APPLICATIONS OF HIGH-THROUGHPUT SEQUENCING DATA ANALYSIS IN TRANSCRIPTIONAL STUDIES

A Dissertation

by

ZHENGYU GUO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Aniruddha Datta
Committee Members,	Ulisses Braga-Neto
	Xiaoning Qian
	Alan Dabney
Head of Department,	Miroslav M. Begovic

December 2017

Major Subject: Electrical Engineering

ABSTRACT

High-throughput sequencing has become one of the most powerful tools for studies in genomics, transcriptomics, epigenomics, and metagenomics. In recent years, HTS protocols for enhancing the understanding of the diverse cellular roles of RNA have been designed, such as RNA-Seq, CLIP-Seq, and RIP-Seq. In this work, we explore the applications of HTS data analysis in transcriptional studies. First, the differential expression analysis of RNA-Seq data is discussed and applied to a sheep RNA-Seq dataset to examine the biological mechanisms of the sheep resistance to worm infection. We develop an automatic pipeline to analyze the RNA-Seq dataset, and use a negative binomial model for gene expression analysis. Functional analysis is conducted over the differentially expressed genes, and a broad range of mechanisms providing protection against the parasite are identified in the resistant sheep breed. This study provides insights into the underlying biology of sheep host resistance. Then, a deep learning method is proposed to predict the RNA binding protein binding preferences using CLIP-Seq data. The proposed method uses a deep convolutional autoencoder to effectively learn the robust sequence features, and a softmax classifier to predict the RBP binding sites. To demonstrate the efficacy of the proposed method, we evaluate its performance over a dataset containing 31 CLIP-Seq experiments. This benchmarking shows that the proposed method improves the prediction performance in terms of AUC, compared with the existing methods. The analysis also shows that the proposed method is able to provide insights to identify new RBP binding motifs. Therefore, the proposed method will be of great help in understanding the dynamic regulations of RBPs in various biological processes and diseases. Finally, a database is created to facilitate the reuse of the public available mouse RNA-Seq dataset. The metadata of the publicly available mouse RNA-Seq datasets is manually curated and is served by a

well-designed website. The database can be scaled up in the future to serve more types of HTS data.

DEDICATION

To my family.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Aniruddha Datta for his kind help on this work and my stay at Texas A&M University. I thank Dr. Ulisses Braga-Neto, Dr. Xiaoning Qian, and Dr. Alan Dabney, for serving on my committee and for their support. I thank my fellow students and friends, Yue Gan, Yukun Tan, and Xingde Jiang for their friendships, memories, and supports. Most importantly, I would like to thank my family for their encouragement, love and help.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Aniruddha Datta, Dr. Ulisses Braga-Neto and Dr. Xiaoning Qian of the Department of Electrical and Computer Engineering and Dr. Alan Dabney of the Department of Statistics.

The content of Chapter 2 was completed by the student in collaboration with Dr. Robert Li. Jorge Francisco González and Dr. Robert Li conceived and designed the experiment. Biological experiments were performed by Julia Hernandez, Tom McNeilly, Yolanda Corripio-Miyar, David Frew, Jorge Francisco González and Dr. Robert Li. The manuscript partly forms Chapter 2 was written by Dr. Li. In Chapter 4, the curation was conducted in part by Boriana Tzvetkova, Jennifer M. Bassik, Tara Bodziak and Brianna Wojnar from Dr. Bo Hua Hu's team. The website was built in part by Wei Qiao and Md Obaida. Dr. Peng Yu supervised the work of Chapter 2 and Chapter 4.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

The student was supported by TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE) at Texas A&M University.

This work was made possible in part by startup funding from the ECE department and Texas A&M Engineering Experiment Station/Dwight Look College of Engineering at Texas A&M University, NIDCD R01DC010154, AGL2009/09985 and Fondo Social Europeo (FSE).

NOMENCLATURE

AUC	Area Under the receiver operating characteristic Curve
CHB	Canaria Hair Breed
CLIP-Seq	High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
CAE	Convolutional Autoencoder
CNN	Convolutional Neural Network
CS	Canaria Sheep
DEG	Differentially Expressed Gene
dpi	Days post infection
FDR	False Discovery Rate
GO	Gene Ontology
HTS	High Throughput Sequencing
NB	Negative Binomial
PWM	Positional Weight Matrix
RBP	RNA Binding Protein
ReLU	Rectified Linear Unit
RNA-Seq	RNA Sequencing

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGMENTS v

CONTRIBUTORS AND FUNDING SOURCES vi

NOMENCLATURE vii

TABLE OF CONTENTS viii

LIST OF FIGURES x

LIST OF TABLES xii

1. INTRODUCTION 1

    1.1 DNA, RNA and Protein 1
    1.2 The Central Dogma of Molecular Biology 4
        1.2.1 DNA Replication 4
        1.2.2 Transcription 5
        1.2.3 Translation 6
    1.3 High-Throughput Sequencing 6
    1.4 High-Throughput Sequencing for Transcriptomics 9
    1.5 Dissertation Outline 11

2. COMPUTATIONAL DIFFERENTIAL EXPRESSION ANALYSIS FOR UNDERSTANDING THE MOLECULAR BASIS OF SHEEP HOST RESISTANCE 13

    2.1 Introduction 13
        2.1.1 Sheep Host Resistance 13
        2.1.2 Comparative Analysis of Indigenous Sheep from Canary Islands ... 14
    2.2 Materials and Methods 15
        2.2.1 RNA Extraction and Sequencing 15

2.2.2 Computational Analysis of Sheep Host Resistance RNA-Seq Data 17
2.2.3 Data Preprocessing and Alignment 17
2.2.4 Differential Expression Analysis 18
2.2.5 Gene Ontology Analysis 21
2.3 Results 23
2.3.1 *Haemonchus* Infection Induced Distinctly Different Transcriptome Patterns in the Abomasal Mucosa of CHB and CS Breeds 23
2.3.2 Gene Ontology Implicated in Host Resistance 32
2.4 Discussion 36

3. A CONVOLUTIONAL AUTOENCODER BASED METHOD FOR PREDICTING RNA PROTEIN INTERACTIONS 43

3.1 Introduction 43
3.2 Materials and Methods 46
3.2.1 Dataset 46
3.2.2 RNA Sequence Encoding 46
3.2.3 Convolutional Layers 47
3.2.4 Model Architecture 49
3.2.5 Model Implementation 52
3.3 Results 52
3.3.1 Model Evaluation 52
3.3.2 Insight in Motif Discovery 54
3.4 Conclusion 58

4. CREATION OF A DATABASE AND WEB-SERVER FOR METADATA OF PUBLICLY AVAILABLE MOUSE RNA-SEQ DATA SETS 60

4.1 Introduction 60
4.2 Methods 62
4.3 Results 64
4.4 Discussion 66

5. CONCLUSIONS 67

REFERENCES 69

LIST OF FIGURES

FIGURE		Page
1.1	Double-Stranded DNA and the Base Pairing Rule	2
1.2	The Structure of a Gene	3
1.3	General Transfers of the Central Dogma of Molecular Biology	5
1.4	Illustration of Illumina Sequencing by Reversible Termination	10
2.1	Analysis Workflow of the Sheep Host Resistance RNA-Seq Dataset	18
2.2	Differences in Worm Counts Between Resistant and Susceptible Sheep Breeds Under Experimental *Haemonchus contortus* Infection	22
2.3	Venn Diagram of Differentially Expressed Genes and Scatter Plot of Log2 Ratio (Fold Change) vs Mean	23
2.4	Differences in Worm Counts Between Resistant and Susceptible Sheep Breeds Under Experimental *Haemonchus contortus* Infection	24
2.5	Real-Time RT-PCR Analysis (qPCR) of Selected Genes	33
2.6	Linear Regression Analysis of Fold Changes Calculated from qPCR and RNA-Seq Analysis	34
2.7	Gene Ontology Lineage Relations.	37
3.1	An Example of Convolutional Network	47
3.2	Architecture of the Proposed Model	51
3.3	The AUC Distribution of All Experiments for the Three Methods	53
3.4	AUC Comparison Between the Proposed Method and the Baseline Methods	56
3.5	The Proposed Method Recovers the Known Motifs in cisBP-RNA.	57
3.6	The Proposed Method Recovers the Known Motifs in the Literature	58

4.1 The Database Schema of RNASeqMetaDB 62
4.2 Basic Searching Functionality of RNASeqMetaDB 63
4.3 Some Statistics of RNASeqMetaDB 65

LIST OF TABLES

TABLE		Page
1.1	Information Transfers of the Central Dogma of Molecular Biology	4
1.2	Genetic Codes for Translation from Nucleotide Sequences to Amino Acids	7
1.3	Some High-Throughput Sequencing Platforms and Their Statistics	8
2.1	Genes Significantly Impacted by *Haemonchus contortus* Infection in Both CHB and CS Breeds	25
2.2	41 Extracellular Matrix (ECM) Related Genes Significantly Affected by *Haemonchus contortus* Infection in the Abomasal Mucosa of the Canaria Hair Breed Sheep	27
2.3	100 Cell-Cycle Related Genes Significantly Affected by *Haemonchus contortus* Infection in the Resistant Breed	29
2.4	Gene Ontology (GO) Biological Processes (BP) Significantly Enriched in Both Resistant (CHB) and Susceptible (CS) Breeds	32
2.5	Selected Gene Ontology Terms Significantly Impacted by *Haemonchus contortus* Infection in the Resistant Breed	34
3.1	Mean AUC of the Three Methods for All Experiments	55

1. INTRODUCTION

1.1 DNA, RNA and Protein

Deoxyribonucleic acid (DNA) is the genomic material in cells that encodes the genetic information used in development and functioning of all living organisms [1]. In cells, DNA has two chains of monomers called nucleotides. A nucleotide is composed of three subunits: a deoxyribose sugar, a phosphate group, and a nitrogenous base (one from cytosine/C, guanine/G, adenine/A, or thymine/T). The nucleotide chain is formed by connecting the deoxyribose sugar of one nucleotide with the phosphate group of the next with phosphodiester bonds. This sugar-phosphate region is referred as the backbone of the DNA molecule, which is on the outside of the double-stranded DNA. The nitrogenous bases are on the inside of the double-stranded DNA, and the order of the bases encodes the genetic information carried by DNA molecules. The two strands of a double-stranded DNA molecule are held together by hydrogen bonds between complementary nitrogenous bases of different strands to have a double helix structure. The base pairing strictly follows the complementary rule: A pairs with T through a double intermolecular hydrogen bond, and C pairs with G through a triple intermolecular hydrogen bond. This rule ensures the strict alignment of the two strands of the DNA molecule. One strand contains the full information encoded in the double-stranded DNA, and each strand can act as a template for the other one during replication. Figure 1.1 shows the structure of a double-stranded DNA and the base pairing rule. In eukaryotic cells, long double-stranded DNA molecules are packaged into long structures called chromosomes. A chromosome is made up of a single DNA molecule that is tightly coiled around specialized proteins called histones. The chromosomes are stored in the nuclei of the cells, and they are duplicated and evenly distributed into the two daughter cells during cell division.

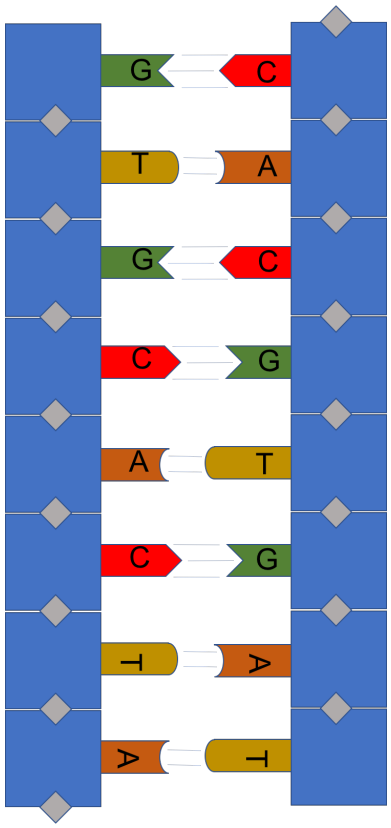


Figure 1.1: Double-stranded DNA and the base pairing rule. A, T, C, and G are adenine, thymine, cytosine, and guanine, respectively. Blue squares represent deoxyribose sugars. Gray diamond shapes represent phosphate groups. The backbone of double-stranded DNA is formed by alternative deoxyribose sugars and phosphate groups connected by phosphodiester bonds. The nitrogenous bases follow the base pairing rule: A vs. T, C vs. G.

The hereditary unit of the encoded genetic information in a DNA molecule is a gene [1]. A gene occupies a particular locus on a chromosome and affects an organism's traits by either encoding instructions to produce functional products (e.g., proteins) or regulating the production of them. Figure 1.2 shows the structure of a eukaryotic protein-coding gene. It contains regulatory sequences, exons, and introns. Regulatory sequences, such as promoters and enhancers, are required for a gene to express. They specify when and how a gene is transcribed to RNA for protein production. Exons are coding regions which encode the amino acid sequences of proteins. Introns are not translated into proteins.

In any organism, a vast amount of genetic information is encoded in very long DNA sequences that contain thousands of genes. For example, the total length of the human genome is approximately three billion base pairs. It is composed of 23 pairs of chromosomes which contain around 20,000 genes in total.

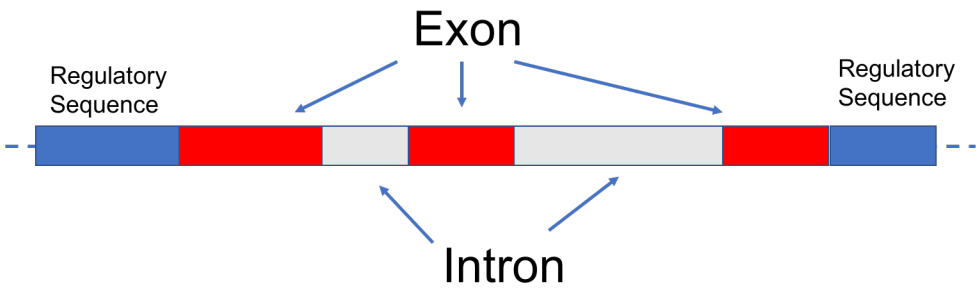


Figure 1.2: The structure of a gene. A gene consists of exons, introns and regulatory sequences. Red rectangles represent exons, and grey rectangles represent introns. Blue rectangles represent regulatory sequences.

Ribonucleic acid (RNA) is another kind of biological molecule that can carry genetic information [1]. Similar to DNA, an RNA molecule is assembled as a chain of four types of nucleotides linked by phosphodiester bonds. However, RNA differs from DNA in a few ways. First, it uses ribose rather than deoxyribose in DNA. Second, it contains Uracil (U) in place of Thymine (T) in DNA. Besides, it is usually single-stranded. The single-stranded RNA can loop back on itself to produce more complex secondary and tertiary structures. In living cells, RNA can be classified into many types according to their biological functions. For example, messenger RNA is the template for protein synthesis. Ribosomal RNA is a component of the ribosome. Transfer RNA carries the amino acid to the growing polypeptide during protein production.

Proteins are essential building blocks of organisms [1]. They are indispensable in

virtually all processes within cells. A protein consists of one or more long chains of amino acids called polypeptides. These polypeptides, are composed of amino acids connected by peptide bonds. In nature, there are in total 20 different amino acids for producing proteins in organisms.

1.2 The Central Dogma of Molecular Biology

Table 1.1: Information transfers of the central dogma of molecular biology [2]. There are in total nine transfers, which are classified into three categories.

General transfer	Special transfer	Unknown transfer
DNA $\rightarrow$ DNA	RNA $\rightarrow$ DNA	Protein $\rightarrow$ DNA
DNA $\rightarrow$ RNA	RNA $\rightarrow$ RNA	Protein $\rightarrow$ RNA
RNA $\rightarrow$ Protein	RNA $\rightarrow$ Protein	Protein $\rightarrow$ Protein

The central dogma of molecular biology [2], stated by Francis Crick in 1958, describes the flow of genetic information among DNA, RNA, and protein in biological systems. There are in total nine possible information transfers, which are classified into three groups. Table 1.1 shows the nine transfers and their classifications.

General transfers are normal flows occur in cells. Special transfers are uncommon flows occur only in some viruses or a laboratory. Unknown transfers are believed never to occur. Here only the general transfers are discussed (Figure 1.3).

The three information flows in general transfers describe three primary biological processes in living organisms: DNA replication, transcription, and translation.

1.2.1 DNA Replication

Genetic information transfers from DNA to DNA. In the process of DNA replication, a DNA molecule is copied, and two DNA molecules identical to the original DNA are produced. In living organisms, DNA replication occurs when cells divide. During cell

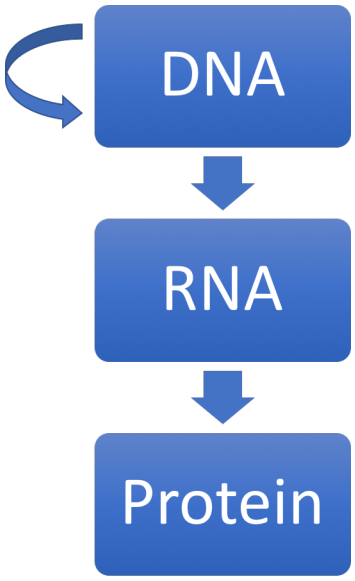


Figure 1.3: General transfers of the central dogma of molecular biology. Information can flow from DNA to DNA (replication), DNA to RNA (transcription), and RNA to protein (translation).

division, the complementary strands of DNA are separated and serve as the templates for producing the new complementary strands. The new strands are synthesized according to the base-pairing rule with nucleotides, and thus the two final products are identical to the original DNA. The two copies of DNA will then be distributed into two daughter cells. This process ensures the daughter cells contain identical copies of genetic information of their parent cell. Therefore, it is required in cell division and serves as the basis of inheritance. It is arguably considered as the fundamental step in the central dogma.

1.2.2 Transcription

Genetic information transfers from DNA to RNA. It is an essential step in gene expression. In the process of transcription, the sequence of a segment of DNA (for example a gene) is copied into RNA (mostly mRNA). Since DNA and RNA are both nucleic acids, the DNA serves as the template and the RNA is synthesized according to the base pairing

rule. The transcription is initiated by the binding of protein machinery consisting of RNA polymerase and transcription factors to the promoter sequences of DNA. RNA polymerase reads the DNA sequence and adds RNA nucleotides to produce a complementary RNA sequence. The produced RNA is call pre-mRNA, as it contains both exons and introns and is not ready for translation. The pre-mRNA will then be processed by 5' capping (adding a 5' cap), polyadenylation (adding a poly-A tail), and splicing (removing the introns) to produce the mature mRNA.

1.2.3 Translation

Genetic information transfers from RNA to protein. Translation is a process in which linear polypeptides of proteins are produced based on the RNA sequences. In eukaryotic cells, the mature mRNA exported from the nucleus is translated in the ribosome. Each time the ribosome will read three non-overlapping bases (a codon) and map them to a particular amino acid, by base pairing the anticodon sequences of the tRNA carrying amino acids. The amino acid is then linked to the growing peptide chain. There are in total 64 codons that can be formed by three bases, while there are only 20 amino acids. Therefore, most amino acids can be encoded by more than one codon. Besides, several codons have special meanings: UAA, UAG, and UGA are called stop codons as they are used to end the polypeptide production; AUG is the start codon, and it also encodes the amino acid methionine (Met). Table 1.2 shows all the codons and the amino acids they encode.

1.3 High-Throughput Sequencing

DNA sequencing [3] is a class of technologies for determining the precise order of nucleotides that make up a DNA molecule. It was first developed by Sanger and Coulson [4] and Maxam and Gilbert [5] in 1977. With the developments in the following years, Sanger's method gradually evolved into an automated DNA sequencing procedure referred as "First Generation Sequencing." Sanger Sequencing was used to complete the human

Table 1.2: Genetic codes for translation from nucleotide sequences to amino acids [1]. All the three-nucleotide codons and the corresponding amino acids are listed. Most amino acids are represented by more than one codon. There is also a start codon (annotated by a star symbol) and three stop codons.

First Base	Second Base								Third Base
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met*	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

genome sequence draft in 2001 [6] and the genome sequences of several other model organisms. However, Sanger sequencing has several limitations such as high cost and low throughput, which greatly restrict the application of DNA sequencing in research and clinics.

To overcome the limitations of Sanger Sequencing, cheaper and faster techniques were developed. These methods are referred as High-Throughput Sequencing (HTS) [7], as they can sequence a massive number of DNA sequences of a sample in parallel and produce millions of sequences concurrently with low costs and in short time. The emerge of HTS has enabled the broad application of sequencing. It has become an essential tool for research in genomics, transcriptomics, epigenomics, and metagenomics. Currently, many

HTS platforms are commercially available, such as Roche/454, Illumina, ABI SOLiD, Pacific BioSciences and IonTorrent. These platforms have distinct read length, run time, throughput and cost. Table 1.3 shows statistics of some platforms [8].

Table 1.3: Some HTS platforms and their statictics [8].

Platform Instrument	Year	Reads per run	Read length	Bases per run (GB)
454 GS Junior+	2014	100000	700	0.07
IonTorrent Proton PI	2012	50000000	200	10
Illumina HiSeq 2000	2010	2000000000	100	200
Illumina HiSeq 2500 RR	2014	600000000	250	300
Illumina MiSeq	2013	30000000	300	15
SOLiD 5500xl W	2013	3000000000	75	320
PacBio RS II P6 C4	2014	660000	13500	9.000

In general, most HTS sequencing approaches share some commonalities in their procedures [9]. First, the DNA fragments are amplified on a solid surface to form thousands of DNA colonies, each of which consists of identical copies of a single DNA fragment. The DNA colonies are critical for the sequencing because a large amount of identical DNA fragments in a restricted area ensure that the signal intensity is enough to be detected in the following steps. Then the sequences of DNA colonies are read out by massive parallel sequencing with two different approaches depending on the platforms: sequencing by ligation (SBL) or sequencing by synthesis (SBS). In SBL, probe sequences are labeled with fluorescent dyes according to the bases to be sequenced. The labeled probes are ligated to the anchor sequence when the bases of probe sequences match the unknown DNA. Then the sequence at the position can be inferred from the fluorescence produced by the molecule. In SBS, DNA polymerase is used for adding the bases complementary to the unknown DNA. The sequences are determined by either the fluorescences produced by the fluorophores attached to the added nucleotides or the ionic concentration changes gener-

ated by adding the nucleotides. In both approaches, the sequencing machine can process millions of these reactions simultaneously, and thus determine the sequence of millions of DNA fragments in parallel.

Illumina HTS platforms [9] are the most commonly used ones among all HTS platforms. Since the release of Illumina Genome Analyzer II in 2006, Illumina has developed a series of platforms for different sequencing requirements and has tremendously increased the sequencing throughput and reduced the costs. Currently, Illumina machines have dominated the HTS market. Illumina's sequencing starts with solid-phase bridge amplification, which creates colonies of identical DNA fragments bound to the flow cell. Then the DNA fragments are sequenced by reversible termination using reversible terminator (RT) nucleotides (Figure 1.4). RT nucleotides are labeled with fluorescent dyes. DNA polymerase incorporates one RT nucleotide into the synthesized strand per cycle. Then the flow cell is imaged to read out the added nucleotides. Finally, the fluorescent dyes are removed, and the RT nucleotides are de-protected to enable the next cycle of sequencing.

1.4 High-Throughput Sequencing for Transcriptomics

With the falling of sequencing costs, HTS has been widely applied in various areas for more comprehensive understanding of the landscapes of genomics, epigenomics, transcriptomics, and metagenomics. To enhance the understanding of the diverse cellular roles of RNA, HTS technologies have been developed to characterize the expression profile, RNA structure, RNA-protein interaction, and RNA localization [7, 9]. The HTS technology that can directly sequence various RNA species, such as mRNA, microRNA, snoRNA, is called RNA Sequencing (RNA-Seq). People have designed sophisticated tools for RNA-Seq data alignment, transcriptome assembly, differential gene expression, and alternative splicing analysis. These efforts together make RNA-Seq a powerful method for profiling the RNA species of interest across the entire transcriptome. RNA-Seq for

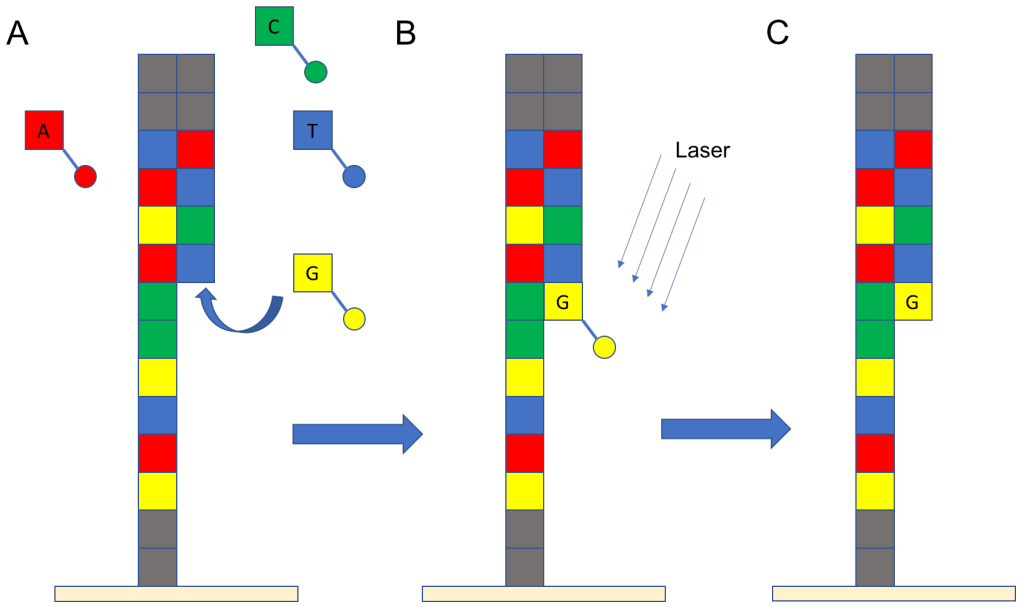


Figure 1.4: Illustration of Illumina sequencing by reversible termination. The squares represent the nucleotide bases. The circles represent the fluorescent dyes. A: RT-nucleotides labeled by fluorescent dyes are added to the flow cell. One RT nucleotide is added to the synthesizing strand of the DNA molecule. B: The added nucleotide is read by imaging. C: The fluorescent dye is removed and the RT nucleotide is de-protected for the next cycle of sequencing.

microRNA, snoRNA, and other RNA species provides a systematic method for studying these RNA species, including expression measurement, and new variant detection. There are some HTS variants that are applied to RNA sequences interacting with other biological molecules, (e.g., proteins), for studying the interaction between RNA and those molecules. For example, RNA Immunoprecipitation Sequencing (RIP-Seq) [10] uses immunoprecipitation of an RNA-binding protein (RBP) to detect the RNA fragments coupled with the RBP, and then uses HTS to read out all those sequences. Similarly, high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (CLIP-Seq) [11] utilizes UV cross-linking with immunoprecipitation to extract protein binding regions or RNA modification regions of RNA sequences to sequence the RNA-protein interaction regions.

These approaches enable the genome-wide studies of protein-RNA binding (binding sites of RNA-binding proteins) and RNA modification (e.g., N6-Methyladenosine(m6A) in mRNA). The use of HTS in transcriptomics has provided researchers powerful tools to understand the gene expression, transcriptional regulation, and post-transcriptional regulation in a comprehensive manner.

1.5 Dissertation Outline

The dissertation is organized as follows:

In chapter 2, the differential expression analysis of RNA-Seq data is discussed and applied to a sheep RNA-Seq dataset to study the biological mechanisms of the sheep resistance to worm infection. The differentially expressed genes are identified by the RNA-Seq analysis workflow based on a negative binomial model. Gene Ontology analysis is performed over the identified genes to reveal the biological pathways contributing to sheep host resistance.

In chapter 3, a deep learning method to predict RBP binding sites is proposed. Deep convolutional autoencoder model is used for robust extraction of the sequence features, and the compact representation of the input data generated by the autoencoder is then used for binding site prediction with a feed-forward neural network. The performance of the method is evaluated on a benchmark dataset and compared with other state-of-the-art methods. The proposed model yields better performance. The ability to discover binding motifs using the convolutional kernels learned from the data is also investigated.

In chapter 4, a database for publicly available mouse RNA-Seq datasets is created. The purpose of building the database is to facilitate more convenient data reuse, as the amount of data generated by HTS technologies is growing incredibly fast. The metadata of the publicly available mouse RNA-Seq datasets are manually curated and served by a well-designed website. We believe that the website will be beneficial for the reuse of a

significant number of RNA-Seq datasets in various research activities.

In chapter 5, conclusions and future work are discussed.

2. COMPUTATIONAL DIFFERENTIAL EXPRESSION ANALYSIS FOR UNDERSTANDING THE MOLECULAR BASIS OF SHEEP HOST RESISTANCE*

2.1 Introduction

2.1.1 Sheep Host Resistance

The intestinal worm *Haemonchus contortus* is arguably the most economically important helminth parasite for small ruminant production in many regions of the world. As a voracious blood feeder residing in the mucosal layer of the abomasum, *H. contortus* causes anemia and hyper-gastrinemia and alters abomasal secretion. *H. contortus* infection results in reduced growth, compromised reproduction, and elevated mortality, due to its ubiquitous distribution and severe pathogenicity. Consequently, *H. contortus* parasitism represents the primary constraint to profitable production of sheep and goats worldwide.

Over the past years, the rapid emergence of drug-resistant *H. contortus* strains and increasing demands by consumers for inexpensive organic meat and milk products with fewer drug residues have spurred research on the development of anthelmintic-independent parasite control strategies, such as vaccines [12] and novel biologics, nutrient supplements and bioactive compounds, and selective breeding. Among them, selectively breeding sheep and goats with abilities to better resist parasitic infections appears to be a solution to sustainable small ruminant production.

*Part of this chapter is reprinted from "Possible mechanisms of host resistance to *Haemonchus contortus* infection in sheep breeds native to the Canary Islands," by Z. Guo, J. F. González, J. N. Hernandez, T. N. McNeilly, Y. Corripio-Miyar, D. Frew, T. Morrison, P. Yu, and R. W. Li, *Scientific reports*, vol. 6, 2016, Copyright[2016] by Nature Publishing Group, licensed under a Creative Commons Attribution 4.0 International License.

2.1.2 Comparative Analysis of Indigenous Sheep from Canary Islands

Differences in resistance and susceptibility to parasitic infections between sheep breeds have been long documented [13]. Over the decades, comparative studies have identified at least 19 sheep breeds displaying varying degrees of resistance to parasitic infections [14]. For example, St. Croix lambs shed significantly fewer eggs and harbor 99% fewer worms in the abomasum than the age-matched Dorset lambs during both natural and experimental infections [15]. Locally-adapted breeds such as Santa Ines sheep of Brazil have significantly reduced worm burdens and fewer nodular lesions under natural infections than Suffolk and Ile de France lambs on the same pasture [16]. In Europe, resistance against *H. contortus* is better developed in Merinoland sheep than in Rhon sheep [17]. Red Maasai sheep have been shown to be more resistant to *Haemonchus* infection than the South African Dorper breed during natural exposure to parasites in Kenya [18]. Moreover, resistance to parasite infection has a significant genetic component. The contribution of the host genome and genetics has been estimated. For example, additive genetic variation accounts for approximately 30% of the overall variation for parasitic infection [19]. The resistance traits are often polygenic in nature and not influenced by genes with major effects [20]. Nevertheless, estimates of heritability for parasite indicator traits in small ruminants are phenotype-dependent, ranging from 0.11 to 0.40 for transformed fecal egg counts (EPG) and 0.19 to 0.26 for packed cell volume (PCV) in German Rhon sheep [21]. In addition, the host age plays a role. A good example is that in Scottish Blackface lambs at the end of the first grazing season, the heritability of adult worm length is very strong at 0.628. While many efforts have been made to identify genetic variants associated with parasite resistance and tolerance in sheep breeds [22, 23, 24], molecular mechanisms and biological pathways underlying host resistance to parasitic infections in sheep remain largely unknown.

Due to unique geographical characteristics of the Canary Islands, indigenous sheep breeds have been exploited by local farmers for centuries. Among them, the Canaria Hair Breed (CHB) and Canaria sheep (CS) are predominately raised for the production of meat and milk, respectively. Previous studies demonstrate that CHB constantly displays better resistance phenotypes to *H. contortus* infection than CS, including significantly lower levels of fecal egg counts, fewer adult worm counts, lower number of eggs in utero and female worm stunting [25]. Further studies [26] identified significant negative correlations between two effector cells, eosinophils and $\gamma\delta$/WC1+ T cells, and parasite fecundity in CHB, suggesting that inter-breed difference in regulating immune responses affects *Haemonchus* infection. In this study, we conducted a RNA-seq based comparative transcriptome analysis in the two indigenous breeds and attempted to understand the molecular basis underlying host resistance.

2.2 Materials and Methods

2.2.1 RNA Extraction and Sequencing

Male lambs of CHB (11 animals) and CS breeds (12 animals) were obtained from local farms in the Gran Canaria Island (Spain), weaned, and kept in pens at the Faculty of Veterinary Science, University of Las Palmas de Gran Canaria until they were approximately one year old. The animals were fed with a commercial pelleted sheep ration ad libitum and had free access to water throughout the experimental period. The animals were drenched upon arrival with levamisole (Cyber, Fort Dodge, Spain) at the recommended dose (1 ml/10 kg bodyweight) and remained free of parasites (as determined by fecal egg counts) until experimental parasite inoculation. Seven CHB and eight CS animals were inoculated intraruminally with 20,000 *H. contortus* infective L3 larvae. Four age-matched animals of each breed remained uninfected and served as controls. The experimental infection was allowed to progress for 20 dpi. The time point chosen for this study was based on the results

from a previous report that the difference in resistance phenotypes, especially mean EPG values, is most profound between the two breeds [25]. At 20 dpi, both infected and control animals were sacrificed. The fundic abomasum tissue was then sampled and snap frozen in liquid nitrogen prior to storage at $-80\,^{\circ}\mathrm{C}$ until total RNA was extracted. The *Haemonchus* strain used in this trial was initially donated by Drs. Knox and Bartley (Moredun Research Institute, Edinburgh, Scotland) and passaged through successive inoculations in sheep at the premises of the Faculty of Veterinary Science, University of Las Palmas de Gran Canaria (Spain). During the experiment, all animal protocols were approved by the Animal Care and Use Committee of University of Las Palmas per the Institutional Animal Care and Use Committee (IACUC) guidelines. All experimental procedures were carried out in accordance with the approved protocols.

Total RNA from fundic abomasal samples of both CHS and CS sheep breeds was extracted using Trizol (Invitrogen, Carlsbad, CA, USA) followed by DNase digestion and Qiagen RNeasy column purification (Qiagen, Valencia, CA, USA), as previously described [27, 28]. The RNA integrity was verified using an Agilent BioAnalyzer 2100 (Agilent, Palo Alto, CA, USA). High-quality RNA (RNA integrity number or RIN > 7.5) was processed using an Illumina TruSeq RNA sample prep kit following the manufacturer's instructions (Illumina, San Diego, CA, USA). Pooled RNAseq libraries were sequenced at 2×101 bp / sequence read using an Illumina HiSeq 2000 sequencer, as described previously [29]. Approximately 56 million paired-end sequence reads per sample (mean $\pm$ SD $= 55,945,621 \pm 41,305,493.24$; $N = 23$) were generated. The metadata and raw sequences files related to this project were deposited in the NCBI Sequence Read Archive (Accession #SRP059627).

2.2.2 Computational Analysis of Sheep Host Resistance RNA-Seq Data

The computational analysis pipeline is depicted in Figure 2.1. First, RNA-Seq data will be preprocessed by the steps including quality control, read mapping, and feature counting. We will trim off the low-quality nucleotides from all raw RNA-Seq reads and then align the trimmed reads against the ovine reference genome Oar_v3.1. The uniquely mapped read will be used to count against the Ensembl annotation Oar_v3.1 for calculating the number of reads per gene. Then for the differential analysis, the RNA-Seq read counts will be normalized to eliminate the variations resulting from different library sizes. The normalized counts of each gene will then be used to estimate the expression level of the gene and test the differential expression between conditions. With the list of differentially expressed genes (DEGs), Gene Ontology (GO) analysis will be conducted to identify potential functions of the DEGs for understanding the sheep host resistance mechanisms. As the functional annotation of sheep genes is not available, we will first find the homologs of all sheep genes in human using homology search and then perform the functional analysis using human GO annotations. Finally, we will inspect the differential analysis results and functional analysis results to explain the biological meanings in sheep host resistance.

2.2.3 Data Preprocessing and Alignment

The quality of raw sequence reads was first checked using FastQC (`http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/`). The effect of trimming of low-quality nucleotides on genome alignment was examined using sickle (`https://github.com/najoshi/sickle`) and STAR algorithm [30]. Raw sequence reads (FASTQ files) of 23 samples were mapped against the ovine reference genome Oar_v3.1 using STAR (v2.3.1t) with default parameters. The uniquely mapped reads were used to count against the Ensembl annotation Oar_v3.1 using customized program for calculating the number of reads per gene. The counts of all samples were tabulated.

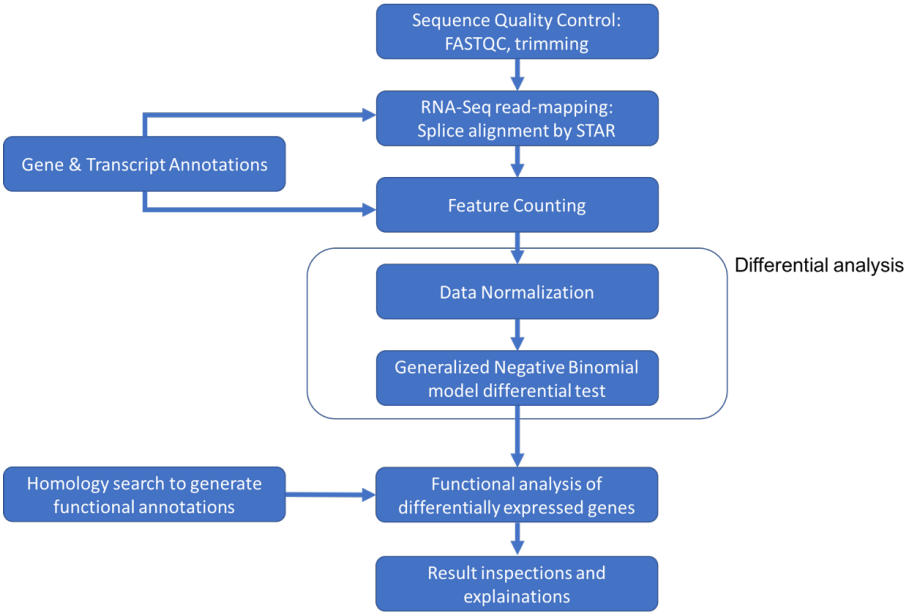


Figure 2.1: Analysis workflow of the sheep host resistance RNA-Seq dataset.

2.2.4 Differential Expression Analysis

2.2.4.1 Modeling

Differential expression analysis [31, 32] tests whether the observed read count difference between two biological groups is significantly larger than random variations for a given gene. It is reasonable to assume that the sequenced reads are independently sampled. As the compositions of genes are fixed, the read counts would follow a multinomial distribution, which can be well approximated by a Poisson distribution for statistical simplicity. Poisson distribution has been used to model the read counts in some studies [33, 34]. However, due to the sample heterogeneity and other technical and biological reasons, the read counts of RNA-Seq samples usually show larger variations than what is predicted by the Poisson distribution [35, 36, 37]. This phenomenon is widely noted as overdispersion. As a result, using Poisson distribution for RNA-Seq differential expression analysis will not control the type-I error well.

In this study, the RNA-Seq read counts are modeled with the negative binomial (NB) distribution, which has been proposed to resolve the overdispersion problem in RNA-Seq count data[31, 32]. The read count of gene i in sample j, n_{ij}, is assumed to follow a negative binomial distribution,

$$n_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2), \tag{2.1}$$

where μ_{ij} is the mean and σ_{ij}^2 is the variance.

2.2.4.2 Nomalization

As the sequenced RNA fragments are uniformly sampled from a transcript pool, the number of reads aligned to a particular gene is determined by

- The expression level of the gene.
- The length of the gene.
- The library size, or the number of mapped short reads obtained in sequencing a given library.

Since the length of each gene is identical across samples, it is necessary to eliminate the variation of the cDNA library size before comparing the gene expression between two conditions. The raw read counts of each gene need to be scaled by sample-specific factors to normalize the library size effects. This normalization procedure removes systematic effects that are not associated with the biological differences and is crucial for proper differential expression analysis of high-throughput sequencing data.

Different normalization methods have been proposed and evaluated[38, 39], such as total count, upper quartile [36], median, DESeq normalization [31], trimmed mean of M-values [40], quantile normalization [41] and RPKM [42]. All these methods behave

differently in normalizing the high-throughput sequencing counts. It has been reported that the DESeq normalization method and trimmed mean of M-values method perform better in terms of robustness to the different library sizes and compositions, while the trimmed mean of M-values method is more sensitive to the filtering strategy [38, 39]. Therefore, the DESeq normalization method is chosen to normalize the RNA-Seq data in this study.

To incorporate the normalization, the parameter μ_{ij} in Equation. 2.1, that is, the expected read count of gene i in sample j, is determined by

$$\mu_{ij} = q_{ij} s_j, \tag{2.2}$$

where q_{ij} is proportional to the true fraction of reads originated from gene i in sample j, and s_j is a normalization factor that represents the library size of sample j.

The nomalization factor s_j can be estimated [31] by

$$\hat{s}_j = \underset{i}{median} \left(\frac{n_{ij}}{\prod_{k=1}^{k=m} n_{ik}} \right)^{\frac{1}{m}}. \tag{2.3}$$

Equation 2.3 computes the median of ratios between read counts of all genes and a reference count that is simulated by the geometric mean of counts of all genes. The assumption of this normalization is that most genes are not differentially expressed and thus should have similar read counts between samples. Therefore, the median of the ratios of all genes estimates the sample-wise normalization factor to fulfill the hypothesis.

2.2.4.3 Generalized Linear Model for Differential Expression Test

Based on Equation. 2.1 and Equation. 2.2, the read count n_{ij} can be modeled by a negative binomial generalized linear model with a logarithmic link [31]:

$$\begin{aligned} n_{ij} &\sim NB(\mu_{ij}, \sigma_{ij}^2) \\ mu_{ij} &= s_j q_{ij} \\ \log q_{ij} &= \sum_r x_{ir}\beta_{ir}, \end{aligned} \tag{2.4}$$

where x_{jr} is the design matrix element and β_{ir} is the coefficient.

2.2.5 Gene Ontology Analysis

Gene Ontology (GO) [43] provides evidence-supported annotations to associate genes with biological terms and describe the functional roles of genes by classifying them using ontologies. The Gene Ontology consortium [44] has developed three structured ontologies, namely molecular functions, biological processes and cellular components for different model organisms. It not only addresses the need for comprehensive coverage and consistent description of the gene products but also provides community-wide agreed annotation for the gene function descriptions across organisms.

The development of genome-wide or transcriptome-wide technologies has made GO analysis an important analytical method for interpreting these large-scale analysis results [45]. For example, the differential expression analysis of RNA-Seq data will produce a set of DEGs whose biological functions can be difficult to interpret. GO analysis enables the understanding of these large gene sets using the prior functional knowledge from the GO annotations. It answers an important question: what molecular functions or biological processes or cellular components are related with a set of genes.

The following method is adopted to identify GO terms statistically over-represented

in a given gene set, compared to the reference gene set. Suppose for a given GO term, a given gene set and reference gene set (for example, all the expressed genes in a dataset), all genes can be classified into four categories:

- a genes in the set that are annotated by the GO term;
- b genes are not annotated by the GO term;
- c genes are not in the gene set but are annotated by the GO term;
- d genes are neither in the gene set nor annotated by the GO term.

Then X, the number of genes within the gene set and annotated by the GO term, follows a hyper-geometric distribution.

$$Pr(X) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} \tag{2.5}$$

The significance of the enrichment of a given gene set in a GO term could be calculated by Fisher's exact test.

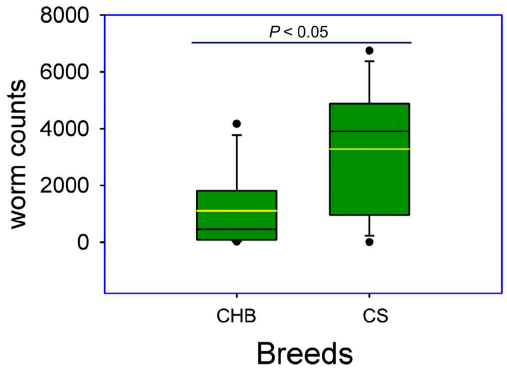


Figure 2.2: Differences in worm counts between resistant and susceptible sheep breeds under experimental *Haemonchus contortus* infection. Boxes denote the inter-quartile range between the 1st and 3rd quartiles (25 and 75%, respectively). Black line: mean; Yellow line: median. CHB: Canaria Hair Breed (resistant). CS: Canaria sheep (susceptible).

2.3 Results

2.3.1 *Haemonchus* Infection Induced Distinctly Different Transcriptome Patterns in the Abomasal Mucosa of CHB and CS Breeds

The total worms recovered from the infected groups of CHB and CS were $1,109.75$ ($\pm 1,547.73$, SD) and $3,280.50$ ($\pm 2,398.03$, SD), respectively. The difference is statistically significant ($P < 0.05$, Figure 2.2). Neither *Haemonchus* worms nor fecal eggs were recovered from the uninfected group of either breed, as expected. EPG values detected from infected CS sheep were 262.50 ± 287.54 (mean $\pm$ SD) while no fecal eggs were detectable in the infected group of CHB sheep at 20 days post infection (dpi). No parasite eggs in either group prior to the experimental challenge were observed.

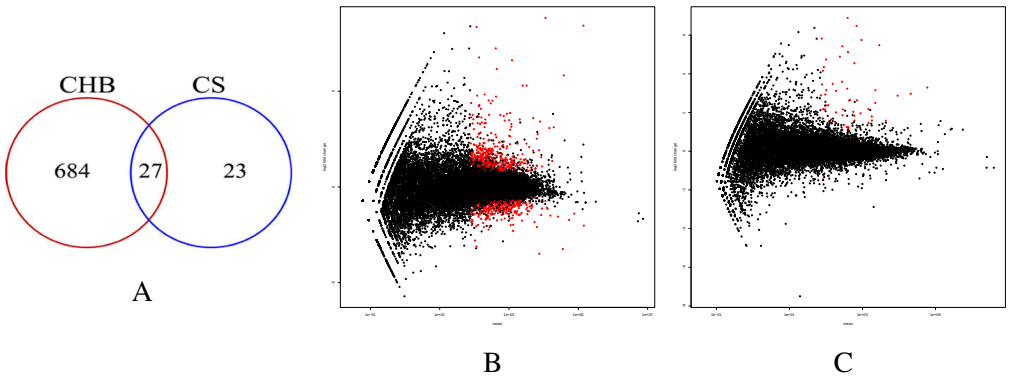


Figure 2.3: Venn Diagram of differentially expressed genes and scatter plot of Log2 ratio vs Mean. (A) Venn Diagram showing the number of genes with significant differences in transcript abundance induced by infection in two sheep breeds compared to their respective uninfected controls at a false discovery rate (FDR) cutoff < 0.05. (B,C) Scatter plot of log2 ratio (fold change) vs mean. The red color indicates genes detected as differentially expressed between the infected group and uninfected controls at a false discovery rate (FDR < 0.05) in CHB (B) and CS (C).

In this study, approximately 79.91% of raw reads ($\pm 7.08\%$; SD) were uniquely mapped

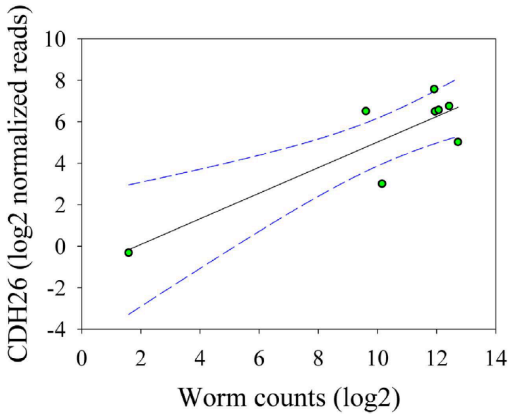


Figure 2.4: Differences in worm counts between resistant and susceptible sheep breeds under experimental *Haemonchus contortus* infection. Nonlinear regression between worm counts and normalized transcript abundance per million mapped reads of the gene cadherin 26 (CDH26) in susceptible Canaria Sheep (CS). Dotted lines: 95% confidence interval.

to the ovine genome. Compared to their respective uninfected controls, the numbers of genes significantly impacted by infection in CHB and CS breeds at a stringent cut-off value (false discovery rate or FDR < 0.05), were 711 and 49, respectively (Figure 2.3). The abundance of 27 genes was significantly changed by infection in both breeds (Table 2.1). Among them, 25 genes, such as arachidonate 15-lipoxygenase (ALOX15), collagen, type VI, α 5 (COL6A5), and serglycin (SRGN), were significantly upregulated while the expression of transthyretin (TTC) was repressed by infection. Intriguingly, the transcript abundance of cadherin 26 (CDH26) was significantly induced by infection in both breeds (adjusted P value or FDR $< 1.63 \times 10^{10}$); and is strongly correlated with worm counts only in CS (Figure 2.4). However, infection had a bidirectional impact on the transcript abundance of a uncharacterized gene containing a unknown microRNA (ENSOARG00000023771), which was significantly upregulated in CHB but downregulated in CS. The genes significantly impacted by infection only in CS included mast cell proteinase-3, γ-glutamyltransferase 5 (GGT5), CD163 as well as those involved in

smooth muscle contraction, such as tropomyosin (TPM2), myosin, light chain 9, regulatory (MYL9), and calponin 1, basic, smooth muscle (CNN1).

Table 2.1: Genes significantly impacted by *Haemonchus contortus* infection in both CHB and CS breeds.

Gene ID	Symbol	Fold Change		FDR	
		CHB	CS	CHB	CS
ENSOARG00000000338	ABCA2	2.86	2.40	2.28%	1.29%
ENSOARG00000008480	ALOX15	10.02	11.45	2.06%	0.00%
ENSOARG00000015249	CDH26	150.19	54.19	0.00%	0.00%
ENSOARG00000018133	CFTR	4.33	2.38	1.21%	3.67%
ENSOARG00000014842	COL6A5	20.79	6.28	1.67%	0.04%
ENSOARG00000007787	FCER1A	20.69	18.23	0.00%	0.00%
ENSOARG00000019163	HBBB	11.47	43.75	1.63%	0.00%
ENSOARG00000008994	IGHE	22.73	35.02	0.00%	0.00%
ENSOARG00000013111	IL1RL1	9.73	5.26	0.00%	0.00%
ENSOARG00000016842	MCTP1	3.89	2.95	1.26%	1.03%
ENSOARG00000002234	SLC2A3	4.39	3.06	0.00%	0.03%
ENSOARG00000005322	SRGN	4.25	3.60	0.45%	0.00%
ENSOARG00000012855	ST3GAL4	2.44	3.69	0.84%	0.94%
ENSOARG00000009990	SYNM	2.16	6.18	2.99%	0.09%
ENSOARG00000005941	TNC	3.00	4.01	2.80%	3.19%
ENSOARG00000014689	TPSAB1	6.96	9.43	0.00%	0.00%
ENSOARG00000006342	TTR	0.28	0.37	0.00%	3.17%
ENSOARG00000000857		8.41	7.84	0.00%	1.29%
ENSOARG00000002036		38.84	44.81	0.00%	0.00%
ENSOARG00000002629		13.34	23.51	0.00%	0.00%
ENSOARG00000002942		7.16	7.83	0.00%	0.00%
ENSOARG00000002964		12.61	29.65	0.01%	0.00%
ENSOARG00000006087		13.49	26.69	0.00%	0.00%
ENSOARG00000013005		3.00	3.53	4.13%	0.94%
ENSOARG00000013263		71.33	91.67	0.00%	0.01%
ENSOARG00000017398		5.73	7.34	2.15%	0.65%
ENSOARG00000023771		0.46	3.00	0.19%	1.55%

Among the genes significantly impacted by infection in CHB sheep, several cytokine

receptors and chemokines were strongly upregulated. Notable, the transcript of IL17 receptor beta (IL17RB) was 14.4 fold higher in infected animals than uninfected controls in CHB. IL2 receptor beta (IL2B) was also upregulated. Similarly, chemokine CXC ligand 12 (CXCL12) and chemokine (CXC motif) receptor 6 (CXCR6) were upregulated by infection in CHB. Among the well-known Th2 cytokines, the expression of IL6, IL10 and IL13 was upregulated by infection in both breeds. Moreover, while the extent of upregulation of IL6 by infection remained similar in both breeds (~ 6.8 fold), overexpression of both IL10 and IL13 mRNA molecules was more profound in the resistant breed (CHB) than in CS. On the other hand, the IL5 mRNA was upregulated by infection in CS but barely detectable in CHB at the sequencing depth in this study. The IL4 expression followed the similar trend: it was upregulated approximately 9 fold by infection in CS but was barely detectable in CHB. However, the IL9 mRNA level remained unchanged by infection in both breeds.

Several genes involved in arachidonic acids metabolism, including eicosanoids metabolism, were significantly impacted by infection, such as arachidonate 5-lipoxygenase (ALOX5) and its activating protein (ALOX5P), prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase) (PTGS1, COX1), prostaglandin reductase 1 (PTGR1), and thromboxane A synthase 1 (TBXAS1), were all strongly upregulated by infection in CHB. In addition, at least 11 genes implicated in complement activation were significantly impacted by infection in CHB, such as complement factor properdin (CFP, 2.8 fold), complement component 7 (C7, 4.2 fold), and complement factor I (CFI, 12.1 fold). Other known genes involved in protective immunity to helminth infection strongly upregulated by infection in CHB included amphiregulin (AREG, 2.2 fold), granzyme genes A and B (GZMA and GZMB, 6.8 and 12.9 fold, respectively).

41 of the 711 genes significantly impacted by infection in CHB are related to extracellular matrix (ECM, Table 2.2). Of them, fibronectin 1 (FN1) was strongly upregulated. At

least ten collagen genes were significantly upregulated, such as those from Type I, Type III, Type V, Type VI, and Type XII (Table 2.2). For example, the expression of collagen, type VI, alpha 5 (COL6A5) and collagen, type XII, alpha 1 (COL12A1) was increased 20.8 and 2.4 fold, respectively in CHB, compared to the uninfected controls. Likewise, matrix metallopeptidase 1 (MMP1), MMP2, and MMP14 were significantly up-regulated while the transcript of MMP11 was repressed by infection. Furthermore, several cell adhesion molecules, including integrins, lectins, and cadhesion, were strongly upregulated by infection in CHB, such as conglutinin-like (COLEC8, 451.7 fold), integrin, α 11 (ITGA11, 3.4 fold), and lectin, galactoside-binding, soluble, 15 (LGALS15, 340.1 fold).

Table 2.2: 41 extracellular matrix (ECM) related genes significantly affected by *Haemonchus contortus* infection in the abomasal mucosa of the Canaria Hair Breed sheep.

GeneID	Symbol	Fold change	P-value	FDR
ENSOARG00000013782	ALB	0.20	0.0000	0.10%
ENSOARG00000008507	ALPL	10.59	0.0000	0.00%
ENSOARG00000005139	APLP1	0.54	0.0001	0.49%
ENSOARG00000018738	BMP2	1.87	0.0004	1.11%
ENSOARG00000012877	CFP	2.78	0.0000	0.19%
ENSOARG00000004871	COL1A1	2.18	0.0005	1.24%
ENSOARG00000001508	COL1A2	1.87	0.0029	4.42%
ENSOARG00000016476	COL3A1	2.09	0.0000	0.21%
ENSOARG00000002129	COL5A1	2.39	0.0002	0.55%
ENSOARG00000016440	COL5A2	1.93	0.0001	0.28%
ENSOARG00000012810	COL6A1	3.72	0.0000	0.00%
ENSOARG00000012880	COL6A2	3.25	0.0000	0.17%
ENSOARG00000019080	COL6A3	2.65	0.0003	0.83%
ENSOARG00000014842	COL6A5	20.79	0.0008	1.67%
ENSOARG00000006410	COL12A1	2.37	0.0005	1.26%
ENSOARG00000009670	CPXM2	2.10	0.0007	1.63%
ENSOARG00000017328	F3	3.03	0.0000	0.06%
ENSOARG00000019404	FBLN1	1.53	0.0022	3.54%
ENSOARG00000017189	FBN2	0.41	0.0000	0.01%
ENSOARG00000016733	FGA	0.38	0.0009	1.93%
ENSOARG00000019329	FN1	4.77	0.0000	0.07%
ENSOARG00000018483	ITGA11	3.37	0.0015	2.75%

Table 2.2: Continued

GeneID	Symbol	Fold change	P-value	FDR
ENSOARG00000016642	ITGB7	1.93	0.0032	4.70%
ENSOARG00000010344	LTBP1	2.34	0.0000	0.16%
ENSOARG00000005315	MMP1	11.03	0.0001	0.47%
ENSOARG00000013161	MMP11	0.41	0.0000	0.21%
ENSOARG00000019414	MMP14	1.85	0.0001	0.25%
ENSOARG00000018035	MMP2	1.99	0.0000	0.12%
ENSOARG00000008537	NAV2	0.57	0.0022	3.52%
ENSOARG00000010519	OLFML2B	2.26	0.0004	1.03%
ENSOARG00000006153	PDGFA	0.51	0.0009	1.85%
ENSOARG00000005685	PLOD2	3.06	0.0011	2.15%
ENSOARG00000010041	POSTN	6.73	0.0000	0.00%
ENSOARG00000005275	PXDN	1.94	0.0010	1.98%
ENSOARG00000004813	SDC2	1.95	0.0006	1.39%
ENSOARG00000005209	SDC4	0.57	0.0009	1.93%
ENSOARG00000006391	SERPINB5	3.99	0.0000	0.00%
ENSOARG00000020413	SERPINE2	1.78	0.0003	0.80%
ENSOARG00000015081	TGFBI	3.12	0.0000	0.13%
ENSOARG00000005941	TNC	3.00	0.0016	2.80%
ENSOARG00000008334	VEGFA	0.44	0.0000	0.05%

Of note, approximately 15% of the genes significantly impacted by infection are cell-cycle related. The expression of these cell cycle related genes was predominantly enhanced by *Haemonchus* infection in CHB. As Table 2.3 shows, at least 92 genes were significantly upregulated by infection, such as cyclin A2 (CCNA2), cyclin B3 (CCNB3), various centromere proteins (CENPL, CENPN, CENPT, and CENPW) and kinesin family (KIF) members, and at least 5 minichromosome maintenance complex (MCM) components (MCM3, MCM4, MCM5, MCM6, and MCM10). Nevertheless, the infection was also able to repress cell cycle related genes, such as cyclin G1 (CCNG1), regulator of cell cycle (RGCC), and synaptonemal complex protein 3 (SYCP3). Moreover, at least five transcription factors, such as the oncogene MYB, SMAD family members 6 and 9 (SMAD6 and SMAD9), and histone decetylase 5 (HDAC5), were significantly affected by

infection in CHB.

Table 2.3: 100 cell-cycle related genes significantly affected by *Haemonchus contortus* infection in the resistant breed. Fold is expressed as infected/uninfected controls. FDR: false discovery rate.

Gene ID	Locus (chr:start:end)	Gene symbol	Fold change	FDR
ENSOARG00000004361	4:60803768:60856028	ANLN	5.01	0.0022
ENSOARG00000019052	11:27395740:27400878	AURKB	2.94	0.0031
ENSOARG00000012875	18:20942195:21007732	BLM	2.46	0.0185
ENSOARG00000018738	13:48462231:48472599	BMP2	1.87	0.0111
ENSOARG00000004835	11:42540868:42607539	BRCA1	3.09	0.0152
ENSOARG00000020126	7:32810618:32861855	BUB1B	3.38	0.0037
ENSOARG00000020247	7:33203355:33258031	CASC5	3.52	0.0442
ENSOARG00000014176	6:3657736:3663110	CCNA2	3.95	0.0000
ENSOARG00000009908	X:51894389:51938870	CCNB3	4.54	0.0269
ENSOARG00000010352	3:209738381:209761224	CCND2	2.44	0.0060
ENSOARG00000004318	25:16420008:16437119	CDC2	3.96	0.0089
ENSOARG00000020542	1:17815061:17818632	CDC20	3.07	0.0140
ENSOARG00000001274	13:50726485:50734786	CDC25B	1.82	0.0476
ENSOARG00000014063	11:40114667:40125179	CDC6	4.24	0.0000
ENSOARG00000009851	2:39869389:39909559	CDCA2	2.74	0.0218
ENSOARG00000005652	3:207550908:207552937	CDCA3	2.65	0.0015
ENSOARG00000019830	1:12275289:12289262	CDCA8	3.70	0.0029
ENSOARG00000011059	6:21798087:21861426	CENPE	3.68	0.0008
ENSOARG00000012529	12:53214185:53224480	CENPL	2.72	0.0158
ENSOARG00000008206	14:7100459:7127206	CENPN	3.59	0.0029
ENSOARG00000003186	14:34675201:34680944	CENPT	2.71	0.0276
ENSOARG00000007744	8:12047574:12055213	CENPW	3.21	0.0049
ENSOARG00000003158	22:14531509:14550251	CEP55	4.11	0.0002
ENSOARG00000009704	5:17254276:17278480	CHAF1A	2.25	0.0059
ENSOARG00000009289	10:21841538:21861932	CKAP2	3.27	0.0403
ENSOARG00000007721	2:23987476:23992086	CKS2	3.46	0.0001
ENSOARG00000019542	1:10450581:10480905	CLSPN	2.00	0.0269
ENSOARG00000021089	7:64546752:64587073	DLGAP5	3.93	0.0147
ENSOARG00000017620	1:186409:193575	DTYMK	1.92	0.0339
ENSOARG00000007334	2:242210461:242229702	E2F2	3.47	0.0019
ENSOARG00000008807	21:25034831:25052009	E2F8	4.39	0.0000
ENSOARG00000020622	3:199025742:199151626	EPS8	2.36	0.0025
ENSOARG00000005908	X:61406551:61410223	ERCC6L	4.54	0.0001

Table 2.3: Continued

Gene ID	Locus (chr:start:end)	Gene symbol	Fold change	FDR
ENSOARG00000014845	2:100985329:101014711	ESCO2	3.40	0.0022
ENSOARG00000000644	13:67321805:67342479	FAM83D	3.04	0.0037
ENSOARG00000005211	19:16642527:16690458	FANCD2	2.84	0.0015
ENSOARG00000003968	8:76466657:76471151	FBXO5	3.65	0.0030
ENSOARG00000015633	21:39647404:39648523	FEN1	2.10	0.0191
ENSOARG00000011054	3:210883196:210893395	FOXM1	3.42	0.0158
ENSOARG00000007717	13:52921512:52934511	GINS1	3.70	0.0006
ENSOARG00000004495	22:15498816:15537211	HELLS	3.82	0.0000
ENSOARG00000019189	1:6959524:6970726	HJURP	3.11	0.0051
ENSOARG00000020743	7:42727941:42735567	KIAA0101	4.06	0.0001
ENSOARG00000004780	19:16405186:16463300	KIF15	3.29	0.0413
ENSOARG00000015211	15:56874125:56943420	KIF18A	2.92	0.0295
ENSOARG00000015873	5:46994289:47001867	KIF20A	3.56	0.0028
ENSOARG00000005591	24:26450684:26467845	KIF22	2.77	0.0030
ENSOARG00000018647	7:16078138:16122235	KIF23	2.92	0.0288
ENSOARG00000001102	1:19196065:19216520	KIF2C	3.45	0.0079
ENSOARG00000009637	20:7678567:7686289	KIFC1	2.48	0.0139
ENSOARG00000020216	7:33016321:33025738	KNSTRN	2.15	0.0292
ENSOARG00000009349	17:52444861:52510658	KNTC1	2.66	0.0208
ENSOARG00000015347	11:48337064:48345443	KPNA2	2.30	0.0044
ENSOARG00000015665	6:5610960:5622911	MAD2L1	2.76	0.0062
ENSOARG00000005416	13:26912745:26937031	MCM10	2.02	0.0487
ENSOARG00000014143	20:24477536:24494074	MCM3	2.06	0.0017
ENSOARG00000012797	9:32400919:32411149	MCM4	2.57	0.0001
ENSOARG00000018527	3:178690575:178707822	MCM5	2.32	0.0010
ENSOARG00000010614	2:173834090:173868287	MCM6	2.25	0.0019
ENSOARG00000011541	2:51686232:51755300	MELK	3.87	0.0002
ENSOARG00000009575	18:53870504:53915925	MIS18BP1	3.74	0.0499
ENSOARG00000014562	22:46439178:46468099	MKI67	3.15	0.0149
ENSOARG00000014901	8:60237288:60271515	MYB	3.06	0.0014
ENSOARG00000003547	13:71800636:71832436	MYBL2	3.10	0.0000
ENSOARG00000004016	6:37256547:37333851	NCAPG	3.36	0.0004
ENSOARG00000007995	4:118730937:118799615	NCAPG2	2.64	0.0025
ENSOARG00000009604	23:37208584:37244969	NDC80	3.71	0.0003
ENSOARG00000011466	12:69915815:69927483	NEK2	3.45	0.0019
ENSOARG00000011189	1:113098396:113135220	NUF2	2.47	0.0199
ENSOARG00000005282	1:26531082:26561083	ORC1	2.82	0.0021

Table 2.3: Continued

Gene ID	Locus (chr:start:end)	Gene symbol	Fold change	FDR
ENSOARG00000014858	2:101015056:101043656	PBK	4.32	0.0008
ENSOARG00000017133	13:46615170:46619285	PCNA	2.05	0.0043
ENSOARG00000010890	12:52045755:52079590	PDPN	2.08	0.0264
ENSOARG00000015691	17:29540624:29557094	PLK4	1.97	0.0338
ENSOARG00000020607	7:39763421:39793166	POLE2	2.83	0.0096
ENSOARG00000012267	18:20787183:20802103	PRC1	3.88	0.0198
ENSOARG00000020254	7:33275919:33306843	RAD51	3.12	0.0031
ENSOARG00000010707	24:15349788:15350439	RAN	1.79	0.0300
ENSOARG00000017221	13:65583237:65642844	RBL1	2.12	0.0338
ENSOARG00000017883	13:58047850:58061508	RBM38	2.35	0.0464
ENSOARG00000015333	3:19185137:19191043	RRM2	5.12	0.0000
ENSOARG00000017493	6:91484955:91532036	SEPT11	2.18	0.0063
ENSOARG00000018302	2:239211366:239212113	SFN	2.06	0.0420
ENSOARG00000007399	2:18870299:18918858	SMC2	2.60	0.0324
ENSOARG00000001001	11:19659072:19677569	SPAG5	2.45	0.0058
ENSOARG00000017725	5:13248115:13254108	SPC24	2.98	0.0089
ENSOARG00000002888	16:1518544:1544794	SPDL1	1.95	0.0442
ENSOARG00000011578	18:20124947:20166662	TICRR	2.83	0.0286
ENSOARG00000001419	13:60597535:60651659	TPX2	2.52	0.0149
ENSOARG00000016302	7:3740229:3789354	TRIM36	10.24	0.0000
ENSOARG00000007151	8:6885831:6925302	TTK	3.59	0.0006
ENSOARG00000018884	3:136887716:136929972	TUBA4A	2.02	0.0042
ENSOARG00000006520	13:74125916:74129104	UBE2C	3.13	0.0008
ENSOARG00000008530	5:16805815:16842579	UHRF1	3.99	0.0000
ENSOARG00000004351	8:81449297:81450503		1.84	0.0185
ENSOARG00000006991	19:59822300:59844027		2.08	0.0084
ENSOARG00000005764	2:240064407:240068660		2.37	0.0015
ENSOARG00000006571	21:17310180:17310609		2.66	0.0003
ENSOARG00000005759	16:10486354:10496192		3.81	0.0058
ENSOARG00000000647	22:13731293:13771253		3.92	0.0106

Intriguingly, four genes known to regulate abomasal acid secretion and gastric function [46] were downregulated by *Haemonchus* infection in CHB, including ATPase, H+ /K+ exchanging, alpha polypeptide (ATP4A), progastricsin (pepsinogen C, PGC), appetite-regulating hormone precursor (GHRL), and forkhead box A2 (FOXA1). However, the

transcript abundance of these four genes remained unchanged by infection in CS.

The RNAseq results of selected genes were validated by real-time RT-PCR (Figure 2.5). For example, the expression of CFI, CXCR6, LGALS15, and MMP1 was significantly upregulated while TFF2 mRNA level was significantly repressed by infection only in the resistant breed (CHB), in a good agreement with the RNAseq analysis. A strong correlation in log2 transformed fold values between the two platforms, qPCR and RNAseq, was evident (a correlation coefficient $R = 0.946$; Figure 2.6).

2.3.2 Gene Ontology Implicated in Host Resistance

Table 2.4: Gene Ontology (GO) biological processes (BP) significantly enriched in both resistant (CHB) and susceptible (CS) breeds.

GO ID	Description	Z Score		P-value	
		CHB	CS	CHB	CS
GO:0002250	adaptive immune response	4.65	5.09	4.59×10^{-05}	7.37×10^{-04}
GO:0046456	eicosanoid biosynthetic process	6.92	6.84	3.81×10^{-07}	6.59×10^{-04}
GO:0006691	leukotriene metabolic process	7.22	9.67	9.37×10^{-07}	9.52×10^{-05}
GO:1901570	fatty acid derivative biosynthetic process	6.92	6.84	3.81×10^{-07}	6.59×10^{-04}
GO:0006636	unsaturated fatty acid biosynthetic process	7.02	6.43	2.12×10^{-07}	9.16×10^{-04}

Among 477 and 16 GO terms significantly enriched in CHB and CS at a P value cutoff 1.0×10^{-4}, respectively, five were significantly enriched in both breeds (Table 2.4). Select GO terms that may be implicated in the development of host resistance to *Haemonchus* infection are listed in Table 2.5. Several GO related to complement activation (both classical and alternative pathways) and its regulation were significantly enriched only in CHB. Numerous cell cycle related GO were significantly enriched as well (Figure 2.7). GO related to secretory granule and gastric acid secretion were also enriched, suggesting that

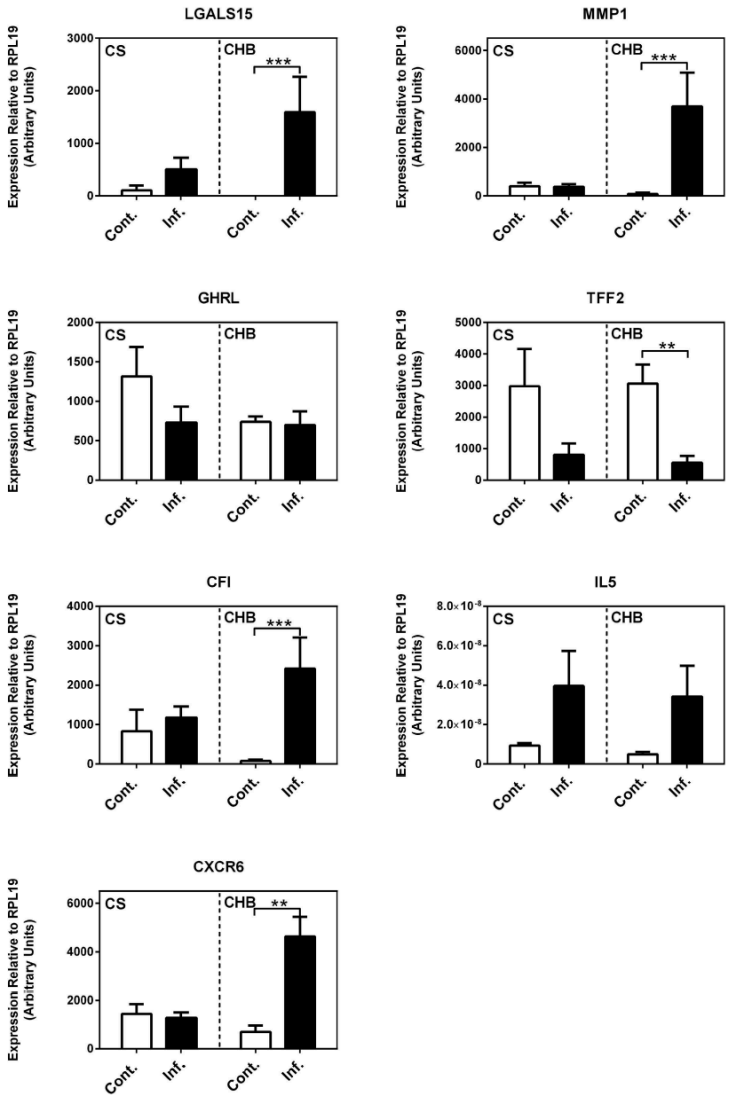


Figure 2.5: Real-Time RT-PCR analysis (qPCR) of selected genes. Relative expression levels calculated from standard curves were normalized to the endogenous control gene RPL19. Numbers represent mean values plus standard error. Cont.: uninfected controls; Inf.: 20 days post infection by Haemonchus contortus. CS: Canaria Sheep; CHB: Canaria Hair Breed. CFI: Complement factor I; CXCR6: Chemokine (C-X-C motif) receptor 6; GHRL: Ghrelin/obestatin prepropeptide; LGALS15: Galectin 15; IL5: Interleukin 5. MMP1: Matrix metallopeptidase 1; TFF2: Trefoil factor 2. **P < 0.001; ***P < 0.0001.

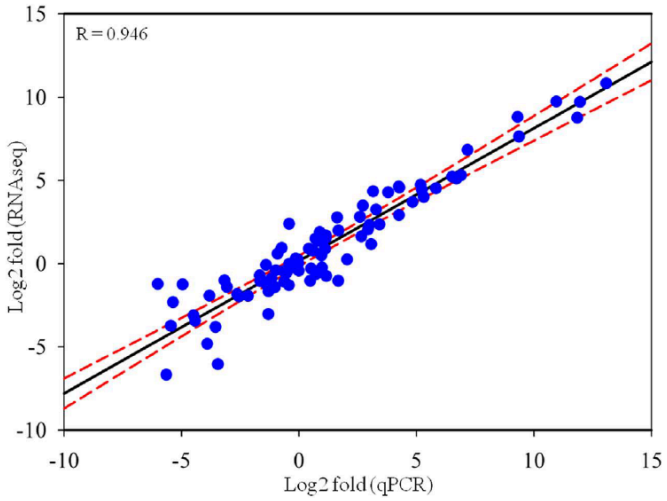


Figure 2.6: Linear regression analysis of fold changes calculated from qPCR and RNA-Seq analysis. Blue dots represent log2 transformed fold change values of a single gene in an infected sample obtained from qPCR (X-axis) and RNAseq analysis (Y-axis). Dashed lines: 99% Confidence Interval. R: correlation coefficient.

the ability to regulate secretory and gastric function of the host may be involved in the development of host resistance. Furthermore, the regulation of inflammation at the site of infection (mucosa), including arachidonic acid metabolism, cyclooxygenase pathway, and positive regulation of MAPK cascade, as well as leukocyte migration were also implicated in host resistance. On the other hand, four of the 11 GO unique to CS were related to muscle contraction.

Table 2.5: Selected Gene Ontology (GO) terms significantly impacted by *Haemonchus contortus* infection in the resistant breed (CHB). BP = Biological processes. MF = Molecular functions. CC = Cellular components.

GO id	Ontology	Description	Observed /Total	Z Score	P-Value
GO:0019369	BP	arachidonic acid metabolic process	10/36	4.95	1.13×10^{-04}
GO:0002673	BP	regulation of acute inflammatory response	15/38	7.93	1.16×10^{-08}

Table 2.5: Continued

GO id	Ontology	Description	Observed /Total	Z Score	P-Value
GO:0002675	BP	positive regulation of acute inflammatory response	8/13	7.78	4.69×10^{-07}
GO:0050727	BP	regulation of inflammatory response	24/135	5.01	1.53×10^{-05}
GO:0050900	BP	leukocyte migration	30/183	5.11	7.54×10^{-06}
GO:0043410	BP	positive regulation of MAPK cascade	30/197	4.65	3.28×10^{-05}
GO:0006956	BP	complement activation	14/33	8.06	1.18×10^{-08}
GO:0030449	BP	regulation of complement activation	11/23	7.75	9.91×10^{-08}
GO:0006957	BP	complement activation, alternative pathway	7/10	7.87	7.26×10^{-07}
GO:0006958	BP	complement activation, classical pathway	11/25	7.33	2.87×10^{-07}
GO:0031714	MF	C5a anaphylatoxin chemotactic receptor binding	5/5	8.21	1.55×10^{-06}
GO:0031715	MF	C5L2 anaphylatoxin chemotactic receptor binding	4/4	7.35	2.25×10^{-05}
GO:0007049	BP	cell cycle	119/1137	5.00	1.53×10^{-06}
GO:0022402	BP	cell cycle process	97/865	5.21	8.16×10^{-07}
GO:0008283	BP	cell proliferation	111/1013	5.34	3.84×10^{-07}
GO:0051301	BP	cell division	77/504	7.59	1.83×10^{-11}
GO:0051302	BP	regulation of cell division	23/140	4.47	8.26×10^{-05}
GO:0051321	BP	meiotic cell cycle	15/87	3.82	8.23×10^{-04}
GO:0031577	BP	spindle checkpoint	11/40	5.15	5.72×10^{-05}
GO:0007088	BP	regulation of mitosis	15/80	4.20	3.26×10^{-04}
GO:0042555	CC	MCM complex	5/11	5.05	5.03×10^{-04}

Table 2.5: Continued

GO id	Ontology	Description	Observed /Total	Z Score	P-Value
GO:0044818	BP	mitotic G2/M transition checkpoint	6/11	6.24	3.63×10^{-05}
GO:0007186	BP	G-protein coupled receptor signaling pathway	33/207	5.18	4.97×10^{-06}
GO:0043627	BP	response to estrogen	16/83	4.46	1.48×10^{-04}
GO:0001676	BP	long-chain fatty acid metabolic process	11/53	3.99	8.27×10^{-04}
GO:0033500	BP	carbohydrate homeostasis	15/81	4.14	3.75×10^{-04}
GO:0006865	BP	amino acid transport	14/71	4.27	2.97×10^{-04}
GO:0030141	CC	secretory granule	28/172	4.89	1.72×10^{-05}
GO:0051048	BP	negative regulation of secretion	12/57	4.22	4.23×10^{-04}
GO:0071229	BP	cellular response to acid	19/96	5.00	2.49×10^{-05}
GO:0001696	BP	gastric acid secretion	6/7	8.23	7.00×10^{-07}

2.4 Discussion

Parasite resistance refers to the ability of the host to avert infection, resulting in reduced worm burden [47]. Numerous factors affect this trait. Among them, host genetics play a predominant role in controlling the development of resistance while host sex, age, and prior exposure are also important [48]. Differences in parasite resistance and susceptibility existing in various sheep breeds have been long recognized [14]. Moreover, inter- and intra-host variations in resistance are evident in certain sheep populations [48]. Identifying genetics components controlling inter-, and intra-breed differences in parasite resistance has both pragmatic and theoretical implications. Towards this end, numerous efforts have been made over the decades to unravel genes and/or genetic variants responsible for resistance, partially driven by strong desires to breed farm animals with strong resis-

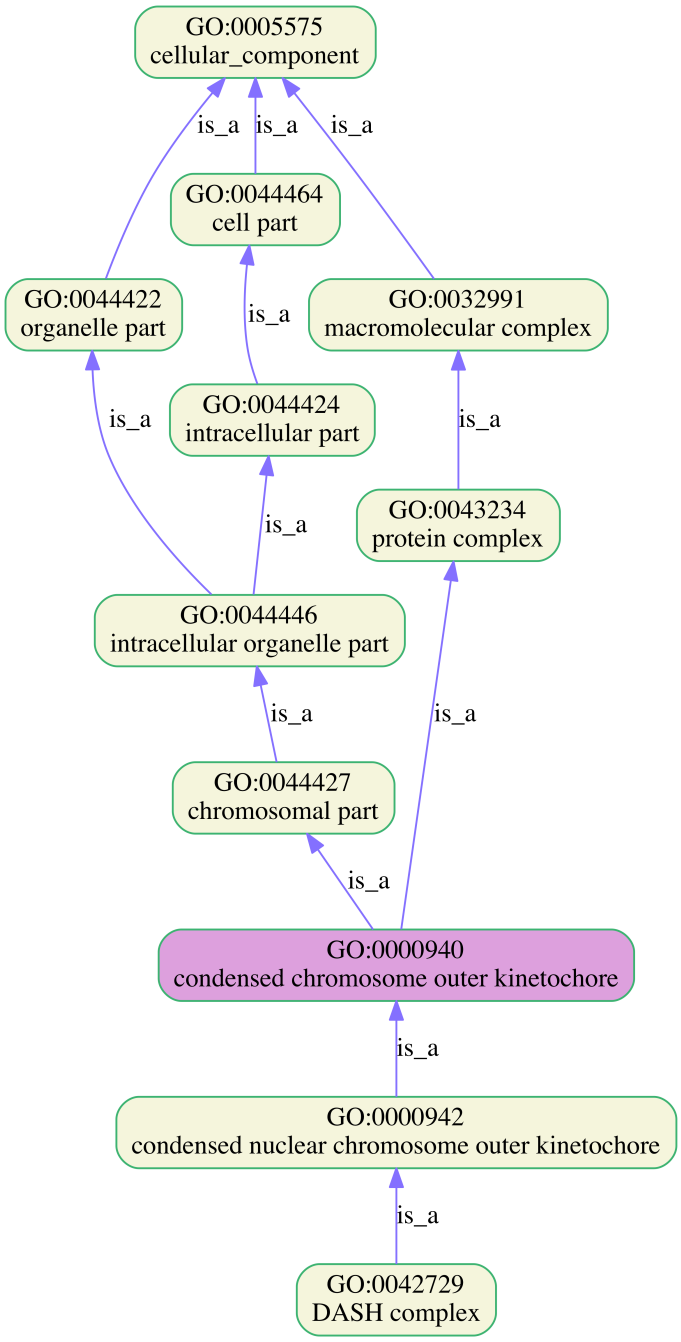


Figure 2.7: Gene Ontology lineage relations. The Cellular Component Ontology term GO:0000940 (condensed chromosome outer kinetochore) significantly enriched in resistant Canaria Hair Breed (CHB, P-value$< 2.30 \times 10^4$).

tance traits. Traditional QTL analysis and Genome-wide Association Studies (GWAS) have led to reports of dozens of QTL or markers on almost every ovine chromosome that are associated with various resistance phenotypes, such as fecal egg counts, packed cell volume, and parasite-specific antibody titers [20, 22, 24]. Nevertheless, the development of parasite resistance relies upon the precise control of expression of the host genome. Understanding these regulatory elements will be crucial towards unraveling their functional relevance. As a result, while much progress has been made to identify genes associated with nematode resistance in sheep during the past few years [49, 50], an in-depth comparison and characterization of transcriptome responses of various breeds and populations, especially those local indigenous breeds harboring varying degrees of parasite resistance and susceptibility, is urgently needed.

The two indigenous breeds of sheep native to the Canary Islands, CHB and CS, display unique and distinct differences in parasite resistance and susceptibility. When co-grazing together on the same pasture under natural infections, differences in fecal trichostrongylid egg counts between CHB and CS are consistently observed [25]. Under experimental infections with *H. contortus*, CHB has a significantly lower, by approximately 50%, worm burden than CS, a undeniable trait of parasite resistance [25, 26], which is confirmed in this study. Moreover, worms recovered in CHB tend to have significantly shorter body length than those in CS. A significantly lower EPG value is consistently observed in the feces of CHB sheep than those of CS animals during experimental infection. For example, at 27dpi, the mean EPG in CS is 5 fold higher than in CHB [25]. CHB sheep not only shed significantly fewer parasite eggs but also tend to have a delayed egg production, indicating an anti-fecundity effect of the immune response in this breed. The results from this study show that at 20 dpi, no parasite eggs were recovered in the feces of infected CHB animals while EPG in the feces of infected CS sheep reached 262.50 (± 287.54, SD). This observation is in agreement with the previous findings [25]. *Haemonchus contortus*

infection generally elicits a potent Th2 immune response in small ruminants. A strong upregulation of several well-known Th2 cytokines by infection in CHB were observed in this study. Previous studies in the Canary Island breeds suggest that divergence in immune response mechanisms exist between CHB and CS. Among various immune cells, abomasal eosinophil numbers are 2 fold higher in CHB than in CS, suggesting that CHB sheep may have developed abilities for enhanced recruitment of eosinophils to the site of infection (abomasal mucosa). Furthermore, CHB sheep have evolved mechanisms attacking the adult stage of the *Haemonchus* parasite, especially its reproduction, as evidenced by the fact that fecundity is negatively correlated with eosinophils and $\gamma\delta$ T cells in the abomasal mucosa [26]. However, the precise molecular mechanisms of the parasite resistance in CHB breed remain largely unclear.

In this study, we identified a total of 477 and 16 Gene Ontology (GO) terms that are significantly enriched in the transcriptome of resistant and susceptible sheep breeds in responses to *Haemonchus* infection, respectively. Among them, only five enriched GO were shared by both breeds. These GO, including leukotriene metabolic process, eicosanoid biosynthesis process, adaptive immune response, and unsaturated fatty acid biosynthesis, likely represents the basic mechanisms of host immune responses to helminth infection in sheep. Indeed, local inflammatory responses are known to be involved in the development of host resistance [51]. The enriched GO unique to the susceptible CS breed were predominantly muscle contraction-related. In cattle, our previous results suggest that smooth muscle hypercontractility induced by primary infection of the intestinal worm Cooperia oncophora represents an important aspect of host responses [27], as in several other host-parasite systems [52]. In rodent models, helminth infection results in an increase in thickness of jejunal smooth muscle layers. Other studies also support the idea that enhanced muscle contractility appears to be associated with more rapid worm expulsion and stronger host immune responses [53]. In addition, granzyme-mediated apoptotic

signaling pathway (GO:0008626) may play an important role in protecting the host from *H. contortus* infection in the susceptible CS breed.

Complement activation as one of the earliest events in host immune responses to helminth infection plays an important role in the development of host resistance [54]. At least 11 complement related genes, such as CFI and C7, were significantly impacted by infection in the resistant CHB breed compared to uninfected controls while none of these genes were affected by infection in the susceptible breed. As a result, both classical and alternative complement pathways appeared to be activated in the resistant breed. Furthermore, two GO molecular functions related to C5a (GO:0031714) and C5L2 anaphylatoxin chemotactic receptor binding (GO:0031715) were significantly enriched in the resistant breed. It is conceivable that these peptides play a critical role in subsequent recruitment of effector cells, such eosinophils and mast cells, to the site of infection.

Intriguingly, approximately 15% of the 711 genes whose transcript abundance were significantly altered by infection in the resistant breed were cell cycle related. The vast majority of these genes were significantly upregulated (Table 2.3). These genes included several cyclins, minichromosome maintenance complex components, and various kinesin family members (Table 2.3). In addition, a large class of genes significantly impacted in the resistant breed was ECM related (Table 2.2). ECM related genes are required during the classical stages of wound repair, inflammation, new tissue formation, and remodeling [55]. Previous studies identified essential roles of Th2 cytokines in limiting tissue damage during helminth infection in rodent models, especially the involvement of IL17 in the early stage of tissue repair via its role in neutrophil recruitment [56]. In this study, the transcript abundance of IL17RB was increased approximately 14 fold by *Haemonchus* infection only in the resistant breed. Of note, upregulation of Th2 cytokines IL10 and IL13 by infection was more profound in CHB than in CS. Together, our findings suggest that the accelerated tissue repair ability, likely mediated by Th2 cytokines, has evolved in the resistant CHB

breed.

Recently, a significant SNP marker on ovine chromosome (OAR) 6 was reported to affect one of key resistant phenotypes in sheep, EPG [20]. This maker explains approximately 4% of the variance observed for EPG. It is suggested that there may exist up to 3 QTL within the 5 Mb region of this locus ($73.1 - 78.3$ Mb), in addition to a fourth QTL at $55.9 - 62.6$ Mb on OAV6. Several earlier reports also indicate the presence of the QTL lined to EPG in various sheep breeds [57, 58]. Among 21 DEGs located on OAV6 identified in our study in CHB breeds, at least 5 genes are located within 15 Mb of this marker. The expression of four genes were significantly induced whereas one, albumin, was repressed by infection. Of note, mast cell stem cell factor (SCF) receptor KIT gene (chr6:70189728:70234612) is the closest to the SNP marker. Two major receptors, KIT and the high affinity receptor for IgE, are responsible for regulating various mast cell functions, including chemotaxis, proliferation, apoptosis, and cytokine releases [59]. The critical roles of mast cells in host immune responses to helminth infection have long been recognized [60]. Neutralization of KIT and its ligand, SCF, using monoclonal antibodies completely abrogates the mast cell hyperplasia generated by T. spiralis infection in mice, resulting in drastically delayed worm expulsion and a reduced mucosal eosinophilia [61]. This finding suggests that KIT plays an important in host-parasite interaction. In the past few years, increasing evidence suggests that the epidermal growth factor like molecule, amphiregulin (AREG), plays critical roles in regulating immunity and inflammation as well as in enhancing host resistance to helminth parasites [62, 63]. In rodent models, *T. suis* infection increases AREG expression, in parallel with the expression of Th2 cytokines IL4 and IL1333. Furthermore, worm clearance is significantly delayed at 14 dpi in AREG deficient mice, which correlates with reduced proliferation of colonic epithelial cells. Recent studies show that AREG is critical for efficient regulatory T cell function [64] and may play an important role in orchestrating immunity, inflammation, and tissue

repair [63]. In this study, AREG transcript abundance was significantly enhanced by infection only in the resistant breed, suggesting that this gene may play an important role in the development of host resistance. It would be intriguing to identify SNPs in both coding and promoter regions of the genes located within or closer to the QTL related to parasite resistance on OAV6, including AREG, and correlate the observed genetic variations with various resistant phenotypes. Moreover, dissecting mechanisms of transcriptional regulation of AREG and understanding how it promotes epithelial cell proliferation and regulates host immunity in the gastrointestinal tract warrant further investigation.

In conclusions, the two sheep breeds native to the Canaria Island displayed a distinct difference in several *Haemonchus contortus* resistant phenotypes under both natural and experimental infections. CHB tends to have significantly reduced worm burden, delayed egg production, and decreased fecal egg yield (counts) than the susceptible Canaria Sheep. A broad range of mechanisms have evolved in resistant CHB to provide protection against *H. contortus*. Readily inducible acute inflammation responses, complement activation, accelerated cell proliferation and subsequent tissue repair, and innate and acquired immunity directly against worm fecundity are likely to contribute to the development of host resistance to gastrointestinal nematode infection in the CHB breed.

3. A CONVOLUTIONAL AUTOENCODER BASED METHOD FOR PREDICTING RNA PROTEIN INTERACTIONS

3.1 Introduction

RNA binding proteins (RBPs) have been shown to play crucial roles in various biological processes, especially in the post-transcriptional regulation of RNAs. They are important regulators of RNA splicing, polyadenylation, and localization. Some RBPs have been studied previously, such as splicing regulators SRSF1 [65], NOVA1 [66], PTBP2 [67], RBFOX1 [68], RBFOX2 [69] and ESRP1/ESRP2 [70], polyadenylation regulator CPSF3 [71], and RNA localization regulator ZBP1 [72]. There are also studies showing that the mutation of these RBPs can result in diseases such as cancers, muscular atrophies, and neuropathies [73]. RBPs specifically recognize their binding targets by RNA-binding domains. Most of the RBPs contain several RNA-binding domains, e.g., different types of RNA-recognition motifs, double-stranded RNA-binding motifs or zinc-finger motifs [74]. These domains have high specificities in recognizing RNA sequences and structures to facilitate the binding and functioning of RBPs. Therefore, studies of the RBP binding specificities will be of great help in understanding the dynamic regulations of RBPs in various biological processes and diseases.

Since RBPs function through the specific binding to RNA sequences, such regulatory relationships can be identified by methods designed for assaying protein-RNA interactions, such as RNA immunoprecipitation (RIP) [75] and cross-linking immunoprecipitation (CLIP) [66]. With the development of high-throughput sequencing technologies, researchers have designed multiple sequencing protocols to study the RBP binding specificity on a transcriptome-wide scale, including RNA Immunoprecipitation with high-throughput sequencing (RIP-Seq) [10] and cross-linking immunoprecipitation cou-

pled with high-throughput sequencing (CLIP-Seq) [11]. Variants of CLIP-Seq with improved resolution, signal-to-noise ratio and sequence yield, such as PAR-CLIP [76], iCLIP [77], and eCLIP [78], are developed and widely used for the RBP research.

The wide application of HTS technologies in RNA-protein interaction identification has produced a large amount of data, and the reliable RBP binding sites identified by the HTS technologies have become invaluable resources for understanding the RBP binding specificities [76, 77, 79]. However, CLIP-Seq-based methods also have some shortcomings. One of the most severe ones is that the binding sites often have high false-negative rates [80]. As CLIP-Seq sequences the RNA molecules that are bound by the RBPs, it is highly sensitive to the gene expression level [81]. If a gene is not expressed in a sample, there is no way for CLIP-Seq based methods to identify the potential binding sites within that gene. Furthermore, the false negative rate it also inflated by the mapabilty difficulty of the sequences spanning the splicing junctions [82]. These together prevent the CLIP-Seq method from becoming an universal tool for identifying all possible RBP binding sites across the transcriptome.

To overcome the shortcomings of CLIP-Seq, many computational methods have been designed for predicting the RBP binding sites. These methods use RBP binding sites identified by CLIP-Seq experiments as the training and testing data to build models for predicting whether an unknown sequence in the transcriptome can be bound by RBPs. For instance, RNAcontext [83] builds a model that is trained by primary sequences and predicted secondary structural context features to predict the binding strength of RBPs. GraphProt [80] uses a graph-kernel strategy to encode the sequence and structural features of the binding preferences and trains a support vector machine (SVM) model to predict RBP binding sites. Some methods not only use sequence and structural information but also incorporate other data into the model to improve the prediction performances. For example, iONMF [84] predicts RNA-protein interaction sites by training a multi-modal

predictor using different types of data, such as kmer sequence, secondary structure, RBP co-binding, Gene Ontology, and binding region type.

Recently, deep learning [85, 86] has become a powerful tool for analyzing complex data. It has been used for various learning problems such as image classification, natural language processing, audio recognition and machine translation. Studies have demonstrated the superior performance of deep learning over the traditional machine learning methods [87]. To address the problem of predicting DNA- and RNA-protein interactions, researchers have designed multiple deep learning models and obtained promising results. For example, DeepBind has been developed to learn the binding patterns from million of reads produced by high-throughput technologies and predict the binding targets [88]. DeeperBind attempts to improve DeepBind by adding a recurrent layer after the convolutional layers [89]. DeepMotif combines convolutional layers and a highway network to learn and predict binding specificities, and uses input optimization to generate typical binding motifs [90]. [91] evaluated the performances of different convolutional neural network (CNN) based models and proposed a CNN architecture for predicting the protein binding specificities. [92] developed a deep learning model trained with the primary sequence, secondary structure and tertiary structure features of the RBP binding sites to predict the RNA binding specificities. iDeep uses a CNN and several deep belief networks to build a multimodal deep learning model using various data resources for RBP binding site predictions [93].

In this chapter, we propose a novel deep learning method for predicting the RBP binding sites and learning the RBP binding motifs. The method consists of a convolutional autoencoder that learns the high-level sequence features of the RBP binding sites, and a softmax classifier that uses the learned features to predict the RBP binding specificities for given input sequences. To evaluate the proposed method, we compare our model with existing methods and show that the performance of the proposed model surpasses the ex-

isting methods as evaluated using the AUC metric. In addition, we investigate the motifs captured by the convolutional autoencoder to prove the capability of the proposed method for finding the motifs in an unsupervised manner.

The rest of this chapter is organized as follows: Section 3.2 describes the proposed method. Section 3.3 evaluates the performance of the proposed method using the AUC metric and discusses its motif discovery ability. Section 3.4 concludes the chapter.

3.2 Materials and Methods

3.2.1 Dataset

The dataset we used in this study is from iONMF [84] (available at `https://github.com/mstrazar/ionmf`). It contains positive RBP binding sequences from CLIP-Seq data and negative RBP binding sequences sampled from genomic regions with no RBP binding in the CLIP-Seq experiments. This dataset is also used by iDeep [93] for benchmarking. The negative sequences uniformly sampled from the non-interactive genomic regions are more representative of the true sequence composition of the genome compared with the randomly shuffled sequences used in some studies. Therefore, the model trained on this dataset will likely provide better predictions of the RBP binding sites. The positive sequences are binding sites with the highest cDNA counts in the CLIP-Seq experiments. The dataset consists of 31 CLIP-Seq experiments for 19 RBPs, and we use all of the 31 experiments for benchmarking. For each experiment, 18,000 positive binding sequences and 18,000 negative binding sequences are used in the performance evaluation.

3.2.2 RNA Sequence Encoding

In this study, similar to other methods [88, 93, 91], RNA sequences are one-hot encoded and fed into the proposed model. The one-hot encoding of a given RNA sequence

$s = (s_1, s_2, \ldots, s_{l-1}, s_l)$ is a matrix S such that

$$S_{i,j} = \begin{cases} 1, & \text{if } (s_i, j) \in \{(A,1),(U,2),(C,3),(G,4)\} \\ 0, & \text{otherwise} \end{cases}, \tag{3.1}$$

where L is the length of the RNA sequence and $s_i \in \{A, U, C, G, N\}$, $1 \leq i \leq L$, is the ith base of the sequence.

Therefore, each sequence in the input data is encoded as a $L \times 4$ matrix, where length L is 101 in our experiments. The encoding process is shown in Figure 3.1.

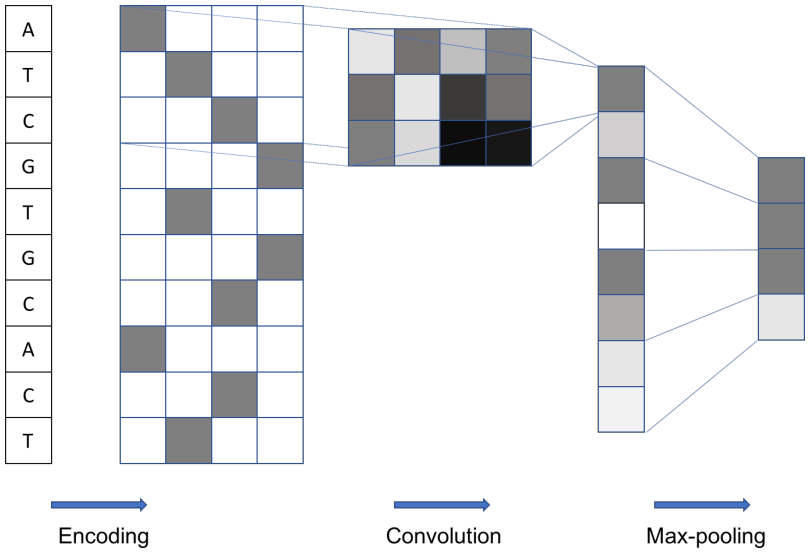


Figure 3.1: An example of convolutional network. The genomic sequence is one-hot encoded. Then the encoded sequence is convolved with a convolutional kernel of size 3. Finally, a max-pooling of size 2 is applied.

3.2.3 Convolutional Layers

Convolutional Neural Network (CNN) is a well-established deep learning architecture inspired by the organization of the animal visual cortex. In recent years, many deep CNN

architectures were designed, especially in the area of image classification, such as AlexNet [94], VGGNet [95], GoogleNet [96] and ResNet [97]. The evolution of the state-of-the-art CNN architectures reveals a trend that the networks are becoming deeper. For example, ResNet, the champion of ILSVRC 2015, is about 20 times deeper than AlexNet and eight times deeper than VGGNet. The increase in depth improves the ability of the network to approximate the target function and produce better feature representations, and thus, greatly enhances the performance of the network, compared with the shallow networks of the same number of parameters. Currently, deep CNN has been widely applied to image classification, video recognition, recommendation systems, natural language processing, text classification, scene labeling, and speech recognition.

Typically, a CNN model consists of several different types of layers, i.e., convolutional layers, activation function layers, and pooling layers. Different models will use different numbers of these layers and stack them in different orders. Figure 3.1 shows a simple CNN architecture with one convolutional layer, one activation layer, and one max pooling layer.

The key component of a CNN is the convolutional layer [94], which applies convolution operation to the input of the layer and passes the convolved input to the next layer. For RNA sequence data, each nucleotide is one-hot encoded as a vector of length four as described in Section 3.2.2. Then a sequence of length L is encoded as a matrix $S \in R^{L\times 4}$. And a convolutional kernel can be defined as $W \in R^{l\times 4}$, where l is the length of the convolutional kernel. Then the convolution operation of W over the one-hot encoded RNA sequence S is defined as

$$S_{out}(i) = \sum_{m=1}^{l} \sum_{j=1}^{4} S_{i-m,j} W_{m,j}. \tag{3.2}$$

The activation function layer applies activation function to its input. As the activation

functions are usually non-linear, they introduce non-linearities to CNNs to detect non-linear features. Assume $a(\cdot)$ denotes the activation function. The activation value A_i of convolution feature $S_{out}(i)$ is

$$A_i = a(S_{out}(i)). \tag{3.3}$$

Typically in CNN, Rectified Linear Unit (ReLU) [98] is used as the activation function:

$$Relu(x) = \max(0, x). \tag{3.4}$$

In a pooling layer, for example, max pooling, the input is split into non-overlapping subregions of a fixed length, and a max filter is applied to each of the subregions. Suppose the input to the pooling layer is A and the output is O. Then we have

$$O_k = \max\{A_{j,k}, \ldots, A_{j+m_p-1,k}\}, \tag{3.5}$$

where k is the index of the pooling output, j is the start position of the kth pooling, and the m_p is the pooling size.

3.2.4 Model Architecture

Autoencoder [99] is an unsupervised neural network which aims to learn the efficient representation of the input data. The simplest architecture of an autoencoder is a feed-forward neural network with the output layer having the same dimension as the input layer. It extracts hierarchical features and generates a compact representation of the input. Using the compact representation, the decoder can reconstruct the input reliably.

An autoencoder contains an encoder ϕ and a decoder ψ, in which the encoder transforms the input data $x \in \mathcal{X}$ into a compact representation $\phi(x) \in \mathcal{F}$ and the decoder reconstructs the input using the compact representation generated by the encoder $\psi(\phi(x)) \in \mathcal{X}$.

$$\begin{aligned} \phi &: \mathcal{X} \to \mathcal{F} \\ \psi &: \mathcal{F} \to \mathcal{X} \\ \phi, \psi &= \arg\min_{\phi,\psi} \|X - (\psi \circ \phi)X\|^2 \end{aligned} \tag{3.6}$$

Convolutional AutoEncoders (CAEs) [100] are autoencoders constructed using convolutional layers. Compared with autoencoders with fully connected layers which force the features to be global, it can extract localized features while capturing the original input structure.

In this study, we propose to use convolutional autoencoder to learn the important features of the RBP binding sites and use the compact representation which contains the most important features of the input sequences to predict the RBP binding sites. Figure 3.2 shows the general architecture of the convolutional autoencoder and the softmax classifier of the proposed model.

The input sequences are one-hot encoded as described in Section 3.2.2. Then three Convolution-ReLU-Max-Pooling layers are applied to the encoded sequences to extract the local sequence composition features as well as the high-level features. Then two fully connected layers are used to transform the features into the compact representations of the input sequences. The decoder reconstructs the input sequence by using fully connected layers and Convolution-ReLU-Max-Pooling layers. The classification part of the model is a two-layer feedforward neural network with softmax output. It takes the learned compact representations of the input sequences to predict if the sequences are binding sites of a particular RBP.

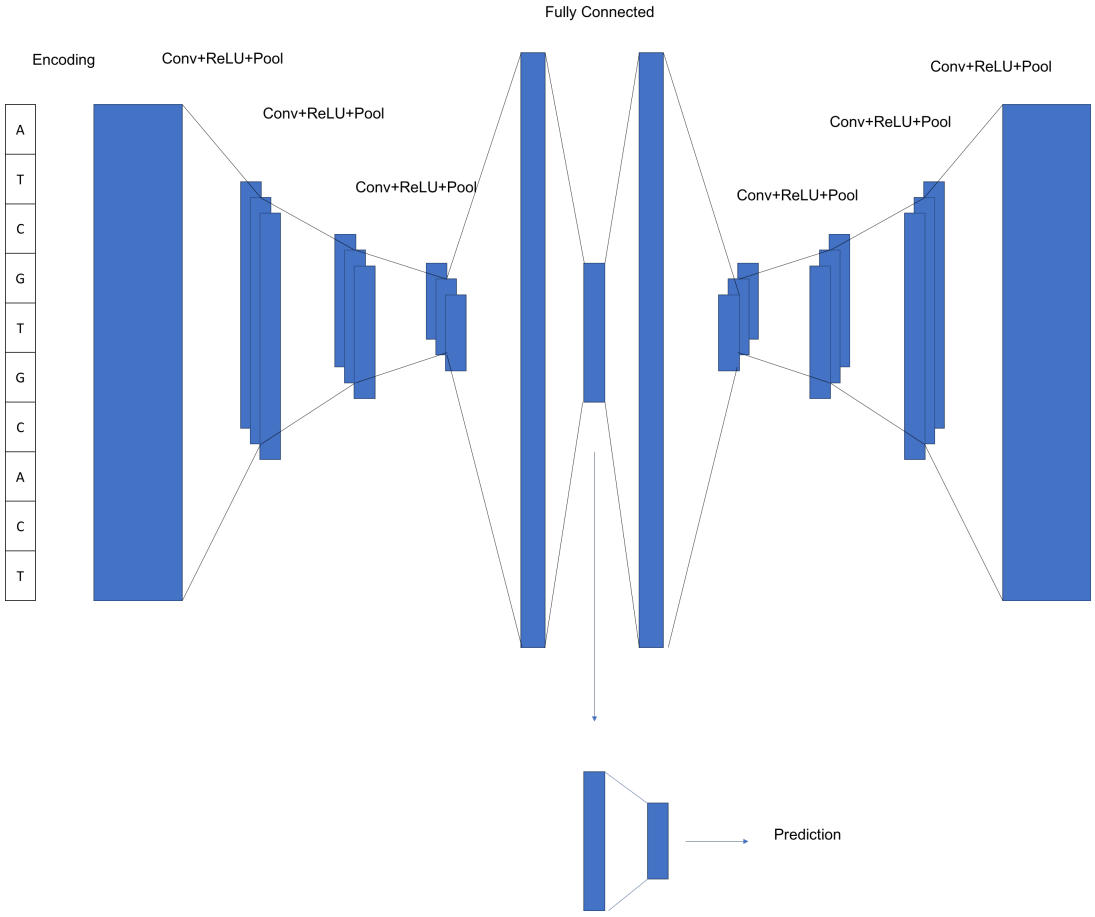


Figure 3.2: Architecture of the proposed model. Input is one-hot encoded and fed into the autoencoder. The encoder consists of three Convolution-Relu-MaxPooling layers and two fully connected layers. The decoder takes the output of the encoder and reconstructs the input by one fully connected layer and three Convolution-Relu-MaxPooling layers. The softmax classifier takes the output the learned encoder to predict the RBP binding sites.

3.2.5 Model Implementation

The proposed model is implemented using Keras with TensorFlow backend in Python. In the autoencoder, the size of the convolutional kernels is set to 10 and the size of the max-pooling is set to 2. The maximum number of epochs is set to 100, and the batch size is set to 100. To control the over-fitting of the model, we use early stopping, which monitors the validation loss and stops the training when the validation loss converges. The early stopping tolerance is set to 5 epochs. Other techniques for controlling over-fitting are also applied, such as batch normalization [101] and dropout [102]. The autoencoder is trained by back-propagation to minimize the reconstruction loss, i.e., the mean squared error for all bases, using RMSprop optimizer. The softmax classifier is trained by back-propagation to minimized the binary cross-entropy loss for all sequences using RMSprop optimizer. The implementation of the model is available at `http://github.com/zhengyuguo/deepclip`.

3.3 Results

In this section, the performance of the proposed method for predicting RBP binding sites using CLIP-Seq data is evaluated and compared with the performances of the state-of-the-art methods [88, 93]. In addition, we discuss the ability of the proposed method to discover the binding motifs of the RBPs.

3.3.1 Model Evaluation

We evaluate the performance of the proposed method and compare it with two baseline models. One baseline model is iDeep [93], which is a multi-modal deep learning method using both RBP binding sequence information and other information such as region type, clip-cobinding, structure, and motif. It has been shown to have better performance compared with methods such as GraphProt, iOMNF, etc. To facilitate a fair comparison, we

compare the proposed method with the modal for primary sequence in iDeep, which is a CNN with binding sequence input. Another is DeepBind [88], and it has been shown by [88] that the model outperforms the traditional methods and some other CNN architectures. We reimplement the model using the exact parameters provided by the author and compare our proposed model with it.

We use five-fold cross-validation to accurately assess the performance of the models. The data of each experiment is evenly divided into five subsets. For each fold, four subsets are used as training set and one subset is held out for testing purposes. The performance for each fold is measured using the area under the receiver operating characteristic curve (AUC). For each experiment, we take the average AUC over all folds.

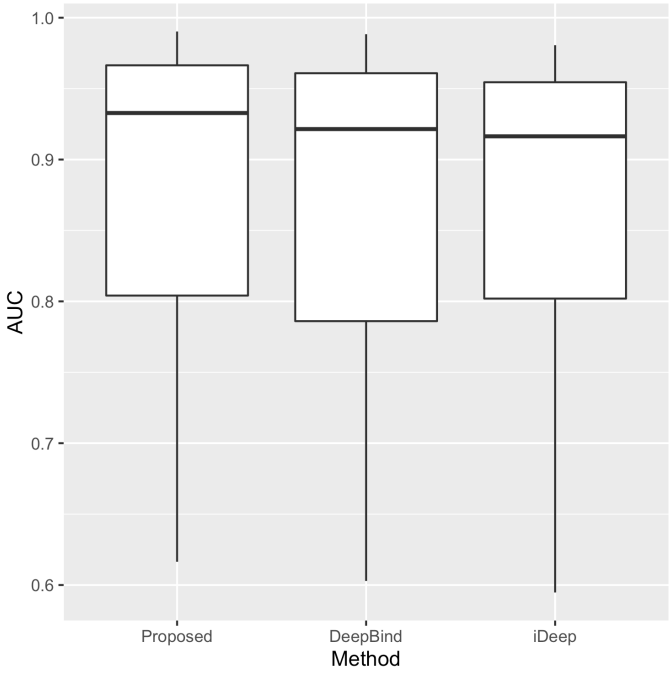


Figure 3.3: The AUC distribution of all experiments for the three methods.

Figure 3.3 shows the distributions of AUCs of the three methods for all 31 experiments.

In general, our proposed method has better performance than both DeepBind and iDeep. The proposed method achieves higher median, first quantile, and third quantile AUCs compared with the other two methods. DeepBind slightly outperforms iDeep under the experiment setup.

Table 3.1 lists the AUCs of the three methods for all experiments. For each method and each experiment, the AUC value is the average AUC of all five folds of the cross-validation. For each experiment, the highest AUC value is in bold. Among the 31 experiments, our proposed method outperforms DeepBind over 20 experiments (our proposed method and DeepBind have the same AUC over one of the experiments). Among all the experiments, the proposed method outperforms DeepBind for at least 0.01 in terms of AUC in 17 experiments, while DeepBind outperforms the proposed method for at least 0.01 in terms of AUC in only two experiments. For some experiments, such as 18 HNRNPL, 19 HNRNPL, 20 HNRNPL, 22 NSUN2, and 30 U2AF2, the proposed method outperforms DeepBind by at least 0.05 in terms of AUC. Therefore, the proposed method produces better predictions than DeepBind in general. We can draw the same conclusion from Figure 3.4a. Similarly, the proposed method has superior performance compared with iDeep. Among the 31 experiments, the proposed method gets higher AUC values in 28 of them (Figure 3.4b).

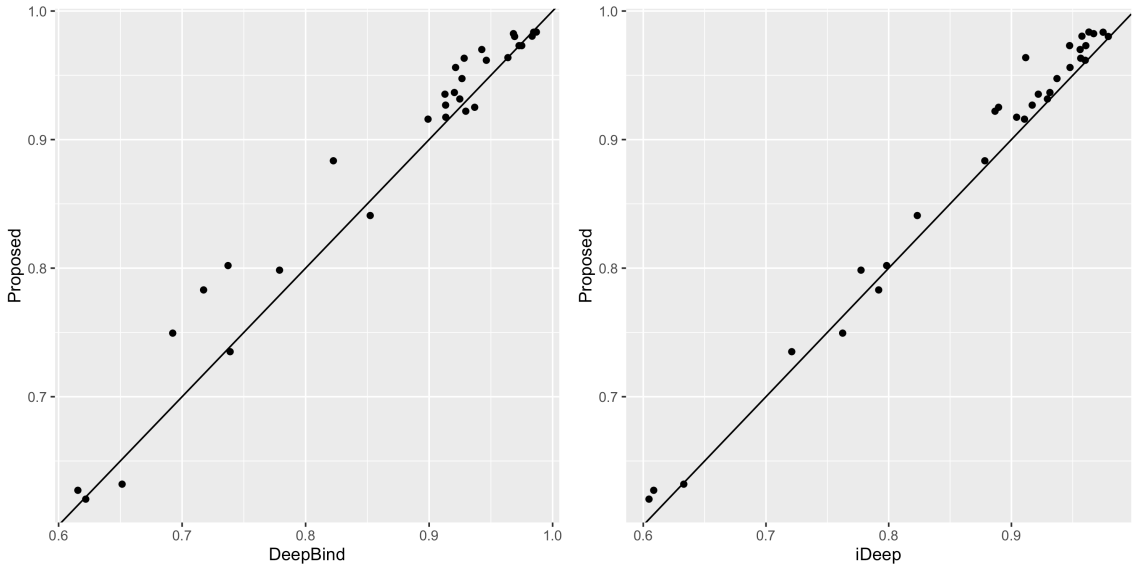
3.3.2 Insight in Motif Discovery

The convolutional kernels can be seen as feature extractors of the input. Especially for the kernels of the first convolutional layer, they scan the input sequences and activate when they detect some specific types of sequence patterns. By training the model, the convolutional kernels will learn the enriched sequence patterns within the positive binding sites compared with the background sequences with no binding signals. Therefore, the learned convolutional kernels act as the motif detectors, which provide biological insights of the

Table 3.1: Mean AUC of the three methods for all experiments.

ID	RBP	Proposed Method	DeepBind	iDeep
1	AGO/EIF	**0.7984**	0.7789	0.7774
2	AGO2	0.6203	**0.6219**	0.6046
3	AGO2	0.9251	**0.9368**	0.8895
4	AGO2	0.9220	**0.9296**	0.8866
5	AGO2	0.6319	**0.6514**	0.6330
6	EIF4A3	0.9804	**0.9834**	0.9574
7	EIF4A3	0.9836	**0.9868**	0.9631
8	ELAVL1	**0.9268**	0.9133	0.9169
9	ELAVL1	**0.6272**	0.6155	0.6085
10	ELAVL1A	0.9637	0.9637	0.9116
11	ELAVL1	0.9730	**0.9749**	0.9474
12	EWSR1	**0.9353**	0.9128	0.9219
13	FUS	**0.9560**	0.9214	0.9477
14	FUS	**0.9700**	0.9426	0.9559
15	IGF2BP1	0.7350	**0.7388**	0.7209
16	HNRNPC	**0.9616**	0.9463	0.9601
17	HNRNPC	**0.9802**	0.9692	0.9789
18	HNRNPL	**0.8020**	0.7371	0.7983
19	HNRNPL	0.7830	0.7173	**0.7918**
20	HNRNPL	0.7494	0.6923	**0.7624**
21	MOV10	0.8409	**0.8523**	0.8233
22	NSUN2	**0.8835**	0.8224	0.8783
23	PUM2	**0.9730**	0.9728	0.9606
24	QKI	0.9835	**0.9846**	0.9747
25	SFRS1	**0.9173**	0.9134	0.9042
26	TAF15	**0.9824**	0.9682	0.9670
27	TDP43	**0.9315**	0.9248	0.9293
28	TIA1	**0.9366**	0.9204	0.9313
29	TIAL1	**0.9158**	0.8991	0.9106
30	U2AF2	**0.9632**	0.9283	0.9564
31	U2AF2	**0.9474**	0.9266	0.9371

RBP binding preferences. The convolutional kernels of the higher layers will learn higher level features of the input sequences. However, because of the difficulty in interpreting those kernels in a biologically meaningful way, we will only focus on the kernels of the



(a) Proposed Method v.s. DeepBind (b) Proposed Method v.s. iDeep

Figure 3.4: AUC comparison between the proposed method and the baseline methods.

first convolutional layer.

To investigate what motifs are learned from the data by the proposed model, we construct the position weight matrices (PWM) for all the learned convolution kernels of the first convolutional layer. The PWMs are aligned against a known human RBP motif database cisBP-RNA [103] using Tomtom [104] from MEME Suite [105]. The database contains 102 known motifs.

Figure 3.5 shows four RBPs for which the proposed method successfully learned their known motifs in cisBP-RNA from the data. For HNRNPC, TIA1, and U2AF2, the output of the proposed method resembles the U-rich consensus motifs that closely match the known motifs in cisBP-RNA. For QKI, the output of the proposed method resembles the UAAY motif that matches a known QKI motif in cisBP-RNA. Figure 3.6 shows four RBPs for which the proposed method recovers motifs discovered by some other CLIP-Seq based experiments. For example, TDP43 has been shown to bind to GU-rich sequnces [106],

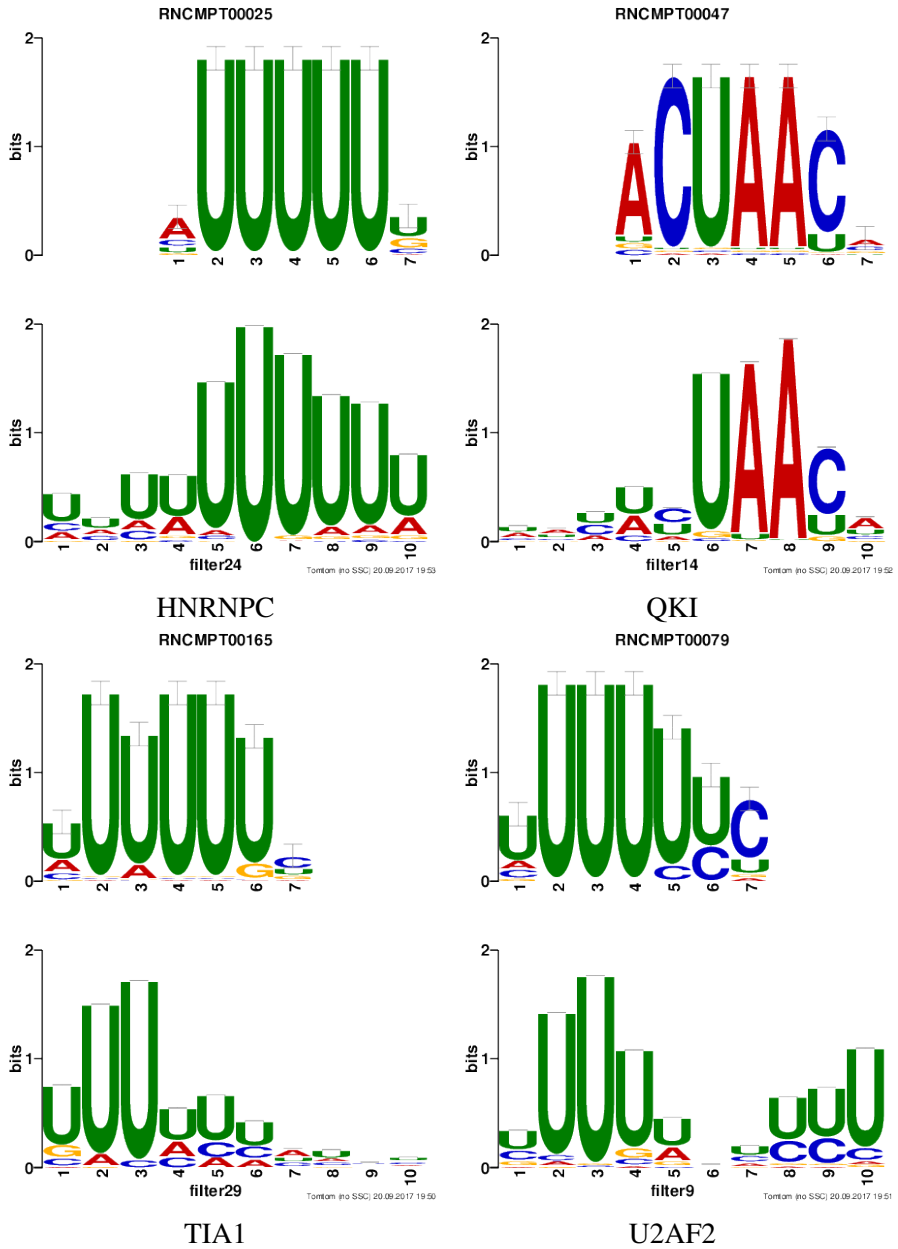


Figure 3.5: The proposed method recovers the known motifs in cisBP-RNA. Four example RBPs showing that the proposed method can successfully recover the known motifs in cisBP-RNA.

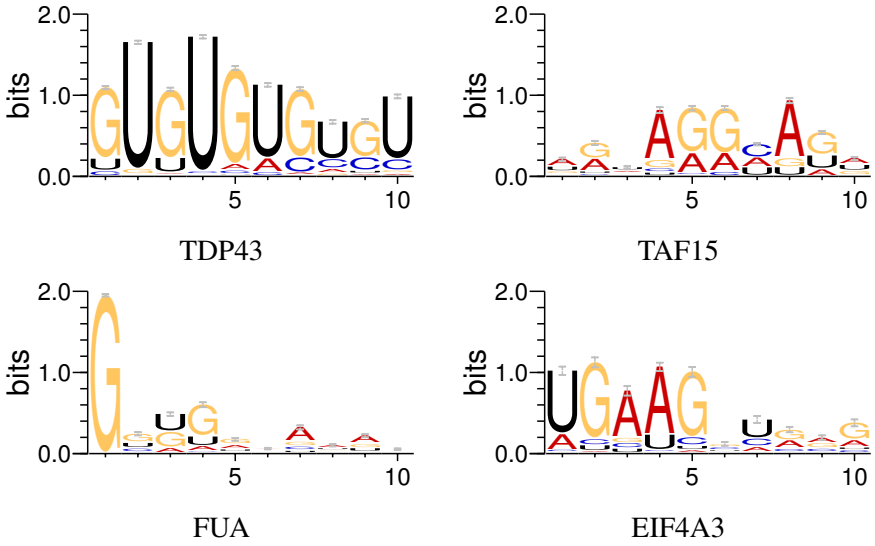


Figure 3.6: The proposed method recovers the known motifs in the literature. Four example RBPs showing that the proposed method can successfully recovers the known motifs in literature.

and the proposed method recovers the GU-rich binding motif of TDP43. The proposed method also recovers the (A/G)GGUA motif for TAF15 [107], the GGUG motif for FUS [107], and the GAAGA motif for EIF4A3 [108]. Besides motifs that match the currently known ones, there are many new motifs that are identified by the proposed method. As the RBP motif database is not complete and there are limitations in understanding the binding mechanisms of the RBPs, these new motifs are likely to provide important information to guide future directions.

3.4 Conclusion

The rapid development of HTS technologies has enabled the transcriptome-wide study of the RBP binding preferences using CLIP-Seq and its variants. These experiments have generated a huge amount of data from which a large amount of the reliable RBP binding sites have been identified. Due to the intrinsic property of CLIP-Seq, the false negative problem greatly limited the application of CLIP-Seq experiments in finding all possible

RBP binding sites. Therefore, computational methods are increasingly important in understanding the RBP mechanism.

In this chapter, we propose an accurate deep learning method to predict the RBP binding preferences using CLIP-Seq data. The proposed method uses a convolutional autoencoder to learn the robust features from the data, and uses a softmax classifier for the prediction. Our proposed model is able to learn the robust representations of the input data in an unsupervised manner, and use the robust representations for accurate classification of the RBP binding sites. We evaluate the performance of the proposed method over a large CLIP-Seq dataset containing 31 experiments from [84]. Performance test over the benchmarking dataset shows that the proposed method outperforms the state-of-the-art methods in most of the experiments, and reaches comparable performance in the remaining ones. In addition, the convolutional nature of the proposed method enables the interpretation and visualization of the learned sequence motif features. The proposed method successfully recovers the known RBP motifs, while providing more candidate motifs that may be important for future understanding of the RBP mechanisms.

4. CREATION OF A DATABASE AND WEB-SERVER FOR METADATA OF PUBLICLY AVAILABLE MOUSE RNA-SEQ DATA SETS*

4.1 Introduction

Gene targeting [109], a powerful technique used to manipulate a specific locus in the genome of an organism, is an indispensable tool for assessing in vivo functions of specific gene products. This technique provides great flexibility in manipulating the genome of an organism since it can be used to delete a gene or an exon, to introduce an exogenous gene, or to create point mutations. Moreover, gene targeting not only can introduce permanent mutations but can also conditionally change targeted genes. Thousands of genetically engineered mice have been generated using this technique. They provide valuable models for studying mechanisms of human diseases.

RNA-Seq [110], a high-throughput sequencing (HTS) method for transcriptome analysis, has been successfully used on many of these mouse models, enabling global analyses of specific genomic alterations at a high sequencing depth with a reasonable accuracy. As RNA-Seq becomes increasingly popular, hundreds of RNA-Seq data sets have been generated and have been released to the public. These data are currently available from online repositories, such as, Gene Expression Omnibus (GEO) [111], ArrayExpress [112], Sequence Read Archive (SRA) [113], and European Nucleotide Archive (ENA) [114], whose primary purposes are to store raw and processed HTS data from a wide variety of organisms. However, a data submitter typically provides only limited metadata for each data set sufficient to get the data set accepted into the public repository. There is currently no stringent and uniform quality check of submitted metadata. This results in inconsis-

*Part of this chapter is reprinted with permission from "RNASEQMETADB: a database and web server for navigating metadata of publicly available mouse rna-seq datasets," by Z. Guo, B. Tzvetkova, J. M. Bassik, T. Bodziak, B. M. Wojnar, W. Qiao, M. A. Obaida, S. B. Nelson, B. H. Hu, and P. Yu, *Bioinformatics*, vol. 31, no. 24, pp. 40384040, 2015, Copyright[2015] by Oxford University Press

tency and ambiguity in data set annotation. For example, nonofficial gene symbols are used in some of the data sets. Other public databases such as InSilico DB [115] also suffer from the same problems, making searching for data sets in these databases inefficient.

The recent HTS data explosion has motivated researchers to create several metadata databases. However, these databases, e.g., CistromeMap [116], focus primarily on ChIP-Seq data. To fill the gap for RNA-Seq data, we collected RNA-seq metadata from all the publicly available data sets that were generated using mouse models mostly with targeted mutations and curated a database called RNASeqMetaDB. Haynes et al. recently suggested that measuring transcription factor binding might not be the best way to decipher transcriptional regulatory networks [117]. Instead, their work showed that gene expression data could be of greater value in revealing functional gene regulatory relations. Therefore, RNASeqMetaDB may be a helpful resource for researchers trying to build gene regulatory networks.

We developed a web server to provide a user-friendly query interface for locating relevant RNA-Seq data sets based on targeted gene names, disease names, tissue types, keywords, publications, and accession IDs, etc. An ontological search function is also offered that allows users to find the data sets related to, but not necessarily annotated to, the exact search term. This helps ensure search sensitivity (see the help page on the website for examples). This database can help biomedical scientists navigate the complex landscape of mouse genetic experiments and can provide rich contexts for these data sets. Using this database, users will be able to find related data sets for further analyses easily. For example, RNA-Seq data can be used to infer splicing isoform functions [118] and information extracted from existing RNA-Seq data can be used as prior knowledge for causal reasoning on biological networks [119]. Moreover, RNA-Seq data can be integrated with sequence- and structure-binding preferences of RNA-binding proteins (RBPs) learned with computational methods such as GraphProt [80], which can increase our understanding of the

mechanisms of posttranscriptional regulation.

4.2 Methods

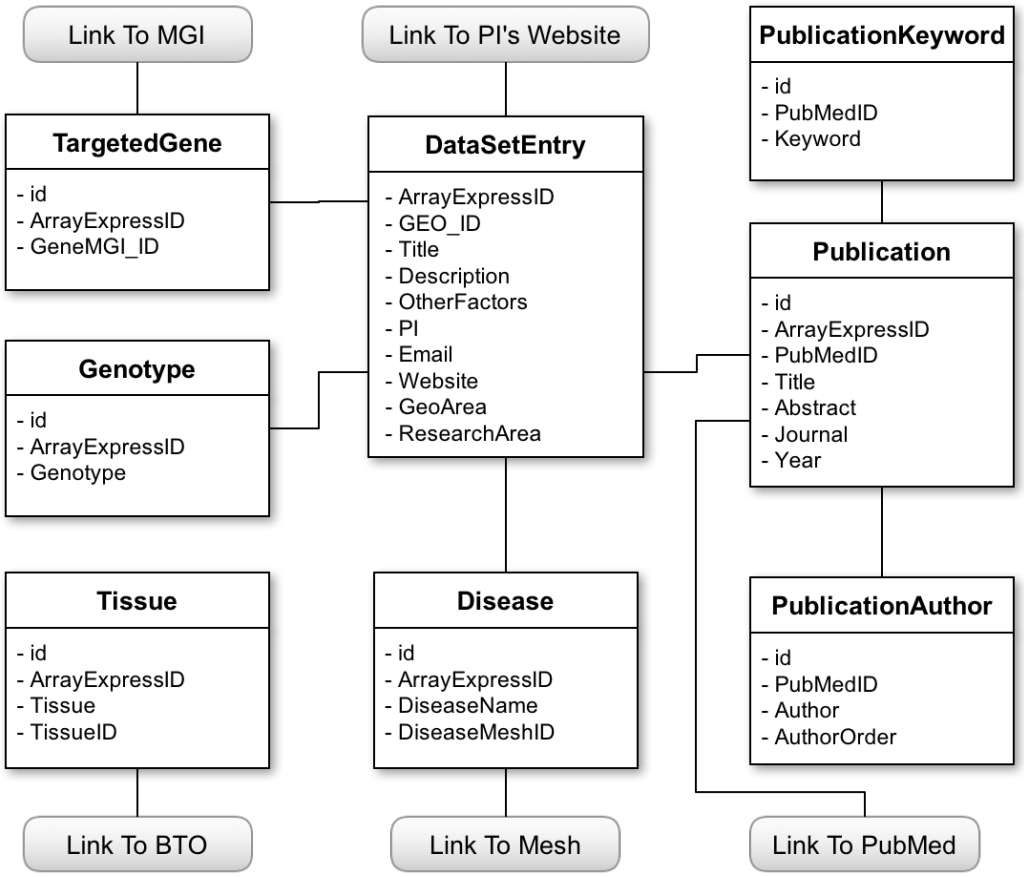


Figure 4.1: The database schema of RNASeqMetaDB. For each RNA-Seq data set, the metadata of gene symbol, genotype, reference (including title, authors, abstract, PubMed ID), disease, tissue type, corresponding author and authors website link are systematically organized in a relational database.

We collected raw annotations of publicly available mouse RNA-Seq data sets from high-throughput sequencing data repositories including GEO, ArrayExpress, and ENA. At the time of writing, RNASeqMetaDB contains 306 experiments in total. The following metadata were systematically annotated for each RNA-Seq data set: gene symbol,

genotype, reference (including title, authors, abstract, PubMed ID), disease, tissue type, corresponding author and author's website link (Figure 4.1). Genotype and disease annotations were manually curated and extracted from the original publications. For consistency, genes, alleles, diseases and tissue types were annotated using official symbols or controlled vocabularies from online resources including Mouse Genome Informatics (MGI) [120], Medical Subject Headings (MeSH) [121] and BRENDA Tissue Ontology (BTO) [122].

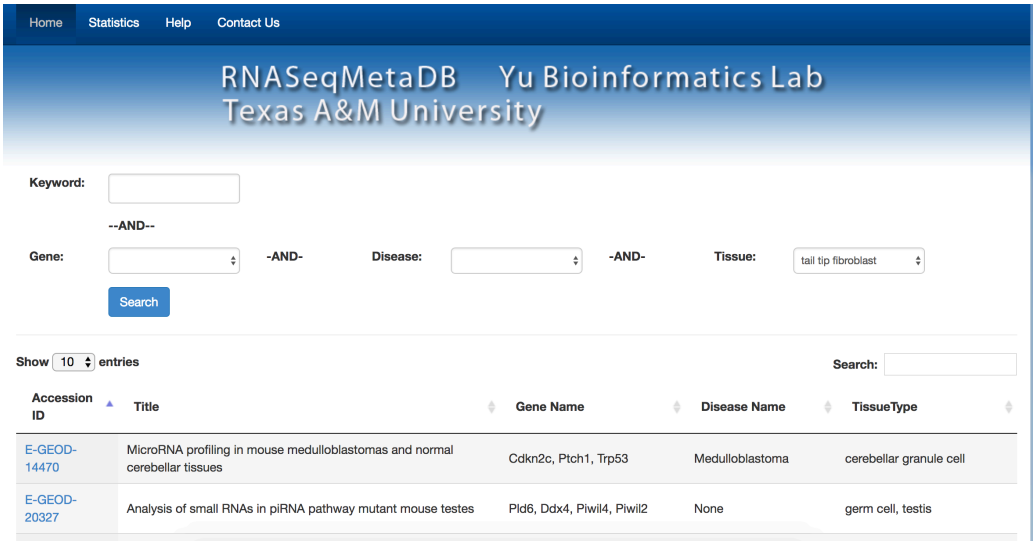


Figure 4.2: Basic searching functionality of RNASeqMetaDB. The database supports searching by gene symbols, disease names, tissue types, or keywords.

To facilitate querying and data retrieval, we implemented a web server (`http://rnaseqmetadb.ece.tamu.edu`). All search functionalities are integrated within a single web interface (Figure 4.2). Users can search for one or multiple data sets using gene symbols, disease names, tissue types, or keywords. The search generates a list of matched data sets containing accession IDs, titles, mutated genes, related diseases and tissue types. To support more general database queries, ambiguous keyword search is

provided by the server. Users can type one or multiple terms that they are interested in into the Keyword search box. Both the typed words and their synonyms defined by the Experimental Factor Ontology (EFO) [123] will be searched in the database, and then the query results will be displayed. If a more targeted search is needed, users are allowed to use additional terms in the Search text box above the result table to refine the results. This retrieves only the data sets whose titles match these additional terms. When the accession ID of a data set is clicked, RNASeqMetaDB displays a summary table of all the metadata available for that data set. The links to other databases and websites like ArrayExpress, GEO, MGI, PubMed and MeSH, and the PIs' lab websites are provided so that users can easily obtain additional information related to the data sets. Users can also bulk download data after registering an account on the website. To keep the data updated, the website allows users to submit requests for adding data entries for newly published RNA-seq data sets.

4.3 Results

RNASeqMetaDB permits efficient searching of its database containing comprehensive information for all public RNA-Seq data sets on mice with genotype as a factor. It contains metadata for a total of 306 experiments targeting 298 different genes. These experiments are from 264 different research groups, among which 154 are from the United States, and 76 are from Europe (Figure 4.3A). For journals publishing the studies using these data sets, Nature ranks at the top with the greatest number of studies, followed by Proceedings of the National Academy of Sciences of the United States of America and Genes & Development (Figure 4.3B). One interesting observation is that the number of publications on RNA-Seq studies has been increasing exponentially since 2008 (Figure 4.3C). This indicates that RNA-Seq experiments on gene-targeted mouse models have become more popular in recent years. Summary statistics of the datasets is also available at the database website

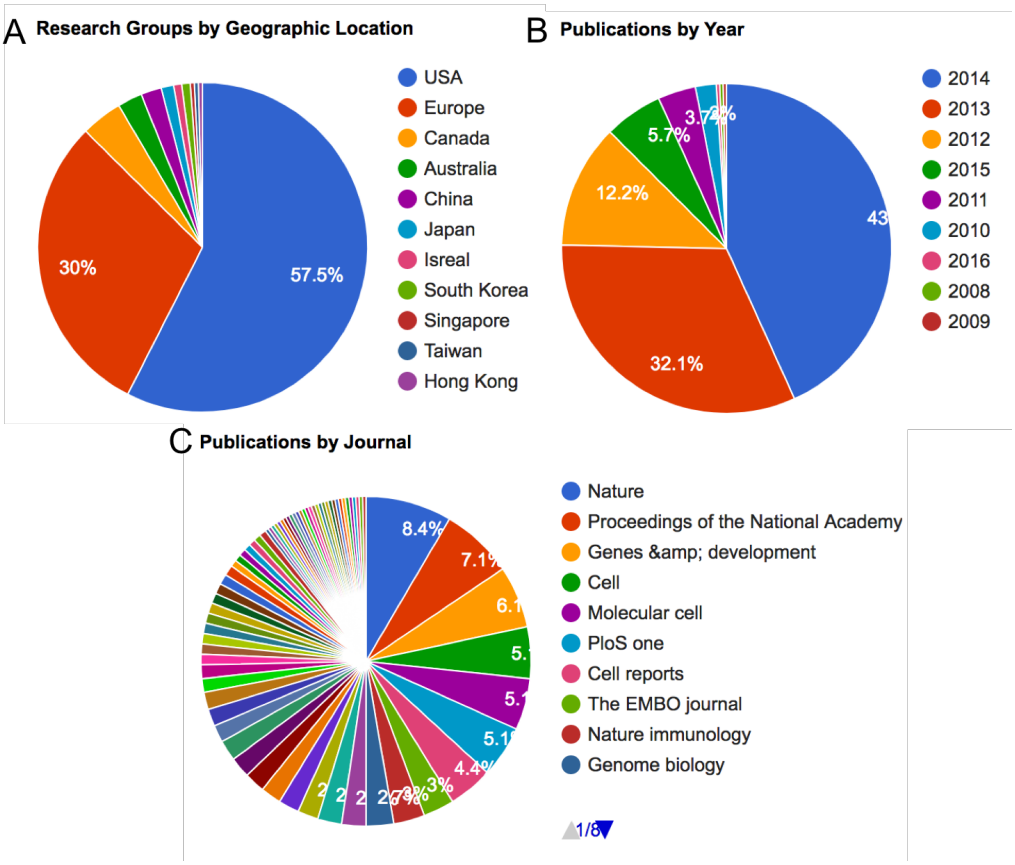


Figure 4.3: Some statistics of RNASeqMetaDB. A: Geographic locations of the research groups generating mouse RNA-Seq datasets. B: The publication years of the papers based on the mouse RNA-Seq datasets. C: The journals publishing the papers based on mouse RNA-Seq datasets.

4.4 Discussion

RNASeqMetaDB broadens the use of these data sets by providing well-curated metadata and efficient query functionalities. Scientists can easily find the data sets that they are interested in and retrieve detailed information to enable more comprehensive understanding of the experiments. In the future, we will develop additional functionality and import more data sets into the database. We believe that RNASeqMetaDB will be a valuable tool for the biomedical research community.

5. CONCLUSIONS

With the development of technologies, high-throughput sequencing has become one of the most powerful tools for studies in genomics, transcriptomics, epigenomics, and metagenomics. Compared with the first generation sequencing technology called Sanger sequencing, HTS technologies are much cheaper and faster: they have the ability to sequence a massive number of genomic reads in parallel and produce millions of sequences concurrently with low costs and in a short time. In recent years, various HTS technologies for different purposes have been designed, such as RNA-Seq, Ribo-Seq, ChIP-Seq, CLIP-Seq, and RIP-Seq. With the wide application of these technologies in various research areas, thousands of HTS datasets have been generated and publicly deposited in databases such as SRA, GEO, and ENA. These datasets provide invaluable resources for more comprehensive understanding of the genomic landscapes.

HTS technologies have been designed for enhancing the understanding of the diverse cellular roles of RNA. RNA-Seq is able to sequence various RNA species, such as mRNA, microRNA, snoRNA, etc. It is a powerful method for precisely and comprehensively measure the gene expression level. Utilizing UV cross-linking with immunoprecipitation to extract protein binding regions or RNA modification regions from RNA sequences, CLIP-Seq is widely used for studying protein-RNA interactions and RNA modifications. These HTS technologies have enabled researchers to understand the gene expression, transcriptional regulation and post-transcriptional regulation in a comprehensive manner.

In this dissertation, we present HTS applications for transcriptional studies. First, the differential expression analysis of RNA-Seq data is discussed and applied on a sheep RNA-Seq dataset to study the biological mechanisms of the sheep resistance to worm infection. The RNA-Seq dataset is analyzed with an automatic pipeline and the gene

expression is modeled by a negative binomial distribution. The study provides insights into the underlying biology of sheep host resistance. Then, the deep learning method is used to model the RBP binding sites and predict the RBP binding preferences using CLIP-Seq data. In the proposed method, deep convolutional autoencoder is used for robust feature extraction, and a softmax classifier is used for binding site prediction. This method can be used for studying the RBP regulations, and it can also provide insights for identifying the RBP binding motifs. Finally, a database is created to facilitate the reuse of the publicly available RNA-Seq dataset. The metadata of the publicly available RNA-Seq datasets are manually curated and are served by a well-designed website. Furthermore, the database can be scaled up in the future to serve more types of HTS data, such as CLIP-Seq data and ChIP-Seq data.

In summary, this dissertation describes a set of HTS applications for gene expression and post-transcriptional regulation. We believe this dissertation has reached the goal of improving the HTS applications for transcriptional studies and provided some knowledge to the research community. Along the lines of this dissertation, there are additional research problems such as modeling differential alternative splicing and validating the predicted RBP targets, and it is our hope that these problems will be addressed in the future.

REFERENCES

[1] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. C. Raff, K. Roberts, P. Walter, J. Wilson, and T. Hunt, *Molecular biology of the cell. Sixth edition. Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, Peter Walter; with problems by John Wilson, Tim Hunt.* New York: Garland Science, [2015], 2015.

[2] Wikipedia, "Central dogma of molecular biology — wikipedia, the free encyclopedia," 2017. [Online; accessed 21-September-2017].

[3] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, no. 2, pp. 105–111, 2009.

[4] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors," *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[5] A. M. Maxam and W. Gilbert, "A new method for sequencing dna," *Proceedings of the National Academy of Sciences*, vol. 74, no. 2, pp. 560–564, 1977.

[6] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al.*, "Initial sequencing and analysis of the human genome," 2001.

[7] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-throughput sequencing technologies," *Molecular cell*, vol. 58, no. 4, pp. 586–597, 2015.

[8] L. Nederbragt, "developments in NGS," 7 2016.

[9] M. L. Metzker, "Sequencing technologies–the next generation," *Nature reviews. Genetics*, vol. 11, no. 1, p. 31, 2010.

[10] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond, "Stem cell transcriptome profiling via massive-scale mrna sequencing," *Nat Methods*, vol. 5, no. 7, pp. 613–9, 2008.

[11] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell, "Hits-clip yields genome-wide insights into brain alternative rna processing," *Nature*, vol. 456, no. 7221, pp. 464–9, 2008.

[12] B. Besier, J. Lyon, D. Michael, G. Newlands, and D. Smith, "Towards a commercial vaccine against haemonchus contortus-a field trial in western australia," in *Proc. Australian Sheep Vet Conf*, vol. 2012, pp. 14–18, 2012.

[13] T. Sreter, T. Kassai, and E. Takacs, "The heritability and specificity of responsiveness to infection with haemonchus contortus in sheep," *Int J Parasitol*, vol. 24, no. 6, pp. 871–6, 1994.

[14] F. Alba-Hurtado and M. A. Munoz-Guzman, "Immune responses associated with resistance to haemonchosis in sheep," *Biomed Res Int*, vol. 2013, p. 162158, 2013.

[15] H. R. Gamble and A. M. Zajac, "Resistance of st. croix lambs to haemonchus contortus in experimentally and naturally acquired infections," *Vet Parasitol*, vol. 41, no. 3-4, pp. 211–25, 1992.

[16] A. F. Amarante, P. A. Bricarello, R. A. Rocha, and S. M. Gennari, "Resistance of santa ines, suffolk and ile de france sheep to naturally acquired gastrointestinal nematode infections," *Vet Parasitol*, vol. 120, no. 1-2, pp. 91–106, 2004.

[17] M. Gauly, M. Kraus, L. Vervelde, M. A. van Leeuwen, and G. Erhardt, "Estimating genetic differences in natural resistance in rhon and merinoland sheep following experimental haemonchus contortus infection," *Vet Parasitol*, vol. 106, no. 1, pp. 55–67, 2002.

[18] J. M. Mugambi, R. K. Bain, S. W. Wanyangu, M. A. Ihiga, J. L. Duncan, M. Murray, and M. J. Stear, "Resistance of four sheep breeds to natural and subsequent artificial haemonchus contortus infection," *Vet Parasitol*, vol. 69, no. 3-4, pp. 265–73, 1997.

[19] M. J. Stear, B. Boag, I. Cattadori, and L. Murphy, "Genetic variation in resistance to mixed, predominantly teladorsagia circumcincta nematode infections of sheep: from heritabilities to gene identification," *Parasite Immunol*, vol. 31, no. 5, pp. 274–82, 2009.

[20] M. V. Benavides, T. S. Sonstegard, S. Kemp, J. M. Mugambi, J. P. Gibson, R. L. Baker, O. Hanotte, K. Marshall, and C. Van Tassell, "Identification of novel loci associated with gastrointestinal parasite resistance in a red maasai x dorper backcross population," *PLoS One*, vol. 10, no. 4, p. e0122797, 2015.

[21] M. Gauly and G. Erhardt, "Genetic resistance to gastrointestinal nematode parasites in rhon sheep following natural infection," *Vet Parasitol*, vol. 102, no. 3, pp. 253–9, 2001.

[22] K. Periasamy, R. Pichler, M. Poli, S. Cristel, B. Cetra, D. Medus, M. Basar, K. T. A, S. Ramasamy, M. B. Ellahi, F. Mohammed, A. Teneva, M. Shamsuddin, M. G. Podesta, and A. Diallo, "Candidate gene approach for parasite resistance in sheep–variation in immune pathway genes and association with fecal egg count," *PLoS One*, vol. 9, no. 2, p. e88337, 2014.

[23] M. V. Silva, T. S. Sonstegard, O. Hanotte, J. M. Mugambi, J. F. Garcia, S. Nagda, J. P. Gibson, F. A. Iraqi, A. E. McClintock, S. J. Kemp, P. J. Boettcher, M. Malek,

C. P. Van Tassell, and R. L. Baker, "Identification of quantitative trait loci affecting resistance to gastrointestinal parasites in a double backcross population of red maasai and dorper sheep," *Anim Genet*, vol. 43, no. 1, pp. 63–71, 2012.

[24] K. Marshall, J. M. Mugambi, S. Nagda, T. S. Sonstegard, C. P. Van Tassell, R. L. Baker, and J. P. Gibson, "Quantitative trait loci for resistance to haemonchus contortus artificial challenge in red maasai and dorper sheep of east africa," *Anim Genet*, vol. 44, no. 3, pp. 285–95, 2013.

[25] J. F. Gonzalez, A. Hernandez, J. M. Molina, A. Fernandez, H. W. Raadsma, E. N. Meeusen, and D. Piedrafita, "Comparative experimental haemonchus contortus infection of two sheep breeds native to the canary islands," *Vet Parasitol*, vol. 153, no. 3-4, pp. 374–8, 2008.

[26] J. F. Gonzalez, A. Hernandez, E. N. Meeusen, F. Rodriguez, J. M. Molina, J. R. Jaber, H. W. Raadsma, and D. Piedrafita, "Fecundity in adult haemonchus contortus parasites is correlated with abomasal tissue eosinophils and gammadelta t cells in resistant canaria hair breed sheep," *Vet Parasitol*, vol. 178, no. 3-4, pp. 286–92, 2011.

[27] R. W. Li and S. G. Schroeder, "Cytoskeleton remodeling and alterations in smooth muscle contractility in the bovine jejunum during nematode infection," *Funct Integr Genomics*, vol. 12, no. 1, pp. 35–44, 2012.

[28] R. W. Li and L. C. Gasbarre, "A temporal shift in regulatory networks and pathways in the bovine small intestine during cooperia oncophora infection," *Int J Parasitol*, vol. 39, no. 7, pp. 813–24, 2009.

[29] R. W. Li, M. Rinaldi, and A. V. Capuco, "Characterization of the abomasal transcriptome for mechanisms of resistance to gastrointestinal nematodes in cattle," *Vet Res*, vol. 42, p. 114, 2011.

[30] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[31] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol*, vol. 11, no. 10, p. R106, 2010.

[32] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–40, 2010.

[33] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res*, vol. 18, no. 9, pp. 1509–17, 2008.

[34] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "Degseq: an r package for identifying differentially expressed genes from rna-seq data," *Bioinformatics*, vol. 26, no. 1, pp. 136–8, 2010.

[35] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by rna sequencing," *Science*, vol. 320, no. 5881, pp. 1344–9, 2008.

[36] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments," *BMC Bioinformatics*, vol. 11, p. 94, 2010.

[37] M. D. Robinson and G. K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, vol. 23, no. 21, pp. 2881–7, 2007.

[38] Y. Lin, K. Golovnina, Z. X. Chen, H. N. Lee, Y. L. Negron, H. Sultana, B. Oliver, and S. T. Harbison, "Comparison of normalization and differential expression anal-

yses using rna-seq data from 726 individual drosophila melanogaster," *BMC Genomics*, vol. 17, p. 28, 2016.

[39] M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, F. Jaffrezic, and C. French StatOmique, "A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis," *Brief Bioinform*, vol. 14, no. 6, pp. 671–83, 2013.

[40] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of rna-seq data," *Genome Biol*, vol. 11, no. 3, p. R25, 2010.

[41] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–93, 2003.

[42] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nat Methods*, vol. 5, no. 7, pp. 621–8, 2008.

[43] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nat Genet*, vol. 25, no. 1, pp. 25–9, 2000.

[44] C. Gene Ontology, "Gene ontology consortium: going forward," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D1049–56, 2015.

[45] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack, "Gene ontology analysis for rna-seq: accounting for selection bias," *Genome Biol*, vol. 11, no. 2, p. R14, 2010.

[46] G. R. Ordonez, L. W. Hillier, W. C. Warren, F. Grutzner, C. Lopez-Otin, and X. S. Puente, "Loss of genes implicated in gastric function during platypus evolution," *Genome Biol*, vol. 9, no. 5, p. R81, 2008.

[47] A. D. Hayward, D. H. Nussey, A. J. Wilson, C. Berenos, J. G. Pilkington, K. A. Watt, J. M. Pemberton, and A. L. Graham, "Natural selection on individual variation in tolerance of gastrointestinal nematode infection," *PLoS Biol*, vol. 12, no. 7, p. e1001917, 2014.

[48] A. D. Hayward, "Causes and consequences of intra- and inter-host heterogeneity in defence against nematodes," *Parasite Immunol*, vol. 35, no. 11, pp. 362–73, 2013.

[49] C. Diez-Tascon, O. M. Keane, T. Wilson, A. Zadissa, D. L. Hyndman, D. B. Baird, J. C. McEwan, and A. M. Crawford, "Microarray analysis of selection lines from outbred populations to identify genes involved with nematode parasite resistance in sheep," *Physiol Genomics*, vol. 21, no. 1, pp. 59–69, 2005.

[50] H. N. Kadarmideen, N. S. Watson-Haigh, and N. M. Andronicos, "Systems biology of ovine intestinal parasite resistance: disease gene modules and biomarkers," *Mol Biosyst*, vol. 7, no. 1, pp. 235–46, 2011.

[51] R. W. Li, T. S. Sonstegard, C. P. Van Tassell, and L. C. Gasbarre, "Local inflammation as a possible mechanism of resistance to gastrointestinal nematodes in angus heifers," *Vet Parasitol*, vol. 145, no. 1-2, pp. 100–7, 2007.

[52] F. Antignano, S. C. Mullaly, K. Burrows, and C. Zaph, "Trichuris muris infection: a model of type 2 immunity and inflammation in the gut," *J Vis Exp*, no. 51, 2011.

[53] B. A. Vallance, P. A. Blennerhassett, and S. M. Collins, "Increased intestinal muscle contractility and worm expulsion in nematode-infected mice," *Am J Physiol*, vol. 272, no. 2 Pt 1, pp. G321–7, 1997.

[54] R. W. Li, Y. Hou, C. Li, and L. C. Gasbarre, "Localized complement activation in the development of protective immunity against ostertagia ostertagi infections in cattle," *Vet Parasitol*, vol. 174, no. 3-4, pp. 247–56, 2010.

[55] G. C. Gurtner, S. Werner, Y. Barrandon, and M. T. Longaker, "Wound repair and regeneration," *Nature*, vol. 453, no. 7193, pp. 314–21, 2008.

[56] F. Chen, Z. Liu, W. Wu, C. Rozo, S. Bowdridge, A. Millman, N. Van Rooijen, J. Urban, J. F., T. A. Wynn, and W. C. Gause, "An essential role for th2-type responses in limiting acute tissue damage during experimental helminth infection," *Nat Med*, vol. 18, no. 2, pp. 260–6, 2012.

[57] D. Beraldi, A. F. McRae, J. Gratten, J. G. Pilkington, J. Slate, P. M. Visscher, and J. M. Pemberton, "Quantitative trait loci (qtl) mapping of resistance to strongyles and coccidia in the free-living soay sheep (ovis aries)," *Int J Parasitol*, vol. 37, no. 1, pp. 121–9, 2007.

[58] B. Gutierrez-Gil, J. Perez, L. Alvarez, M. Martinez-Valladares, L. F. de la Fuente, Y. Bayon, A. Meana, F. San Primitivo, F. A. Rojo-Vazquez, and J. J. Arranz, "Quantitative trait loci for resistance to trichostrongylid infection in spanish churra sheep," *Genet Sel Evol*, vol. 41, p. 46, 2009.

[59] S. P. Huntley, M. Davies, J. B. Matthews, G. Thomas, J. Marshall, C. M. Robinson, J. W. Eveson, I. C. Paterson, and S. S. Prime, "Attenuated type ii tgf-beta receptor signalling in human malignant oral keratinocytes induces a less differentiated and more aggressive phenotype that is associated with metastatic dissemination," *Int J Cancer*, vol. 110, no. 2, pp. 170–6, 2004.

[60] H. R. Miller, "Mucosal mast cells and the allergic response against nematode parasites," *Vet Immunol Immunopathol*, vol. 54, no. 1-4, pp. 331–6, 1996.

[61] L. E. Donaldson, E. Schmitt, J. F. Huntley, G. F. Newlands, and R. K. Grencis, "A critical role for stem cell factor and c-kit in host protective immunity to an intestinal helminth," *Int Immunol*, vol. 8, no. 4, pp. 559–67, 1996.

[62] D. M. Zaiss, L. Yang, P. R. Shah, J. J. Kobie, J. F. Urban, and T. R. Mosmann, "Amphiregulin, a th2 cytokine enhancing resistance to nematodes," *Science*, vol. 314, no. 5806, p. 1746, 2006.

[63] D. M. Zaiss, W. C. Gause, L. C. Osborne, and D. Artis, "Emerging functions of amphiregulin in orchestrating immunity, inflammation, and tissue repair," *Immunity*, vol. 42, no. 2, pp. 216–26, 2015.

[64] D. M. Zaiss, J. van Loosdregt, A. Gorlani, C. P. Bekker, A. Grone, M. Sibilia, P. M. van Bergen en Henegouwen, R. C. Roovers, P. J. Coffer, and A. J. Sijts, "Amphiregulin enhances regulatory t cell-suppressive function via the epidermal growth factor receptor," *Immunity*, vol. 38, no. 2, pp. 275–84, 2013.

[65] V. Goncalves and P. Jordan, "Posttranscriptional regulation of splicing factor srsf1 and its role in cancer cell biology," *Biomed Res Int*, vol. 2015, p. 287048, 2015.

[66] J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell, "Clip identifies nova-regulated rna networks in the brain," *Science*, vol. 302, no. 5648, pp. 1212–5, 2003.

[67] Q. Li, S. Zheng, A. Han, C. H. Lin, P. Stoilov, X. D. Fu, and D. L. Black, "The splicing regulator ptbp2 controls a program of embryonic splicing required for neuronal maturation," *Elife*, vol. 3, p. e01201, 2014.

[68] L. T. Gehman, P. Stoilov, J. Maguire, A. Damianov, C. H. Lin, L. Shiue, J. Ares, M., I. Mody, and D. L. Black, "The splicing regulator rbfox1 (a2bp1) controls neuronal excitation in the mammalian brain," *Nat Genet*, vol. 43, no. 7, pp. 706–11, 2011.

[69] S. M. Weyn-Vanhentenryck, A. Mele, Q. Yan, S. Sun, N. Farny, Z. Zhang, C. Xue, M. Herre, P. A. Silver, M. Q. Zhang, A. R. Krainer, R. B. Darnell, and C. Zhang, "Hits-clip and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism," *Cell Rep*, vol. 6, no. 6, pp. 1139–52, 2014.

[70] C. C. Warzecha, S. Shen, Y. Xing, and R. P. Carstens, "The epithelial splicing factors esrp1 and esrp2 positively and negatively regulate diverse types of alternative splicing events," *RNA Biol*, vol. 6, no. 5, pp. 546–62, 2009.

[71] C. R. Mandel, S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley, and L. Tong, "Polyadenylation factor cpsf-73 is the pre-mrna 3'-end-processing endonuclease," *Nature*, vol. 444, no. 7121, pp. 953–6, 2006.

[72] H. H. Kim, S. J. Lee, A. S. Gardiner, N. I. Perrone-Bizzozero, and S. Yoo, "Different motif requirements for the localization zipcode element of beta-actin mrna binding by hud and zbp1," *Nucleic Acids Res*, vol. 43, no. 15, pp. 7432–46, 2015.

[73] K. E. Lukong, K. W. Chang, E. W. Khandjian, and S. Richard, "Rna-binding proteins in human genetic disease," *Trends Genet*, vol. 24, no. 8, pp. 416–25, 2008.

[74] B. M. Lunde, C. Moore, and G. Varani, "Rna-binding proteins: modular design for efficient function," *Nat Rev Mol Cell Biol*, vol. 8, no. 6, pp. 479–90, 2007.

[75] L. A. Selth, C. Gilbert, and J. Q. Svejstrup, "Rna immunoprecipitation to determine rna-protein associations in vivo," *Cold Spring Harb Protoc*, vol. 2009, no. 6, p. pdb prot5234, 2009.

[76] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, J. Ascano, M., A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl, "Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip," *Cell*, vol. 141, no. 1, pp. 129–41, 2010.

[77] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule, "iclip reveals the function of hnrnp particles in splicing at individual nucleotide resolution," *Nat Struct Mol Biol*, vol. 17, no. 7, pp. 909–15, 2010.

[78] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo, "Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip)," *Nat Methods*, 2016.

[79] Y. C. Yang, C. Di, B. Hu, M. Zhou, Y. Liu, N. Song, Y. Li, J. Umetsu, and Z. J. Lu, "Clipdb: a clip-seq database for protein-rna interactions," *BMC Genomics*, vol. 16, p. 51, 2015.

[80] D. Maticzka, S. J. Lange, F. Costa, and R. Backofen, "Graphprot: modeling binding preferences of rna-binding proteins," *Genome Biol*, vol. 15, no. 1, p. R17, 2014.

[81] B. J. Blencowe, S. Ahmad, and L. J. Lee, "Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes," *Genes Dev*, vol. 23, no. 12, pp. 1379–86, 2009.

[82] T. Derrien, J. Estelle, S. Marco Sola, D. G. Knowles, E. Raineri, R. Guigo, and P. Ribeca, "Fast computation and applications of genome mappability," *PLoS One*, vol. 7, no. 1, p. e30377, 2012.

[83] H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, and Q. Morris, "Rnacontext: a new method for learning the sequence and structure binding preferences of rna-binding proteins," *PLoS Comput Biol*, vol. 6, p. e1000832, 2010.

[84] M. Strazar, M. Zitnik, B. Zupan, J. Ule, and T. Curk, "Orthogonal matrix factorization enables integrative analysis of multiple rna binding proteins," *Bioinformatics*, vol. 32, no. 10, pp. 1527–35, 2016.

[85] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput*, vol. 18, no. 7, pp. 1527–54, 2006.

[86] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[87] L. Deng, D. Yu, *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[88] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna- and rna-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–+, 2015.

[89] H. R. Hassanzadeh and M. D. Wang, "Deeperbind: Enhancing prediction of sequence specificities of dna binding proteins," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pp. 178–183, IEEE, 2016.

[90] J. Lanchantin, R. Singh, Z. Lin, and Y. Qi, "Deep motif: Visualizing genomic sequence classifications," *arXiv preprint arXiv:1605.01133*, 2016.

[91] H. Y. Zeng, M. D. Edwards, G. Liu, and D. K. Gifford, "Convolutional neural network architectures for predicting dna-protein binding," *Bioinformatics*, vol. 32, no. 12, pp. 121–127, 2016.

[92] S. Zhang, J. Zhou, H. Hu, H. Gong, L. Chen, C. Cheng, and J. Zeng, "A deep learning framework for modeling structural features of rna-binding protein targets," *Nucleic acids research*, vol. 44, no. 4, pp. e32–e32, 2015.

[93] X. Pan and H. B. Shen, "Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach," *BMC Bioinformatics*, vol. 18, no. 1, p. 136, 2017.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[95] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[96] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[97] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[98] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

[99] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[100] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," *Artificial Neural Networks and Ma-*

chine Learning–ICANN 2011, pp. 52–59, 2011.

[101] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

[102] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[103] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, *et al.*, "A compendium of rna-binding motifs for decoding gene regulation," *Nature*, vol. 499, no. 7457, p. 172, 2013.

[104] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," *Genome biology*, vol. 8, no. 2, p. R24, 2007.

[105] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "Meme suite: tools for motif discovery and searching," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W202–W208, 2009.

[106] C. Colombrita, E. Onesto, F. Megiorni, A. Pizzuti, F. E. Baralle, E. Buratti, V. Silani, and A. Ratti, "Tdp-43 and fus rna-binding proteins bind distinct sets of cytoplasmic messenger rnas and differently regulate their post-transcriptional fate in motoneuron-like cells," *Journal of Biological Chemistry*, vol. 287, no. 19, pp. 15635–15647, 2012.

[107] K. Kapeli, G. A. Pratt, A. Q. Vu, K. R. Hutt, F. J. Martinez, B. Sundararaman, R. Batra, P. Freese, N. J. Lambert, S. C. Huelga, *et al.*, "Distinct and shared functions of als-associated proteins tdp-43, fus and taf15 revealed by multisystem analyses," *Nature communications*, vol. 7, 2016.

[108] J. Saulière, V. Murigneux, Z. Wang, E. Marquenet, I. Barbosa, O. Le Tonquèze, Y. Audic, L. Paillard, H. R. Crollius, and H. Le Hir, "Clip-seq of eif4aiii reveals transcriptome-wide mapping of the human exon junction complex," *Nature structural & molecular biology*, vol. 19, no. 11, pp. 1124–1131, 2012.

[109] M. R. Capecchi, "The new mouse genetics: altering the genome by gene targeting," *Trends Genet*, vol. 5, no. 3, pp. 70–6, 1989.

[110] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nat Rev Genet*, vol. 10, no. 1, pp. 57–63, 2009.

[111] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "Ncbi geo: archive for functional genomics data sets–update," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D991–5, 2013.

[112] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. Pedro Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, and U. Sarkans, "Arrayexpress update–trends in database growth and links to data analysis tools," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D987–90, 2013.

[113] Y. Kodama, M. Shumway, R. Leinonen, and C. International Nucleotide Sequence Database, "The sequence read archive: explosive growth of sequencing data," *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D54–6, 2012.

[114] S. Brunak, A. Danchin, M. Hattori, H. Nakamura, K. Shinozaki, T. Matise, and D. Preuss, "Nucleotide sequence database policies," *Science*, vol. 298, no. 5597, p. 1333, 2002.

[115] A. Coletta, C. Molter, R. Duque, D. Steenhoff, J. Taminau, V. de Schaetzen, S. Meganck, C. Lazar, D. Venet, V. Detours, A. Nowe, H. Bersini, and D. Y. Weiss Solis, "Insilico db genomic datasets hub: an efficient starting point for analyzing genome-wide studies in genepattern, integrative genomics viewer, and r/bioconductor," *Genome Biol*, vol. 13, no. 11, p. R104, 2012.

[116] B. Qin, M. Zhou, Y. Ge, L. Taing, T. Liu, Q. Wang, S. Wang, J. Chen, L. Shen, X. Duan, S. Hu, W. Li, H. Long, Y. Zhang, and X. S. Liu, "Cistromemap: a knowledgebase and web server for chip-seq and dnase-seq studies in mouse and human," *Bioinformatics*, vol. 28, no. 10, pp. 1411–2, 2012.

[117] B. C. Haynes, E. J. Maier, M. H. Kramer, P. I. Wang, H. Brown, and M. R. Brent, "Mapping functional transcription factor networks from gene expression data," *Genome Res*, vol. 23, no. 8, pp. 1319–28, 2013.

[118] R. Eksi, H. D. Li, R. Menon, Y. Wen, G. S. Omenn, M. Kretzler, and Y. Guan, "Systematically differentiating functions for alternatively spliced isoforms through integrating rna-seq data," *PLoS Comput Biol*, vol. 9, no. 11, p. e1003314, 2013.

[119] L. Chindelevitch, D. Ziemek, A. Enayetallah, R. Randhawa, B. Sidders, C. Brockel, and E. S. Huang, "Causal reasoning on biological networks: interpreting transcriptional changes," *Bioinformatics*, vol. 28, no. 8, pp. 1114–21, 2012.

[120] J. A. Blake, C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and G. Mouse Genome Database, "The mouse genome database: integration of and access to knowledge about the laboratory mouse," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D810–7, 2014.

[121] F. B. Rogers, "Medical subject headings," *Bull Med Libr Assoc*, vol. 51, pp. 114–6, 1963.

[122] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg, "The brenda tissue ontology (bto): the first all-integrating ontology of all organisms for enzyme sources," *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D507–13, 2011.

[123] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, "Modeling sample variables with an experimental factor ontology," *Bioinformatics*, vol. 26, no. 8, pp. 1112–8, 2010.