional coordinate separated spleen but not adherent cell samples, which suggest a greater effect of Leishmania infection in the whole spleen samples than in the adherent spleen cells.

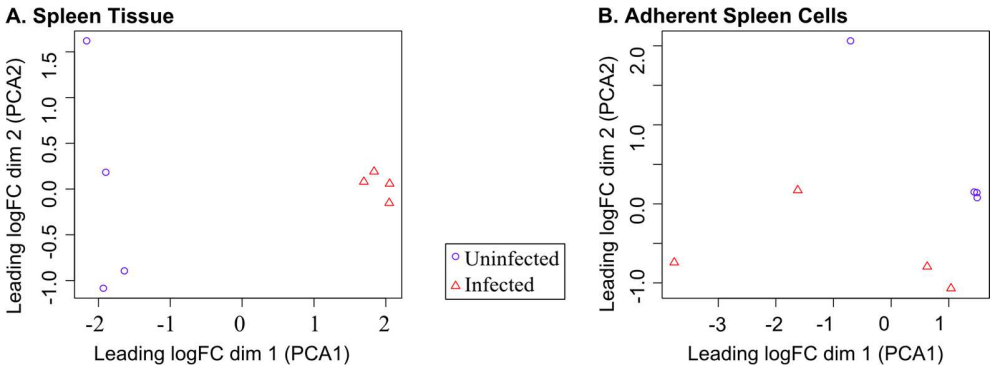

Figure 4.3 MDS plot of hamster VL samples. We examined the clustering patterns of hamster VL samples in spleen tissue (A) and splenic adherent cells (B) using classic Torgerson metric MDS plot, which also known as PCA plot.

To identify differentially expressed transcripts in each experiment, we performed differential expression analysis, using two R BioConductor packages *edgeR* and *DESeq2*. In total, three different approaches were applied: classic exact test and generalized linear model with likelihood ratio test in BioConductor packages *edgeR* [69], and Wald test in *DESeq*2 [70]. Only transcripts with at least one count per million mapped reads (CPM) in at least three out four samples in control and/or experimental group were used for analysis. We considered a transcript to be significant when it was detected by all three different approaches as differentially expressed transcript. We set the cutoff as FDR < 0.01 and identified 4,360 differentially expressed transcripts in the spleen samples, which included 2,340 (53.7%) up-regulated and 2,020 (46.3%) down-regulated genes (Figure 4.4A). At this FDR cutoff, splenic adherent cells had substantially fewer differentially expressed transcripts than the whole spleen tissue: 692 transcripts including 449 (64.9%) up-regulated transcripts and 243 (35.1%) down-regulated transcripts (Figure 4.4C). A number of differentially expressed transcripts were common to both spleen tissue and splenic adherent cells (240 up-regulated and 64 down-regulated) (Figure 4.4A, C). A smear plot and a heat map (Fig 4.4B, D) representation of these results are also included. The number of differentially expressed transcripts in the spleen tissue and splenic adherent cells were decreased to 2778 and 363 by tightening the FDR to < 0.001. Only the *edgeR* likelihood ratio test (LRT) results were listed and used for pathway analysis.

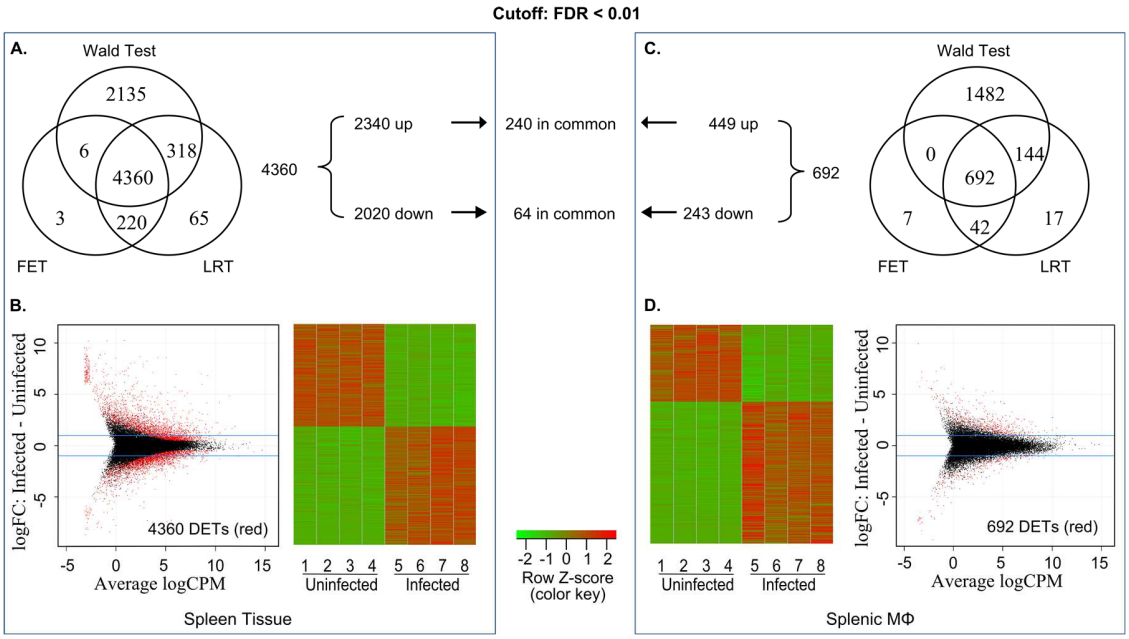

Figure 4.4 RNA-Seq expression results of hamster VL. We here presented the RNA-Seq differential expression analysis results of spleen (left panel) and splenic adherent cells (right panel) during VL. The numbers of differentially expressed transcripts from three different approaches – likelihood ratio test (LRT), Fisher's exact test (FET) and Wald test – were showed in top panel (A, B). The differentially expressed transcripts were the overlap of all three approaches. We colored these differentially expressed transcripts in red in the smear plots and also examined them using heat map (C, D).

We evaluated the functional significance of differentially expressed genes associated with Leishmania infection using QIAGEN's Ingenuity® Pathway Analysis suite (IPA®, QIAGEN Redwood City,www.qiagen.com/ingenuity) [54]. We uploaded the DEG data set (FDR < 0.001) into IPA, and then performed an IPA Core Analysis to identify the most significant canonical pathways. We further enriched genes in each pathway by relaxing the FDR to < 0.01. The top 10 pathways (ranked by –log(p-value)) identified in the 28-day infected spleen tissue and splenic adherent cells (shown to be splenic macrophages, i.e. MΦ, see section 4.5.1a) are shown in Figure 4.5A, B,

respectively. Four out of the top ten pathways identified in whole spleen tissue were also identified in splenic macrophages, supporting the central importance of macrophages in the immunopathogenesis of the splenic infection.

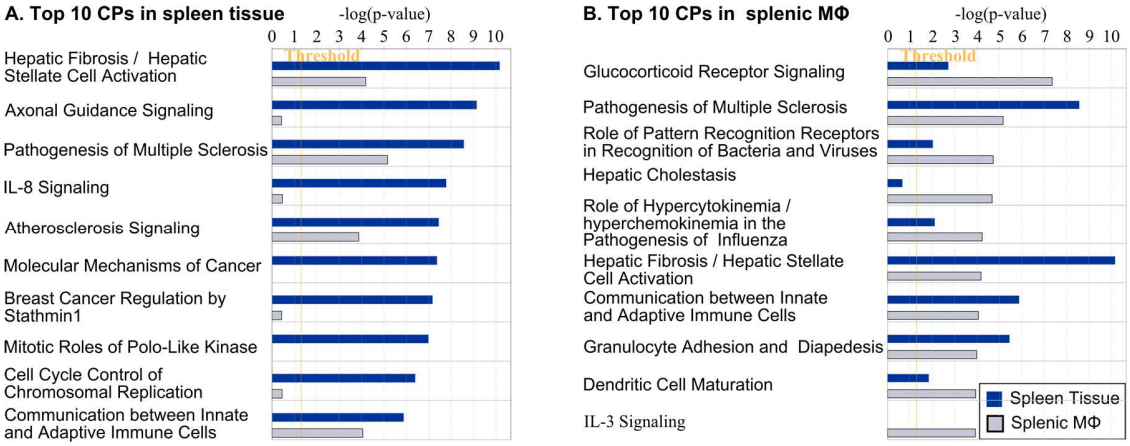

Figure 4.5 Top canonical pathways in hamster VL. We presented top 10 canonical pathways (CPs) in the spleen tissue (A) and splenic macrophages (B), and the comparison between them.

Moreover, we also performed gene set enrichment analysis (GSEA) using software from the Broad Institute (http://www.broadinstitute.org/gsea) and the MSigDB C5: GO gene set collection (1453 gene set available) (v4.0). A GSEA determines the significance of a pre-defined gene set by first calculating the correlation between the expression of genes inside the gene set and the class distinction, then comparing against noisy populations generated through random permutations [71]. We carried out 1000 random gene set permutations to create noise and set the significance threshold as FDR<0.1. The GSEA results indicated that infected spleen had enrichment of up-regulated genes associated with the inflammatory response: chemokines cell migration, cell proliferation, cell cycle and mitochondrial metabolism. Enrichment of down-

regulated genes associated with tissue morphogenesis and structure, receptor-mediated signaling and extracellular matrix, was also observed (Figure 4.6).

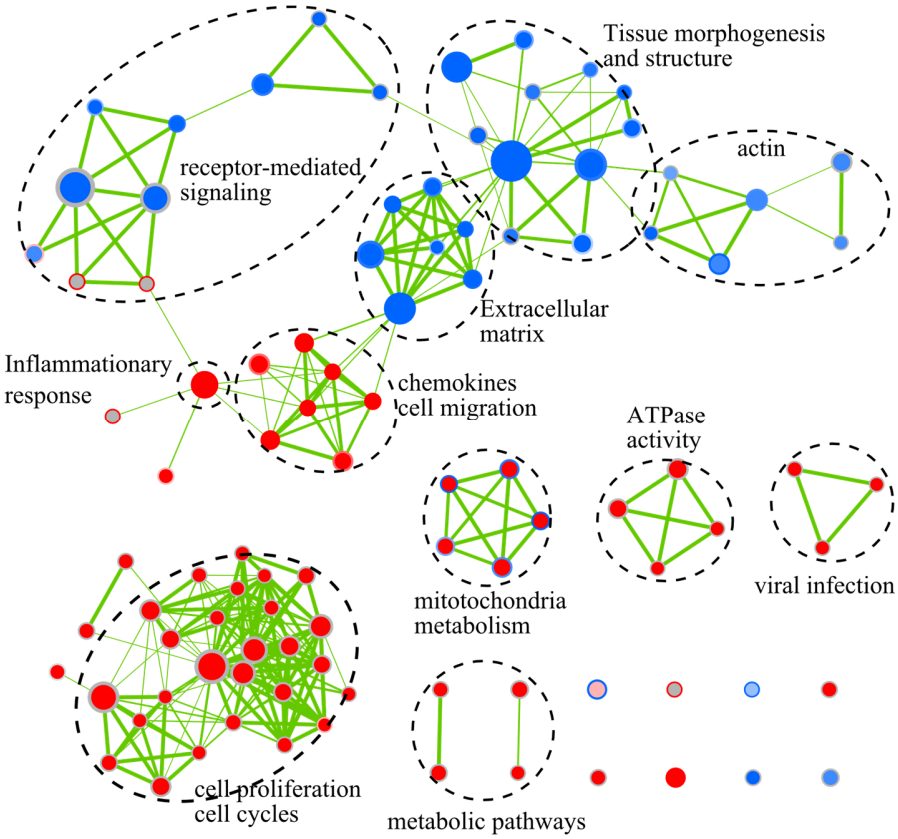

Figure 4.6 GSEA results in hamster VL. We visualized GSEA results using the software Cytoscape and Enrichment Map. The cut off was *p-value* < 0.1. The analysis results of hamster spleen and splenic adherent cell were indicated as the inner and outer of the cycles, respectively. Each cycle represented a gene set. The gene sets enriched with up-regulated and down-regulated genes were colored in red and blue, respectively. The grey color means the gene set was not significant.

4.3.3 Expression of leishmania reads in hamster host

We next investigated the expression levels of Leishmania genes during infection. We mapped the raw RNA-Seq reads against the Leishmania *donovani* BPK282A1

genome (NCBI BioProject PRJEA61817). The infected spleen and adherent cells samples were highly enriched with Leishmania transcripts (up to 61 and 7 folds, respectively), when compared to the uninfected controls (Table 4.1). This confirmed the infectivity of the treatment group.

	Infected Samples				Uninfected Controls			
Spleen	1.22	0.98	0.84	0.74	0.02	0.02	0.02	0.02
Adherent	0.31	0.20	0.17	0.16	0.04	0.04	0.04	0.03

Table 4.1 Successful alignment rates (%) mapped to Leishmania genome.

The results showed that whole spleen tissue had higher parasite reads than the splenic adherent cells. We can explain this phenomenon mainly by two reasons. First, other cell populations present in the whole spleen are also target cells for parasite infection like neutrophils and fibroblasts [59, 72]. Second, even though macrophages are the main target cell of the parasite, we have observed that high parasite burden inhibits macrophage adherence (Melby PC et al., personal communication). Thus, highly infected macrophages are less adhesive; therefore the detection of parasites in the adherent cells is decreased. The observation revealed that sequences from hamsters shared high homology with Leishmania sequences, such as actin, tubulin, calmodulin and unc104-like kinesin. All contributed to the fact that the uninfected controls failed to have a successful alignment rate of 0.

Further analysis of Leishmania sequences from infected samples allowed us to identify unique parasite genes expressed during infection (Table 4.2). Many of these genes encode for proteins that have been characterized as pathogenic factors or tested as vaccine / diagnostic candidates. The amastin-like protein, 60S ribosomal protein L22, and 40S ribosomal protein S19 were evident to protect against murine Leishmania major

using the vaccine screening approach. The vaccine for Leishmania histone $H1^{+}$ dendritic cells leads to a protective phenotype in murine visceral leishmaniasis. The recombinants of H2A, H2B, H3 and H4 proteins were highly immunogenic and offered optimum prophylactic efficacy against Leishmania challenge in hamsters. The elongation factor 1-alpha (ef-1-alpha) and the activated protein kinase c receptor (LACK) have been proposed as targets for drug or vaccine development. The cathepsin L-like protease is a potential diagnostic maker. The delivery of kinetoplastid membrane protein-11 or cysteine peptidase C with nanoparticles induces parasite killing or protective immunity against infection. All the discussed proteins appeared in our list as products of the highly expressed parasite genes. Thus, identification of these highly expressed parasite genes opens an opportunity to study their role during infection and their potential usage as vaccine candidates or diagnostic markers.

Rank	Protein Id	mRNA product	Vaccine Candidate Ref
1	CBZ31914.1	amastin-like protein	C.B. Stober, et al., 2006
2	CBZ33166.1	elongation factor 1-alpha	M. Lopes, et al., 2007
3	CBZ33929.1	histone H2A, putative	R.K. Baharia, et al., 2014
4	CBZ33928.1	histone H2A	R.K. Baharia, et al., 2014
5	CBZ34196.1	hypothetical protein LDBPK_191680	
6	CBZ38090.1	60S ribosomal protein L5, putative	
7	CBZ32099.1	histone H2B	R.K. Baharia, et al., 2014
8	CBZ35700.1	activated protein kinase c receptor (LACK)	S. Sinha, et al., 2013
9	CBZ31585.1	histone H4	R.K. Baharia, et al., 2014
10	CBZ35841.1	ribosomal protein L1a, putative	
11	CBZ31772.1	60S ribosomal protein L7a, putative	
12	CBZ35267.1	histone H1, putative	M. Agallou, et al., 2012
13	CBZ34604.1	40S ribosomal protein S8, putative	
14	CBZ31937.1	cathepsin L-like protease, partial	P.A. Ortiz, et al., 2009
15	CBZ32384.1	ATP-binding cassette protein subfamily A, member 2, putative	
16	CBZ34112.1	hypothetical protein LDBPK_220670, partial	
17	CBZ32611.1	40S ribosomal protein S4, putative	
18	CBZ31625.1	60S ribosomal protein L19, putative	
19	CBZ31909.1	hypothetical protein, unknown function	
20	CBZ36778.1	40S ribosomal protein S2	
21	CBZ32210.1	histone H3	R.K. Baharia, et al., 2014
22	CBZ35406.1	unnamed protein product	
23	CBZ33040.1	histone H3, putative, partial	R.K. Baharia, et al., 2014
24	CBZ37966.1	60S ribosomal protein L18a, putative	
25	CBZ36392.1	ribosomal protein L15, putative	
26	CBZ37821.1	60S ribosomal protein L21, putative	
27	CBZ34037.1	60S ribosomal protein L11 (L5, L16)	

28	CBZ38122.1	60S ribosomal protein L12, putative	
29	CBZ38781.1	60S ribosomal protein L22, putative	C.B. Stober, et al., 2006
30	CBZ38642.1	nucleoside transporter 1, putative	
31	CBZ34905.1	60S ribosomal protein L7, putative	
32	CBZ37237.1	40S ribosomal protein S3, putative	
33	CBZ38740.1	40S ribosomal protein S24e	
34	CBZ37952.1	40S ribosomal protein S3A, putative	
35	CBZ32545.1	40S ribosomal protein S12, putative	
36	CBZ38281.1	60S ribosomal protein L27A/L29, putative	
37	CBZ38910.1	60S ribosomal protein L18, putative	
38	CBZ38434.1	polyadenylate-binding protein 1, putative	
39	CBZ33941.1	60S ribosomal protein L9, putative	
40	CBZ32928.1	tryparedoxin peroxidase	C. Carson, et al., 2009
41	CBZ32374.1	60S ribosomal protein L28, putative	
42	CBZ37820.1	hypothetical protein, pseudogene	
43	CBZ38830.1	60S ribosomal protein L34, putative	
44	CBZ37649.1	amastin-like surface protein, putative	
45	CBZ38504.1	ubiquitin/ribosomal protein S27a, putative	
46	CBZ37739.1	40S ribosomal protein S19 protein, putative	C.B. Stober, et al., 2006
47	CBZ34752.1	ribosomal protein S25	
48	CBZ38124.1	kinetoplastid membrane protein-11	D.M. Santos, et al., 2013
49	CBZ38105.1	60S ribosomal protein L32	
50	CBZ38283.1	60S ribosomal protein L23, putative	S. Das, et al., 2013
51	CBZ34076.1	40S ribosomal protein S15, putative	
52	CBZ37471.1	40S ribosomal protein S13, putative	
53	CBZ37012.1	ribosomal protein L27, putative	
54	CBZ31131.1	ribosomal protein S7, putative	
55	CBZ37701.1	ribosomal protein l35a, putative	
56	CBZ36360.1	60S ribosomal protein L9, putative, partial	
57	CBZ36811.1	RNA binding protein, putative	
58	CBZ34392.1	60S ribosomal protein L17, putative	
59	CBZ35110.1	ribosomal protein L38, putative	
60	CBZ32381.1	40S ribosomal protein S15A, putative	
61	CBZ38543.1	40S ribosomal protein S10, putative, partial	
62	CBZ35849.1	tryparedoxin	
63	CBZ38569.1	40S ribosomal protein S9, putative	
64	CBZ35510.1	40S ribosomal protein S14	
65	CBZ32527.1	hypothetical protein, conserved	
66	CBZ34201.1	peroxidoxin	
67	CBZ35641.1	ribosomal protein S29, putative	
68	CBZ38652.1	chaperonin HSP60, mitochondrial precursor	
69	CBZ31166.1	eukaryotic initiation factor 4a, putative	
70	CBZ31415.1	nascent polypeptide associated complex subunit-like protein, copy 1	
71	CBZ31498.1	trypanothione reductase	
72	CBZ34218.1	endoribonuclease L-PSP (pb5), putative	
73	CBZ32640.1	60S ribosomal protein L44, putative	
74	CBZ34225.1	ABC-thiol transporter, partial	
75	CBZ34704.1	eukaryotic initiation factor 5a, putative	
76	CBZ34724.1	cyclophilin a	
77	CBZ36377.1	S-adenosylmethionine synthetase	
78	CBZ37651.1	amastin-like surface protein, putative, partial	
79	CBZ36345.1	ribosomal protein S26, putative	
80	CBZ38972.1	40S ribosomal protein SA, putative, partial	
81	CBZ34101.1	3'a2rel-related protein	
82	CBZ38544.1	40S ribosomal protein S10, putative	
83	CBZ32803.1	myo-inositol-1-phosphate synthase	
84	CBZ38833.1	basic transcription factor 3a, putative	
85	CBZ33102.1	60S ribosomal protein L39, putative	
86	CBZ38793.1	ribosomal protein L29, putative	
87	CBZ31130.1	ribosomal protein S7, putative, partial	

88	CBZ39091.1	glucose transporter, lmgt1	
89	CBZ33891.1	60S Ribosomal protein L36, putative	
90	CBZ36113.1	surface protein amastin, putative	
91	CBZ35810.1	cysteine peptidase C (CPC)	D. Doround, et al., 2011
92	CBZ35049.1	40S ribosomal protein S33, putative	
93	CBZ32934.1	60S acidic ribosomal protein P2	
94	CBZ33458.1	P-type H+-ATPase, putative	
95	CBZ34602.1	60S ribosomal protein L26, putative, partial	
96	CBZ32363.1	pyruvate phosphate dikinase, putative	
97	CBZ38570.1	fructose-1,6-bisphosphate aldolase	
98	CBZ33614.1	oxidoreductase-like protein	
99	CBZ32061.1	elongation factor-1 gamma	
100	CBZ36403.1	aquaglyceroporin	

Table 4.2 Top 100 parasite genes specifically expressed in the infected samples

4.4 RESULTS AND DISCUSSION

Our objective in the current project is to understand the VL pathogenesis mechanisms. The Syrian golden hamster was used as an animal model because it closely mimics the chronic visceral leishmaniasis in human. In this section, we will discuss our findings from the hamster model. The analysis is on-going so the knowledge we can discover from the current project is not limited to what we discussed below.

4.4.1 Spleen adherent cells are enriched with macrophages

Macrophages are the main target cells for the Leishmania parasite. The number of splenic macrophages increases during chronic VL (data not shown) leading to a disease-promoting phenotype [73-75]. We were interested in the gene alteration specific to macrophages in addition to the whole spleen tissue. Currently, no antibody for hamster macrophages is available for purification. We isolated splenic adherent cells from whole spleen cells using a plastic culture dish. The splenic adherent cells are believed to be macrophage enriched. To assure that the adherent spleen cells are splenic macrophages, we examined the adherent spleen cells of their morphology and their expression of cell lineage markers. The adherent spleen cells had typical macrophage morphology and were

positive for the intracellular macrophage marker CD68 determined by immunohistochemistry (data not shown). We further compared the expression levels of several key markers of specific cell lineage between control splenic adherent cells and the control whole spleen cells (Figure 4.7). The results indicated macrophage related markers were highly enriched in the splenic adherent cells. Conversely, the markers for other cell populations including T cells, B cells, neutrophils, dendritic cells and fibroblasts, were all highly enriched in the whole spleen cells with prolyl 4-hydroxylase, beta polypeptide (P4HB) as one exception. P4HB is a fibroblast marker. Its enrichment in the adherent spleen cells is most likely the consequence of expression by inflammatory macrophages [76, 77], but we cannot exclude the possible presence of a small number of fibroblasts among the adherent macrophages. Collectively, these data indicated that the splenic adherent cells were splenic macrophages.

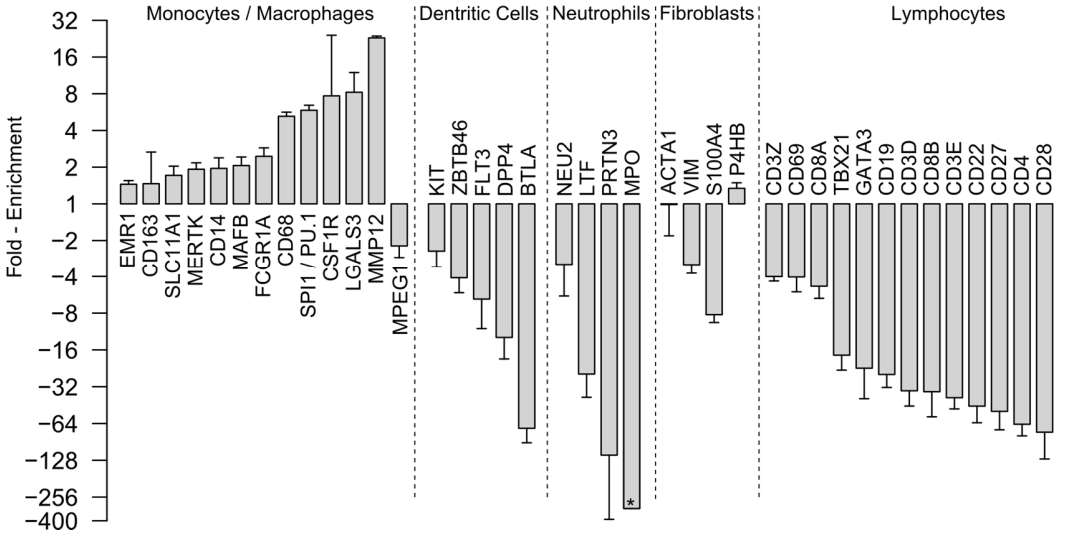

Figure 4.7 Cell lineage check in hamster samples. We plot the enrichment (i.e. CPM) fold changes of markers of several different cells by comparing hamster splenic adherent cell samples to spleen samples. (*: the CPM expression value in hamster splenic cells was 0)

4.4.2 Highly proinflammatory environment in experimental VL

A common feature shared by the majority of the top canonical pathways in spleen and splenic macrophages was the up-regulation of inflammatory cytokines, chemokines and their receptors. The significance of chemokines cell migration was also confirmed by GSEA. At least four out of the top ten enriched gene sets in spleen and splenic macrophages were associated with production, signaling and receptor activity of cytokines and chemokines. We therefore examined the differentially expressed cytokines, chemokines, and their receptors (Figure 4.8).

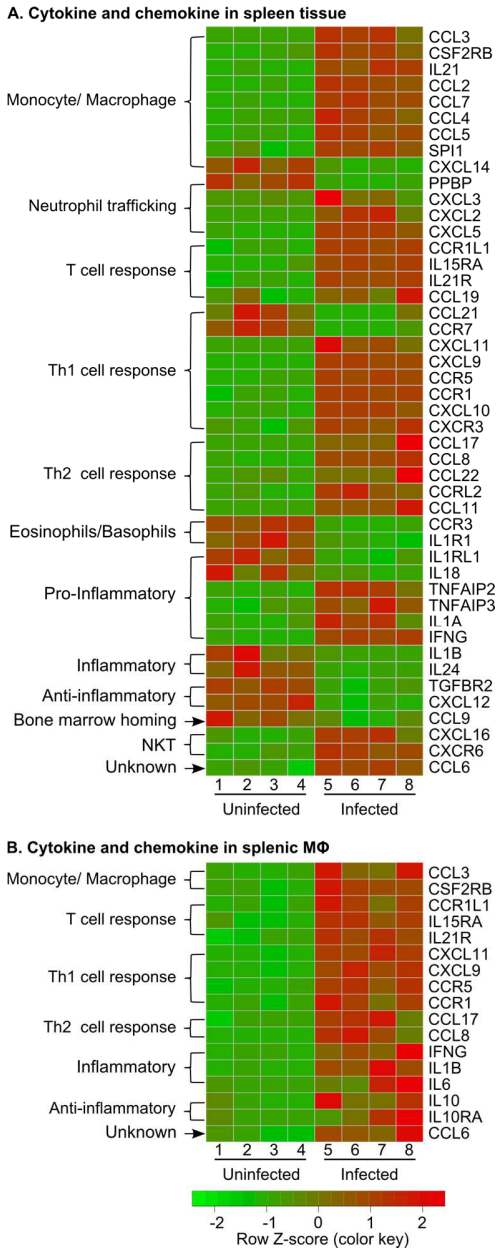

Figure 4.8 Regulation of cytokines and chemokines. Heat map of cytokine, chemokines and receptors in hamster spleen (A) and splenic macrophage (B) samples.

In the whole spleen tissue, we found that many pro-inflammartory genes (e.g. TNFAIP2, TNFAIP3, IL1α, and IFNγ) were up-regulated and many anti-inflammaroty genes (e.g. il1ß, and IL24) were down-regulated during VL. Additionally, many

transcription factors that drive inflammation were up-regulated or predicted to be actived during infection. The up-regulated transcription factor mRNAs included STAT1, STAT2, STAT3, IRF1, IRF7, TBX21, XBP1, LITAF and MHC (XBP1, NLRC5). All are involved in interferon signaling and cytokine responses. Additionally, many transcription factors related to inflammatory response were predicted to be activated in the infected spleen tissue, including NFκB complex (REL, RELA, RELB), NFATC2 (T cell activation), STAT4, IRF3/5, IFI16, HMGB1, BCL10 (NFκB activator), CBP/P300, and DDIT3 (caspase activation, cytokine expression). As a result, we concluded the spleen environment was enriched with inflammatory signal during chronic VL.

Notably, all differentially expressed inflammatory genes in splenic macrophages were up-regulated during VL (Figure 4.8B). This finding distinctly contrasts with data from *in vitro* infected mouse [78, 79] and human [80] macrophages. The published studies indicated that Leishmania infection had a silent or suppressive, rather than activated, effect on macrophage inflammatory gene expression. The difference is possibly the result of the complex inflammatory signal environment of the whole spleen, which would be absent from *in vitro* infected macrophages.

The spleen environment has considerable influence on the activation status of splenic macrophages. In particular, the increased expression of the macrophage-activating cytokines IFNγ, IL1α, TNF, and IL-21 are likely to play a key role in determining the macrophage phenotype. The splenic macrophages from infected animals also showed increased expression of toll-like receptor-4 (TLR4) and the cytokine receptors IL-15Rα, CSF2Rβ/IL-5Rβ (common subunit of the IL-3, IL-5, and GM-CSF receptors) and IL-21R that would amplify the effect of the proinflammatory environment.

4.4.3 Chemokines associated with myeloid cells migration

The immune cells are critical for the generation of innate and adaptive immune responses. The spleen environment was enriched in chemokines and their receptors, which could influnce the immune cells of their migration and positioning [81].

The chemokines CCL2, CCL3, CCL4, CCL5, CCL6 and CCL7 were highly expressed in spleen during VL. They act to recruit monocytes/macrophages. Meanwhile, the receptors of these chekmokines including CCR1 and CCR5 were significantly increased in the spleen and splenic macrophages. They could contribute to the accumulation of monocytes / macrophages. The macrophage recruitment capability of CCL2 had been experimentally demonstrated in mice infected with Leishmania *chagasi* [82]. The effects of these chemokines can be further amplified by the T cell response in infected animals [83]. We experimentally identified an increase in macrophages in the spleen during VL (see section 4.5.1a).

Neutrophile could also be recruited by the increased expression of chemokins CXCL2, CXCL3, CXCL5, and CCL3 in the spleen and/or splenic macrophages (Figure 4.8A). The up-regulated eosinophil chemoattractant CCL11 in the infected spleen may also contribute to the recruitment of neutrophils. When neutrophils quickly localize to the site of Leishmania inoculation, they phagocytose and kill the parasite, or become apoptotic. The apoptosis could promote the subsequent infection of resident or inflamatory macrophages [84, 85]. The signalings of neutrophil chemoatractants and recruitment were activated. However, we did not find any significant increase in neutrophils in the spleen during VL (data not shown). Therefore, neutrophils are more

likely to become apoptotic if they could migrate to the infected spleen, and further promote the parasite infection.

Additionally, chemokines with roles in Th1 (CXCL9, CXCL10, CXCL11), and Th2 (CCL8, CCL17, CCL22) cell recruitment were significantly increased in the infected spleen [81]. CXCL16 and its receptor CXCR6, known to recruit NKT cells and innate lymphoid cells, were significantly up-regulated in the spleen during VL.

The findings above demonstrate that the diverse chemokine expression in infected spleen contribute to the accumulation of immune cells in the spleen. Additionaly, we found enriched and/or up-regulated transcription factors (Egr2, SFPI1, IRF8 and AP1) in the splenic macrophage, which promote myelopoiesis [86, 87]. Therefore, the local environment may also contribute to the accumulation of myeloid cells in the spleen.

4.4.4 Mixed polarized/activated splenic macrophages

Macrophages have dual roles during Leishmania infection: mediating parasite killing and controlling tissue damage and repair [88]. Macrophages exhibit considerable plasticity in their activation state, which depends on cues received from the local environment [89]. Distinct polarization states are evident when purified macrophage populations are exposed to defined activation stimuli [90]. At the extremes of the polarization spectrum, M1 macrophages are important for the clearance of intracellular pathogens including Leishmania parasites [91], while M2 macrophages are protective against helminths and have anti-inflammatory and tissue repair functions [88, 92]. However, accumulating evidence indicates that in complex biological systems the polarization of macrophages does not always fit neatly within the dichotomous M1-M2 classification system [90].

We determined the splenic macrophage activation phenotype by evaluating the expression of associated genes. Splenic macrophages from hamsters with VL had a significantly increased expression of genes characteristic of both M1 (CXCL9, CXCL11, IL1B, IL6, FCGR1A, IDO, IRG1, IFNγ, STAT1, CCL3, CCL5) (Figure 4.9) and M2 (see below) polarization. The fold changes of the M1-associated genes were generally higher than those of M2-associated genes. Moreover, the majority of M1-associated genes were up-regulated and less than a half of M2-associated genes were up-regulated. All of these data suggested splenic macrophages had an M1 dominant phenotype. The up-regulation of IL-1β, IFNγ and IL-6 in splenic macrophages (Figure 4.9) indicated an additional inflammatory effect on the macrophages through paracrine or autocrine activation. IFNγ, IL1β and IL6 are supposed to increase activation of other M1 markers such as NOS2, CXCL13, which, in fact, were not significantly expressed in the infected spleen or the splenic macrophages.

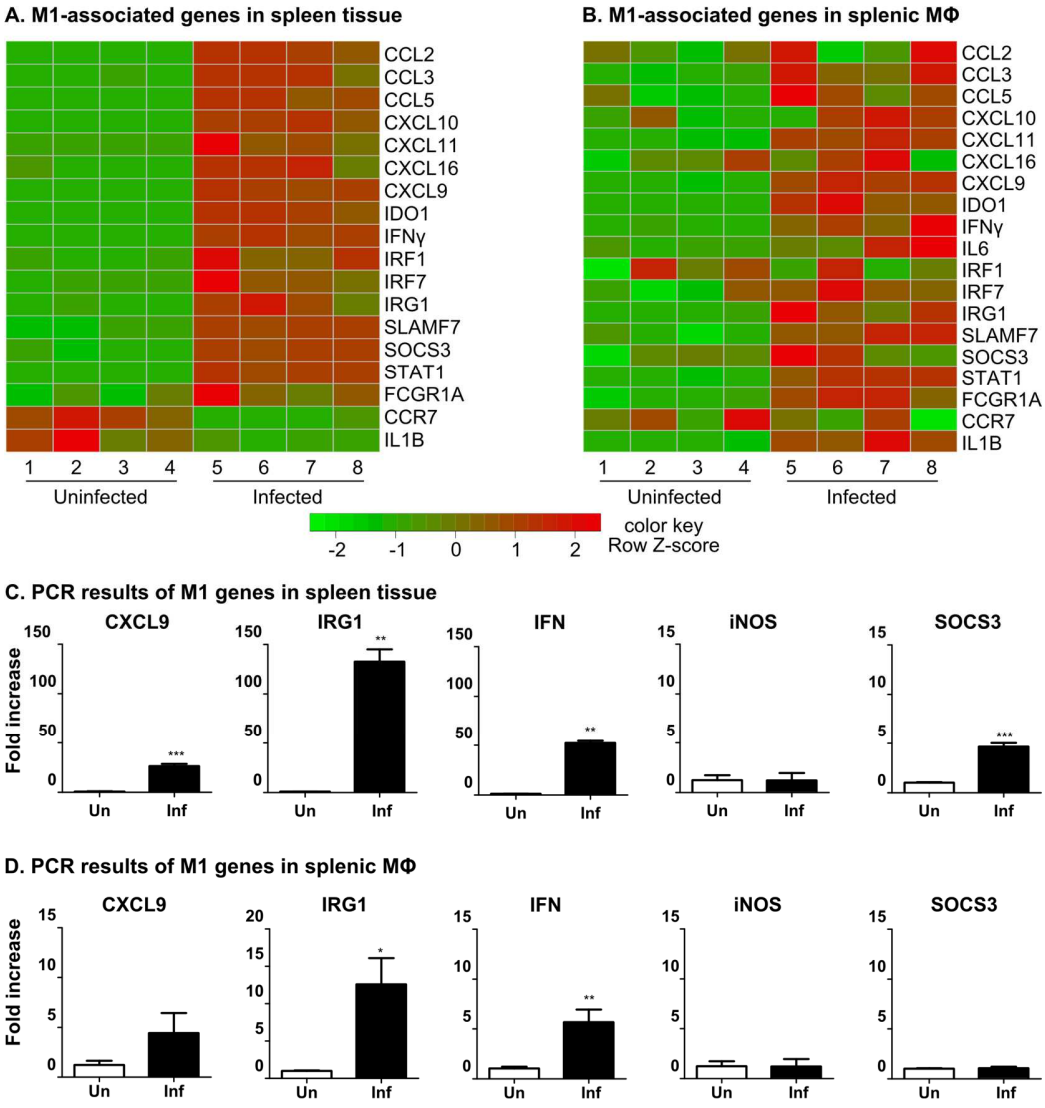

Figure 4.9 Regulation of M1 genes. We showed heat maps for M1 genes in hamster spleen (A) and splenic macrophage (B) samples. The regulations of selected M1 genes were further confirmed using PCR in spleen (C) and splenic macrophage (D).

The splenic macrophages also showed increased expression of M2-associated transcripts, including Arg1, IL-10, SOCS2, CCL17, and Chi3L1 (Figure 4.10). The Arg1 expression can be driven by IL-4, growth factor receptor signaling and parasite-induced STAT6 activation [74, 93]. The induction of arginase is likely to be a result of the

increased expression of IL-10 and IL-10R in splenic macrophages, which could suppress T cell or macrophage effector function [94, 95]. All of these functioned to dampen the effects of the proinflammatory cytokines and promote infection. In contrast, some M2-associated markers, such as CD163 and MSR1 (Figure 4.10), were down-regulated. The angiogenic factors (VEGFA, EPHB1/4, DLL4, LYVE1, ANGPT1, NRP1) displayed no up-regulation in splenic macrophages, which are characteristic of M2 activation. These findings suggest that the splenic macrophages were only partialy M2-like macrophages.

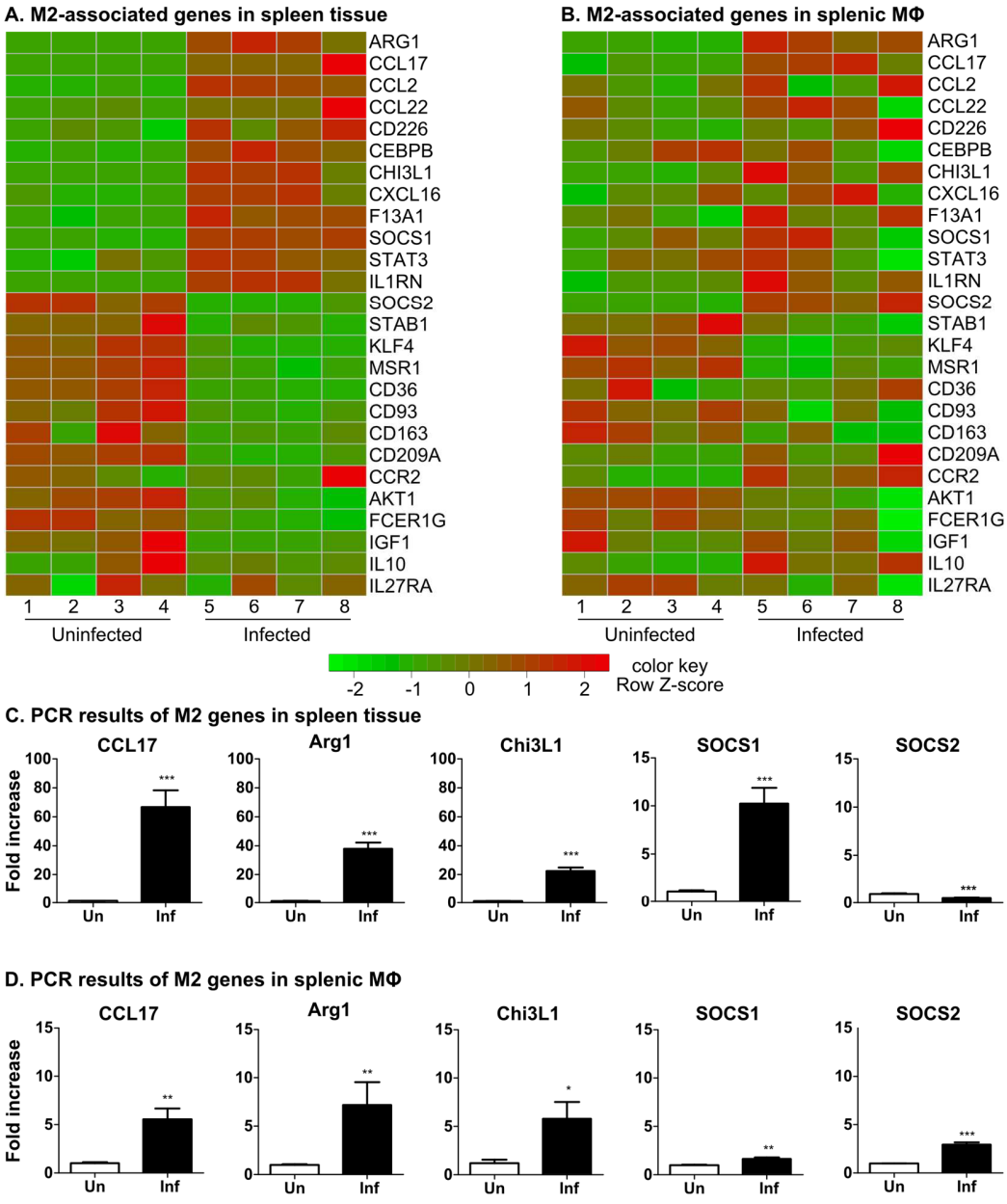

Figure 4.10 Regulation of M2 genes. We showed heat maps for M2 genes in hamster spleen (A) and splenic macrophage (B) samples. The regulations of selected M2 genes were further confirmed using PCR in spleen (C) and splenic macrophage (D).

In some cases M1 and M2 markers were discordant between the whole spleen tissue and splenic macrophages. An example of this is the SOCS family of proteins that

regulate macrophage polarization [96]. SOCS1, which is a critical determinant of IL-4-induced M2 polarization [97], was up-regulated in both the infected spleen tissue and splenic macrophages. However, SOCS2, which is also associated with M2 responses [96], was down-regulated in infected spleens but up-regulated in splenic macrophages. SOCS3, an M1-associated regulator [96], was induced in spleen tissue but not in splenic macrophages.

All the above data suggested that the hamster splenic macrophages, after 28 days of Leishmania infection, were characterized with a mixed M1 and M2 phenotype. We also performed RNA view experiments to examine this hypothesis. We utilized IDO1 and CXCL9 as markers for the M1 macrophage activation and Arg1 as the M2 marker. The results showed the splenic adherent cells had all four different combinations: double negative, single positive for M1 or M2 marker, and double positive. This finding is consistent with our hypothesis that splenic macrophages presented mixed polarization and activation during VL.

4.4.5 Regulators of splenic macrophage polarization in VL

To better understand the macrophage polarization signaling, we investigated the M1/M2-relevant specific upstream transcription factors that were predicted by IPA to be activated or inhibited.

In the whole spleen tissue, many transcription factors associated with either M1 and/or M2 were up-regulated and also were predicted to be activated, including TBX21, STAT2, IRF7, IRF1, STAT1, STAT3 and so on. Some down-regulated transcription factors were predicted to be inhibited including NKX2-3, KLF4 and KDM58. We also found some up-regulated transcription factors were predicted to be inhibited, including

CDKN2A, SARCB1, RBL1, and NUPR. Additionally, we identified many transcription factors without differentially expressed genes that were also predicted to be activated or inhibited. We examined the relations among all of the differentially expressed transcription factors and M1/M2 genes (Figure 4.11). The results indicated that IRF7 was connected only to genes associated with M1 activation but many transcription factors had dual roles in the regulation of the M1 and M2 genes. The most dominant hubs regulating both M1 and M2 markers were STAT1 and STAT3. All were up-regulated and also predicted to be activated.

STAT1 and STAT3 both are members of the signal transducers and activators of the transcription family of transcription factors. STAT1 typically promotes inflammation and inate immunity through poloarizing macrophages into M1 phenotype. STAT3 is believed to enhance cell proliferation, motility and immune tolerance, through polarizing the macrophage into M2 phenotype. Although the functions of STAT1 and STAT3 appear oppose each other, they can be activated by or can themselves activate many common cytokines and growth factor receptors. For example, both Type I and II IFNs have STAT1 as a central mediator and have the ability to acitivate STAT3. Both STAT1 and STAT3 can be actiavted by IL-6. The common downstream regulated M1/M2 genes for STAT1 and STAT3 include CCL5, IFNγ, FAS, IL1B, PSMB9, TNSF10, CXCL10, SOCS3, SOCS1, OASL, CCL2, AKT1, FCER1G, ARG1, and IL1R1. The balance of STAT1 and STAT3 in the expression or phosphorylataion levels may switch cytokine/growth factors between inflammatory and ani-inflammatory, leading to the mixed phenotype of macrophages.

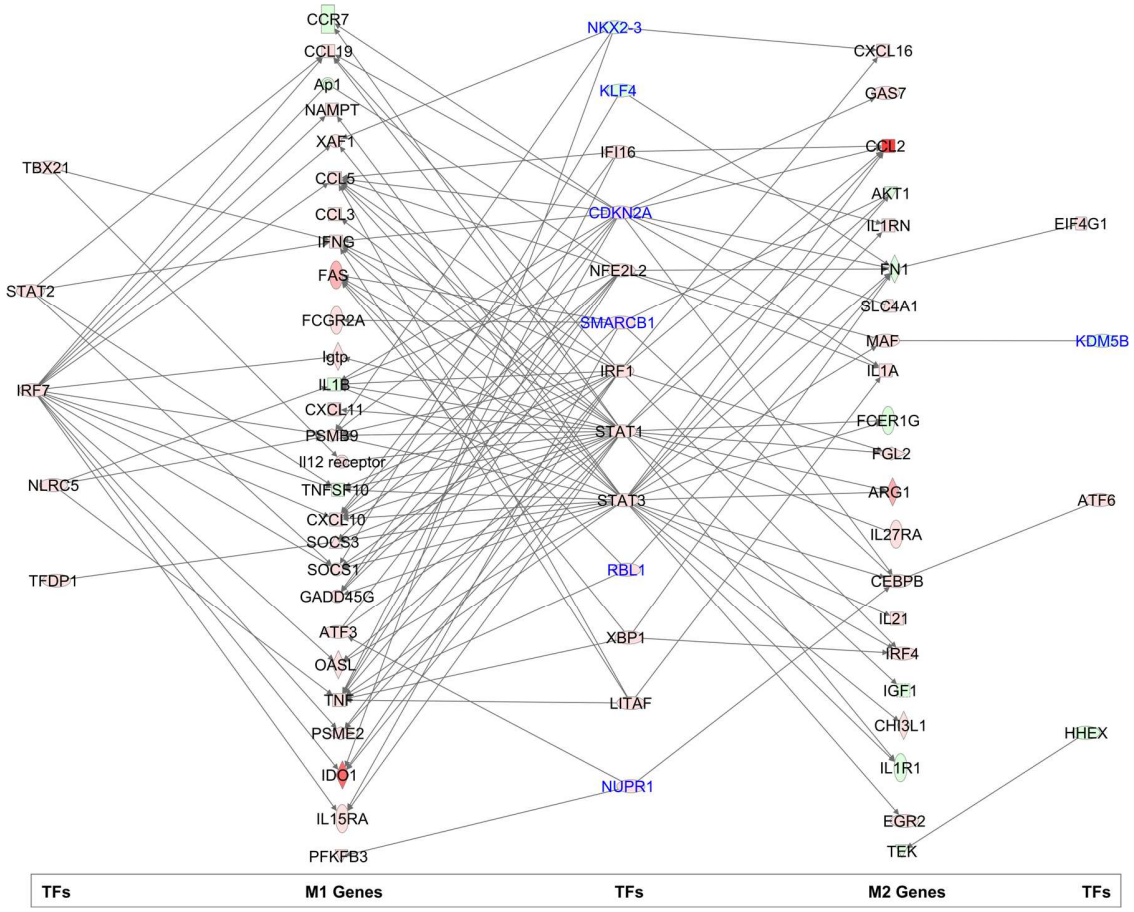

Figure 4.11 Activities of transcription factors. We examined M1 and M2 genes, and the associatied transcription factors (TFs) in spleen. The transcription factors are colored in blue if they were predicted to be inhibited. Otherwise, they were predicted to be activated. Genes colored in red or green indicated whether they were up-regulated or down-regulated during VL.

4.4.6 Role of IFNγ in macrophage polarization

Consistent with previous studies in the hamster [75, 98, 99] and human model [100], IFNγ was highly up-regulated in the infected spleen tissue (FC=52.15, FDR<0.001) and splenic macrophages (FC=11.12, FDR<0.001). Many well-known IFNγ-responsive genes (e.g. CXCL9, CXCL10, CXCL11, IDO, IRG1) (Figure 4.9A)

were also up-regulated in VL spleen. These up-regulated M1-associated genes normally lead to activation of NOS2, which was notably absent in the VL hamster. This suggested a specific deficit in macrophage effector function in progressive VL [99, 101].

IFNγ is mainly secreted by T cells and NK cells, but murine and human macrophages may express IFNγ in response to IL12, IL18, LPS, IFNγ, *M. tuberculosis,* and *Streptococcus pyogenes* [102-106]. The IFNγ-producing macrophages were characterized as immature myeloid cells with protective ability against in vivo *S. pyogenes* infection [105]. T cell-derived IFNγ in whole spleen tissue aims to control this parasitic infection. As a positive regulatory loop, the T cell-derived IFNγ may induce the splenic macrophages to produce IFNγ. Without NOS expession, the IFNγ-producing macrophages may take on an anti-inflammatory or immunosuppressive function.

Indoleamine 2,3 –deoxygenase (IDO), a IFNγ-induced M1-associated gene, was highly up-regulated in the whole spleen (FC=368.73, FDR<0.001) and splenic macrophages (FC=39.87, FDR<0.001). The increased expression of IDO may increase T tell tolorance [107], suppress host adaptive immunity such as anti-leishmanial T cell response [108] and polarize macrophages into M2 phenotype [109]. Thus the high level of IFNγ expression in VL would be expected to promote the development of macrophage and the killing of parasites, but it may paradoxically promote parasite growth and survival by inducing IDO1. Additional research is needed to understand the regulatary mechanisms.

Immunoresponsive gene 1 (IRG1), another IFNγ induced gene, was also highly up-regulated in the VL whole spleen (FC=365.35, FDR<0.001) and splenic macrophage (FC=7.77, FDR<0.001). IRG1 regulates the fatty acid β-oxidation required for

mitochondrial ROS production [110]. The high expression of IRG1 suggests macrophage metabolism in VL is skewed toward the use of fatty acid oxidation. Peroxisome proliferatior-activated receptor gamma coactivator 1 beta (PPARGC1B) (FC=2.52, FDR=0.008), and acyl-Coenzyme A dehydrogenase (ACADL) (FC=1.49, FDR=0.001) were up-regulated in the spleen. Both can induce the macrophage program of fatty acid oxidation [111]. M2 macrophages in murine, but not in human, use fatty acid oxidation to meet their metabolic needs [112, 113]. The metabolic alteration caused by IRG1 may also promote the M2 macrophage polarization.

4.4.7 Dysregulated tissue repair mechanisms in VL

Fibrosis was observed during chronic VL in humans [114] and dogs [115]. The fibrosis pathway (hepatic firobsis) was highly represented in both the spleen tissue and the splenic macrophages (Figure 4.5). GSEA also revealed a cluster of collagen and extracellular matrix related gene sets enriched in the uninfected control samples (Figure 4.6). We found many genes, related to tissue repair, remodeling and fibrosis, were differentially expressed in both spleen tissue and splenic macrophages (Figure 4.12). Pro-fibrogenic genes, such as TIMP1, CCL2, CCL3, CCL11, CCL12 and Chi3L1 [116-118], were domaintly up-regulated during infection. However, genes associated with fibrosis, such as COLA1, IGFBP3, PDGFRβ; MMP2; IGF-1, FGFR1, VEGF and TGFB-R2, [119-125], were down-regulated during infection. The anti-fibrotic gene MMP9 [126] was highly up-regulated during infection in the spleen. These findings suggested that the spleen displayed no fibrosis at this point during infection (28 days) but with great potential to form fibrosis. Further examination showed that the fibrosis related

proinflammatory cytokines were up-regulated in the VL spleen and splenic macrophages, which may also contribute to the formation of fibrosis in the spleen.

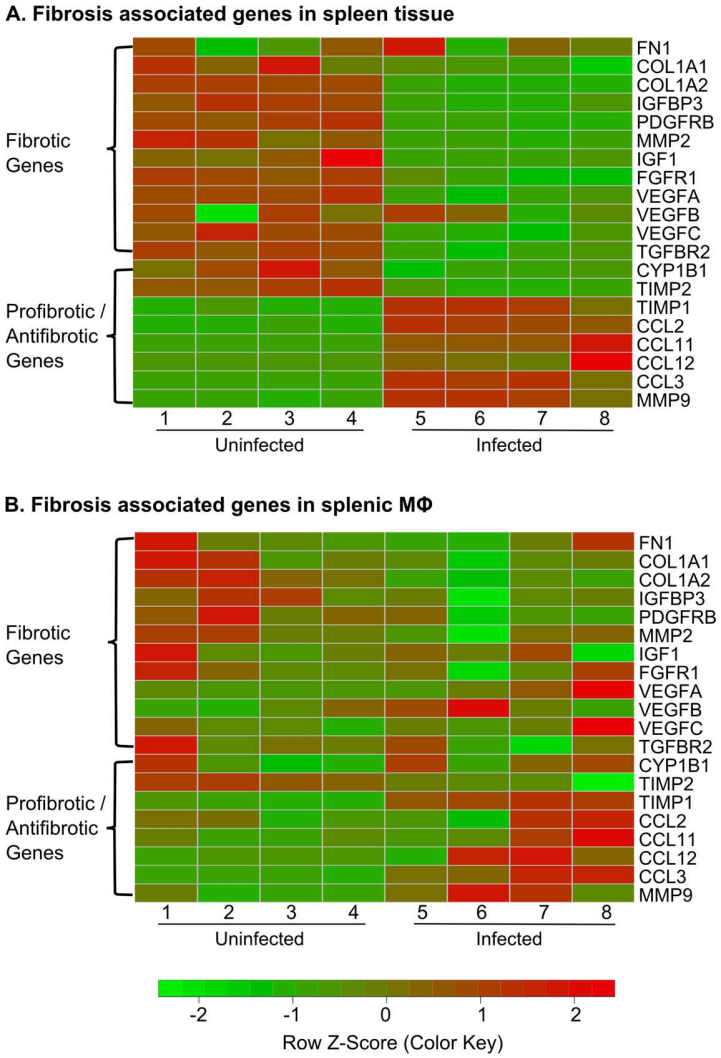

Figure 4.12 Regulation of fibrosis-related genes. We show heat maps for the fibrosis-related genes in hamster spleen (A) and splenic macrophage (B) samples during VL.

4.4.8 Suppressed glucocorticoid receptor signaling in VL

The glucocorticoid receptor (GR) signaling pathway was the most highly represented pathway in the splenic macrophages, which was also enriched in whole

spleen tissue (Figure 4.5). This pathway in the splenic macrophage was characterized by the up-regulation of the heat shock protein family members (HSP90, HSP70) and the down-regulation of DUSP1 (dual specificity phosphatase 1) and SGK1 (Serum and GC regulated kinase) (Fig 4.13). HSP90/70, together with several other proteins, form a heterocomplex that is essential for steroid binding [127]. DUSP1 is the canonical MAPK phosphatase. We identified down-regulation of MAP3K and MAP3K14 in the infected spleen. The MAPKs induced expression of inflammatory mediators [128, 129] may further decrease the anti-inflammatory effects of the GC pathway. Moreover, the anti-inflammatory glucocorticoids have been associated with M2 alternatively activated macrophages [130]. These finding may partially explain why the splenic macrophages appeared to be M1 phenotype dominant (see section 4.5.1d).

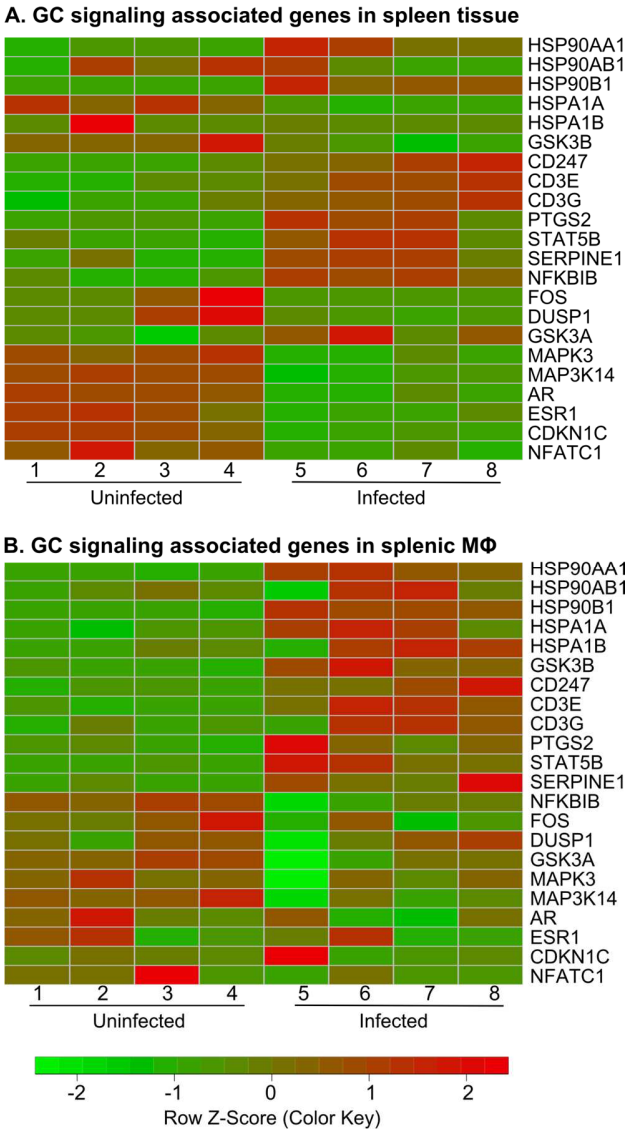

Figure 4.13 Regulation of GR signaling pathway. We show heat maps for the genes related to GR signaling in hamster spleen (A) and splenic macrophage (B) samples during VL.

4.5 CONCLUSION AND LIMITATIONS

In the current project, we have utilized the NGS-based approaches to study VL in Syrian golden hamster. We assembled a *de novo* splenic transcriptome for the hamster and then performed RNA-Seq differential expression analysis. The hamster system

showed dually activated macrophages and broad inflammatory nature. These were expected, but failed, to control the L. *donovani* parasite. These studies provide us many interesting directions for further research on VL.

Across the entire experiment, we had several limitations. First, the RNA-Seq experiments study the organism at the transcriptome level without considering the post-translation modification as discussed in Chapter 3. Second, a strand-specific sequencing approach were not implemented [131]. Third, although adherent splenic cells are enriched with splenic macrophages, their properties may not be representative of the macrophages infected with Leishmania parasites. The acknowledgement and understanding of these limitations will benefit our future research.

Chapter 5. Computational-Aided VEEV Live Attenuated Vaccine Design [3]

5.1 INTRODUCTION

Vaccines are the most cost-effective agents to control and prevent viral infectious diseases [132]. The development of vaccines, unfortunately, is very time consuming and is characterized by low success rates. The average time to develop a vaccine is about 11 years, and even worse, the average probability that it will ever enter the market is only about 6% [133]. It is thus a critical challenge to shorten the preclinical process and increase the success rate, especially considering that many existing and emergent infectious diseases have no useful vaccines [134].

In the current project, we utilized bioinformatics tools to accelerate vaccine development for the Venezuelan Equine Encephalitis Virus (VEEV). VEEV is a positive-sense single-stranded RNA arbovirus in the family Togaviridae, genus Alphavirus [135]. The genome length is 11.4kb including two reading frames that encode for two different polyproteins. The first polyprotein includes four non-structured proteins: nsP1 (negative strand RNA synthesis, RNA capping), nsP2 (helicases, proteinase), nsP3 (RNAsynthesis) and nsP4 (RNA-dependent RNA polymerase). These four proteins in total cover about two thirds of the genome from the 5' end (Figure 5.1). The second polyprotein from the other third of the genome mainly encodes the capsid protein and the envelope glycoproteins (Figure 5.1). As an NIAID category B priority pathogen, VEEV periodically causes epidemics in equids and humans, though it typically circulates between rodents and mosquitoes in an enzootic life cycle [136-140]. VEEV has a

[3]In collaboration with Dr. Naomi Forrester's group at UTMB.

mortality rate of 50-90% in horses, resulting in significant economic impact in VEEV-endemic areas [141]. In humans, VEEV can cause encephalitis and is encountered throughout the Americas [135]. Without specific drugs to treat the VEEV infection, the current treatments are merely supportive [142]. All of these factors make VEEV not just a public health threat, but also a potential bioweapon and a bioterrorist agent [142, 143]. It is thus imperative to develop a safe and effective vaccine against VEEV.

Like most RNA viruses, the replication of VEEV in a host forms a spectrum of virus mutants. The mutation rate has to be well-maintained for the virus to survive. Too many mutations incorporated into the viral progeny results in a greater number of unfit progeny, ultimately leading to a decrease or even extinction of the virus. Conversely, too few mutations results in little variation, thereby reducing the ability of viruses to adapt their overall fitness to changing environments and further impairing the viral transmission [144, 145] A live-attenuated vaccine Tc-83 was created to prevent the VEEV infection [146]. However, 20% of the Tc-83 recipients showed reactogenicity to the vaccine and another 20% showed failure to elicit a positive seroresponse [147]. These problems can be attributed to the fact that live-attenuated vaccines have the potential to revert to wild type or to restore virulence via compensatory mutations. In fact, Tc-83 has been shown to be able to reverse back to wild type in as few as three serial IC passages in infant mice [148]. The viral mutations have been treated as by-products of the RNA-dependent RNA-polymerase (RdRp) because RNA viruses have no proof-reading domain or 3' to 5' exonuclease activity [149]. We proposed to increase the replication fidelity of Tc-83 by introducing specific mutation(s) within the sequence region of RdRp.

5.2 OBJECTIVES AND EXPERIMENTAL DESIGN

To increase the efficacy of the Tc-83 vaccine and to enhance its replication fidelity, we collaborated with Dr. Forrester's group to introduce mutation(s) inside the RdRp sequence and to assess their influence on replication fidelity. The mutant constructs were expected to have fewer single nucleotide polymorphism(s) (SNP) if they are more stable than the Tc-83 strain.

In total, six Tc-83 mutants with one or multiple mutations in the RdRp region were created. The three mutations (G14R, E37G, and A106T) in the 5' end resulted from the point mutations C→G, A→G and G→A, which were created by passaging Tc-83 in the presence of three different mutagens: viz. Ribavirin, 5' Fluorouracil, and Azacytidine (Figure 5.1), respectively. A C488Y mutation at the 3' end of the RdRp was previously identified in the Chikungunya virus [150], a virus from the same family and genus as VEEV. Four of the six Tc-83 mutants were individual mutants while another mutant included all three mutations in the 5' end. The last mutant incorporated all four mutations. Dr. Forrester's laboratory contributed to the mutants' identification and assessment. Please refer to the manuscript (M. Guerbois, et. al., not shown) for detailed information.

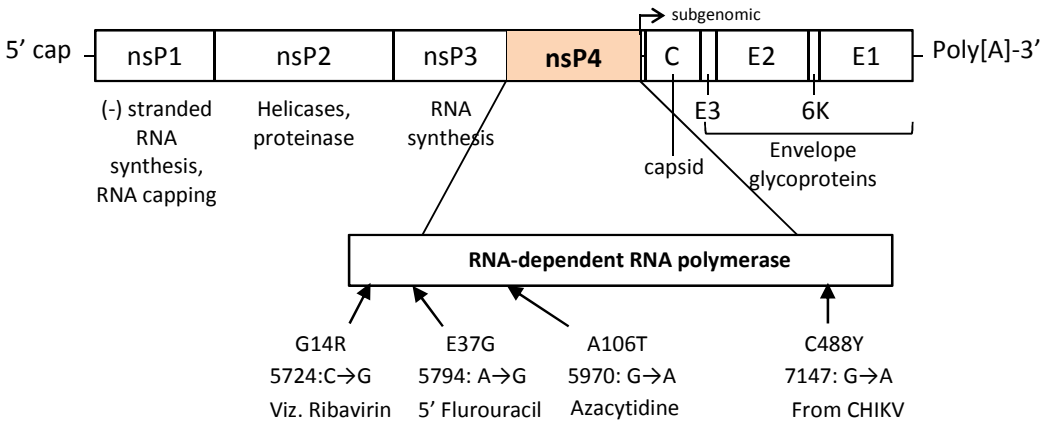

Figure 5.1 Illustration of RdRp mutants. The six mutant constructs are: G14R mutant, E37G mutant, A106T mutant, C488Y mutant, 3x (G14R, E37G and A106T) mutant, and 4x (G14R, E37G, A106T and C488Y) mutant

Next generation sequencing was utilized to assess the genetic stability of these six RdRp mutants by comparing them with the Tc-83 wild type, both *in vitro* and *in vivo*. In one experiment, all six mutants, as well as Tc-83 were passaged once independently (p1) in African green monkey kidney (vero) cells by incubating the cells with the virus for 48 hours before collecting the virus and extracting the viral RNA. In another experiment, each mutant was subjected to five serial intracranial passages (p5) in six-day-old CD1 mice. In each passage, the mice were euthanized 48 hours after inoculation of ca. 10^4 PFU in a 20ul volume per animal. RNA extracted from the mouse brain p5 virus as well as those from the vero p1 experiment above, was sequenced using the UTMB Illumina HiSeq 1000. The cDNA libraries were first prepared using the Illumina TrueSeq RNA Sample Preparation kit under conditions recommended by the manufacturer (Illumina, San Diego, CA). Then TrueSeq PE Cluster Kit v3 and TrueSeq SBS kit v3 were used to form clusters and further sequence the cDNA templates. We used the CASAVA-1.8.2 software to convert base calls to raw RNA-Seq sequence reads.

5.3 INTRA-HOST VARIATION DISCOVERY

The quality of sequence reads from the Illumina HiSeq 1000 was determined using the FastQC v0.10.1 software [151]. In order to obtain overall high quality reads, nucleotides were trimmed when more than 50% of the bases were unresolved at the ends of the sequences. Specifically, because of very high N (unknown) content percentages (> 50%) (Figure 5.2A), we trimmed the first base of the forward reads and the last two bases

of the reverse reads for the sequencing of the vero p1 virus. The sequencing of the mouse brain p5 viruses, however, typically generated a low N content, ranging from 0.00% to 0.15% across the whole position. We were thus able to use all the reads without trimming. We further removed artifact sequences using the FASTX-Toolkit v0.0.13 [152]. This step is not necessary, however, when the reference genome is well-defined and the mutation rate is low because the alignment process requires very high homological similarity to report a successful hit. For example, the default parameters for Bowtie only allow no more than two mismatches in the first 28bp seed. As previously discussed, VEEV has a much higher mutation rate so that the homological similarity would be relaxed when mapping to the Tc-83 genome. The aim of the artifact filter here was to reduce the false discovered mutations introduced by the artifacts.

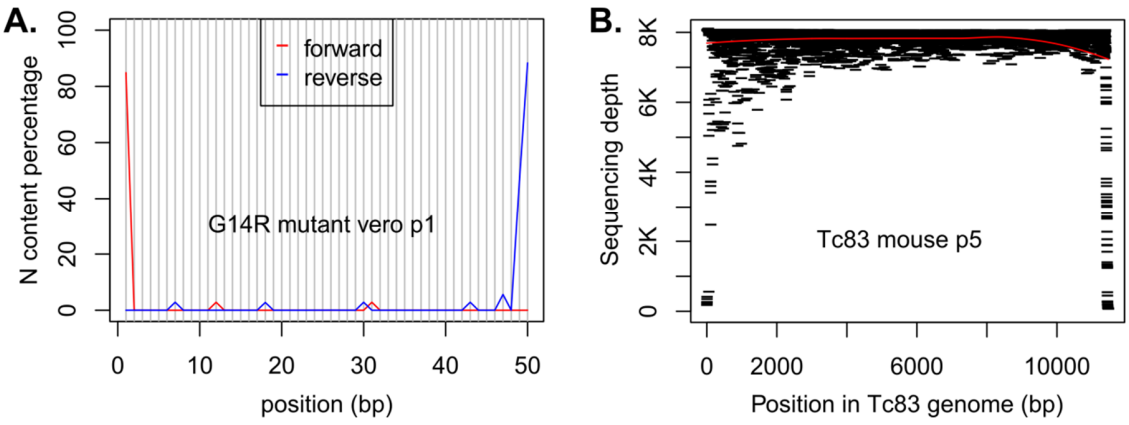

Figure 5.2 RNA-Seq data quality check and sequencing depth examination. Figure A is a representative plot of the N content in the vero p1 viruses. Figure B is a representative plot of the sequencing depth for all vero p1 and mouse p5 viruses.

The acceptable forward and reverse reads after trimming and/or filtering were aligned to the Tc-83 reference genome (GenBank Accession No. L01443) using the

Segemehl v0.1.6 software [153]. Segemehl maps short sequencing reads to a reference genome with detection of both mismatches and gapped matches (insertion and deletion). Compared to other alignment software, such as Bowtie and Tophat, the required homological similarity between the aligned read and the hit region is generally lower. As a test, we utilized the unmapped reads from a human spleen sample classified by Bowtie/TopHat as input reads for Segemehl and found 37.2% of them had at least one hit in the hg38 human genome. The lower homological requirement thus makes Segemehl more suitable for variation detection in quasispecies and viruses with high mutation rates such as VEE. We only reported the best alignments for each sample when mapping the sequence reads against the Tc-83 genome. All mutants and Tc-83 vero p1 viruses returned alignment successful rates larger than 98%, and the mouse p5 viruses had rates ranging from 92.5% to 98.1%. The alignment was then reformatted and checked for coverage using SAMtools 0.1.16 software [154]. The sequencing depth is > 7 K in the whole genome except for two short regions at either end (Figure 5.2B). In fact, each position of the Tc-83 genome was covered by approximately 8k fragments / reads. The sequencing depths from the vero p1 and mouse brain p5 viruses are comparable. In summary, we achieved very deep sequencing of the VEEV samples and Segemehl enabled us to map highly mutated reads to the Tc-83 genes.

The mismatches and gapped matches identified during alignment have three main sources: biological mutations, sequencing errors and alignment algorithms. Our ultimate goal is to distinguish the true biological mutations from errors due to the other two sources. During an RNA-Seq experiment, the genomic materials are randomly fragmented. These fragments are then randomly selected to form clusters and finally, they

are sequenced. In theory, the biological mutations should be randomly distributed across the whole fragment / read with little to no bias in the sequencing positions. In contrast, errors from sequencing and alignment algorithms are more likely to be position dependent because of their sequential replicate and hit procedure. An alignment file usually includes sections of CIGAR string and MD tag. CIGAR string describes the bases alignment (either match/mismatch) with a given reference. Meanwhile, MD tag achieves SNP/INDEL information. We examined the relation between the position and alignment of the mismatches and gapped matches in our samples based upon the CIGAR string and the MD tag. We used the E37G mutant mouse brain p5 virus as a representative example. Among all of the alignments, about 85% of them had perfect matches; about 8% had only mismatches; and the rest had gapped matches (Figure 5.3A). The mismatches / insertions / deletions had a higher frequency at the ends of the reads than in-between (Figure 5.3B, C, D). Additionally, we noticed that the number of mismatches and gapped matches in forward reads and reverse reads were different, even though their patterns are the same (Figure 5.3). We, as well as many other research groups, believe that these non-uniform patterns are mainly associated with the sequencing and alignment procedures. The reliable biological mutations are expected to be in both the forward and reverse reads and could also be identified in multiple regions of the reads.

A

E37G mutant p5	Forward △	Reverse ○
Total hits	27496120	27288143
Alignment match	25739844 (93.61%)	25198252 (92.34%)
Perfect match	23990785 (87.25%)	22658920 (83.04%)

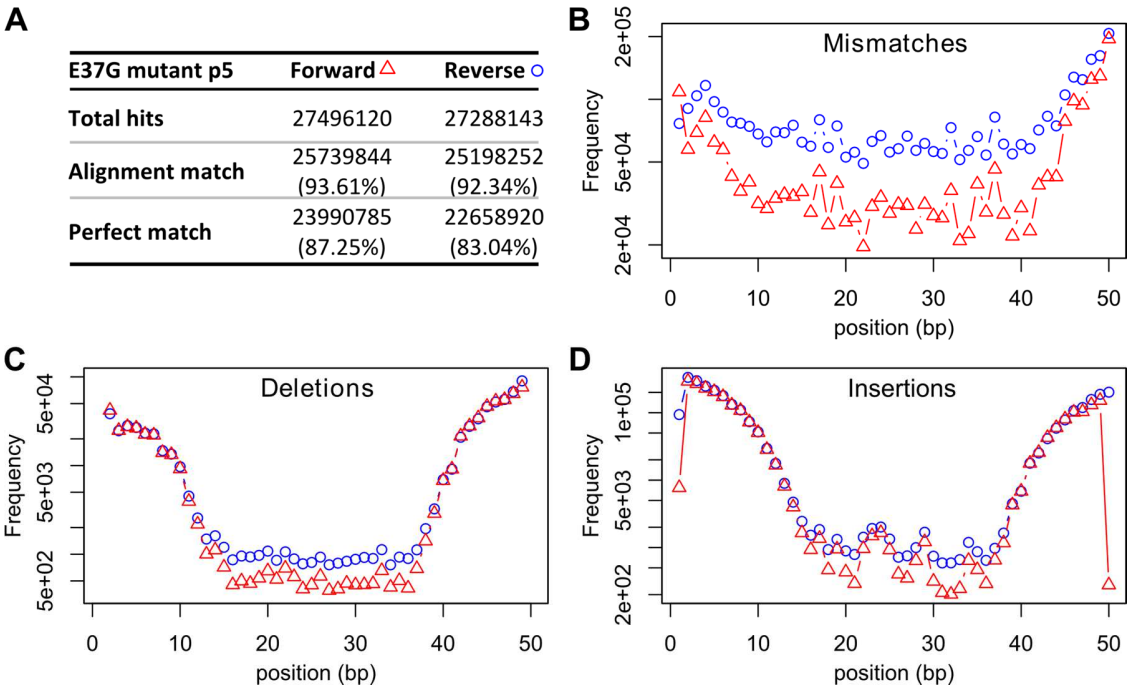

Figure 5.3 Alignment statistics and patterns. We used E37G mutant mouse brain p5 as a representative sample in all four plots. Figure A shows statistics of the alignment results. The "total hits" are all results reported by Segemehl. The "alignment matches" are matches without insertions or deletions including perfect matches and matches with mismatches. The "perfect match" means the whole reads can be perfectly aligned to the reference genome without mismatches. Figure B-D shows the event frequency across the whole reads.

To identify biological variations, the haplotypes (i.e. a set of DNA variations that tend to be inherited together) were constructed using the shotgun mode of local analysis in Shorah v0.6 software [125, 126]. We covered each position with three different windows, so that the position would locate in the first third of the first window, the second third of the second window, and the last third of the last window. In each window, Shorah identified haplotypes using those reads that spanned at least 85% of the region. Typically, more than one haplotype was generated in each window. The first (main) haplotype with the most supported reads always has the largest posterior probability. The

other haplotypes had fewer supported reads and / or smaller posterior probabilities. Variations are defined as the difference between the selected haplotypes and the sequence of the reference genome in the same window.

We first pooled together all of the main haplotypes and compared them with the Tc-83 reference genome (GenBank Accession No. L01443) in order to identify the main variations. The vero p1 Tc-83 virus showed six main variations (Table 5.1). Five of them were synonymous mutations with no change in the encoded amino acid. The only missense mutation, in which a valine was substituted by an alanine, occurred in the E1 protein. These mutations were consistently found in all mutant constructs of vero passage 1 and mouse brain passage 5. As expected, the mutations we introduced to create the construct persisted, indicating that those mutants are stable with no reversion to wild type. Additionally, a couple of other mutations were identified in the mutants. We noticed that all 3x and 4x mutants had a mutation at position 401 (C→G) in nsP1, which is not observed in the individual mutants or the Tc-83 virus. All 3x mutants had two mutations at position 8032 (C→A) in capsid protein and 9760 (T→G) in E2 protein, which were not observed in other constructs. Both 3x and 4x mutants in mouse brain p5 had a unique mutation not shared by others. Excluding these two newly issued mutations, all of the others were synonymous mutations (Table 5.1). The main variation assessment indicated that the mutation pattern may be different from vero cells to mouse brain and that the four point mutations interact with each other. However, no structure for an alphavirus RdRp is available so far. Therefore, we do not know the precise placement of the mutations or how they interact, which limits our understanding of how they work.

All of these main variations were always discovered in all three windows. They were the most favorable nucleotides in the corresponding constructs and they were shared by the majority or even all of the viruses in the colons. To discover the rare mutations, we substituted the identified main variations inside the Tc-83 reference genome in order to re-run Shorah for variation detection. Rather than using only the main haplotypes, we collected all of the haplotypes with a posterior probability larger than 0.9. All variations were collected. Using the A106T mutant mouse p5 virus as an example, we detected from the forward (reverse) reads 12,501 (13,554) variations, including mismatches, insertions and deletions, from the first windows, 324 (353) variations from the second windows, and 292 (317) variations from the third (last) windows. Among them only 77 (100) variations appeared in two or three different windows. As discussed previously, the mutations from only one window were more likely to be derived from sequencing and alignment errors. To reduce the effects of the sequencing and alignment errors, we only passed those variations that were detected in two or three different windows for statistical tests including strand bias and the Benjamini-Hochberg multiple testing correction [127]. We used FDR < 0.05 as a cut off. A variation is believed to be a real biological variation when the haplotype is found in both the forward and the reverse reads. The number of final variations (both SNP and INDEL) and SNP only at vero p1 and mouse brain p5 are shown in Table 5.2. In the vero p1 virus, the 3x mutant had the fewest variations and zero SNPs, indicating the highest replication fidelity. The 4x mutant had the most variations, many more than the Tc-83 wild type. The variations in other mutants were comparable to Tc-83. In contrast, all mutants in mouse brain p5 showed fewer variations and SNPs than Tc-83.

We further checked the main mutations of their ability to revert back. All vero p1 viruses, except the 3x mutant, had some sequences with T in position 10356 (Figure 5.1). This is the only main and missense mutation detected in the Tc-83 virus. This mutation changed a valine to an alanine in E1 protein. E1 becomes activated when E2 binds to host cells. The activated E1 allows the virtual genome to escape from the endosome/virus particle and enter into the cytoplasm. We also found that three individual mutants of vero p1 virus had wild type nucleotide in position 1616 (Figure 5.1). The A106T mouse p5 virus had reverted mutations. The difference between the vero p1 and mouse brain p5 viruses indicated that the mutations may be environment dependent. Moreover, we found that the A106T mutant vero p1 viruses had a rare G14R mutation; and the G14R mutant vero p1 viruses had a rare E37G mutation (Figure 5.1). This, from another perspective, indicated that these three mutations can have interactions.

Protein		nsP1				nsP4					Capsid	E2		E1		
Position (bp)		401	1613	1616	1619	5724	5794	5970	7147	7208	8032	8805	9760	10356	10673	10900
Start		399	1611	1614	1617	5724	5793	5970	7146	7206	8030	8804	9758	10355	10673	10899
Ref		C	A	C	T	G	A	G	G	T	C	A	T	T	A	A
Ref Codons		CTC	GAA	GCC	GAT	GGT	GAA	GCA	TGC	CCT	ATC	CAA	CCT	GTT	AAA	GAA
Ref Amino Acid		Leu	Glu	Ala	Asp	Gly	Glu	Ala	Cys	Pro	Ile	Gln	Pro	Val	Lys	Glu
Var		G	G	A	C	C	G	A	A	C	A	T	G	C	C	G
Var Codon		CTG	GAG	GCA	GAC	CGT	GGA	ACA	TAC	CCC	ATA	CTA	CCG	GCT	CAA	GAG
Var Amino Acid		Leu	Glu	Ala	Asp	Arg	Gly	Thr	Tyr	Pro	Ile	Leu	Pro	Ala	Gln	Glu
vero p1	**G14R**		✔	✔	✔	✔	A→G							✔		✔
	E37G		✔	✔	✔		✔							✔		✔
	A106T		✔	✔	✔	G→C		✔			C→A			✔		✔
	C488Y		✔	✔	✔				✔					✔		✔
	3x	✔	✔	✔	✔	✔	✔	✔			✔		✔	✔		✔
	4x	✔	✔	✔	✔	✔	✔	✔	✔					✔	A→C	✔
	Tc-83		✔	✔	✔									✔		✔
mouse p5	**G14R**		✔	✔	✔	✔								✔		✔
	E37G		✔	✔	✔		✔							✔		✔
	A106T		✔	✔	✔			✔						✔		✔
	C488Y		✔	✔	✔				✔					✔		✔
	3x	✔	✔	✔	✔	✔	✔	✔			✔	✔	✔	✔		✔
	4x	✔	✔	✔	✔	✔	✔	✔	✔					✔	✔	✔
	Tc-83									✔				✔		

Table 5.1 Main variations examination. Information about the main variation is listed in the top part of the table. The corresponding box will be checked (✔) if a construct has that main variation. We use under line to indicate that the wild type nucleotide and main variation are co-existing. We clearly show the mutations if one construct has the main variation detected from other constructs as rare mutations.

A. Variations in vero P1 viruses

	Variants	**Fold change from Tc-83**	**SNP's**	**Fold change from Tc-83**
Tc-83 G14R	8	-1.13	4	1
Tc-83 E37G	9	1	6	1.5
Tc-83 A106T	12	1.33	8	2
Tc-83 C483Y	11	1.22	6	1.5
Tc-83 3x	3	-3	0	NA
Tc-83 4x	24	2.67	19	4.75
Tc-83	9	1	4	1

B. Variations in mouse brain P5

	Variants	Fold change from Tc-83	SNP's	Fold change from Tc-83
Tc-83 G14R	73	-1.22	44	-1.39
Tc-83 E37G	61	-1.46	42	-1.45
Tc-83 A106T	65	-1.37	40	-1.53
Tc-83 C483Y	70	-1.27	47	-1.30
Tc-83 3x	62	-1.44	42	-1.45
Tc-83 4x	73	-1.22	52	-1.17
Tc-83	89	1	61	1

Table 5.2 Number of variants and SNPs identified during NGS sequencing of the constructs and the fold changes from the wild-type Tc-83.

5.4 RESULTS AND DISCUSSION

The VEEV is an arthropod-borne virus transmitted by mosquitos. To reduce the burden of the disease, it is more effective to protect the population via vaccine than to control the mosquito population. The current vaccine for VEEV is a live attenuated vaccine Tc-83, which could potentially revert to wild type or restore virulence via a compensatory mutation. To overcome this severe drawback, we validated a series of high-fidelity mutations of Tc-83 in the model systems.

The six RdRp mutant constructs and Tc-83 wild type were examined for their resistance to 5FU (Figure 5.4) on vero cells. The 3x mutant had a similar replication pattern to Tc-83, but the fold change compared to no treatment controls, was markedly smaller than those of the Tc-83 wild type. In other words, the 3x mutant showed increased resistance to the mutagen, which is consistent with our previous finding that the 3x mutant is the most stable construct in the vero p1 virus. C488Y mutant showed no difference from the wild type in fold change, which suggested no obvious resistance to mutagens 5FU. The fold change of the 4x mutant was between that of the 3x and C488Y mutants, similar to the pattern in the main variation examination, but not a simple superposition or subtraction. The patterns in the other three individual mutants were not consistent with those of the Tc-83, C488Y, 3x or 4x mutants. This could also be explained by our conclusion from the previous main haplotype examination: all of the individual mutations (G14R, E37G and A106T) were stable, and each had the potential to introduce the other two mutations. All three mutations function in concert. Thus, the resistance pattern of the three individual mutants was not as clear as the 3x mutant.

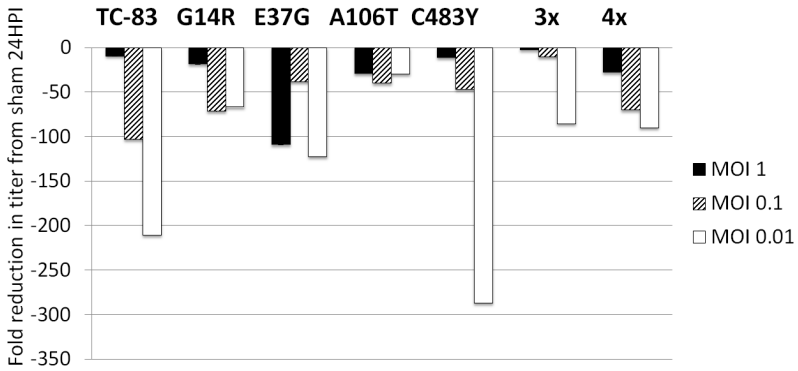

Figure 5.4 Examination of the resistance to 5' fluorouracil treatment (5FU). Each construct was grown in the presence of three different concentrations of 5FU and at three different multiplicities of infection (MOI). Plot shows the fold changes compared to no 5FU treatment under the same conditions. This experiment was designed, performed and analyzed by Dr. Forrester's research group.

For the viruses passing in mouse brain, we utilized the RNA-Seq approach to discover variations in p5 colons. Furthermore, we verified, using Sanger sequencing, the changes in the nucleotide sequence in passage 10 (p10) colons. In theory, with deep sequencing, the RNA-Seq approach should be more sensitive than the Sanger sequencing. In other words, for the same sample, the RNA-Seq approach should detect more and possibly even all mutations when compared to the Sanger sequencing approach. In all p5 and p10 viruses, we identified a set of three synonymous mutations in the nsP1 in contiguous amino acids at nucleotide positions 1613, 1616 and 1619. We noticed that p5 and p10 viruses also had some consistent mutations. For instance, p5 and p10 3x mutants had an A to T mutation in nucleotide position 8805, which changed a glutamate to a leucine in the capsid protein. P5 and p10 4x mutants had an A to C mutation in nucleotide position 10673, which changed an alanine to a lysine in the E1 protein. These two mutations were unique to the corresponding constructs and both appeared only in the

mouse brain but not in the vero passages. However, not all mutations found in the p10 mouse by Sanger sequencing were found in the mouse p5. The p5 virus also included some mutations not found in the p10 virus. Considering the sensitivity of these two approaches, we draw the conclusion that the mutation occurred progressively and may disappear after a number of replications.

All mutant constructs had fewer variants and SNPs compared to the Tc-83 in the mouse brain passing 5 (Table 5.2B). We hypothesized that the RdRp mutant has high-fidelity, which would be more adaptable to the brain environment but could lead to attenuation in mouse models due to their inability to move easily between tissues. The constructs were validated in a lethal mouse model for Tc-83. The viruses were injected into 6-day old mice via the intra-cranial route and the sub-cutaneous route. In the intra-cranial route, the 3x mutant showed more virulence with increasing mortality than the Tc-83 wild type, while the other constructs were similar to Tc-83 but with a slightly shorter mean time to death. In the sub-cutaneous route, the 3x, 4x and C483Y constructs exhibited significantly reduced mortality compared to the wild-type Tc-83. These findings support our previous hypothesis that the RdRp mutant with high fidelity can replicate in the mouse brain, but reduces the ability of the virus to travel from the inoculation site into other tissues and, finally, into the brain where it causes death.

We further tested the effectiveness of Tc-83 and the RdRp mutants as vaccines in an established adult model of VEEV with 7-week old CD-1 mice. Given a lethal challenge with wild-type VEEV (strain 3908 subtype intra-cranial), these animals exhibited no weight loss or viremia illness. Every vaccinated mouse exhibited a strong neutralizing antibody response in four weeks post-vaccination. In fact, all of the

constructs showed a higher neutralizing antibody response than the wild type Tc-83. Considering the report of failure to elicit a positive seroresponse by Tc-83, the higher antibody titer confirms that Tc-83 RpRd mutants can enhance the effectiveness of wild type Tc-83 as a vaccine.

5.5 CONCLUSION AND LIMITATIONS

In the current project, we assessed the replicate fidelity of VEEV mutant constructs from NGS data through a computer-aided approach. The same method can also be used to discover the essential mutations that contribute to the virus transmission across multiple environments. For example, we could distinguish VEEV populations that are isolated in the midgut of the mosquito from the ones that are able to traverse to the mosquito salivary gland. These approaches will help us to understand more about the VEEV intra-host variation, and further accelerate new live attenuated vaccine development. Moreover, this methodology is extensible to live attenuated vaccine development in other RNA viruses.

The current project also has several limitations. We discovered that when making the RdRp mutants, we accidentally used a different clone of Tc-83 which already had three mutations at 1613, 1616, and 1619. Second, we utilized the Illumina HiSeq platform for sequencing. This platform uses the clonal amplification template, which may introduce replication mutations during the sequencing library preparation. This type of error could be reduced if the single molecule template were used. In other words, the results would likely be more reliable if sequencing platforms such as PacBio or Oxford nanopore were used instead.

Chapter 6. Dirichlet Process Mixture Integrated Ensemble Methods

6.1 INTRODUCTION

A learning algorithm is built by summarizing rules from a given training dataset. It then functions to predict an output value given some new input values. The prediction usually cannot be solved exactly, which is known as inductive bias. Due to this inductive bias, it is often difficult to determine whether a single machine learning model is overfitting or underfitting the data [155]. Overfitting occurs when the learning model is excessively complex relative to the amount of data available and thus performs much better for the training dataset than a testing dataset. Meanwhile, underfitting performs better for a test dataset than the training dataset. Ensemble methods incorporate multiple individual machine learning models by assuming that their expertized predictive spaces are different and can be complemented by each other [156]. With their constituent individual models, ensemble methods create an integrated model, which is characterized by high performance and a low risk of selecting a poor model. Ensemble methods excel in building models for data with relatively small sample sizes, high-dimensionality and complexity patterns [157].

Ensemble methods are expected to perform better than their component single machine learning models because of the diversity of the individual models within the ensemble. The influence of model diversity on prediction abilities is illustrated in the following example. We will predict a new dataset using three models. Each model has an accuracy rate of 0.8, which corresponds to an 80% confidence level for correct prediction. If the models have maximum diversity (i.e. all the models are entirely

independent of each other), we can take the majority prediction as the final result. This corresponds to an accuracy rate of 0.8 * 0.8 * 0.8 + 3 * 0.2 * 0.8 * 0.8 = 0.896. The performance of ensemble methods generally exceeds that of any single machine learning algorithm. If the models have the lowest diversity, (i.e. all the models are the same), we can only obtain an accuracy rate of 0.8, the same as any single model. Therefore, the effectiveness of an ensemble method is largely dependent upon the diversity of its component models.

Bagging and adaptive boosting (AdaBoost) are two of the most widely used ensemble algorithms. Both bagging and AdaBoost create various individual models by using different training datasets, which are randomly selected in bagging but are biased to the previously misclassified samples in AdaBoost. All of the individual models form a model committee, which are averaged or weighted averaged to generate the final prediction. Similar to many other established ensemble methods (e.g. random forest), all of these generated individual models are included in the final committee pool without assessment of their diversity. As discussed previously, the key to a better predictive ability for an ensemble method lies in its diversity. Here, we proposed to increase the predictive ability of an ensemble method by enhancing the diversity of the model pool before the final vote.

To enhance model diversity, we cluster the original model pool of classic ensemble methods into different groups then filter out models with high degrees of similarity. This idea was motivated by our NGS-based quasispecies analysis in Chapter 5. We utilized the Bayesian nonparametric Dirichlet Process Mixture (DPM) algorithm to cluster reads into various haplotypes (i.e. groups). Each read can be treated as a string

with a specific number of characters. Similarly, we can characterize each model by the prediction resulting from a predefined input dataset. Each read has some sequencing errors. Analogously, each model has its inductive bias, which would cause some prediction errors in the prediction results. In the NGS field, DPM has been successfully applied to construct different haplotypes that can be associated with one gene by clustering many sequencing reads. Due to the similarity mentioned above, we predict that DPM could be an optimal method for model clustering. In the following, we will introduce the DPM integrated ensemble methods of Bagging and AdaBoost and then evaluate them by making a comparison with the classic approach without diversity assessment.

6.2 METHODOLOGY

The DPM algorithm [158-166] describes models with an infinite number of mixture groups [167]. As a Bayesian nonparametric clustering method, DPM clusters data into groups without knowing the number of mixing groups beforehand. Newly arrived data will be assigned to either a previously existing group, or a new instantiated group whose probability is controlled by the parameter α. The groups are determined by different parameters $\theta_1, \theta_2, \theta_3, \ldots\ldots$. Each data point is drawn from one of those components. The parameters θ_i $(i = 1, 2, 3, \ldots)$ are generated from a distribution G. We use the Dirichlet Process (DP) [158] to characterize the distribution G. It uses a positive scaling factor α and a base distribution G_0. Thus

$$G|\alpha, G_0, n \sim DP(\alpha, G_0),$$

$$\theta_n|G \sim G, n = 1, 2, 3, \ldots,$$

$$x_n|\theta_n \sim F(x_n|\theta_n), n = 1, 2, 3, \ldots.$$

where '~' means "is distributed as" in mathematics. Using the first two distributions, we can integrate G, and obtain a joint distribution:

$$\theta_i|\theta_{j\neq i}, \alpha, G_0 = \frac{1}{N-1+\alpha}\sum_{j\neq i}\delta_{\theta_j,\theta_i} + \frac{\alpha}{N-1+\alpha}G_0$$

$$= \sum_{\theta}\frac{N_{i,\theta}}{N-1+\alpha} + \frac{\alpha}{N-1+\alpha}G_0$$

In other words, the probability that DPM assigns new data to an already populated or a new class is formularized by using an α controlled prior as:

$$p(\theta_i = \theta|\theta_j, j\neq i) = \begin{cases} \frac{N_{i,\theta}}{N-1+\alpha} & if\ class\ \theta\ is\ already\ populated \\ \frac{\alpha}{N-1+\alpha} & if\ a\ new\ class\ with\ \theta\ drawn\ from\ G_0\ is\ instantiated \end{cases}$$

where θ_i is the parameter associated with the class of subject i; N is the total number of subjects; and $N_{i,c}$ is the number of subjects that have been assigned to class θ before subject i. In other words, the prior probability of a new subject joining a cluster is proportional to the number of subjects that are already in that cluster.

The discussion above mainly applies to a one-dimensional procedure such as the Urn model and the Chinese Restaurant model. In our case, the predictive result of a given dataset from each model is of high dimension. To cluster the models using the DPM approach, we update the assignment probability as bellow:

$$p(\theta_i = \theta|\theta_j, j\neq i) = \begin{cases} \frac{N_{i,\theta}}{N-1+\alpha}\, p(r_i|c_\theta) & if\ class\ \theta\ is\ already\ populated \\ \frac{\alpha}{N-1+\alpha}\, p(r_i|c_0) & if\ a\ new\ class\ with\ \theta\ drawn\ from\ G_0\ is\ instantiated \end{cases} \quad (1)$$

Here, c_θ represents the voting of all models inside the class θ and c_0 represents the voting of all models in all groups. The value $p(r_i|c_k)$ describes the probability that r_i comes from the component c_k. It can be calculated by:

$$p(r_i|c_k) = \epsilon^{ki}(1-\epsilon)^{kc}$$

where ki and kc are the number of inconsistent and consistent predictions of r_i and c_k, respectively. ϵ is the error parameter, which follows the beta distribution. We updated the beta distribution of its parameters α and β at each iteration:

$$\alpha = \epsilon_1 \epsilon_2 nJ \quad and \quad \beta = \epsilon_2 nJ(1 - \epsilon_1)$$

where ϵ_1 and ϵ_2 are estimation of the mean and variance of the prior error parameter ϵ, respectively, and, n and J are the number models and samples being investigated, respectively. Considering that $p(kit, kct|\epsilon)$ follows a binomial distribution, we estimated ε in each iteration using:

$$p(\epsilon|kit, kct) \sim Beta(kct + \alpha, kit + \beta)$$

where *kit* and *kct* are the total number of inconsistent and consistent predictions of all models and their assigned groups, respectively.

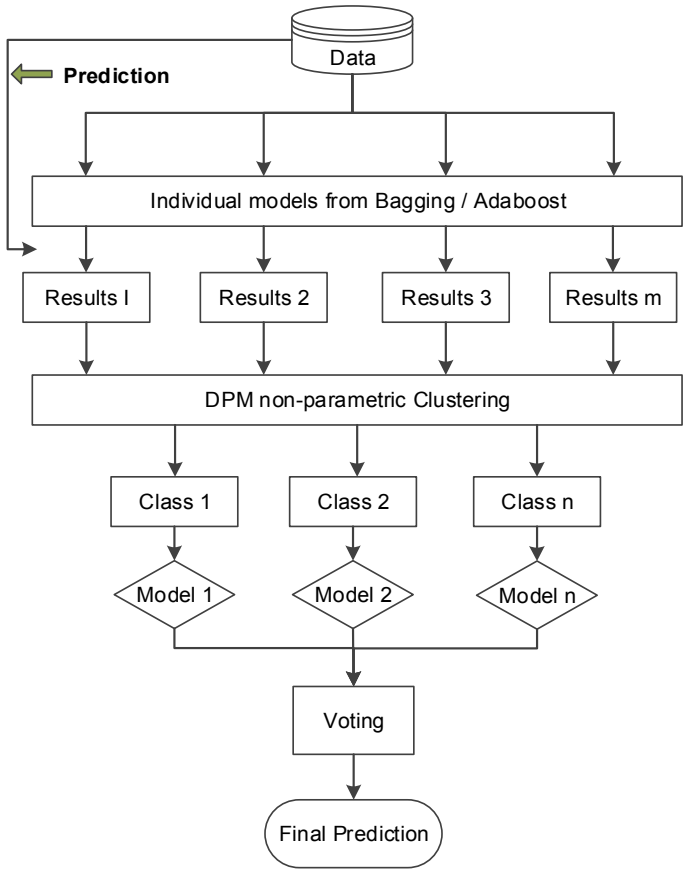

Figure 6.1 Illustration of DPM integrated ensemble methods

We developed the DPM integrated ensemble method as illustrated in Figure 6.1. The whole process includes two steps: 1) create a model pool using classic ensemble methods, such as Bagging and Adaboost; and 2) enhance the model pool diversity using DPM.

In the first step, we utilized the *bagging* or *boosting* function from the R package *adabag* to create an ensemble model. Its component individual models, also known as basis classifiers, form the model pool. In general, the greater the number of the basis classifiers, the better the predictive ability will be. However, the predictive ability can become worse if the classifiers have no spreading diversity.

To increase the diversity of a model pool, we include a DPM clustering in the second step to assess the similarity of all the individual models based upon their predictive results for the whole training datasets. Initially, all models are randomly assigned to a class. We then update the assignments by calculating, over many iterations, the conditional probability of a model being from an existing or a new group given the prior clustering results. The DPM algorithm has been investigated for the probability calculation, which is the key to success in our approach. The entire model pool will be clustered into several groups.

Each group from the DPM clustering has at least one model. If some groups only have one model assigned to them, we pool all of these models together as an ensemble method to predict on the original dataset. The predictive results are set as the reference results. Otherwise, we will use the predictive results from all of the individual models as the reference result. To maximize the prediction space and the diversity of the model

pool, we only select one model from each group. The selected model deviates more from the reference result compared to the other models in the same cluster. The final model pool of DPM ensemble methods will only contain the selected models including the models which form a group by themselves.

We coded the DPM clustering algorithm in C++ based upon the original Shorah code to create DPM integrated bagging and boosting as above. We additionally utilized the R packages *RCpp* and *RCppGSL* to make the DPM algorithm in C++ callable by R. The pseudo code for the whole procedure is listed in Table 6.1.

Input:

md: bagging or boosting object from *R* package *adabag*;
data: train data set
K: number of initial groups (default=10);
T: number of iterations (default=1000);
J: number of record iteration (default=100)

Output:

md_dpm: bagging or boosting object

Pseudo Code:

Generate model pools and its characterized predictive results

1. Create a ensemble model using R package *adabag*
2. **For** each constitute individual model (i) of the ensemble model
3. r_i = predictive results of train data from model i

DPM to cluster models into groups

1. Initialize: $p_i(n) = 1/K$ for i=1, 2, …, N and n=1, 2, … K
2. for i in 1→N; do
3. sample group for r_i base up $p_i(n)$
4. end for
5. **for** t in 1→T; do
6. **for** i in 1→N; do
7. filter out group with size 1
8. update $p_i(n)$ based upon **Formula (1)**
9. re-sample group for r_i base up $p_i(n)$
10. if (t > (T-J)) record;
11. **end for**
12. **end for**

13. **for** k in in 1→N; do
14. assign model r_k to group c_k based upon record
15. **if** r_k has been assigned to multiple groups
16. set c_k as the group with largest frequency
17. **end for**
18. filter empty groups

Select diverse models to form new model committee

1. new_model_committee = NULL
2. **for** each group c_k
3. **if** c_k only consist one model
4. **push** its component model into new_model_committee
5. **if** new_model_committee = NULL
6. reference = predictive results of the classic ensemble method
7. **else** reference = predictive results of models in new_model_committee
8. for each group c_k with two or more models
9. push the model most deviated from reference into new_model_committee
10. Update the ensemble model with the new_model_committee

Table 6.1 Pseudo code for DPM integrated ensemble method

6.3 EVALUATION

The novel DPM integrated bagging and AdaBoost algorithm, referred to below as DPM bagging and DPM AdaBoost, has been evaluated on a representative collection of datasets from the data repository of knowledge extraction based upon evolutionary learning (KEEL). The forty-one datasets, summarized in Table 6.2, have considerable diversity in size, number of classes and number of types of attributes. Their inputs include both categorical and numeric variables. In fact, these datasets, together with the iris dataset, have long been considered as representative datasets for machine learning algorithm evaluation [168, 169]. Here we excluded the iris data because the predictive results from all individual models of the ensemble method were the same and cannot be clustered by the DPM algorithm. In other words, a single classification tree is enough for the iris classification problem.

Dataset	# Variables	Positive		Negative		Imbalance Rate
		Labels	#Events	Labels	#Events	
Abalone9	8	19	32	remainder	4142	0.0077
Abalone9vs18	8	18	42	9	689	0.0575
Ecoli0137vs26	7	pp, imL	54	cp, im, imU, imS	257	0.1736
Ecoli0vs1	7	im	77	cp	143	0.3500
Ecoli1	7	im	77	remainder	259	0.2292
Ecoli2	7	pp	52	remainder	284	0.1548
Ecoli3	7	imU	35	remainder	301	0.1042
Ecoli4	7	om	20	remainder	316	0.0595
Glass0	9	1	70	remainder	144	0.3271
Glass0123vs456	9	5, 6, 7	51	1, 2, 3, 4	163	0.2383
Glass016vs2	9	3	17	1, 2, 7	175	0.0885
Glass016vs5	9	6	9	1, 2, 7	175	0.0489
Glass1	9	2	76	remainder	138	0.3551
Glass2	9	3	17	remainder	197	0.0794
Glass4	9	5	13	remainder	201	0.0607
Glass5	9	6	9	remainder	205	0.0421
Glass6	9	7	29	remainder	185	0.1355
Haberman	3	positive	81	negative	225	0.2647
New-thyroid1	5	2	35	remainder	180	0.1628
New-thyroid2	5	3	30	remainder	185	0.1395
Page-blocks0	10	2, 3, 4, 5	559	1	4913	0.1022
Page-blocks13vs2	10	3	28	2, 4	416	0.0631
Pima	8	tested_positive	268	tested_negative	500	0.3490
Segment0	19	1	330	remainder	1980	0.1429
Vehicle0	18	van	199	remainder	647	0.2352
Vehicle1	18	saab	217	remainder	629	0.2565
Vehicle2	18	bus	218	remainder	628	0.2577
Vehicle3	18	opel	212	remainder	634	0.2506
Vowel0	13	0	90	remainder	900	0.0909
Wisconsin	9	4	239	2	444	0.3499
Yeast05679vs4	8	me2	51	mit, me3, exc, vac, erl	477	0.0966
Yeast1	8	nuc	429	remainder	1055	0.2891
Yeast1289vs7	8	vac	30	nuc,cyt,pox,erl	917	0.0317
Yeast1458vs7	8	vac	30	nuc, me2, me3, pox	663	0.0433
Yeast1vs7	8	vac	30	nuc	429	0.0654
Yeast2vs4	8	me2	51	cyt	463	0.0992
Yeast2vs8	8	pox	20	cyt	463	0.0414
Yeast3	8	me3	163	remainder	1321	0.1098
Yeast4	8	me2	51	remainder	1433	0.0344
Yeast5	8	me1	44	remainder	1440	0.0296
Yeast6	8	exc	35	remainder	1449	0.0236

Table 6.2 Summary of 41 datasets from KEEL for evaluation

For each dataset, we downloaded the partitions using a five-fold distribution optimally balanced stratified cross-validation (5DOBSCV) from the KEEL official website [170]. As a result, we had five training and testing dataset couples for each data. The whole data set could be obtained by combining a training dataset with its corresponding test dataset or by combining all of the five testing datasets. The sizes of all the training datasets are approximately the same, and the sizes of the test datasets are approximately the same as well.

We created the bagging and AdaBoost ensemble models using the commands *bagging* and *boosting* in the R package with the default parameters for each training dataset. 100 trees, as default, should be generated if an ensemble model was successfully created. Although some training datasets failed to create an ensemble model with the default parameters, we did no fine tuning of the parameters for evaluation purposes. For each ensemble method, we further integrated the DPM clustering to enhance the diversity of the committee models as discussed previously. The weight of each model remained the same as that in the original ensemble model. Thus, the DPM integrated ensemble can only have equal or fewer individual models than the original ensemble model.

We characterized the performance of the classic and the DPM integrated ensemble methods by two values: error rate and area under the receiver operating characteristic (ROC) curve (AUC). The error rate measures the frequency of misclassifications for a model. The lower the error rate is, the better the predictive models will be. The ROC curve plots the predictive results of a model in such a way that the x axis is the true positive rate measuring the sensitivity and the y axis is the false negative

rate measuring the specificity. AUC is the area under the ROC curve. Therefore, AUC is 0.5 for a truly random classification, and 1.0 for an ideal model. However, ROC curves from real models are usually located somewhere between the random and the ideal curves, so the AUC value is typically smaller than 1.0 but greater than 0.5. A good model should have both high sensitivity and high specificity. Consequently, the larger the AUC value is, the better the model performs. We calculated the AUC values using the R package *ROCR*. The AUC value is believed to be a better measurement of the predictive modeling accuracy than the error rate, especially for an imbalanced dataset, in which the number of negative events is much larger than the number of positive events.

The *bagging* function in the R package *adabag* was successfully implemented to create classic bagging models for 28 out of the 41 datasets using the default parameters (Table 6.3). Each classic bagging model had 100 trees. On average, the DPM bagging reduced the tree numbers to an average of 56, ranging from 8 to 100. 19 DPM bagging models had 4 equivalent and 15 larger AUC values than the classic bagging models. Meanwhile, 18 DPM bagging models had 8 equivalent and 10 smaller error rates. Only four datasets (ecoli1, page-blocks13vs2, pima, vehicle1) returned smaller AUC values and larger error rates when comparing the DPM bagging models to the classic bagging models.

Dataset	Classic Bagging			DPM Bagging		
	Trees	Error	AUC	Trees	Error	AUC
abalone9	-	-	-	-	-	-
abalone9vs18	100	0.047	0.8530	63	0.047	0.8491
ecoli0137vs26	100	0.074	0.9429	38	0.061	0.9461
ecoli0vs1	100	0.041	0.9806	9	0.027	0.9801
ecoli1	100	0.089	0.9378	57	0.092	0.9325
ecoli2	100	0.065	0.9441	47	0.063	0.9444
ecoli3	-	-	-	-	-	-

ecoli4	-	-	-	-	-	-
glass0	100	0.150	0.9114	82	0.145	0.9125
glass0123vs456	100	0.065	0.9706	24	0.065	0.9727
glass016vs2		-	-	-	-	-
glass016vs5		-	-	-	-	-
glass1	100	0.192	0.8596	96	0.168	0.8596
glass2		-	-	-	-	-
glass4		-	-	-	-	-
glass5		-	-	-	-	-
glass6	100	0.033	0.9387	10	0.033	0.9378
haberman	100	0.255	0.7231	96	0.261	0.7225
newthyroid1	100	0.033	0.9913	17	0.023	0.9944
newthyroid2	100	0.023	0.9802	8	0.033	0.9806
page-blocks0	100	0.028	0.9868	91	0.028	0.9869
page-blocks13vs2	100	0.018	0.9977	17	0.020	0.9959
pima	100	0.240	0.8395	100	0.238	0.8394
segment0	100	0.008	0.9929	19	0.008	0.9930
vehicle0	100	0.053	0.9885	89	0.056	0.9889
vehicle1	100	0.209	0.8508	98	0.210	0.8507
vehicle2	100	0.031	0.9904	79	0.030	0.9904
vehicle3	100	0.209	0.8522	97	0.209	0.8522
vowel0	100	0.013	0.9919	39	0.012	0.9923
wisconsin	100	0.034	0.9905	49	0.031	0.9908
yeast05679vs4	100	0.074	0.8844	65	0.080	0.8859
yeast1	100	0.226	0.7854	98	0.226	0.7855
yeast1289vs7	-	-	-	-	-	-
yeast1458vs7	-	-	-	-	-	-
yeast1vs7	-	-	-	-	-	-
yeast2vs4	100	0.039	0.9842	30	0.041	0.9855
yeast2vs8	-	-	-	-	-	-
yeast3	100	0.046	0.9629	63	0.048	0.9632
yeast4	100	0.028	0.8822	58	0.029	0.8829
yeast5	100	0.020	0.9791	36	0.020	0.9780
yeast6	-	-	-	-	-	-

Table 6.3 Evaluation of DPM bagging

We examined the four datasets with decreased AUC and accuracy from the DPM bagging method. The DPM bagging models for ecoli1, page-blocks13vs2, pima, vehicle1 datasets had 57, 17, 100, and 98 trees, respectively. We noticed that the number of trees in the DPM bagging model for page-blocks13vs2 was relatively smaller. Thus, the DPM bagging may have filtered out too many trees and consequently failed to preserve the existing diversity. We, therefore, increased the alpha parameter from 0.001 to 0.1, which

made the proposal of new groups easier. These results showed that the DPM bagging method can improve the classic bagging model by decreasing the error rate (Table 6.4). Meanwhile, the numbers of trees for the other three datasets were relatively large. Therefore, the original model pool from the classic bagging may already present high diversity. We increased the number of trees to 1000 in the classic bagging method and re-ran the DPM bagging procedure. All datasets displayed decreased or equivalent AUC values but fewer trees when we compared the DPM bagging with the classic bagging results (Table 6.4).

Dataset	**Classic Bagging**			**DPM Bagging**		
	Trees	**Error**	**AUC**	**Trees**	**Error**	**AUC**
100 trees in classic bagging, alpha=0.001 (Default)						
ecoli1	100	0.089	0.9378	57	0.092	0.9325
page-blocks13vs2	100	0.018	0.9977	17	0.020	0.9959
pima	100	0.240	0.8395	100	0.238	0.8394
vehicle1	100	0.209	0.8508	98	0.210	0.8507
100 trees in classic bagging, alpha=**0.1**						
page-blocks13vs2	100	0.018	0.9977	20	0.016	0.9969
1000 trees in classic bagging, alpha=0.001						
ecoli1	1000	0.093	0.9383	265	0.090	0.9383
pima	1000	0.236	0.8393	926	0.233	0.8394
vehicle1	1000	0.207	0.8537	661	0.209	0.8564

Table 6.4 Parameter fine tuning for DPM bagging

To understand the reason why the DPM bagging method works, we took the newthyroid1 dataset as an example. In this dataset, the DPM bagging method reduced the number of trees from 100 to 17, but had improvements in both error rate and AUC value. We utilized a PCA plot to visualize the distribution of the model pool of the classic and DPM bagging models (Figure 6.2). We found that many individual models from the classic bagging method were collapsed together into a two-dimensional PCA plot. Thus,

many models were similar to each other, which indicated a low diversity of the model pool. The individual models of the DPM bagging method spread around the true label with high diversity. These findings confirmed our hypothesis that DPM clustering can increase the model diversity and explained why DPM bagging has better predictive ability than the classic bagging method.

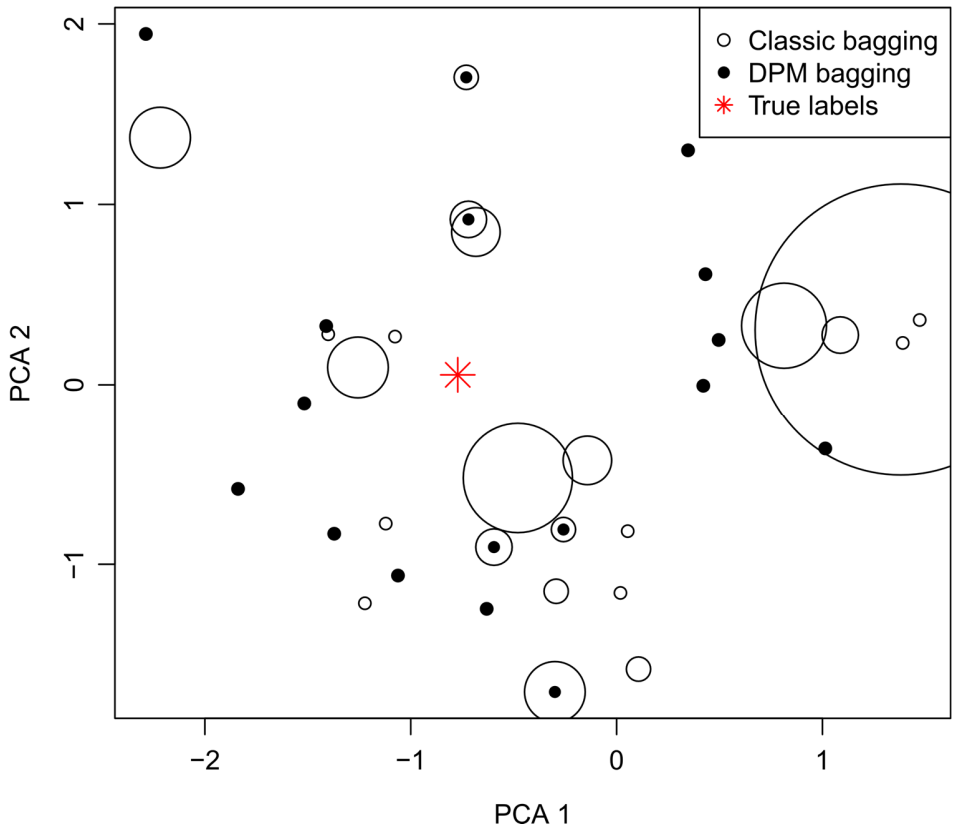

Figure 6.2 DPM bagging evaluation distributions. We visualized the prediction results for whole dataset from each individual model of the classic bagging (open circle, ○) and DPM bagging (solid circle, ●). The size of a circle represents number of models that locate in the center of the circle. The true label given by the dataset is indicated by the star (*).

Furthermore, we assessed, on the same datasets, the integration of DPM clustering into another ensemble method AdaBoost. The *boosting* command in *adabag* with default parameters successfully created classic AdaBoost models for 31 out of the same 41 datasets. Similar to the classic bagging models, each classic AdaBoost model had 100 trees. On average, the DPM AdaBoost reduced about 25% of the trees in the classic AdaBoost method in 31 evaluated datasets. This is understandable because the AdaBoost models are more diverse than those of the bagging technique: 23 of them had 17 equivalent and 6 larger AUC values, while 27 of them had 22 equivalent and 5 smaller error rates. Only three datasets (glass0123vs456, page-blocks13vs2, Pima) returned smaller AUC values and larger error rates when comparing the DPM bagging models to the classic bagging models. The DPM AdaBoost models for these three datasets included 79, 21 and 96 trees. We tuned the parameters of tree number and alpha for model generation (Table 6.6). These results showed that DPM AdaBoost could improve the classic AdaBoost in error rate or AUC value or both.

Dataset	Classic AdaBoost			DPM AdaBoost		
	Tree	Error	AUC	Trees	Error	AUC
abalone9	-	-	-	-	-	-
abalone9vs18	100	0.048	0.8521	100	0.048	0.8521
ecoli0137vs26	100	0.074	0.9669	99	0.074	0.9669
ecoli0vs1	100	0.027	0.9873	61	0.032	0.9891
ecoli1	100	0.107	0.9436	100	0.107	0.9436
ecoli2	100	0.048	0.9536	100	0.048	0.9536
ecoli3	100	0.071	0.9257	99	0.071	0.9248
ecoli4	100	0.024	0.9773	75	0.021	0.9803
glass0	100	0.121	0.9209	100	0.121	0.9209
glass0123vs456	100	0.047	0.9857	79	0.051	0.9856
glass016vs2		0.083	0.8204	99	0.083	0.8185
glass016vs5	-	-	-	-	-	-
glass1	100	0.178	0.8812	100	0.178	0.8812
glass2	-	-	-	-	-	-
glass4	-	-	-	-	-	-

glass5	-	-	-	-	-	-
glass6	100	0.037	0.9734	64	0.028	0.9763
haberman	100	0.324	0.6619	100	0.324	0.6619
newthyroid1	100	0.009	1.0000	34	0.009	1.0000
newthyroid2	100	0.014	0.9982	30	0.014	0.9991
page-blocks0	100	-	-	-	-	-
page-blocks13vs2	100	0.007	1.0000	21	0.009	0.9988
pima	100	0.245	0.7998	96	0.251	0.7976
segment0	100	0.003	0.9999	63	0.003	0.9999
vehicle0	100	0.024	0.9978	100	0.024	0.9978
vehicle1	100	0.197	0.8776	98	0.197	0.8762
vehicle2	100	0.011	0.9992	100	0.011	0.9992
vehicle3	100	0.199	0.8645	97	0.190	0.8643
vowel0	100	0.006	0.9996	69	0.003	0.9996
wisconsin	100	0.040	0.9913	99	0.040	0.9914
yeast05679vs4	100	0.074	0.9099	100	0.074	0.9099
yeast1	-	-	-	-	-	-
yeast1289vs7	-	-	-	-	-	-
yeast1458vs7	-	-	-	-	-	-
yeast1vs7	100	0.048	0.8433	100	0.048	0.8433
yeast2vs4	100	0.041	0.9775	97	0.041	0.9775
yeast2vs8	-	-	-	-	-	-
yeast3	100	0.051	0.9629	100	0.051	0.9629
yeast4	100	0.038	0.9035	100	0.038	0.9016
yeast5	100	0.018	0.9922	70	0.017	0.9925
yeast6	100	0.016	0.9001	100	0.016	0.9001

Table 6.5 Evaluation of DPM AdaBoost

Dataset	**Classic AdaBoost**			**DPM Adaboost**		
	Trees	**Error**	**AUC**	**Trees**	**Error**	**AUC**
100 trees in classic AdaBoost, alpha=0.001 (Default)						
glass0123vs456	100	0.047	0.9857	79	0.051	0.9856
page-blocks13vs2	100	0.007	1.0000	21	0.009	0.9988
pima	100	0.245	0.7998	96	0.251	0.7976
200 trees in classic AdaBoost, alpha=0.001						
glass0123vs456	200	0.051	0.9851	145	0.048	0.9871
pima	200	0.246	0.8088	167	0.233	0.8017
100 trees in classic AdaBoost, alpha=**0.1**						
page-blocks13vs2	100	0.007	1.0000	25	0.007	1.000

Table 6.6 Parameter fine tuning for DPM AdaBoost

6.4 CONCLUSION AND LIMITATIONS

In this project, we have developed a novel ensemble method by integrating a C++ coded DPM approach to enhance the diversity of model committees. We have evaluated the DPM ensemble method using representative online datasets. The results showed that our DPM ensemble method can increase the predictive ability of the majority of datasets when compared to the classic bagging and AdaBoost techniques.

The DPM methods have some limitations, however. Like all machine learning algorithms, they are not applicable to all datasets. Also, the number of trees and alpha parameters must be manually fine-tuned to optimize the final results.

Chapter 7. Meta-analysis to compare intervention strategies of schistosomiasis[4]

7.1 INTRODUCTION

Schistosomiasis is a parasitic disease caused by the blood flukes of the genus Schistosoma. It ranks second after malaria among the global human parasitic diseases in terms of socio-economic and public health importance in tropical and subtropical areas [171]. Worldwide, this neglected tropical disease infects more than 207 million people, with 779 million people in 76 countries at risk [172]. This disease causes 0.2 million deaths [173] and 1.75 to 2 million disability adjusted life lost each year [174].

Three major schistosome species are known to infect humans, including S. haematobium, S. mansoni, and S. japonicum [171]. Schistosomiasis japonica, caused by S. japonicum, is endemic mainly in China, the Philippines, and parts of Indonesia [171]. Concerted controls since the 1950s have dramatically reduced the number of parasites as well as the burden of disease in these endemic areas. However, schistosomiasis japonica remains a major public health concern in China, where it is one of the four priorities for communicable disease control defined by the central government [175]. Currently, the disease remains endemic in the lake regions of five provinces along the middle and lower branch of the Yangtze River, and in some mountainous areas in the Sichuan and Yunnan provinces.

The national strategy for schistosomiasis control has shifted three times in China since its first initiation from a transmission strategy in the mid-1950s to early 1980s, to a

[4] In collaboration with Dr. Wei Wang at the Key Laboratory of Technology for Parasitic Disease Prevention and Control in China

morbidity strategy in the mid-1980s to 2003, then to an integrated strategy from 2004 to present [176]. The morbidity strategy, also known as the conventional strategy, emphasized the synchronous chemotherapy of humans and bovines. The new strategy, developed in 2004, intervenes in the transmission pathway of schistosomiasis japonica mainly through the replacement of bovines with machines for plowing and farming, the prohibition of grazing cattle on the grasslands, improved sanitation, the installation of fecal-matter containers on boats, praziquantel chemotherapy, snail control and health education [177]. This new integrated control strategy has been efficient in reducing the rate of S. japonicum infection in both humans and the intermediate host snails [178-181]. However, the effectiveness of this new integrated strategy varies across earlier reports, and is dependent on its implementation in various endemic regions and different local circumstances [182]. Therefore, we present a systematic literature review and meta-analysis to better evaluate the effectiveness of the new integrated strategy in controlling the transmission of S. japonicum in China.

7.2 DATA QUERY AND EVALUATION

7.2.1 Literature search

We searched all publications pertaining to schistosomiasis control from January 1st, 2000 through December 31st, 2014. Our keywords included “schistosomiasis”, in combination with “integrated control strategy”, “comprehensive control strategy” or “infectious source control measures”. Our electronic databases included PubMed, Web of Science, Embase, Proquest, Cochrane Library, China National Knowledge Infrastructure, the Wanfang Database and the VIP Database. The title and abstract of each publication were read carefully, and the full texts were reviewed.

Both inclusion and exclusion criteria were defined for identifying the publications to be included in our meta-analysis. Inclusion criteria involved: (1) the control measures targeting schistosomiasis japonica; (2) the implementation of the study in China; (3) a detailed description of integrated control interventions with emphasis on control of the infectious source of schistosomiasis; (4) the inclusion of both study and control areas, and an assessment of effectiveness in both groups; (5) the description and evaluation of the prevalence of human S. japonicum infection and snail infection as outcomes of the interventions; and (6) available full text for review. The literature that met the following criteria were excluded: (1) lack of control areas or lack of effectiveness evaluation in control areas; (2) no description of quantitative outcomes of interventions; (3) the original data regarding the outcomes of interventions were not available; and (4) the full text was unavailable.

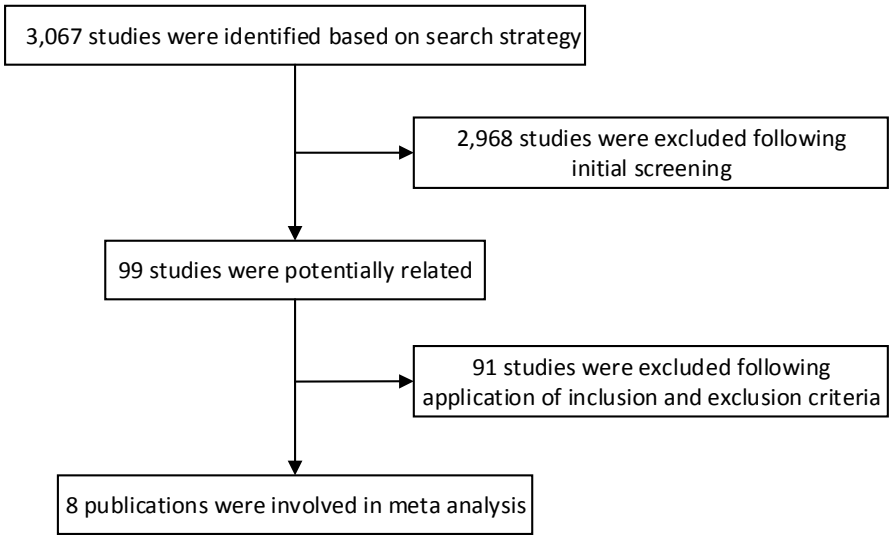

Figure 7.1 Workflow for publication selection. A summary of the publication searching procedures from online searching to final selection.

No.	Study region	Study period	Integrated interventions targeting control of infectious sources	Study measurements	References
1	Anhui province	2002–2003	Replacement of bovines with machines, improvement of sanitation, and building lavatories and latrines	Human *S. japonicum* infection and snail infection	22
2	Mountainous regions of Yunnan province	2006–2007	Improvement of sanitation, and building lavatories and latrines and prohibition of grazing cattle in the grasslands	Human *S. japonicum* infection and snail infection	23
3	Poyang Lake region	2005–2007	Removing cattle from snail-infested grasslands, providing farmers with mechanized farm equipment, improving sanitation by supplying tap water and building lavatories and latrines, providing boats with fecal-matter containers, and implementing an intensive health-education program	Human *S. japonicum* infection and snail infection	12
4	Four provinces of Anhui, Hubei, Hunan and Jiangxi	2005–2008	Removing cattle from snail-infested grasslands, providing farmers with mechanized farm equipment, improving sanitation by supplying tap water and building lavatories and latrines, providing boats with fecal-matter containers, and implementing an intensive health-education program	Human *S. japonicum* infection and snail infection	13
5	Xuancheng city of Anhui province	2006–2007	Replacement of bovines with machines, improvement of sanitation, and building lavatories and latrines	Human *S. japonicum* infection and snail infection	24
6	Jingzhou city of Hubei province	2010–2011	Replacement of bovines with machines, and prohibition of grazing cattle in the grasslands	Human *S. japonicum* infection and snail infection	25
7	Gong'an county of Hubei province	2009–2011	Building fences to limit the grazing area for bovines, building safe pastures for grazing, improving the residents' health conditions and facilities	Human *S. japonicum* infection and snail infection	26
8	Jinxian county along Poyang Lake region	2004–2005	Grazing and marshland isolation, replacing bovines with tractors, and improving access to water and sanitation facilities	Human *S. japonicum* infection and snail infection	27

Table 7.1 Characteristics of the studies enrolled in meta-analysis

A total of 3,067 publications were identified of which 99 were potentially relevant according to our initial screening. Following the application of the inclusion and exclusion criteria, 91 more studies were excluded. Finally, eight papers satisfied all our criteria and were included in the following meta-analysis (Figure 7.1). Five of the eight selected papers included two study areas and two control areas. Table 7.1 describes the general characteristics of the studies included in the analysis.

7.2.2 Publication bias assessment

We evaluated the literature quality using a funnel plot. A funnel plot is a scatter plot of the enrolled individual studies of their effect estimates (i.e. logRR in current project) against the standard error of the effect estimation or other measure of each study's size. In other words, a funnel plot examines the dependence of the effect estimate on the sample size. The presence of publication bias causes an asymmetry in the funnel plot. On the other hand, a symmetric funnel plot indicates no publication bias. Our funnel plots showed no obvious asymmetry (Figure 7.2). To obtain quantitative results, we tested funnel plot asymmetry based upon a linear regression method [183] that uses the *metabias* function in the *meta* package of the R software suite [184]. The funnel plots were observed to be symmetric, with all P values of > 0.05 (Figure 7.3). These results indicated no publication bias present in the literatures used in the meta-analysis.

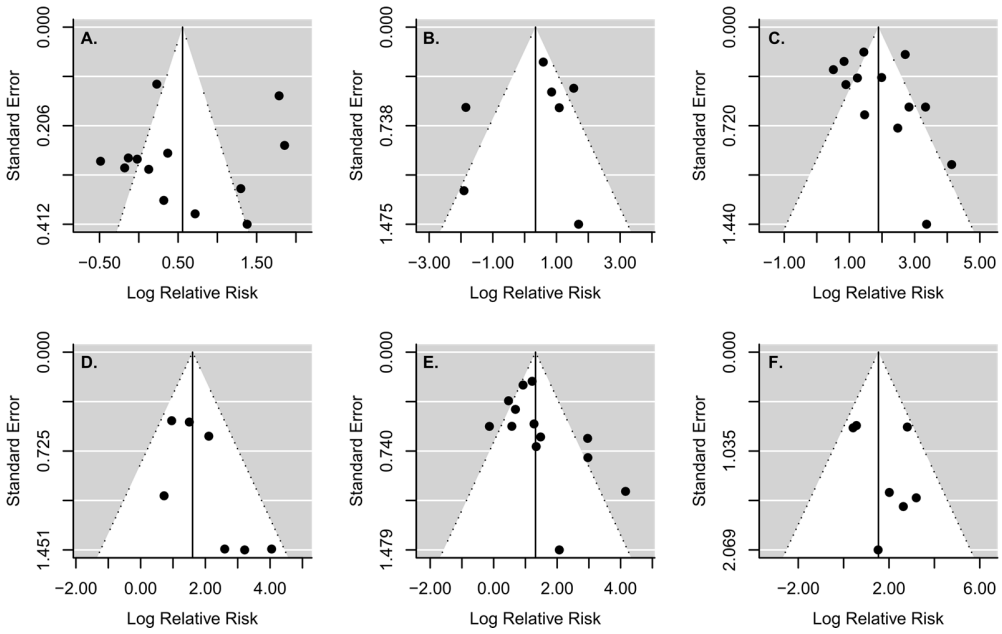

Figure 7.2 Funnel plot. We used the funnel plots to examine the publication bias for studies of the conventional strategy in snail (A) and human (B), the integrated strategy in snails (C) and human (D), and the comparison between these two strategies in snails (E) and human (F).

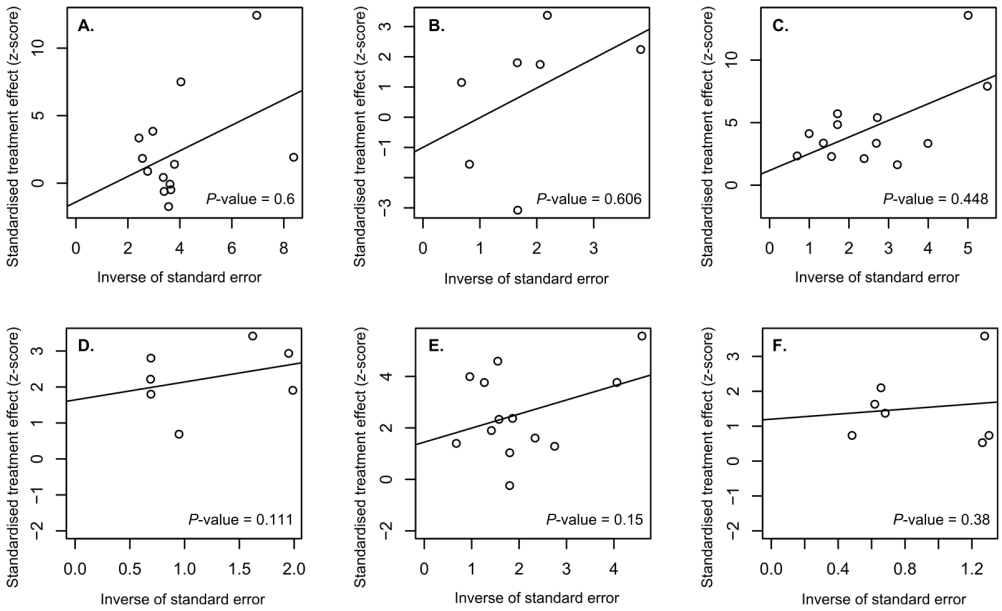

Figure 7.3 Publication bias examination. Linear regression plots were generated to assess the asymmetry of each funnel plot in Figure 7.2.

7.3 META-ANALYSIS

We implemented meta-analysis (fixed- or random-effects models) using the *rma* function of the R package *metafor* [185]. The effect of the new or conventional control strategy in the human/snail study was evaluated with pooled log relative risk (LogRR) (natural logarithm, RR=risk of control without intervention / risk of experiments with intervention) and the corresponding 95% confidence intervals (95% CI). We then calculated the LogRR difference between both strategies and its standard error (SE) as shown below:

$$logRR\ difference\ = logRR\ new\ \ integrated\ strategy - LogRR\ conventional\ strategy\ intervention$$

$$SE(LogRRdifference) = \sqrt{SE(LogRRnewintervention)^2 + SE(LogRRoldintervention)^2}$$

from which we further compared the two strategies with the pooled logRR difference. In all of the analyses, Cochran's Q test and Higgins' I^2 statistics were implemented to measure the heterogeneity between these studies. Cochran's Q, an extensive χ^2 test, examines the significance of difference among multiple studies [186]. Meanwhile, Higgins' I^2 is a transformation of Cochran's Q value and $I^2 > 75\%$ usually indicated unexplained heterogeneity among the investigated studies [187]. A random-effects model was employed if heterogeneity existed in the data source. Otherwise, a fixed-effects model was reported.

All statistical analyses were performed using the *R* software suite, and a *P*-value < 0.05 was considered to be statistically significant.

7.4 RESULTS AND DISCUSSION

Heterogeneity tests revealed the presence of heterogeneity among studies that reported the effects of the conventional strategy on the control of human S. japonicum

infection (I^2 = 90.34, P< 0.001) and snail infection (I^2 = 83.52, P< 0.001), as well as the effects of the new integrated strategy on the control of human infection (I^2 = 86.39, P< 0.001). No heterogeneity was detected among the studies reporting the alteration of snail infection caused by the new integrated strategy (I^2 =10.92, P = 0.361). We then estimated pooled logRR and the corresponding 95% CI using random and fixed effects models, respectively.

We found that the implementation of the conventional strategy caused a reduction in both human S. japonicum infection (logRR = 0.56, 95% CI: 0.12–0.99; Figure 7.4A) and snail infection (logRR = 0.34, 95% CI: -0.69–1.37; Figure 7.4B). Meanwhile, the new integrated strategy significantly reduced both human S. japonicum infection (logRR = 1.89, 95% CI: 1.33–2.46; Figure 7.5A) and snail infection (logRR = 1.61, 95% CI: 1.06–2.15; Figure 7.5B). In other words, the conventional strategy reduced the risk of infection by 1.75-fold (95% CI: 1.13 – 2.69) in humans and 1.40 fold (95% CI: 0.50 – 3.94) in snails. In contrast, the integrated strategy reduced the risk of infection by 6.62-fold (95% CI: 3.78 – 11.70) in humans and by 5.00-fold (95% CI: 2.89 – 8.58) in snails. Further comparison between these two strategies indicated that the integrated strategy was 3.74-fold (95% CI: 2.18 – 6.42) (logRR difference = 1.32, 95% CI: 0.78–1.86; Figure 7.6A) more effective in human infection control and 4.62-fold (95% CI: 2.14 – 10.07) (logRR difference = 1.53, 95% CI: 0.76–2.31; Figure 7.6B) more effective in snail infection.

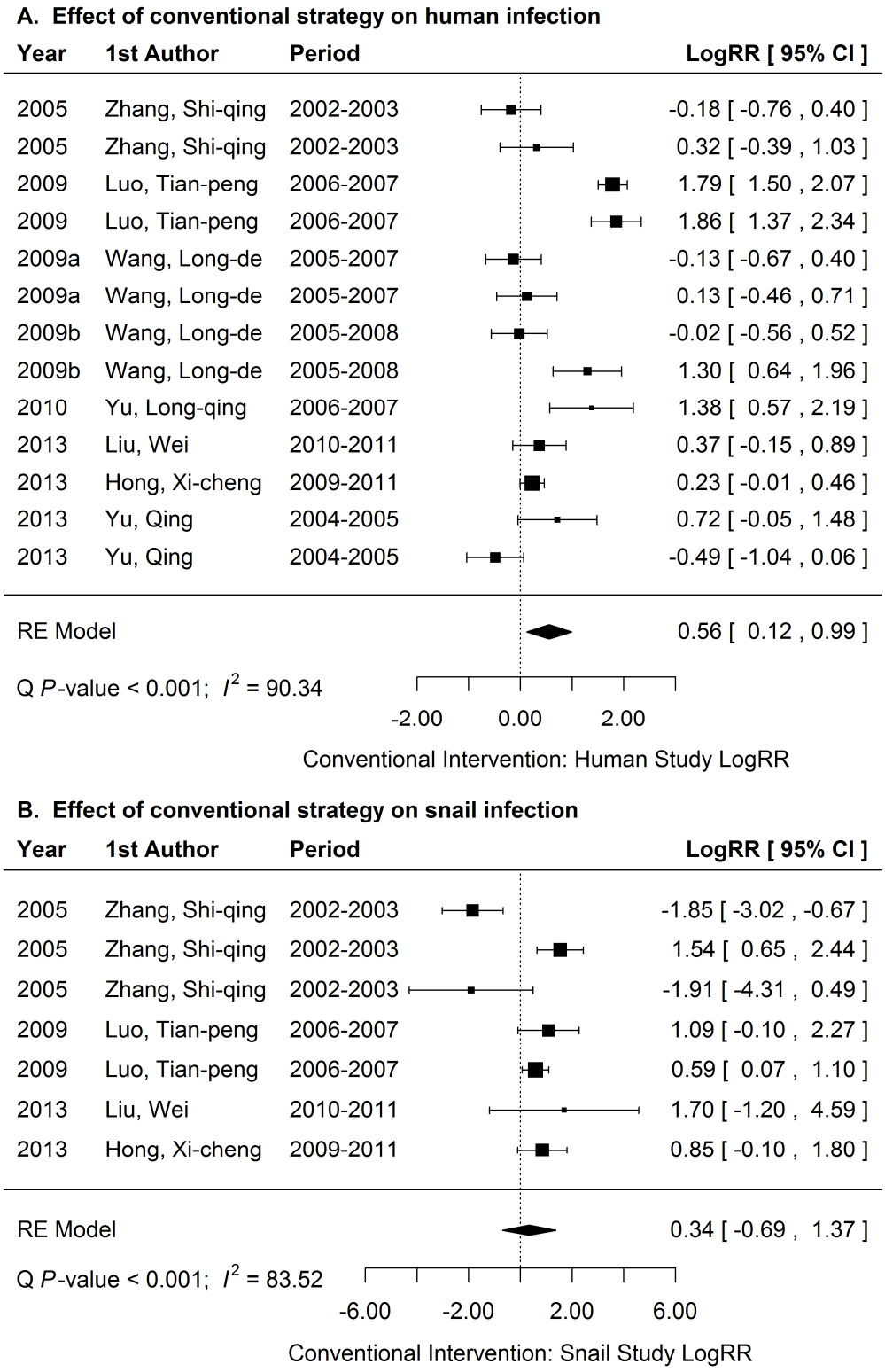

Figure 7.4 Forest plot to evaluate the conventional intervention strategy

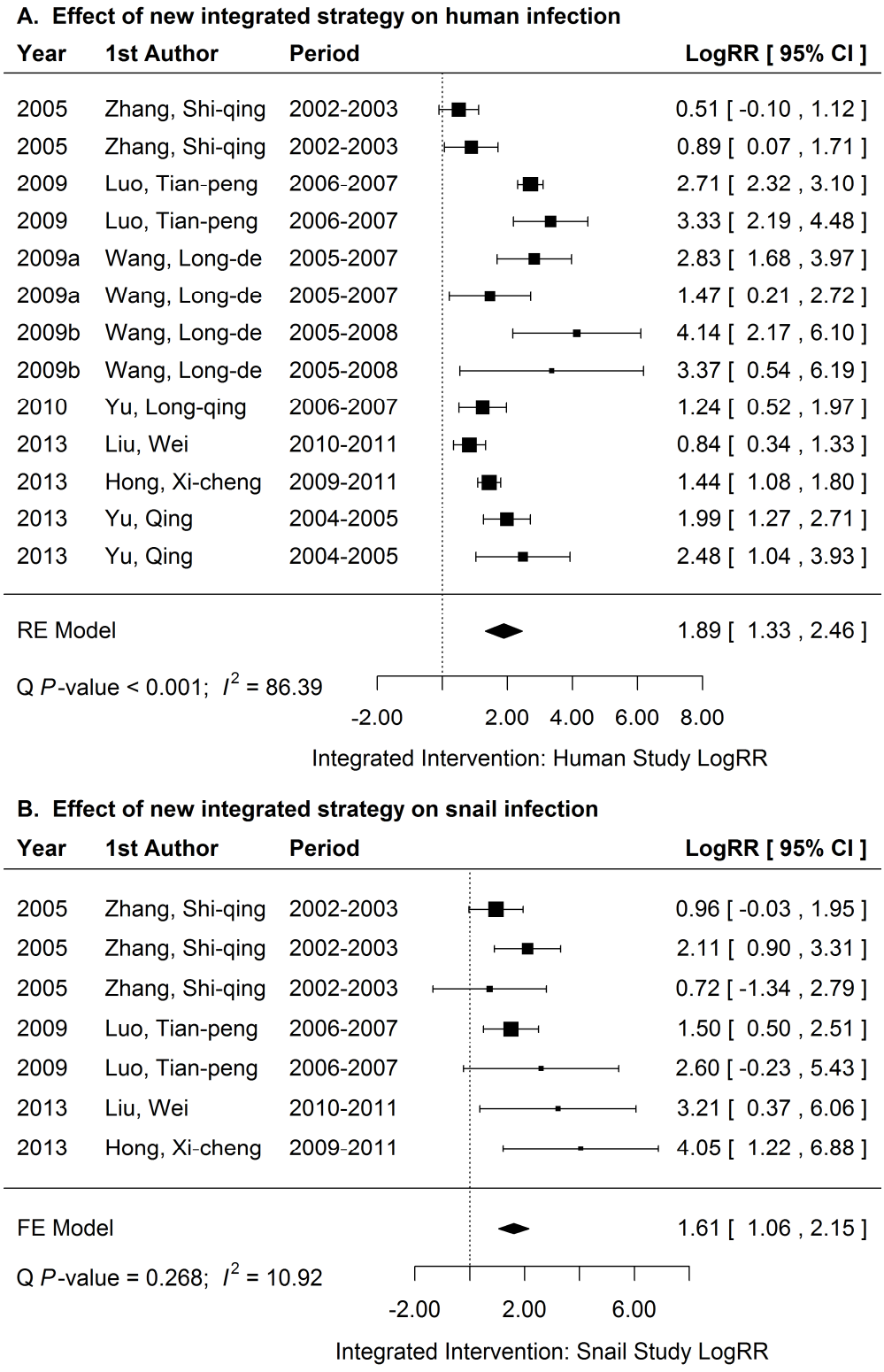

Figure 7.5 Forest plot to evaluate the integrated intervention strategy

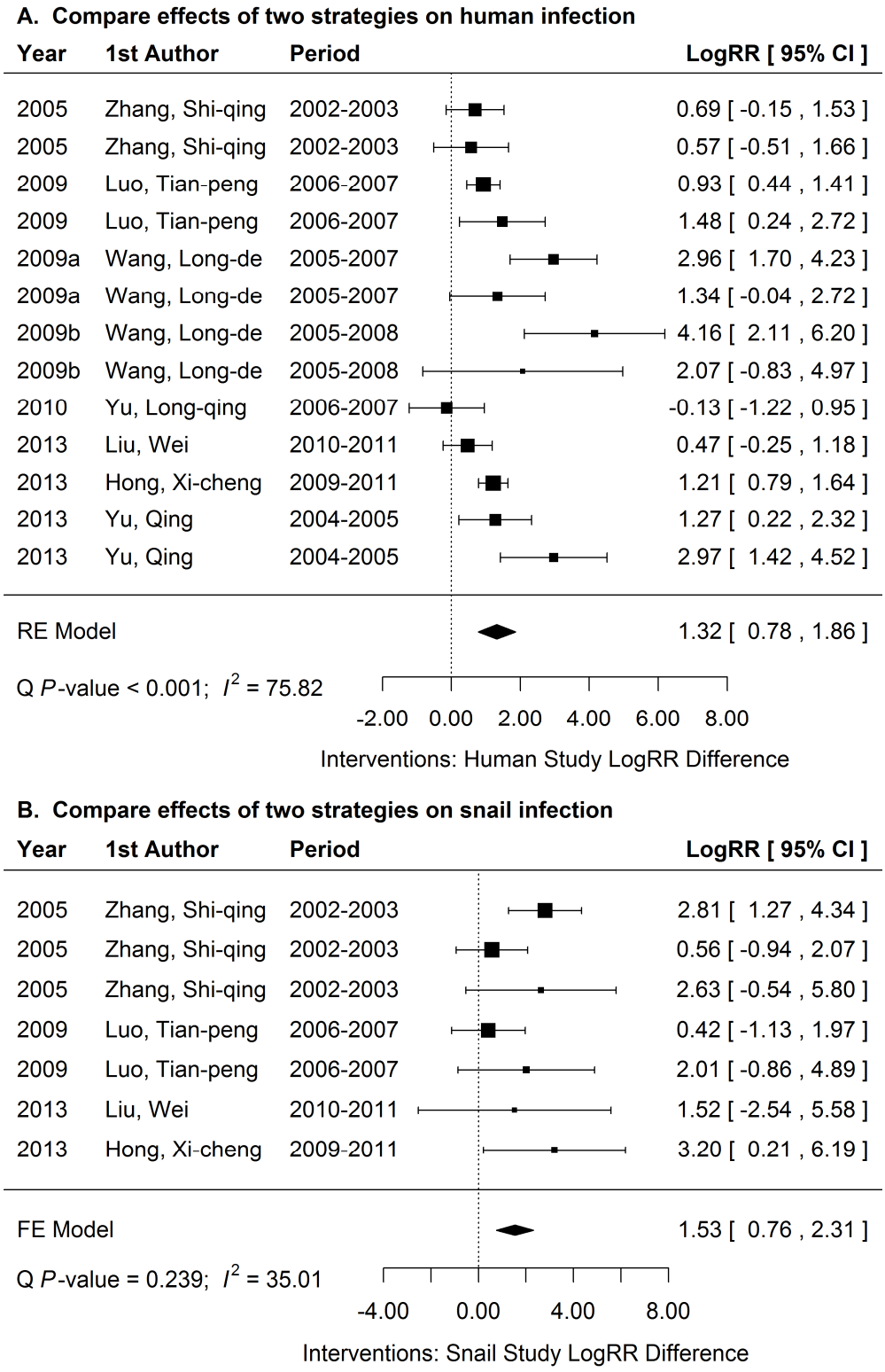

Figure 7.6 Forest plot to compare the integrated with the conventional strategy

The integrated control strategy was designed to reduce the role of bovines and humans as sources of S. japonicum infection. The morbidity (conventional) control strategy mainly involves praziquantel-based chemotherapy, snail control and health education interventions [188]. The praziquantel-based strategy is ineffective in preventing S. japonicum infection and re-infection in China, the Philippines and the African continent [189, 190]. Our findings demonstrate that the new integrated strategy is much more effective than the praziquantel-based conventional strategy in controlling the transmission of S. japonicum in China. We have learned that this expertise is currently being transferred to Africa. The Philippines may also benefit from China's experience and lessons in this important area.

7.5 CONCLUSION AND LIMITATIONS

We analyzed studies from eight eligible publications using meta-analysis to compare the conventional and integrated prevention strategies of schistosomiasis japonica in China. The results showed that the implementation of the new integrated strategy reduced the infection risk by about 3–4 times compared to the conventional strategy. The new integrated strategy is highly effective in controlling the transmission of S. japonicum in China. The same strategy should be recommended to eliminate schistosomiasis japonica in other infected regions, such as Africa and the Philippines.

There are some limitations in this study though. First, only eight eligible studies were enrolled in the meta-analysis. Most of the studies were published in national journals. The research outcomes are more likely to be applicable world-wide if more randomized controlled trials are employed. Second, no stratified analysis was performed. Considering the small number of trials, we were unable to assess the effect related to an

endemic type-specific region, such as the marshland/lake or the hilly/mountainous regions. An optimal control strategy should adapt to each local circumstance in order to facilitate progress towards the elimination of schistosomiasis.

Chapter 8. Conclusions

In this dissertation, we have presented five projects to illustrate the applications of biomedical informatics from multiple perspectives of systems biology, including studies of using genomics, transcriptomics, and clinical data. We have successfully applied the NGS technology to assess the intra-host variation in VEEV and to improve our understanding of how enriched conditions were able to induce a protective phenotype for addiction and visceral leishmaniasis. Moreover, we developed a new ensemble method to improve the predictive ability of classic bagging and AdaBoost data analyses, and validated an integrated intervention strategy for schistosomiasis using meta-analysis.

In chapter 3–5, we successfully utilized NGS techniques to generate unbiased discovery-based transcriptomics and genomics data to investigate addiction phenotype, visceral leishmaniasis, and vaccine development for VEEV, respectively. The first project (Chapter 3) revealed that several pathways play a significant role in addictive behavior and can thus direct the focus on individual differences in susceptibility to addiction. The second project (Chapter 4) was able to identify the differentially expressed genes and altered pathways during visceral leishmaniasis. We found that hamster showed dually activated macrophages and broad inflammatory nature during VL, which was intended to control the L. *donovani* parasite but failed. We have proposed several different future research strategies to better understand this paradox. In the third project (Chapter 5), we investigated the intra-host variations of VEEV in order to enhance the replication fidelity of the live-attenuated vaccine Tc-83. Our results demonstrate that the replication fidelity of Tc-83 can be increased by incorporating three point mutations at the RdRp region. These findings can accelerate the development of the new live

attenuated vaccine for VEEV. All these data analyses were based on NGS techniques to extract additional information and to increase our knowledge and improve our understanding of the complex underlying biology, with a goal to ultimately being able to develop novel prevention strategies, such as vaccines, diagnostic methods and treatment therapies.

In Chapter 6 we developed a novel DPM integrated ensemble classification method to further improve our ability to extract information from biomedical data. It shares the same Bayesian-based approach as our previous intra-host variation project, and thus can be widely applied to classification problems. We evaluated our ensemble method against forty-one online datasets and validated the increased predictive ability of our method. In biomedical informatics, ensemble methods should be particularly useful for identifying novel biomarker panels for diagnosis.

Finally, in Chapter 7 we performed a meta-analysis to compare a new integrated strategy for schistosomiasis control against the established conventional strategy. The result indicates that implementation of the new integrated strategy is a significant improvement that reduces infection risk by ca. three to four times compared to the conventional strategy. This meta-analysis approach is applicable for the evaluation of any new prevention, diagnosis or treatment approaches, such as a new diagnostic method developed using our new DPM ensemble method.

Appendix A. Various Commercial NGS Platforms

The next generation sequencing has been rapidly evolving with increasing accuracy and speed but reducing cost in the past 10 years and will continue growing. According to a 2014 market report, the global NGS market worth was 2.5 billion in 2014 and will increase to $8.7 billion by 2020 [191]. Nowadays, the NGS sequencers are mainly developed, manufactured and sold by a couple companies including Illumina [192], Roche 454 Life Science [193], Life Technologies [194, 195], Pacific Biosciences [196] and Oxford Nanopore Technologies [197]. Here and in the following we exclude the NGS platform from Helicos Biosciences [198], which bankrupted in 2002, as well as a few unpopular platforms such as Polonator [199].

A.1 Illumina NGS platform

Illumina, originally developed by Solexa, currently dominate the NGS platform market with more than 70% market share in 2014. Illumina NGS platform offers 5 series: Genome Analyzer, HiSeq, MiSeq, NextSeq 500 and HiSeq X. Genome Analyzer sequencers were the original Illumina NGS platform, which is no longer available. The HiSeq system includes several instruments: HiSeq 1000, HiSeq 2000, HiSeq 2500, HiSeq 3000, and HiSeq4000 (Table A.1). HiSeq 2000 is the first instrument in this series, which was downgraded to HiSeq 1000 with only single flow cell mode and upgraded to HiSeq 2500 with an additional rapid run mode. HiSeq3000/4000, launched early 2015, adopt the patterned flow cell technology to provide even faster sequencing speed. The HiSeq systems support both single end and paired end sequencing and generate a large number

of read in one run so that they are suitable to analyze large animal or plant genomes. However, the read lengths from HiSeq are relatively small. In contrast, the MiSeq platform returns much longer but fewer reads so they are ideal for small genomes and more appropriate for *de novo* assembly. NextSeq 500, a "HiSeq in a MiSeq" platform, integrates the HiSeq2500 "rapid run" into a MiSeq-sized package. NextSeq system can sequence hundreds of millions of reads in very fast speed. This makes the system suitable for exome, transcriptomics, whole genome and targeted sequencing [200]. The HiSeq X Five / Ten systems consist of 5 / 10 HiSeq X ultra-high-throughput sequencers to enables fast and affordable human whole genome sequencing [201, 202]. (Table A.2)

Sequencer	***HiSeq 1000***	***HiSeq 2000***	***HiSeq 2500***		***HiSeq 3000***	***HiSeq 4000***
Run Mode	N/A	N/A	Rapid Run	High-Output	N/A	N/A
Flow Cells per Run	1	1 or 2	1 or 2	1 or 2	1	1 or 2
Output Range	47-300 Gb	47-600 Gb	10-300 Gb	50-1000 Gb	125-750 Gb	125-1500 Gb
Run Time	1.5-8.5 days	1.5-11 days	7-60 hours	<1-6 days	<1-3.5 days	<1-3.5 days
Reads per Flow Cell†	3 billion	3 billion	300 million	2 billion	2.5 billion	2.5 billion
Maximum Read Length	2 x 100 bp	2 x 100 bp	2 x 250 bp	2 x 125 bp	2 x 150 bp	2 x 150 bp
Launched	2010	2010	2012		2015	2015

Table A.1. Illumina HiSeq platforms adapted from Illumina Webpage [203, 204]

Sequencer	***MiSeq***	***NextSeq 500***		***HiSeq X Five***	***HiSeq X Ten***
Run Mode	N/A	Mid-Output	High-Output	N/A	N/A
Flow Cells per Run	1	1	1	1 or 2	1 or 2
Output Range	0.3-15 Gb	20-39 Gb	30-120 Gb	900-1800 Gb	900-1800 Gb
Run Time	5-55 hours	15-26 hours	12-30 hours	<3 days	<3 days

Reads per Flow Cell†	25 million‡	130 million	400 million	3 billion	3 billion
Maximum Read Length	2 x 300 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp
Launched	2011	2014		2014	2014

Table A.2 Other Illumina platforms adapted from Illumina Webpage [204]

A.2 Roche 454 platform

The Roche 454 platform provided by the 454 life science has experienced 6 systems: GS 20, GS FLX, GS FLX Titanium, GS Junior, GS FLX+ and GS Junior+ System (Table A.3). GS 20 was the first NGS sequencer and is out of data now. All the other available Roche 454 platforms return relatively long read length, which makes them favorable to *de novo* assemblies of microbial genomes, bacterial artificial chromosome and plastids, and examination of 16S variable regions and other targeted amplicon sequences [205, 206]. The overall output is not as high as Illumina platform so they are less cost-effective for transcriptome or larger genome studies.

Sequencer	***GS20***	***GS FLX***	***GS FLX Titanium***	*GS FLX+*	***GS Junior***	***GS Junior+***
Typical Throughput	~20 Mb	~100 Mb	450 Mb	700 Mb	~35 Mb	~70 Mb
Run Time	5.5 hours	8 hours	10 hours	23 hours	10 hours	18 hours
Reads per Run	~20,000‡	~400,000	~1 million shotgun, 700,000 amplicon	~1 million shotgun	~100,000 shotgun, 70,000 amplicon	~100,000 shotgun, 70,000 amplicon
Read Length	~100 bp	≤ 300 bp	≤ 600 bp	≤ 1000 bp	~400 bp	~700 bp
Launched	2005	2006	2008	2010	2011	2014

Table A.3 Roche / 454 platforms adapted from 454 Webpage [207, 208]

A.3 SOLiD and Ion Torrent platforms

Life technologies have two NGS platforms SOLiD (Table A.4) and Ion Torrent (Table A.5). The early platforms SOLiD1/2/3 are out of date. The PI and 4hq platforms have been gradually substituted by the 5500 and 5500xl platforms, which can be upgraded to the 5500 Wildfire or 5500xL Wildfire platforms. Similar to Illumina HiSeq platforms, the SOLiD platforms can generate a large number of reads per run but the read length is relatively small. This makes the SOLiD platforms suitable for differential transcript expression and re-sequencing for large genome but difficult for *de novo* assembly. The Ion Torrent platform has three systems: Ion Personal Genome Machine (PGM), Ion Proton and Ion PGM Dx. Both PGM and Proton experience several versions of chips, aiming to provide optional run time and output. Considering their limited throughput, the Ion PGM is ideal for ideal for sequencing amplicons, small genomes or targeting of small regions within a genome, while the Proton is suitable for sequencing transcriptome, exome and medium sized genomes. Both are designed for basic research. The PGM Dx system is a class II medical device for clinical use proved by the U.S. Food and Drug Administration (FDA) in Sep 2014.

Sequencer	*Typical Throughput*	*Run Time*	*Max Reads per Run*	*Read Length*	*Launched Year*
1	~3 Gb	--	~40 million	25bp	2007
2	3-6 Gb	6-10 days	~115 million	2 x 25 bp	2008
3	~20 Gb	72 hours	~320 million	2 x 50 bp	2008
4	~100 Gb	3-12 days	~1.4 billion	2 x 50 bp	2010
4hq	~300 Gb	3-14 days	~2.4 billion	2 x 75 bp	2010
PI	~50 Gb	1 day	~800 million	2 x 75 bp	2010
5500	~48 Gb	6 days	~400 million	75 bp, 2 x 60 bp	2010
5500xl	~95 Gb	6 days	~800 million	75 bp, 2 x 60 bp	2011
5500 W	~120 Gb	10 days	~1.2 billion	75 bp, 2 x 50 bp	2012

5500xl W	~240 Gb	10 days	~2.4 billion	75 bp, 2 x 50 bp	2012

Table A.4 SOLiD platforms from Life Technologies [209, 210]

Sequencer	*Typical Throughput*	*Run Time*	*Max Reads per Run*	*Read Length*	*Launched Year*
PGM	≤2 Gb	2.3-7.3 hours	5.5 million	35-400 bp	2010
Proton	≤10 Gb	2-4 hours	60-80 million	≤200 bp	2012
PGM Dx	≤1 Gb	<4.5 hours	--	≤200 bp	2014

Table A.5 Ion Torrent platforms from Life Technologies [211]

A.4 PacBio platform

Pacific Biosciences provides two single molecule sequencing systems PacBio RS and PacBio RSII. Both generate extraordinary long reads and extremely high accurate reads in a couple hours (Table 6), making them ideal for *de novo* assembly. Their unique sequencing mechanism also enable them the capability to study base modifications such as characterization of genetic variation, methylation analysis, microbiology studies, etc.

Sequencer	*Throughput*	*Run Time*	*Reads per SMRT Cell*	*Average Read Length*	*Launched Year*
PacBio RS	102 Mb	2 hours	22,375	~5k bp	2011
PacBio RS II	500Mb-1Gb	4 hours	~500,000	>10k bp	2013

Table A.6 PacBio platforms adapted from Product Brochure [212]

A.5 Oxford nanopore platform

Oxford Nanopore Technologies is another company that provides single molecule sequencers using protein nanopores. They introduced the GridION platform as well the portable single-molecule sequencer MinION in 2012 and announced PromethION in late

2014 [213]. All three platforms have no fix run time that can be stopped flexibly according to the data demands.

The NGS data output has increased at least twice every year since it was invented, which outpaces the Moore's law. Meanwhile, the sequencing cost has been dropping faster than the Moore's law (Figure 2). Nowadays, the sequencing of a human whole-genome cost no more than $1000 by using the Illumina HiSeq X Ten system, which only takes a couple weeks from the sample preparation to the final result after data analysis [201]. In comparison, the first human genome project completed in 2003, which had experienced about 15 years since it was first articulated in 1988 and had cost about $3 billion over this period [214]. Additionally, the size of NGS sequencers has becoming smaller and smaller but the output read length is longer and longer. The size MinION from the Oxford Nanopore Technologies is so small that comparable to a packet of chewing gum [215] and the PacBio RS II system from Pacific Bioscience can generate reads with more than 20k base pairs. Due to all these benefits, the NGS sequencers have become prominent tools in biological and biomedical research.

Bibliography/References

1. Health, N.I.o., *"Talking Glossary of Genetic Terms."*, National Human Genome Research Institute: http://www.genome.gov/glossary/
2. Kadakkuzha, B.M. and S.V. Puthanveettil, *Genomics and proteomics in solving brain complexity.* Mol Biosyst, 2013. **9**(7): p. 1807-21.
3. HL, W., *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche*. 1920, Jena: Verlag Fischer.
4. Jou, W.M., et al., *Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein.* Nature, 1972. **237**(5350): p. 82-88.
5. Padmanabhan, R. and R. Wu, *Nucleotide sequence analysis of DNA. IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis.* Biochem Biophys Res Commun, 1972. **48**(5): p. 1295-302.
6. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
7. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
8. Morozova, O. and M.A. Marra, *Applications of next-generation sequencing technologies in functional genomics.* Genomics, 2008. **92**(5): p. 255-264.
9. Bishop, C.M., *Pattern Recognition and Machine Learning*. 2006: Springer.
10. Nilsson, N.J., *Introduction to Machine Learning*. 2015, Stanford AI Lab: Unpublished.
11. O'Rourke, K., *An historical perspective on meta-analysis: dealing quantitatively with varying study results.* Journal of the Royal Society of Medicine, 2007. **100**(12): p. 579-582.
12. Glass, G.V., *Primary, Secondary, and Meta-Analysis of Research.* Educational Researcher, 1976. **5**(10): p. 3-8.
13. Haidich, A.B., *Meta-analysis in medical research.* Hippokratia, 2010. **14**(Suppl 1): p. 29-37.
14. Nose, M., et al., *Clinical interventions for treatment non-adherence in psychosis: meta-analysis.* Br J Psychiatry, 2003. **183**: p. 197-206.
15. Inagaki, M., et al., *Interventions to prevent repeat suicidal behavior in patients admitted to an emergency department for a suicide attempt: A meta-analysis.* J Affect Disord, 2014. **175c**: p. 66-78.
16. McKay, D., et al., *Efficacy of cognitive-behavioral therapy for obsessive-compulsive disorder.* Psychiatry Res, 2014.
17. Barnard, N.D., S.M. Levin, and Y. Yokoyama, *A Systematic Review and Meta-Analysis of Changes in Body Weight in Clinical Trials of Vegetarian Diets.* J Acad Nutr Diet, 2015.
18. Bassuk, E.L., M.K. Richard, and A. Tsertsvadze, *The Prevalence of Mental Illness in Homeless Children: A Systematic Review and Meta-Analysis.* J Am Acad Child Adolesc Psychiatry, 2015. **54**(2): p. 86-96.e2.
19. Hanson, S. and A. Jones, *Is there evidence that walking groups have health benefits? A systematic review and meta-analysis.* Br J Sports Med, 2015.

20. Farinelli, L., E. Kawashima, and P. Mayer, *Method of nucleic acid amplification*. 1998, Google Patents.
21. Farinelli, L., E.H. Kawashima, and P. Mayer, *Method of nucleic acid sequencing*. 1998, Google Patents.
22. Ronaghi, M., et al., *Real-Time DNA Sequencing Using Detection of Pyrophosphate Release.* Analytical Biochemistry, 1996. **242**(1): p. 84-89.
23. Brenner, S., et al., *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nat Biotechnol, 2000. **18**(6): p. 630-4.
24. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.
25. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.
26. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
27. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
28. Wetterstrand, K., *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. 2014: www.genome.gov/sequencingcosts.
29. Knierim, E., et al., *Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing.* PLoS ONE, 2011. **6**(11): p. e28240.
30. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
31. Metzker, M.L., *Emerging technologies in DNA sequencing.* Genome Res, 2005. **15**(12): p. 1767-76.
32. Sequencing, T.S.I., *Illumina Two-Channel SBS Sequencing Technology*. 2014, Technology Spotlight: Illumina Sequencing. p. http://res.illumina.com/documents/products/techspotlights/techspotlight_two-channel_sbs.pdf.
33. *How is genome sequencing done?*, 454 Life Sciences: http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf.
34. *Ion Torrent - Amplicon Sequencing*, Ion Torrent by Life Technologies |: https://www3.appliedbiosystems.com/cms/groups/applied_markets_marketing/documents/generaldocuments/cms_094273.pdf.
35. Eid, J., et al., *Real-Time DNA Sequencing from Single Polymerase Molecules.* Science, 2009. **323**(5910): p. 133-138.
36. Levene, M.J., et al., *Zero-mode waveguides for single-molecule analysis at high concentrations.* Science, 2003. **299**(5607): p. 682-6.
37. Mikheyev, A.S. and M.M.Y. Tin, *A first look at the Oxford Nanopore MinION sequencer.* Molecular Ecology Resources, 2014. **14**(6): p. 1097-1102.
38. *Illumina Sequencing Technology*. 2010, Illumina, Inc: http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.
39. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

40. Green, T.A., B.J. Gehrke, and M.T. Bardo, *Environmental enrichment decreases intravenous amphetamine self-administration in rats: dose-response functions for fixed- and progressive-ratio schedules.* Psychopharmacology (Berl), 2002. **162**(4): p. 373-8.
41. Thiel, K.J., et al., *Environmental living conditions introduced during forced abstinence alter cocaine-seeking behavior and Fos protein expression.* Neuroscience, 2010. **171**(4): p. 1187-96.
42. Lichti, C.F., et al., *Environmental enrichment alters protein expression as well as the proteomic response to cocaine in rat nucleus accumbens.* Front Behav Neurosci, 2014. **8**: p. 246.
43. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Meth, 2008. **5**(7): p. 621-628.
44. Oshlack, A. and M.J. Wakefield, *Transcript length bias in RNA-seq data confounds systems biology.* Biol Direct, 2009. **4**: p. 14.
45. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biology, 2010. **11**(3): p. R25-R25.
46. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**: p. 94.
47. Risso, D., et al., *GC-content normalization for RNA-Seq data.* BMC Bioinformatics, 2011. **12**: p. 480.
48. Risso, D., et al., *Normalization of RNA-seq data using factor analysis of control genes or samples.* Nat Biotechnol, 2014. **32**(9): p. 896-902.
49. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
50. Yu, D., W. Huber, and O. Vitek, *Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size.* Bioinformatics, 2013. **29**(10): p. 1275-82.
51. Li, J. and R. Tibshirani, *Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data.* Stat Methods Med Res, 2013. **22**(5): p. 519-36.
52. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
53. Landau, W.M. and P. Liu, *Dispersion estimation and its effect on test performance in RNA-seq data analysis: a simulation-based comparison of methods.* PLoS One, 2013. **8**(12): p. e81415.
54. Ingenuity® Systems, *Data were analyzed through the use of IPA.* www.ingenuity.com.
55. Larson, E.B., et al., *Over-Expression of CREB in the Nucleus Accumbens Shell Increases Cocaine Reinforcement in Self-Administering Rats.* The Journal of neuroscience : the official journal of the Society for Neuroscience, 2011. **31**(45): p. 16447-16457.
56. Kalsner, S., *Cocaine sensitization of coronary artery contractions: mechanism of drug-induced spasm.* J Pharmacol Exp Ther, 1993. **264**(3): p. 1132-40.

57. World Health Organization, *Leishmaniasis: Situation and trends*, in *Global Health Observatory (GHO)* 2014: World Health Organization, http://www.who.int/gho/neglected_diseases/leishmaniasis/en/.
58. Osorio, Y., et al., *Identification of small molecule lead compounds for visceral leishmaniasis using a novel ex vivo splenic explant model system.* PLoS neglected tropical diseases, 2011. **5**(2).
59. Osorio, E., et al., *Progressive visceral leishmaniasis is driven by dominant parasite-induced STAT6 activation and STAT6-dependent host arginase 1 expression.* PLoS pathogens, 2012. **8**(1).
60. Hammond, S., et al., *Chinese hamster genome database: An online resource for the CHO community at www.CHOgenome.org.* Biotechnology and Bioengineering, 2012. **109**(6): p. 1353-1356.
61. Di Palma, F., et al., *The Draft Genome of Mesocricetus auratus.* Unpublished, Unpublished.
62. Andrews, S., *FastQC A Quality Control tool for High Throughput Sequence Data*: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
63. Downing, T., et al., *Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance.* Genome Res, 2011. **21**(12): p. 2143-56.
64. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Meth, 2012. **9**(4): p. 357-359.
65. Hannon Lab, *FASTX Toolkit*: http://hannonlab.cshl.edu/fastx_toolkit/index.html.
66. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotech, 2011. **29**(7): p. 644-652.
67. Bao, E., T. Jiang, and T. Girke, *BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences.* Bioinformatics, 2013. **29**(10): p. 1250-9.
68. Camacho, C., et al., *BLAST Command Line Applications User Manual.* In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1763/, 2008 Jun 23 [Updated 2013 Jul 30].
69. Robinson, M., D. McCarthy, and G. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics (Oxford, England), 2010. **26**(1): p. 139-140.
70. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome biology, 2010. **11**(10).
71. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.
72. Rousseau, D., et al., *In vivo involvement of polymorphonuclear neutrophils in Leishmania infantum infection.* BMC Microbiol, 2001. **1**: p. 17.
73. Osorio, Y., et al., *Identification of small molecule lead compounds for visceral leishmaniasis using a novel ex vivo splenic explant model system.* PLoS Negl Trop Dis, 2011. **5**(2): p. e962.

74. Osorio, E.Y., et al., *Progressive visceral leishmaniasis is driven by dominant parasite-induced STAT6 activation and STAT6-dependent host arginase 1 expression.* PLoS Pathog, 2012. **8**(1): p. e1002417.
75. Osorio, E.Y., et al., *Growth factor and Th2 cytokine signaling pathways converge at STAT6 to promote arginase expression in progressive experimental visceral leishmaniasis.* PLoS Pathog, 2014. **10**(6): p. e1004165.
76. Osterreicher, C.H., et al., *Fibroblast-specific protein 1 identifies an inflammatory subpopulation of macrophages in the liver.* Proc Natl Acad Sci U S A, 2011. **108**(1): p. 308-13.
77. Mor-Vaknin, N., et al., *Vimentin is secreted by activated macrophages.* Nat Cell Biol, 2003. **5**(1): p. 59-63.
78. Rodriguez, N.E., H.K. Chang, and M.E. Wilson, *Novel program of macrophage gene expression induced by phagocytosis of Leishmania chagasi.* Infect Immun, 2004. **72**(4): p. 2111-22.
79. Buates, S. and G. Matlashewski, *General suppression of macrophage gene expression during Leishmania donovani infection.* J Immunol, 2001. **166**(5): p. 3416-22.
80. Chaussabel, D., et al., *Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites.* Blood, 2003. **102**(2): p. 672-81.
81. Griffith, J.W., C.L. Sokol, and A.D. Luster, *Chemokines and chemokine receptors: positioning cells for host defense and immunity.* Annu Rev Immunol, 2014. **32**: p. 659-702.
82. Rousseau, D., et al., *Sustained parasite burden in the spleen of Leishmania infantum-infected BALB/c mice is accompanied by expression of MCP-1 transcripts and lack of protection against challenge.* Eur Cytokine Netw, 2001. **12**(2): p. 340-7.
83. Cotterell, S.E., C.R. Engwerda, and P.M. Kaye, *Leishmania donovani infection initiates T cell-independent chemokine responses, which are subsequently amplified in a T cell-dependent manner.* Eur J Immunol, 1999. **29**(1): p. 203-14.
84. Peters, N.C., et al., *In vivo imaging reveals an essential role for neutrophils in leishmaniasis transmitted by sand flies.* Science, 2008. **321**(5891): p. 970-4.
85. Vasconcelos, C.O., et al., *Distinct cellular migration induced by Leishmania infantum chagasi and saliva from Lutzomyia longipalpis in a hemorrhagic pool model.* Rev Inst Med Trop Sao Paulo, 2014. **56**(1): p. 21-7.
86. Friedman, A.D., *Transcriptional control of granulocyte and monocyte development.* Oncogene, 2007. **26**(47): p. 6816-28.
87. Friedman, A.D., *C/EBPalpha induces PU.1 and interacts with AP-1 and NF-kappaB to regulate myeloid development.* Blood Cells Mol Dis, 2007. **39**(3): p. 340-3.
88. Murray, P.J. and T.A. Wynn, *Protective and pathogenic functions of macrophage subsets.* Nat Rev Immunol, 2011. **11**(11): p. 723-37.
89. Watkins, S.K., et al., *IL-12 rapidly alters the functional profile of tumor-associated and tumor-infiltrating macrophages in vitro and in vivo.* J Immunol, 2007. **178**(3): p. 1357-62.

90. Murray, P.J., et al., *Macrophage activation and polarization: nomenclature and experimental guidelines.* Immunity, 2014. **41**(1): p. 14-20.
91. Murray, H.W. and C.F. Nathan, *Macrophage microbicidal mechanisms in vivo: reactive nitrogen versus oxygen intermediates in the killing of intracellular visceral Leishmania donovani.* J Exp Med, 1999. **189**(4): p. 741-6.
92. Raes, G., et al., *Macrophage galactose-type C-type lectins as novel markers for alternatively activated macrophages elicited by parasitic infections and allergic airway inflammation.* J Leukoc Biol, 2005. **77**(3): p. 321-7.
93. Ibrahim, M.K., et al., *The malnutrition-related increase in early visceralization of Leishmania donovani is associated with a reduced number of lymph node phagocytes and altered conduit system flow.* PLoS Negl Trop Dis, 2013. **7**(8): p. e2329.
94. Cyktor, J.C. and J. Turner, *Interleukin-10 and immunity against prokaryotic and eukaryotic intracellular pathogens.* Infect Immun, 2011. **79**(8): p. 2964-73.
95. Schreiber, T., et al., *Autocrine IL-10 induces hallmarks of alternative activation in macrophages and suppresses antituberculosis effector mechanisms without compromising T cell immunity.* J Immunol, 2009. **183**(2): p. 1301-12.
96. Spence, S., et al., *Suppressors of cytokine signaling 2 and 3 diametrically control macrophage polarization.* Immunity, 2013. **38**(1): p. 66-78.
97. Whyte, C.S., et al., *Suppressor of cytokine signaling (SOCS)1 is a key determinant of differential macrophage activation and function.* J Leukoc Biol, 2011. **90**(5): p. 845-54.
98. Melby, P.C., et al., *The hamster as a model of human visceral leishmaniasis: progressive disease and impaired generation of nitric oxide in the face of a prominent Th1-like cytokine response.* J Immunol, 2001. **166**(3): p. 1912-20.
99. Perez, L.E., et al., *Reduced nitric oxide synthase 2 (NOS2) promoter activity in the Syrian hamster renders the animal functionally deficient in NOS2 activity and unable to control an intracellular pathogen.* J Immunol, 2006. **176**(9): p. 5519-28.
100. Hailu, A., et al., *Elevated plasma levels of interferon (IFN)-gamma, IFN-gamma inducing cytokines, and IFN-gamma inducible CXC chemokines in visceral leishmaniasis.* Am J Trop Med Hyg, 2004. **71**(5): p. 561-7.
101. Saldarriaga, O., et al., *Identification of hamster inducible nitric oxide synthase (iNOS) promoter sequences that influence basal and inducible iNOS expression.* Journal of leukocyte biology, 2012. **92**(1): p. 205-218.
102. Fultz, M.J., et al., *Induction of IFN-gamma in macrophages by lipopolysaccharide.* Int Immunol, 1993. **5**(11): p. 1383-92.
103. Di Marzio, P., et al., *Interferon gamma upregulates its own gene expression in mouse peritoneal macrophages.* J Exp Med, 1994. **179**(5): p. 1731-6.
104. Fenton, M.J., et al., *Induction of gamma interferon production in human alveolar macrophages by Mycobacterium tuberculosis.* Infect Immun, 1997. **65**(12): p. 5149-56.
105. Matsumura, T., et al., *Interferon-gamma-producing immature myeloid cells confer protection against severe invasive group A Streptococcus infections.* Nat Commun, 2012. **3**: p. 678.

106. Darwich, L., et al., *Secretion of interferon-γ by human macrophages demonstrated at the single-cell level after costimulation with interleukin (IL)-12 plus IL-18.* Immunology, 2009. **126**(3): p. 386-393.
107. Katz, J.B., A.J. Muller, and G.C. Prendergast, *Indoleamine 2,3-dioxygenase in T-cell tolerance and tumoral immune escape.* Immunol Rev, 2008. **222**: p. 206-21.
108. Makala, L.H., et al., *Leishmania major attenuates host immunity by stimulating local indoleamine 2,3-dioxygenase expression.* J Infect Dis, 2011. **203**(5): p. 715-25.
109. Wang, X.F., et al., *The role of indoleamine 2,3-dioxygenase (IDO) in immune tolerance: Focus on macrophage polarization of THP-1 cells.* Cell Immunol, 2014. **289**(1-2): p. 42-48.
110. Hall, C.J., et al., *Immunoresponsive gene 1 augments bactericidal activity of macrophage-lineage cells by regulating beta-oxidation-dependent mitochondrial ROS production.* Cell Metab, 2013. **18**(2): p. 265-78.
111. Vats, D., et al., *Oxidative metabolism and PGC-1beta attenuate macrophage-mediated inflammation.* Cell Metab, 2006. **4**(1): p. 13-24.
112. Namgaladze, D. and B. Brune, *Fatty acid oxidation is dispensable for human macrophage IL-4-induced polarization.* Biochim Biophys Acta, 2014. **1841**(9): p. 1329-35.
113. Huang, S.C.-C., et al., *Cell-intrinsic lysosomal lipolysis is essential for alternative activation of macrophages.* Nat Immunol, 2014. **15**(9): p. 846-855.
114. Veress, B., et al., *Morphology of the spleen and lymph nodes in fatal visceral leishmaniasis.* Immunology, 1977. **33**(5): p. 605-10.
115. Silva, L.C., et al., *Canine visceral leishmaniasis as a systemic fibrotic disease.* Int J Exp Pathol, 2013.
116. Marra, F., et al., *Increased expression of monocyte chemotactic protein-1 during active hepatic fibrogenesis: correlation with monocyte infiltration.* Am J Pathol, 1998. **152**(2): p. 423-30.
117. Iredale, J.P., A. Thompson, and N.C. Henderson, *Extracellular matrix degradation in liver fibrosis: Biochemistry and regulation.* Biochim Biophys Acta, 2013. **1832**(7): p. 876-83.
118. Novo, E., et al., *Cellular and molecular mechanisms in liver fibrogenesis.* Arch Biochem Biophys, 2014. **548c**: p. 20-37.
119. Mannaerts, I., et al., *Gene Expression Profiling of Early Hepatic Stellate Cell Activation Reveals a Role for Igfbp3 in Cell Migration.* PLoS ONE, 2013. **8**(12): p. e84071.
120. Sekiya, Y., et al., *Suppression of hepatic stellate cell activation by microRNA-29b.* Biochemical and Biophysical Research Communications, 2011. **412**(1): p. 74-79.
121. Hori, Y., et al., *Matrix metalloproteinase-2 stimulates collagen-I expression through phosphorylation of focal adhesion kinase in rat cardiac fibroblasts.* Am J Physiol Cell Physiol, 2012. **303**(9): p. C947-53.
122. Honeyman, L., et al., *MicroRNA profiling implicates the insulin-like growth factor pathway in bleomycin-induced pulmonary fibrosis in mice.* Fibrogenesis Tissue Repair, 2013. **6**(1): p. 16.

123. Lin, N., et al., *NP603, a novel and potent inhibitor of FGFR1 tyrosine kinase, inhibits hepatic stellate cell proliferation and ameliorates hepatic fibrosis in rats.* Am J Physiol Cell Physiol, 2011. **301**(2): p. C469-77.
124. Antoniou, K.M., et al., *Expression analysis of angiogenic growth factors and biological axis CXCL12/CXCR4 axis in idiopathic pulmonary fibrosis.* Connect Tissue Res, 2010. **51**(1): p. 71-80.
125. Yang, L. and E. Seki, *Toll-like receptors in liver fibrosis: cellular crosstalk and mechanisms.* Front Physiol, 2012. **3**: p. 138.
126. Cabrera, S., et al., *Overexpression of MMP9 in macrophages attenuates pulmonary fibrosis induced by bleomycin.* Int J Biochem Cell Biol, 2007. **39**(12): p. 2324-38.
127. Pratt, W.B., et al., *Role of molecular chaperones in steroid receptor action.* Essays Biochem, 2004. **40**: p. 41-58.
128. Tchen, C.R., et al., *Glucocorticoid regulation of mouse and human dual specificity phosphatase 1 (DUSP1) genes: unusual cis-acting elements and unexpected evolutionary divergence.* J Biol Chem, 2010. **285**(4): p. 2642-52.
129. Joanny, E., et al., *Anti-inflammatory effects of selective glucocorticoid receptor modulators are partially dependent on up-regulation of dual specificity phosphatase 1.* Br J Pharmacol, 2012. **165**(4b): p. 1124-36.
130. Martinez, F.O., *The transcriptome of human monocyte subsets begins to emerge.* J Biol, 2009. **8**(11): p. 99.
131. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity.* Nature protocols, 2013. **8**(8): p. 10.1038/nprot.2013.084.
132. Centers for Disease Control and Prevention, *A CDC framework for preventing infectious diseases: sustaining the essentials and innovating for the future* 2011.
133. Pronker, E.S., et al., *Risk in Vaccine Research and Development Quantified.* PLoS ONE, 2013. **8**(3): p. e57755.
134. Schlipköter, U. and A. Flahault, *Communicable diseases: achievements and challenges for public health.* Public Health Reviews, 2010. **32**(1): p. 90-119.
135. Atkins, G.J., *The Pathogenesis of Alphaviruses.* ISRN Virology, 2013: p. 22.
136. Lord, R.D., *History and geographic distribution of Venezuelan equine encephalitis.* Bulletin of the Pan American Health Organisation, 1974. **8**: p. 11.
137. Brault, A., et al., *Genetic and antigenic diversity among eastern equine encephalitis viruses from North, Central, and South America.* The American journal of tropical medicine and hygiene, 1999. **61**(4): p. 579-586.
138. Oberste, M., et al., *Association of Venezuelan equine encephalitis virus subtype IE with two equine epizootics in Mexico.* The American journal of tropical medicine and hygiene, 1998. **59**(1): p. 100-107.
139. Rivas, F., et al., *Epidemic Venezuelan equine encephalitis in La Guajira, Colombia, 1995.* The Journal of infectious diseases, 1997. **175**(4): p. 828-832.
140. Sudia, W.D., et al., *Epidemic venezuelan equine encephalitis in north america in 1971: vector studies.* American Journal of Epidemiology, 1975. **101**(1): p. 17-35.
141. Walton, T.E. and M.A. Grayson, *Venezuelan equine encephalomyelitis, p. 203-231. In T. P. Monath (ed.), The Arboviruses: Epidemiology and Ecology, vol. IV. CRC Press, Boca Raton, Florida.* 1988.

142. Sidwell, R. and D. Smee, *Viruses of the Bunya- and Togaviridae families: potential as bioterrorism agents and means of control.* Antiviral research, 2003. **57**(1-2): p. 101-111.
143. Holbrook, M. and B. Gowen, *Animal models of highly pathogenic RNA viral infections: encephalitis viruses.* Antiviral research, 2008. **78**(1): p. 69-78.
144. Pfeiffer, J.K. and K. Kirkegaard, *Increased fidelity reduces Poliovirus fitness and virulence under selective pressure in mice.* Plos Pathogens, 2005. **1**(2): p. e11.
145. Vignuzzi, M., et al., *Quasispecies diversity determines pathogenesis through co-operative interactions in a viral population.* Nature, 2006. **439**: p. 344-348.
146. Berge, T.O., I.S. Banks, and W.D. Tigertt, *ATTENUATION OF VENEZUELAN EQUINE ENCEPHALOMYELITIS VIRUS BY IN VITRO CULTIVATION IN GUINEA-PIG HEART CELLS.* American Journal of Epidemiology, 1961. **73**(2): p. 209-218.
147. Pittman, P.R., et al., *Long-term duration of detectable neutralizing antibodies after administration of live-attenuated VEE vaccine and following booster vaccination with inactivated VEE vaccine.* Vaccine, 1996. **14**(4): p. 337-43.
148. Kenney, J.L., et al., *Stability of RNA Virus Attenuation Approaches.* Vaccine, 2011. **29**(12): p. 2230-2234.
149. Ferrer-Orta, C., et al., *A comparison of viral RNA-dependent RNA polymerases.* Curr Opin Struct Biol, 2006. **16**(1): p. 27-34.
150. Coffey, L.L., et al., *Arbovirus high fidelity variant loses fitness in mosquitoes and mice.* Proc Natl Acad Sci U S A, 2011. **108**(38): p. 16038-43.
151. Andrews, S., *FastQC A Quality Control tool for High Throughput Sequence Data*, in *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.
152. Hannon's_Lab, *FASTX Toolkit*, in *http://hannonlab.cshl.edu/fastx_toolkit/index.html*.
153. Hoffmann, S., et al., *Fast mapping of short sequences with mismatches, insertions and deletions using index structures.* PLoS computational biology, 2009. **5**(9).
154. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics (Oxford, England), 2009. **25**(16): p. 2078-2079.
155. Opitz, D. and R. Maclin, *Popular ensemble methods: An empirical study.* Journal of Artificial Intelligence Research, 1999. **11**: p. 31.
156. Kittler, J., *Combining classifiers: A theoretical framework.* Pattern Analysis & Applications, 1998. **1**(1): p. 18-27.
157. Yang, P., Y. Hwa, and B. B, *A review of ensemble methods in bioinformatics.* Current …, 2010.
158. Ferguson, T.S., *A Bayesian Analysis of Some Nonparametric Problems.* The Annals of Statistics, 1973. **1**(2): p. 21.
159. Blackwell, D. and J.B. MacQueen, *Ferguson Distributions Via Polya Urn Schemes.* 1973: p. 353-355.
160. Berry, D.A. and R. Christensen, *Empirical Bayes Estimation of a Binomial Parameter Via Mixtures of Dirichlet Processes.* The Annals of Statistics, 1979. **7**(3): p. 11.
161. Escobar, M.D. and M. West, *Bayesian Density Estiamtion and Inference Using Mixtures.* Journal of American Statistical Association, 1995. **90**(430): p. 12.

162. Liu, J.S., *Nonparametric hierarchical Bayes via sequential imputations.* 1996: p. 911-930.
163. MacEachern, S.N. and P. Müller, *Estimating Mixture of Dirichlet Process Models.* Journal of Computational and Graphical Statistics, 1998. **7**(2): p. 6.
164. Neal, R.M., *Markov Chain Sampling Methods for Dirichlet Process Mixture Models.* Journal of Computational and Graphical Statistics, 2000. **9**(2): p. 249-265.
165. Rasmussen, C.E., *The Infinite Gaussian Mixture Model.* Neural Information Processing Systems, 2000. **12**.
166. Ishwaran, H. and L.F. James, *Gibbs Sampling Methods for Stick-Breaking Priors.* Journal of the American Statistical Association, 2001. **96**: p. 13.
167. Teh, Y.W., et al., *Hierarchical Dirichlet processes.* Journal of the American Statistical Association, 2007. **101**(476): p. 16.
168. Galar, M., et al., *A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012. **42**.
169. Maciej, Z., et al., *Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients.* Applied Soft Computing, 2014. **14**.
170. Moreno-Torres, J.G., J.A. Saez, and F. Herrera, *Study on the impact of partition-induced dataset shift on k-fold cross-validation.* IEEE Trans Neural Netw Learn Syst, 2012. **23**(8): p. 1304-12.
171. Colley, D.G., et al., *Human schistosomiasis.* Lancet, 2014. **383**(9936): p. 2253-64.
172. Steinmann, P., et al., *Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk.* Lancet Infect Dis, 2006. **6**(7): p. 411-25.
173. Rollinson, D., et al., *Time to set the agenda for schistosomiasis elimination.* Acta Trop, 2013. **128**(2): p. 423-40.
174. King, C.H., *Parasites and poverty: the case of schistosomiasis.* Acta Trop, 2010. **113**(2): p. 95-104.
175. Wang, L., J. Utzinger, and X.N. Zhou, *Schistosomiasis control: experiences and lessons from China.* Lancet, 2008. **372**(9652): p. 1793-5.
176. Xu, J., et al., *Integrated control programmes for schistosomiasis and other helminth infections in P.R. China.* Acta Trop, 2015. **141**(Pt B): p. 332-41.
177. Wang, L.D., et al., *A strategy to control transmission of Schistosoma japonicum in China.* N Engl J Med, 2009. **360**(2): p. 121-8.
178. Wang, L.D., et al., *China's new strategy to block Schistosoma japonicum transmission: experiences and impact beyond schistosomiasis.* Trop Med Int Health, 2009. **14**(12): p. 1475-83.
179. Liu, R., H.F. Dong, and M.S. Jiang, *The new national integrated strategy emphasizing infection sources control for schistosomiasis control in China has made remarkable achievements.* Parasitol Res, 2013. **112**(4): p. 1483-91.
180. Zhou, Y.B., et al., *Spatial-temporal variations of Schistosoma japonicum distribution after an integrated national control strategy: a cohort in a marshland area of China.* BMC Public Health, 2013. **13**: p. 297.
181. Li, S.Z., et al., *Reduction patterns of acute schistosomiasis in the People's Republic of China.* PLoS Negl Trop Dis, 2014. **8**(5): p. e2849.

182. Seto, E.Y., et al., *Toward sustainable and comprehensive control of schistosomiasis in China: lessons from Sichuan.* PLoS Negl Trop Dis, 2011. **5**(10): p. e1372.
183. Sterne, J.A., et al., *Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials.* Bmj, 2011. **343**: p. d4002.
184. Schwarzer, G., *meta: General Package for Meta-Analysis*.
185. Viechtbauer, W., *Conducting meta-analyses in {R} with the {metafor} package.* Journal of Statistical Software, 2010. **36**: p. 1-48.
186. Cochran, W.G., *The Comparison of Percentages in Matched Samples.* Biometrika, 1950. **37**(3/4): p. 256-266.
187. Higgins, J.P. and S.G. Thompson, *Quantifying heterogeneity in a meta-analysis.* Stat Med, 2002. **21**(11): p. 1539-58.
188. Qing-Wu, J., et al., *Morbidity control of schistosomiasis in China.* Acta Trop, 2002. **82**(2): p. 115-25.
189. Ross, A.G., et al., *Road to the elimination of schistosomiasis from Asia: the journey is far from over.* Microbes Infect, 2013. **15**(13): p. 858-65.
190. Colley, D.G., *Morbidity control of schistosomiasis by mass drug administration: how can we do it best and what will it take to move on to elimination?* Trop Med Health, 2014. **42**(2 Suppl): p. 25-32.
191. marketsandmarkets.com, *Next Generation Sequencing (NGS) Market by Platforms (Illumina HiSeq, MiSeq, HiSeqX Ten, NextSeq 500,Thermo Fisher Ion Proton/PGM), Bioinformatics (Exome Sequencing, RNA-Seq, ChIP-Seq), Technology (SBS, SMRT) & by Application (Diagnostics, Personalized Medicine) – Global Forecast to 2020*. 2014, Markets and Markets: http://www.marketsandmarkets.com/Market-Reports/next-generation-sequencing-ngs-technologies-market-546.html.
192. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
193. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
194. Valouev, A., et al., *A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning.* Genome Res, 2008. **18**(7): p. 1051-63.
195. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-52.
196. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
197. Clarke, J., et al., *Continuous base identification for single-molecule nanopore DNA sequencing.* Nat Nanotechnol, 2009. **4**(4): p. 265-70.
198. Harris, T.D., et al., *Single-molecule DNA sequencing of a viral genome.* Science, 2008. **320**(5872): p. 106-9.
199. Shendure, J., et al., *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. **309**(5741): p. 1728-32.
200. Inc., G., *Choosing the Right NGS Sequencing Instrument for Your Study*. 2014, Genohub Inc.: https://genohub.com/ngs-instrument-guide/.

201. *HiSeq X Ten System*. 2015, Illumina, Inc: http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html.
202. *HiSeq X Five System*. 2015, Illumina, Inc: http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html.
203. Illumina, I., *HiSeq™ Sequencing Systems*. 2011, Illumina: http://www.illumina.com/Documents/systems/hiseq/datasheet_hiseq_systems.pdf. p. Pub. No. 770-2010-014
204. Illumina, I., *Sequencing power for every scale*. 2015, Illumnia http://www.illumina.com/systems/sequencing.html.
205. Vera, J.C., et al., *Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.* Mol Ecol, 2008. **17**(7): p. 1636-47.
206. Nyberg, K.G., et al., *Transcriptome characterization via 454 pyrosequencing of the annelid Pristina leidyi, an emerging model for studying the evolution of regeneration.* BMC Genomics, 2012. **13**: p. 287.
207. *454 Sequencing News by 1996-2015 Roche Diagnostics Corporation* 2002-2014, 454 Life Science: http://454.com/resources-support/news.asp. p.
.
208. *454 Products*. 1996-2015, 454 Life Science: http://454.com/products/index.asp.
209. *ABI Announces SOLiD 3, 454 Launches Titanium, Illumina Sets Specs for End '08*. 2008, genomeWeb: https://www.genomeweb.com/sequencing/abi-announces-solid-3-454-launches-titanium-illumina-sets-specs-end-%E2%80%9808. p.
.
210. *Sequencing & Genetic Analyzer Instruments*. 2015, Life Technologies.
211. *PGM & Proton Product Overview by 2015 Thermo Fisher Scientific Inc*. 2015.
212. *Pacific Biosciences Launches the PacBio(R) RS II Sequencing System*. 2013, Pacific Biosciences www.pacb.com/denovo.
213. , Oxford Nanopore Technologies. p. https://nanoporetech.com/news/press-releases/view/39.
214. *The Human Genome Project Completion: Frequently Asked Questions*. 2010, National Human Genome Research Institute: http://www.genome.gov/11006943.
215. Hayden, E.C., *Data from pocket-sized genome sequencer unveiled*. 2014, Nature Publishing Group: Nature News.

VITAE

NAME: Fanping Kong DATE: 6/1/2015

PRESENT POSITION AND ADDRESS**:**

Graduate Assistant
The University of Texas Medical Branch at Galveston
2.118 NMR Dockside Building
301 University Boulevard
Galveston, Texas 77555-1157

BIOGRAPHICAL:

Birthdate: October 25, 1986
Birthplace: Weihui, Henan Province, China
Citizenship: People's Republic of China

EDUCATION:

Aug 2010 – Aug 2015	Doctor of Philosophy in Biomedical Sciences Molecular Biophysics Education Track Department of Biochemistry and Molecular Biology The University of Texas Medical Branch (UTMB)
Aug 2009 – Jul 2010	Graduate Student in Physics Department of Physics University of Houston
Aug 2005 – Jul 2009	Bachelor of Science in Physics University of Science and Technology of China (USTC) Hefei, Anhui Province, China

EXPERIENCE:

RESEARCH ACTIVITIES:

Area of Research 1: Leishmania *donovani* caused visceral leishmaniasis in hamster and human
Mentor: Dr. Bruce A. Luxon Co-mentor: Drs. Peter C. Melby, Heidi M. Spratt

Visceral leishmaniasis in Syrian golden hamsters closely mimics the chronic progression in humans. However, the hamster fails to have a published genome reference. We used a novel approach involving deep sequencing (RNA-Seq) coupled with de novo assembly to identify genes differentially expressed between *L. donovani*-infected and uninfected control hamsters and explore the functions of key proteins and pathways.

Area of Research 2: Development of novel ensemble methods for classification
Mentor: Dr. Bruce A. Luxon Co-mentor: Dr. Heidi M. Spratt

Ensemble methods incorporate multiple individual machine learning models, by assuming that their expertized predictive spaces are different and can be complemented by each other. We employed the *Dirichlet Process Mixture* procedure increase the diversity of the model committee for ensemble methods and further enhance it predictive ability.

Area of Research 3: Enriched environment induced protective phenotype in cocaine addiction
Mentor: Dr. Bruce A. Luxon Co-mentor: Drs. Thomas A. Green, Heidi M. Spratt

Cocaine, a strong central nervous system stimulant, is a powerful addictive drug. Its abuse continues to plague our nation and leads to a persistent public health problem. Research shows that environmental enrichment can cause a protective phenotype against cocaine addiction. We sought to understand the related mechanisms by using multi-factor RNA-Seq experimental designs.

Area of Research 4: Intra–host variations of Venezuelan equine encephalitis virus (VEEV)
Mentor: Dr. Bruce A. Luxon Co-mentor: Drs. Naomi L. Forrester, Heidi M. Spratt

VEEV, a NIAID category B priority pathogen, periodically causes epidemics in equids and humans. To control and prevent such a viral infectious disease, vaccines are the most cost effective agents. We focus on the intra-host variation to determine variations essential for the adaptability of VEEV during transmission and infection. We expect to identify some mutants incapable of being transmitted by mosquitoes, which can be used as candidates to develop live virus vaccines for VEEV.

Area of Research 5: Meta-analysis to compare two prevention strategies of their effectiveness
Mentor: Dr. Bruce A. Luxon

When a new strategy in prevention, diagnosis or treatment is brought up, we need to evaluate its effectiveness by comparing with the established and congenital strategy using systematic literature review and meta-analysis. We evaluated the new integrated and the conventional prevention strategies for schistosomiasis in this study.

PROFESSIONAL TEACHING EXPERIENCE:

Jan 2010 – May 2010	Teaching Assistant, University Physics II Department of Physics University of Houston, TX
Aug 2009 – Dec 2009	Laboratory Instructor, General Physics Laboratory I Department of Physics University of Houston, TX

COMMITTEE RESPONSEIBILITES:

Departmental:

Molecular Biophysics Educational Track Curriculum Committee
Molecular Biophysics Educational Track Recruitment Committee

MEMBERSHIP IN PROFESSIONAL ORGANIZATIONS:

Moody Foundation Traumatic Brain Injury Research Center

HONORS:

Robert Bennett Scholarship, 2014-2015, U of Texas Medical Branch
Jeane B. Kempner Scholar, 2013 – 2014, U of Texas Medical Branch
Excellence Award, undergraduate research project, USTC, 2008
Outstanding Student Scholarship, USTC, 2008
Outstanding Student Scholarship, USTC, 2007

Outstanding Student Scholarship, USTC, 2006

PUBLICATIONS:

- PUBLICATIONS – PUBLISHED:
  C.F. Lichti, X. Fan, R. D. English, Y. Zhang, D. Li, **F. Kong**, M. Sinha, C. R. Andersen, H. M. Spratt, B. A. Luxon, T. A. Green (2014). Front Behav Neurosci 8: 246.

- PUBLICATIONS – IN REVIEW:
  B. Tian, X. Li, M. Kalita, S.G. Widen, J. Yang, S. Bhavnani, B. Dang, A. Kudlicki, M. Sinha, **F. Kong**, T.G. Wood, B. A. Luxon, A. R. Brasier. Analysis Of The TGFβ-Induced Program in Primary Airway Epithelial Cells Shows Essential Role Of NF-κB/RelA Signaling Network In Type II Epithelial Mesenchymal Transition.

  J. Luo, Y. Liang, **F. Kong**, J. Qiu, X. Liu, A. Chen, B.A. Luxon, H.W. Wu, Y. Wang. Schistosoma Antigens-Induced Vascular Endothelial Growth Factor Promotes the Activation of Hepatic Stellate Cells and Modulates the Expression of Fibrosis-Associated Molecules in Mice with Chronic Schistosomiasis, PLOS Neglected Tropical Diseases

- PUBLICATIONS – In Progress:
  F. Kong, O. A. Saldarriaga, H.M. Spratt, B.A. Luxon E. Y. Osorio, B. L. Travi and P.C. Melby. Transcriptional profiling reveals a proinflammatory spleen environment and mixed activation phenotype of disease-promoting splenic macrophages in progressive experimental VL.

  M. Mbuchi, OA. Saldarriaga, A. Muia, C. Magiri, **F. Kong**, H. Kanyi, S. Njenga, H. Spratt, B.A. Luxon, M. Wasunna, P.C. Melby. Exploration of Immunoregulatory Networks and Immunopathogenic Pathways in Human Visceral Leishmaniasis.

  Y. Zhang, **F. Kong**, E.J. Crofton, M. Sinha, D. Li, X. Fan, J.D. Hommel, H.M. Spratt, B. A. Luxon, T.A. Green. Transcriptomic study of environmental enrichm ent and cocaine-taking behavior in rat nucleus accumbens.

  F. Kong, H.M. Spratt and B.A. Luxon. Dirichlet Process Mixture integrated approach enhance predictive ability for bagging and Adaboost.

  M. Guerbois*, T. Kautz*, **F. Kong***, R. Yun, R. Langsjoen, M. D. Alcorn, H. M. Spratt, B. A. Luxon, S. C. Weaver and N. L. Forrester High-fidelity mutations in the vaccine TC-83 increase immunogenicity and attenuation. (* equally contributed)

  W. Wang*, **F. Kong*,** S. Wu, Y. Liang, Z. Jie, H. Wang, J. Dai, YS Liang. Effectiveness of the new integrated strategy to control the transmission of Schistosoma japonicum in China: a systematic review and meta-analysis. (* equally contributed)

  B. Fongang, **F. Kong**, S. Negi, W. Braun, A.S. Kudlicki, A Conserved Structural Signature of the Homeobox Coding DNA Suggests Its Role As a Cis-Regulatory Element in Metazoans.

- ORAL PRESENTATIONS:

Dirichlet Process Mixture Integrated Ensemble Method. Jun 2014, Galveston, TX, UTMB Sealy Center for Structural Biology and Molecular Biophysics Community Building Seminar

RNA-Seq and de novo transcriptome assembly to determine splenic gene expression in hamster visceral leishmania. Feb 2014, Galveston, TX, UTMB BMB Student Seminar
RNA-Seq and de novo transcriptome assembly. Sep 2013, Galveston, TX, UTMB BMB Student Seminar

Dimensionality reduced cortical features and their use in the classification of Alzheimer's disease and mild cognitive impairment. Oct 2012, Galveston, TX, UTMB Biochemistry Journal Club

Analysis of cancer metabolism with high throughput technologies. Feb 2012, Galveston, TX, UTMB Biochemistry Journal Club

- ABSTRACTS – POSTERS:

Zhang Y, Crofton E. J., **Kong F.**, Spratt H. , Sinha M., Andersen C. R., Li D., Fan X., Luxon B., Hommel J. and Green T. Retinoic acid signaling is a novel mechanism of environmental enrichment. Mar. 2015, Behavior, Biology, and Chemistry: Translational Research in Addiction in San Antonio

Saldarriaga O.A, **Kong F**, Mbuchi M, Muia A, Magiri C, Kanyi H, Chelugo A, Muthoni A, Rono R, Gachigi S, Njenga S, Spratt, H, Luxon B, Wasunna M, Melby P., Transcriptome Analysis of Human Visceral Leishmaniasis: Exploring Immunoregulatory Networks and Immunopathogenic Pathways. 3rd Annual Clinical & Translational Research Forum, Galveston TX, March2015

Mbuchi M., Saldarriaga O., Muia A.1, Magiri C., **Kong F.,** Kanyi H., Anderson, Agnes, Ronald , Susan , Njenga S., Spratt, H., Luxon B., Wasunna M., Melby P., Exploration of Immunoregulatory Networks and Immunopathogenic Pathways in Human Visceral Leishmaniasis, 5th Kemri Annual Scientific & Health (Kash) Conference, Feb 2015

O. Saldarriaga, F. Kong, H.M. Spratt, B.A. Luxon, E.Y. Osorio, B.L. Travi, P.C. Melby, Immunoregulatory Networks and Immunopathogenic Pathways in Visceral Leishmaniasis, Nov 2014, the American Society of Tropical Medicine and Hygiene 63rd Annual Meeting

O. Saldarriaga, **F. Kong**, M. Mbuchi, et. al.; Immunopathogenesis of visceral leishmaniasis: from a novel animal model to human disease, Feb 2014, 2nd Annual Clinical & Translational Research Forum

Y. Zhang, **F. Kong**, H.M. Spratt, et. al.; Transcriptomic study of environmental enrichment and cocaine-taking behavior in rat nucleus accumbens. Nov 2013, Society for Neuroscience
F. Kong, Y. Zhang, H.M. Spratt, et al.; Next Generation Sequencing to discover the transcriptomics of cocaine-taking behavior in rats. April 2013, 54th Annual National Student Research Forum

Y. Zhang, **F. Kong**, H.M. Spratt, et al.; The transcriptomics of individual differences in cocaine-taking behavior in rats. April 2013, 54th Annual National Student Research Forum

F. Kong, O. Saldarriaga, H.M. Spratt, B.A. Luxon and P.C. Melby; RNA-Seq and de novo transcriptome assembly to determine splenic gene expression in a novel model of visceral leishmaniasis. April 2013, IHII / McLaughlin Colloquium on Infection & Immunity

O. Saldarriaga, **F. Kong**, H.M. Spratt, B.A. Luxon and P.C. Melby; synergism between parasite and cytokine -induced expression of suppressor of cytokine signaling (socs) in macrophages in experimental visceral leishmaniasis. April 2013, IHII / McLaughlin Colloquium on Infection & Immunity

F. Kong, O. Saldarriaga, H.M. Spratt, B.A. Luxon and P.C. Melby; RNA-Seq and de novo transcriptome assembly to determine splenic gene expression in a novel model of visceral leishmaniasis. Feb 2013, CTSA 1st Poster Session

F. Kong, O. Saldarriaga, H.M. Spratt, B.A. Luxon and P.C. Melby; RNA-Seq and de novo transcriptome assembly to determine splenic gene expression in a novel model of visceral leishmaniasis. Dec 2012, 2nd Annual Institute for Human Infections & Immunity (IHII) Retreat

F. Kong, C. Shumate, and S. Stoilova-McPhie; Cryo-Electron Tomography of Phospholipid Vesicles and Nanotubes for Structure Determination of Membrane-Associate Proteins. Sep 2010, Molecules and Mechanisms 7th Annual Fall Research Retreat

C. Shumate, **F. Kong**, and S. Stoilova-McPhie; Binding of Coagulation Factor VIII to Phospholipid Vesicles: a Combined Cryo-Electron Microscopy and Biophysical Study. Sep 2010, Molecules and Mechanisms 7th Annual Fall Research Retreat