# Characterizing Immune Responses to Marburg Virus Infection in Animal Hosts Using Statistical Transcriptomic Analysis

Albert Lee

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Characterizing Immune Responses to Marburg Virus Infection in Animal Hosts Using Statistical Transcriptomic Analysis

Albert Lee

Marburg virus (MARV)–along with Ebola Virus–comprises *Filoviridae*, a family of virus which causes the life-threatening hemorrhagic fever in human and non-human primates for which there is no clinically approved vaccine. For this reason, this virus can potentially lend itself to pandemic and weapons of bioterrorism. Strikingly, this virus yields asymptomatic responses in its recently discovered host *Rousettus aegyptiacus*. Understanding of the interaction between MARV and different animal hosts will enable the improved understanding of filovirus immunology and the development of effective therapeutic agents.

Although cell lines and primary cells have been used to investigate gene expression analysis of this virus, the transcriptomic view of MARV infection on the tissue samples of animal hosts has been an uncharted territory. The comprehensive analysis of transcriptome in hosts and spillover hosts will shed light on the immune responses on a molecular level and potentially allow the comparative analysis to understand the phenotypical differences.

However, there have been gaps in resources necessary to carry the transcriptome research for MARV. For example, MARV host *Rousettus aegyptiacus* genome and transcriptome had not been available. Furthermore, the statistical machinery necessary to analyze multi-tissue/multi-time data was not available.

In this dissertation, I introduce the two items that fill these gaps and show the appli-

cation of the tools I built for novel biological discovery. In particular, I have built 1) the comprehensive *de novo* transcriptome reference of *Rousettus aegyptiacus* and 2) the Multilevel Analysis of Gene Expression (MAGE) pipeline to analyze the RNA-seq data with the complex experimental design.

I show the application of MAGE in multi-time, multi-tissue transcriptome data of *Macaca mulata* infected with MARV. In this study, 15 rhesus macaques were sequentially sacrificed via aerosol exposure to MARV Angola over the course of 9 days, and 3 types of lymph node tissues (tracheobronchial, mesenteric, and inguinal) were extracted from each sample and sequenced for gene expression analysis.

With MAGE pipeline, I discovered that the posterior median log2FC of genes separates the samples based on day post infection and viral load. I discovered the set of genes such as CD40LG and TMEM197 with interesting trends over time and how similar and different pathways have been influenced in three lymph nodes. I also identified the biologically meaningful clusters of genes based on the topology-based clustering algorithm known as Mapper. Using the MAGE posterior samples, I also determined the genes that are preferentially expressed in tracheobronchial lymph nodes. In addition to new analysis tools and biological findings, I built the gene expression exploration tool for biologists to examine differential gene expression over time in various immune-related pathways and contributing members of the pathways.

In conclusion, I have contributed to the two important components in the transcriptome analysis in MARV research and discovered novel biological insights. The MAGE pipeline is modular and extensible and will be useful for the transcriptome research with the complex experimental designs which are becoming increasingly prevalent with the decrease in

the cost of sequencing.

# *Table of Contents*

# *List of Figures*

# *List of Tables*

# *Acknowledgments*

I would like to first thank my thesis advisor Dr. Raul Rabadan for his generous support and guidance throughout my graduate program. He has been an amazing mentor in both science and life ever since I joined his lab as a naive master student in 2011. Every time I interacted with him, I learned something new and interesting from him whether it be the mathematical formula of Poisson distribution[1], some advanced physics or math concepts, or how to be a good scientist. I must say my Ph.D experience was something I will look back with a smile in the future because I had a wonderful, resourceful, and insightful Ph.D. advisor like Raul. I would like to express my gratitude toward my comittee members: Dr. Aristidis Floratos, Dr. Yufeng Shen, Dr. Hossein Khiabanian, and Dr. Adler Perotte for their support and guidance throughout my Ph.D. career. Their comments and direction as well as inspirations from their work have been tremendously and continually helpful and insightful, and helped me improve this dissertation. I would like to acknowledge faculty and staff members of Department of Biomedical Informatics and the Department of Systems Biology for years of help and support.

I would also like to thank my primary collaborator, Dr. Kirsten Kulcsar. She is a great biologist, a thinker, and a communicator from whom I learned immensely. We col-

[1] Once and for all, it's $Pr(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

laborated on the *Rousettus aegyptiacus* transcriptome project and MARV-rhesus infection study discussed in this dissertation. Just working with her taught me how to reason with biological concepts and convey the findings well in writing. Her positive energy, patience, and perseverance made the science we do all the more fun and meaningful.

I would like to thank Dr. Gustavo Palacios and Dr. Mariano Sanchez-Lockhart, and Sean Lovett from USAMRIID for good discussions and support for projects described in this dissertation. I would also like to acknowledge Dr. Tom Kepler and Stephanie D'Souza from Boston University and Dr. Jenna Kelly and Dr. Jonathan S. Towner from Centers for Disease Control and Prevention for their helpful feedback and insightful comments from the monthly conference call regarding *Rousettus aegyptiacus* projects.

I want to thank current and past Rabadan lab members. In particular, I acknowledge Oliver Elliot, Tim Chu, Dr. Daniel Rosenbloom, Dr. Francesco Abate, Dr. Ioan Filip, and Kernyu Park for the helpful discussions and feedback they gave regarding the projects discussed in this dissertation.

Personally, I would like to thank my wife, Rachel. She has been immensely supportive and helpful in my journey as a graduate student through thick and thin. I thank her for all the support and sacrifices she made. Lastly, I would like to thank my family members, my father(Kyuseok Lee) and mother(Moonhee Kim), and three of my sisters–Janet, Nancy, and Julie–for their continual support and encouragement. Furthermore, I want to express my sincere gratitude for my grandmother, Eul-Ryun Kim, who has always been resourceful financially and emotionally. My sisters and I live internationally pursuing our dreams, and I thank my parents and grandparents for their hard work and support to make it possible. I feel lucky to have an amazing family.

Dedicated to my parents

Kyuseok Lee and Moonhee Kim,

who paved the way for the wonders of life for me.

# *Introduction*

## 0.1 Motivation of this thesis

Marburg virus (MARV) belongs to the viral family *Filoviridae* that causes severe hemorrhagic fever disease in humans and nonhuman primates, but results in little to no pathological consequences in bats (Bean et al. 2013; Towner et al. 2006). Little has been done for the comprehensive and global characterization of MARV pathogenesis in animal hosts due to lack of the data and resources such as the genome and transcriptome of hosts.

The comprehensive analysis of transcriptome in hosts such as bats and spillover hosts such as non-human primates will shed light on the characteristic immune responses due to MARV infection on a molecular level and potentially allow the comparative analysis to understand the phenotypical differences, ultimately increasing our understanding of this virus and chances of developing effective treatments.

However, there have been gaps in resources necessary to carry the transcriptome research for Marburg virus. For example, until recently, Marburg virus host *Rousettus aegyptiacus* genome and transcriptome had not been available. Furthermore, the statistical machinery available to analyze the RNA-seq data with complex experimental design were limited.

To this end, I worked toward bridging this gap in my doctorate program, especially in the space of transcriptomic analysis. I have built the comprehensive *de novo* transcriptome reference of *Rousettus aegyptiacus* and the Multilevel Analysis of Gene Expression (MAGE) framework designed to handle the complex RNA-seq experiments with more than one factor.

Before diving into the detailed contents of my thesis in each chapters that follow, let me briefly discuss the history of gene expression analysis, which will lay the foundation for everything that follows.

## 0.2 History of Gene Expression Analysis and High Throughput Transcriptome Analysis

All organisms–according to the famous cell theory by Theodor Schwann and Matthias Jakob Schleiden (Wikipedia 2004)–consist of one or more cells, which themselves consist of multiple entities. Therefore, to understand any organism, one must study the processes of its cells and the constituents involved, in particular what they are and how they do what they do. In the field of molecular biology, one such activity is the quantification of the expression of a gene, also known as *gene expression analysis*.

Accurate quantification of expressions of individual genes under the particular state of cells is a key to understanding the role of genes in physiology, pathology, and immunology of an organism (Barczak et al. 2012; Liang and Pardee 1992). With the relevant genes quantified, one can perform differential gene expression analysis measuring the change of gene

expression as a function of multiple biological states to infer the molecular mechanisms of biological functions and regulations as well as to predict clinical outcomes.

As the techniques and methodologies of molecular biology become more sophisticated and advanced, the researchers became increasingly interested in quantifying multiple genes simultaneously–also known as *gene expression profile*. Hence, the analysis of transcriptome–a set of all RNA transcripts–has become increasingly popular. There was a huge interest in transcriptomic analysis because by studying expressions of multiple genes together, one can obtain the holistic and systematic view of a particular state,whether it be normal, cancerous, or disease, in cell populations. Furthermore, it can speed up the discovery for the determinants of particular biological states and targets for interventions (Ye et al. 2002).

In early 90s, gene expression of multiple genes was measured by the technologies such as differential display (Liang and Pardee 1992), expressed sequence tags (EST) (Adams et al. 1991), and Serial Analysis of Gene Expression (Velculescu et al. 1995). Between the mid 90s and early 2000s, it was widely performed with microarrays (Brazma et al. 2001; Duggan et al. 1999; Pan 2002; Schena et al. 1995; Van't Veer et al. 2002; Watson et al. 1998). Since thousands of genes were measured simultaneously, the need for Bioinformatics has increased as well. Many statistical tools have been developed to address the challenges that come with how the data were generated (more on this in the next section ). Although microarrays are useful, they also have limitations; one prominent issue was that it could only quantify known genes since it relied on hybridization techniques and this issue is problematic for genes with complicated splicing schema and unknown structure (Mortazavi et al. 2008).

Since the mid 2016, however, the RNA-seq analysis become the *de facto* method to analyze gene expression simultaneously (Mortazavi et al. 2008; Wang, Gerstein, and Snyder 2009). It enabled the high-throughput digital quantification at a single nucleotide level in an accurate way. Compared to microarray, RNA-seq offers higher sensitivity and specificity: It could not only quantify previously unknown genes, did not have cross-hybridization and background normalization issues, and has high correlations among technical replicates (Mortazavi et al. 2008; Wang, Gerstein, and Snyder 2009).

Despite the ease of ability to quantify the global gene expression with the the help of microarray or RNA-seq, it comes with another set of challenges. The data generated from high-throughput sequencing technology is stochastic, and this resulted in the development of various normalization techniques and statistical models. Normalization techniques such as quantile normalization, TMM-normalization, and DESeq normalization have been developed to mitigate the biases that might be caused by differences in library sizes and gene expression composition (Bolstad et al. 2003; Degexp et al. 2010; Oshlack, Robinson, and Young 2010). Another issue is the statistical modeling of the underlying error model based on a platform. For example, with microarray, the unit of measurements is the light intensity (real values) whereas in RNA-seq the data were generated in the form of counts. The former is usually modeled with normal distribution, and the latter Poisson or negative binomial distribution. More importantly, despite that the measurements were done for thousands of genes, the number of experimental units were usually limited to three due to the cost. The small number of replicates posed yet another problems; naively applying the traditional statistical tools such as t-test would lead to exaggerated parameter estimates and result in high false positive results. To address this issues, the tools such as limma,

edgeR, or DESeq (Degexp et al. 2010; Robinson, McCarthy, and Smyth 2010a; Smyth 2005) have been developed to leverage information sharing across genes with similar gene expression level. With these tools, numerous novel and important biological discoveries are continually being made at an increasing speed.

## 0.3 Gene Expression Analysis Workflow

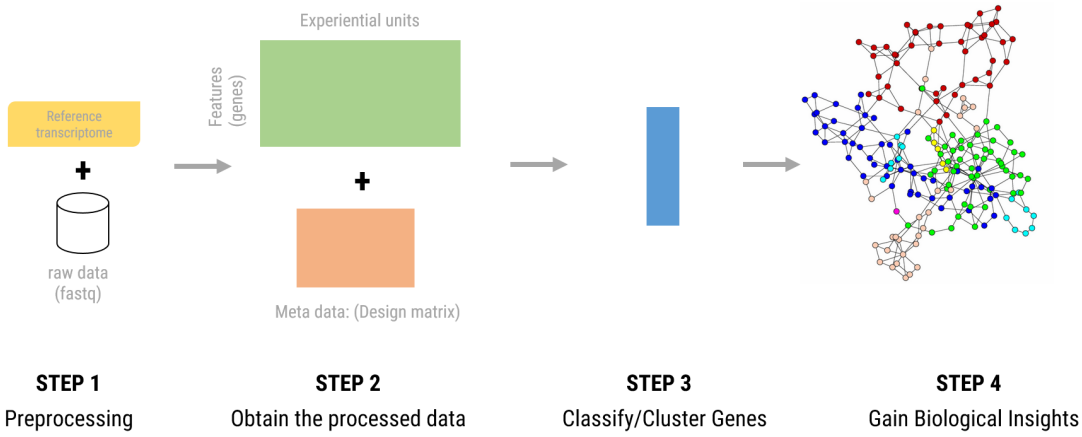


Figure 0.1: **Transcriptomic Analysis Workflow**.

Typical transcriptome analysis workflow consists of the steps shown in Figure 0.1. First, a reference transcriptome and the expression data in the form of RNA-seq or microarray data are obtained (Barczak et al. 2012; Morens, Folkers, and Fauci 2004). The reference transcriptome is a set of all transcript sequences, for each the samples can be quantified against. Important thing to note is that not all organisms have the reference transcriptome available, and when there is no publicly available reference, the investigator must create one using *de novo* transcriptome assembly. In Chapter 1, I describe how I constructed the *de novo* transcriptome reference of *Rousettus aegyptiacus*.

With the reference and the sequence data, the quantification is performed to generate the *NxM* expression matrix where each of N rows correspond to *n*th gene (or transcript or probe, depending on the platform used) and columns correspond to *m*th samples. The choice of quantification software depends on the data type(microarray, RNA-seq, or single cell RNA-seq), the reference type (genome vs transcriptome) and the level of analysis(gene vs transcript). The popular RNA-seq transcriptome quantification programs include, but not limited to, tophat2 (Kim et al. 2013, STAR (Dobin et al. 2013), HISAT2 (Kim, Langmead, and Salzberg 2015), RSEM (Li and Dewey 2011), and Kallisto (Pimentel et al. 2017). The individual cells contain the measurement for a particular gene in a particular sample in the form of counts or expression measurements such as RPKM (Reads Per Kilobase of transcript per Million mapped reads) or TPM (Transcripts Per Million). With the meta information, this expression matrix can then be analyzed via appropriate statistical models. The majority of gene expression analysis tools for RNA-seq such as EdgeR, DESeq2, limma-Voom, and Sleuth employs a generalized linear modeling framework (Degexp et al. 2010; Law et al. 2014; Love, Huber, and Anders 2014; Pimentel et al. 2017; Robinson, McCarthy, and Smyth 2010a) with underlying likelihood model of negative binomial with a factor with $K$ levels in which $k$-th element corresponds to distinct biological states such as normal, cancer, drugged, etc. Typically, the goal is to classify genes into positively or negatively regulated groups in a particular biological state compared to a control group and fold change is the parameter of interest. By observing the fold change, one performs a differential gene expression analysis. The more advanced classification becomes necessary with more meta information become available, which is a topic of Chapter 2 and Chapter 3 in this dissertation. Once the interesting subset of genes are identified, the knowledge base

approach is used by the means of Gene Set Enrichment Analysis(Subramanian et al. 2005; Yu and He 2016) or Pathway Analysis (Krämer et al. 2013) to infer the biological pathways or processes involved.

## 0.4 Organization of this Dissertation

A typical RNA-seq data analysis consists of multiple subcomponents as shown in Figure 0.1 and the granular components can be found more in detail in references such as Conesa et al. 2016, but the two important components are 1) the reference transcriptome and 2) the robust statistical methods to accurately estimate the various parameters of interest while controlling the noise caused by technical and biological variances.

This dissertation revolves around these two themes as well as novel biological findings I made with those two, and has a following progression:

In Chapter 1, I will discuss *de novo* transcriptome assembly of *Rousettus aegyptiacus*–a species recently identified as a natural reservoir host of Marburg virus–which I comprehensively built and characterized. Briefly, using deep RNA-seq data from 11 distinct tissues from one male and one female, colleagues and I sequenced, assembled, and annotated the reference transcriptome using the homology based approach. This transcriptome is a important resource for understanding bat immunology, physiology, disease pathogenesis, and virus transmission.

In Chapter 2, I discuss the Multilevel Analysis of Gene Expression(MAGE) pipeline, which I developed to fully utilize the structure of the data set and perform the direct inference based on the posterior distribution of all the relevant parameters. MAGE allows one

to obtain samples from the posterior distribution of transformed parameters of one's choice whether it be log2Fold Change, tissue specificity, and waveforms, all of which allow one to answer relevant and interesting biological questions.

In Chapter 3, I discuss the application of MAGE on the serially sacrificed rhesus macaques infected with MARV virus. 15 rhesus macaques were sequentially sacrificed via aerosol exposure to Marburg Virus Angola over the course of 9 days, and 3 types of lymph nodes(tracheobrochial, mesenteric, and inguinal) were extracted from each sample and sequenced for gene expression analysis. Multiple interesting biological insights have been learned via posterior inferences supported by MAGE pipeline.

Chapter 1

# *De novo transcriptome reconstruction and annotation of the Egyptian rousette bat*

## 1.1 Background

Bats (order: Chiroptera) constitute an abundant and diverse mammalian lineage comprising approximately 20% of all known mammalian diversity (Wilson and Reeder 2005). Bats have evolved apart from other mammals more than 50 million years (Moratelli and Calisher 2015) and are divided into two major suborders; the Yinpterochiroptera (megachiroptera) and the Yangochiroptera (microchiroptera). Yinpterochiroptera includes the family *Pteropodidae* and genera *Rousettes* and *Pteropus* whereas Yangochiroptera includes the family *Myotidae* and genus *Myotis* (Teeling et al. 2002). Unlike most mammals, bats can fly and this ability enabled their wide geographical range and increased metabolism (Moratelli and Calisher 2015). Interestingly, bats have recently come to the forefront of zoonotic disease research with vast number of pathogens identified in a wide-variety of bat species (ibid.).

Upwards of 85 different viruses, primarily RNA viruses, have been detected and/or isolated from bats (Calisher et al. 2006; Moratelli and Calisher 2015). Amongst these are emerging viruses that cause lethal disease in humans and nonhuman primates in-

cluding Nipah virus (Smith et al. 2011; Young et al. 1996), Hendra virus (Chua et al. 2000), severe acute respiratory syndrome (SARS)-like coronavirus (Li et al. 2005), Middle East respiratory syndrome coronavirus (MERS-CoV) (Groot et al. 2013), Marburg virus (MARV) (Amman et al. 2015; Swanepoel et al. 2007; Towner et al. 2007, 2009), and Ebola virus(EBOV) (Leroy et al. 2005; Ogawa et al. 2015; Saéz et al. 2015). Despite the severe virulence of these viruses in humans, infected bats are often asymptomatic (Amman et al. 2015; Middleton et al. 2007; Paweska et al. 2012, 2015; Swanepoel et al. 1996; Williamson et al. 1998, 2000). Nipah virus and Hendra virus interactions with their natural reservoir hosts, *Pteropus vampyrus* and *Pteropus alecto*, respectively, are well characterized. Experimental infections of bats with high doses of henipaviruses have shown virus replication and shedding with little to no disease (Middleton et al. 2007; Williamson et al. 1998, 2000). Remarkably, the only viruses known to have induced any observable pathology in bats are Rabies virus and Australian bat lyssavirus (Field, McCall, and Barrett 1999; Moratelli and Calisher 2015). Understanding mechanisms of disease and differential responses to infection in asymptomatic reservoir host species compared to species that exhibit severe pathology will help inform the development of novel therapeutics and disease prevention approaches.

*Rousettus aegyptiacus*, commonly known as the Egyptian rousette bat, has been identified as a natural reservoir host for MARV through ecological, epidemiological, and experimental studies (Amman et al. 2012, 2015; Paweska et al. 2012, 2015; Swanepoel et al. 2007; Towner et al. 2009). Furthermore, it has been speculated this bat could be a host of Ebola virus (Feldmann and Geisbert 2011; Olival et al. 2013; Pourrut et al. 2009; Towner et al. 2009), although recent experimental infection studies have shown Ebola virus does

not replicate well in *R. aegeyptiacus* (Jones et al. 2015). The majority of human outbreaks due to MARV have been associated with caves inhabited by *R. aegyptiacus*. Furthermore, epidemiological surveillance of the *R. aegyptiacus* colony located in the Python cave in Uganda revealed a biannual spike in Marburg virus prevalence. This pattern correlated strongly with spillover transmission events in humans (Amman et al. 2012). Initial studies in captive bats evaluated clinical signs, virus dissemination, and virus shedding patterns during experimental infection with a MARV isolate derived from wild bats (Amman et al. 2015). Consistent with a natural reservoir host, the bats showed little to no evidence of disease even though the virus disseminated throughout their body and was actively shed (ibid.). These results were confirmed when bats were infected with MARV Angola, a strain isolated from a lethal human case (Paweska et al. 2015). In the absence of genetic and transcriptomic information for *R. aegyptiacus* and with limited available reagents, studying this reservoir host animal model has been challenging.

The rapid expansion in genomic knowledge for different bat species has facilitated comparative studies that rely on the identification of genes and gene families, and has established a framework for developing necessary reagents. Full genome annotations for *Pteropus vampyrus* (2.63X, Ensembl), *Myotis lucifugus* (7X, Ensembl) *Pteropus alecto* (110x, (Zhang et al. 2013)), *Myotis davidii* (110x, (ibid.)), and *Myotis brandtii* (77.8X, (Seim et al. 2013)) are now available. Additionally, transcriptomic annotations for *Pteropus alecto* (Papenfuss et al. 2012) and *Artibeus jamaicensis* (Shaw et al. 2012) have been published. In particular, the complementary genome and transcriptome annotations for *P. alecto* has aided studies on henipavirus infections in its reservoir host (Papenfuss et al. 2012; Zhang et al. 2013).

Here, we report the transcriptomic annotation of *R. aegyptiacus* from a *de novo* assembly of RNA sequencing data from 11 tissues isolated from a male and a female bat. We identified 24,118 canonical coding transcripts whose expression profiles were consistent with the corresponding tissues of origin. In addition, we identified and validated the novel coding transcripts that do not have any homology with the known sequences. Furthermore, we evaluated annotated immune-related genes and assessed the presence and expression of genes associated with a variety of immune functions.

## 1.2 Results and discussion

### *De novo* transcriptome assembly of *R.aegyptiacus*

We employed a *de novo* assembly approach to generate a comprehensive transcriptome without relying on a genome reference. First, we generated 20 RNA-seq libraries consisting of 11 tissue types (Table 1, Figure 1.1A) each collected from one male and one female *R. aegyptiacus* bat, which yielded approximately 2.1 billion reads.

We then assembled the high quality reads using Trinity (Grabherr et al. 2011) (Figure 1.1B). This process generated 14,796,219 contigs. The assembly had high continuity and coverage with a median number of 718,807 contigs and median N50 of 1,540 across all tissues (Table 1). To comprehensively annotate the contigs, we used the Multiple Species Annotation (MSA) pipeline (Lee et al. 2014), which leverages the homology of known sequences of related species. We assigned gene symbols to contigs when this information was available. This process clustered the contigs into isoform groups (Figure 1.1C).

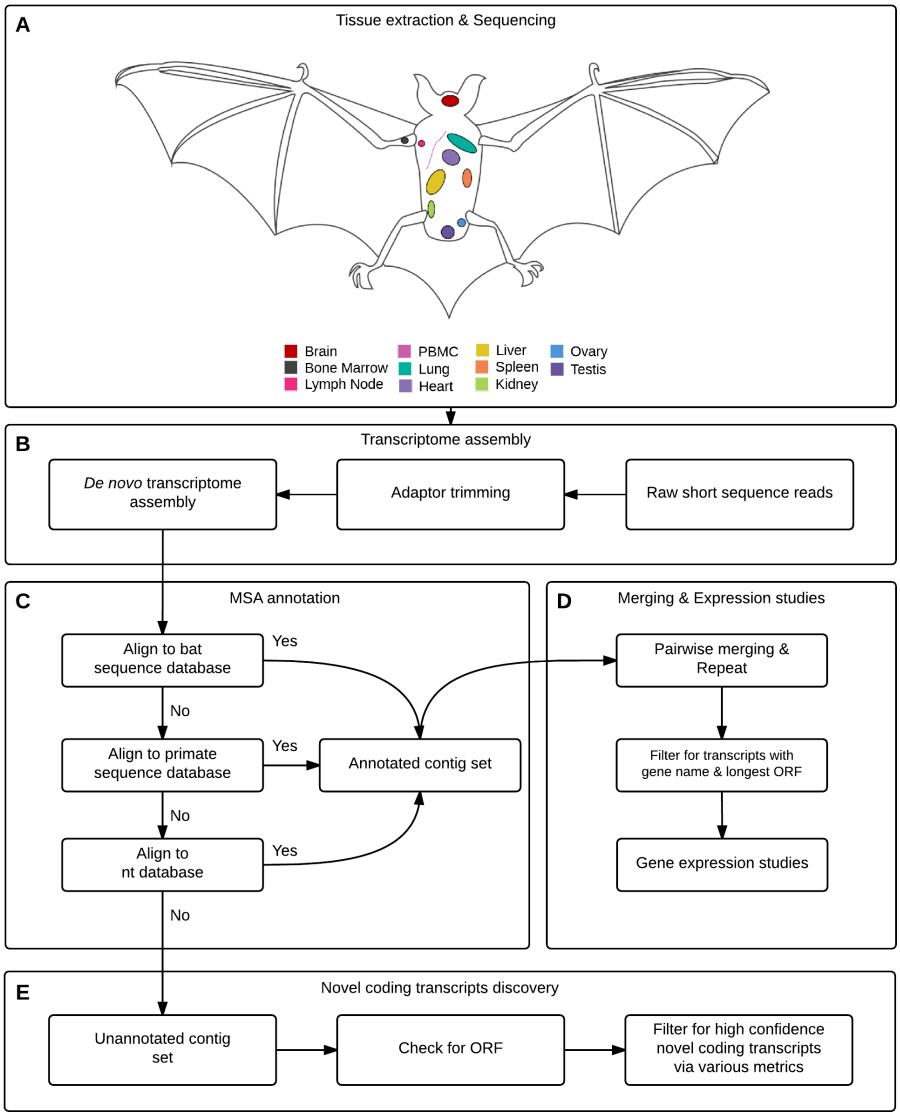


Figure 1.1: **Schematic of the *de novo* transcriptome reconstruction and analysis pipeline**. The pipeline consists of 5 steps. A) Data generation: Multiple tissues are extracted from *R. aegyptiacus* and sequenced. B) *De novo* Transcriptome assembly: Individual samples are first preprocessed to remove adapter sequences and assembled into contigs *de novo*. C) MSA annotation: Once the set of contigs is generated, they are annotated using BLAST against three databases. In each step, unannotated contigs are iteratively annotated using the downstream databases. D) Merging and Expression studies: A nonredundant contig set is obtained by merging the contig set of individual tissues two at a time. This pairwise merging is repeated until only one contig set is left. The subset of this contig can be obtained for the downstream analysis such as gene expression analysis by taking the transcripts with gene symbol and ORF sequence. See Figure 1.2 for details. E) Discovery of Novel Coding Transcripts: Novel coding transcripts can be identified by searching for contigs that failed annotation in the previous steps. Various metrics can be applied to generate high confidence novel coding transcript candidates.

Table 1.1: **Sample Information and Basic Statistics**.

| Bat | Gender | Tissue | Read count | Library | N50 | Number of contigs |
|---|---|---|---|---|---|---|
| BAT01 | F | BM | 67896687 | single | 1736 | 609943 |
| BAT02 | F | BR | 55004118 | single | 884 | 896445 |
| BAT03 | F | HT | 77315750 | single | 1263 | 717588 |
| BAT04 | F | KY | 59782352 | single | 1174 | 720026 |
| BAT05 | F | LG | 77510852 | paired | 1822 | 903831 |
| BAT06 | F | LN | 63170354 | single | 1566 | 638083 |
| BAT07 | F | LV | 89970603 | paired | 1566 | 697125 |
| BAT08 | F | OV | 75051316 | single | 1401 | 875888 |
| BAT09 | F | PB | 56553369 | single | 1890 | 404332 |
| BAT10 | F | SP | 56141808 | single | 1340 | 716771 |
| BAT11 | M | BM | 47988156 | paired | 1808 | 744115 |
| BAT12 | M | BR | 75378417 | paired | 1490 | 1088331 |
| BAT13 | M | HT | 20042200 | paired | 748 | 497729 |
| BAT14 | M | KY | 71478010 | paired | 1514 | 872829 |
| BAT15 | M | LG | 15525010 | paired | 668 | 575991 |
| BAT16 | M | LN | 88471565 | paired | 2186 | 797125 |
| BAT17 | M | LV | 27358079 | paired | 925 | 431513 |
| BAT18 | M | PB | 92707184 | paired | 1745 | 556053 |
| BAT19 | M | SP | 98465277 | paired | 2141 | 873259 |
| BAT20 | M | TT | 96476242 | paired | 1866 | 1179242 |

## *R.aegyptiacus* transcriptome captures a majority of bat transcripts

We compared our assembly to the transcriptomes of three related bat species -- *M. davidii*, *P. alecto*, and *M. brandtii*. Using BLAST, we recovered 90.1% of *M. davidii* transcripts, 89.54% of *M. brandtii* transcripts, and 97.38% of *P.alecto* transcripts. This result is consistent with the evolutionary history of these bats considering that *P. alecto* and *R. aegyptiacus* belong to the same family of *Pteropodidae*.

## Combining the transcriptome to generate nonredundant contigs

Tissue-specific transcriptome assemblies contained different numbers of contigs, due to their different levels of expression and sequencing depth. Without a common ground for

comparison, it was difficult to perform downstream comparative analyses such as differential gene expression analysis; therefore, we combined contigs from all tissues into one unified, nonredundant reference transcriptome (Figure 1.1D). To this end, we iteratively merged the assemblies two at a time, similar to the approach employed in (Robertson et al. 2010) (Figure 1.1D). We obtained 4,746,293 nonredundant contigs. Among the nonredundant contigs, 974,765 (20.54%) of the sequences were annotated by bat sequences, 860,578 (18.13%) by primate sequences, and 104,796 (2.2%) by sequences in nt database (Figure 1.2A).

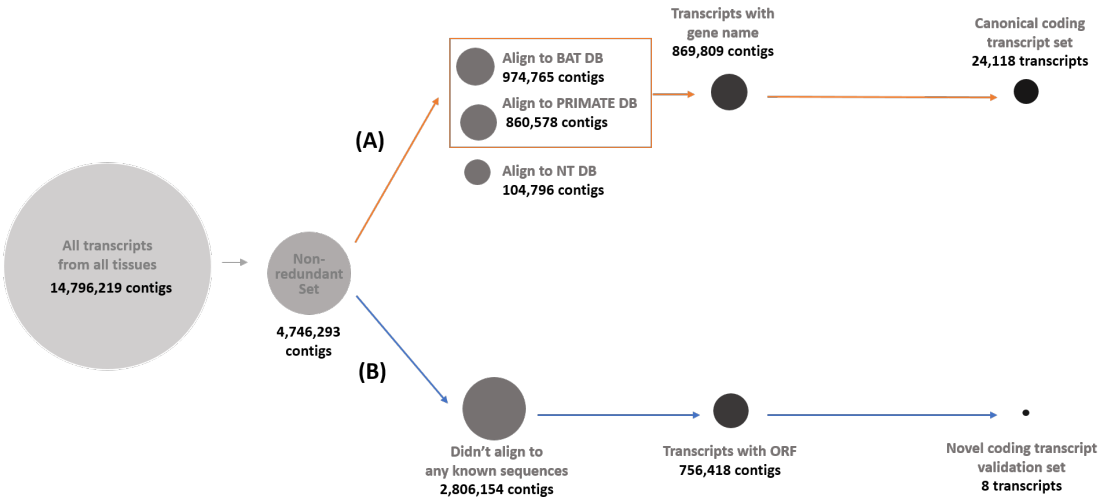


Figure 1.2: **Generation of Nonredundant Contig Set, Canonical Coding Transcript Set, and High Confidence Novel Transcript Set**. From the union of all contigs, we generated the nonredundant set of transcripts by iterative pairwise merging of contig set of all tissues; this yielded 68% reduction of the contig set. A) To generate Canonical Coding Transcript Set, we selected the contigs that are annotated with MSA pipeline. The annotated contigs are further filtered for contigs that have a gene symbol. For an individual gene cluster, we chose a transcript with the longest ORF to represent the corresponding gene (Canonical Coding Transcript Set). B) For unannoated contigs, we selected for expression level, presence of an ORF with both start and stop codons in the CDS, and a minimum length of 400 nucleotides. We identified 8 high-confidence novel coding transcript candidates for validation.

The nonredundant contig set had slightly lower sensitivity, though it still remained

high; 86.60% of *M. davidii*, 85.95% of *M. brandtii*, and 95.30% of *P. alecto* transcripts were recovered. The resulting annotated contigs were assigned gene names and combined using the longest annotated ORF as the transcript. This results in an annotation for *R. aegyptiacus* that contained a total of 24,118 genes. To determine the efficiency of using the MSA pipeline, we determined that 84% (20,207 genes) of the contigs were annotated using the bat database and 16% (3,911 genes) were subsequently annotated using the primate database. These data show that the MSA pipeline, which utilizes known transcripts from related species only, is a sensitive and efficient method for *de novo* transcriptome annotation.

## Biological validity via expression analysis

We evaluated biological validity of the reconstructed transcriptome by analyzing global expression patterns across the different tissues. If the transcriptome assembly and annotations were accurate, the expression profiles of a given tissue should cluster with those of the same tissue origin and segregate from those of different origins (Brawand et al. 2011; Lee et al. 2014). A gene can result in more than one transcript isoform; therefore, to capture the highest amount of information, for each gene, we focused on the transcript with the longest open reading frame (ORF) (Figure 1.2A). After normalizing the expression values, we performed Multidimensional Scaling (MDS) to determine the relationships between the gene expression patterns in different tissues. As expected, MDS showed a clear separation of the samples according to the tissue of origin (Figure 1.3A) and explains 74% of the variance in the data.

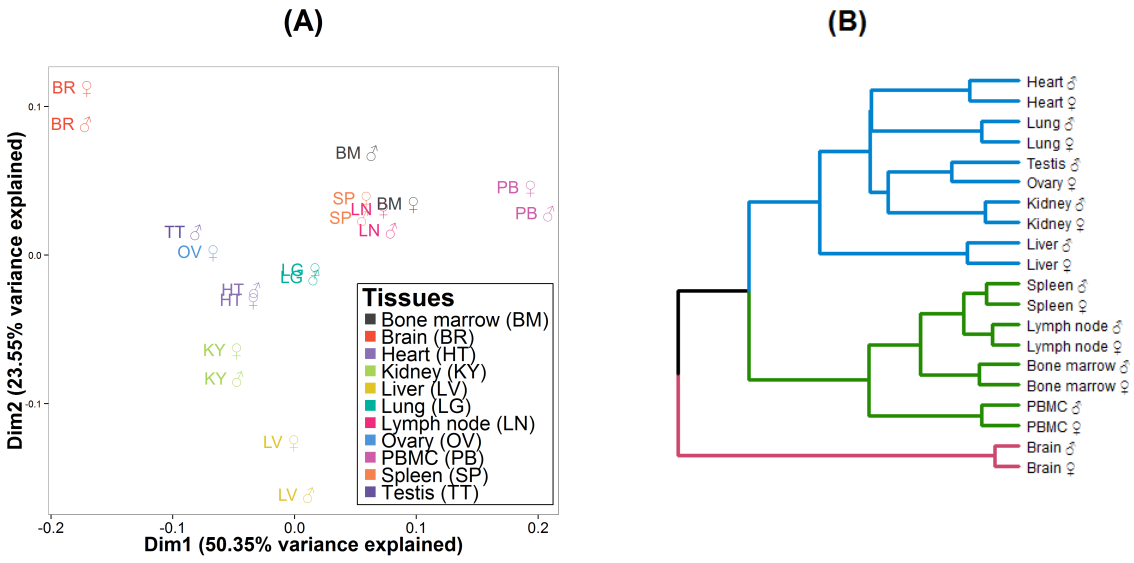


Figure 1.3: **MDS of Gene Expression Profiles of Bat Tissues**. A) We assessed the biological validity and quality of our transcriptome annotations by performing Multidimensional Scaling (using 1-spearman correlation as distance) on gene expression profiles of all tissues using the 22,398 genes as feature vector. The first two coordinates explained 73.9% of the variance in the data. B) We performed hierarchical clustering of expression profiles using 1-spearman correlation as distance. The clustering suggested presence of three groups that correspond to separate developmental origins. Tissues used are Bone (BM), Brain (BR), Heart (HT), Kidney (KY), Liver (LV), Lung (LG), Lymph (LN), Ovary (OV), PBMC (PB), Spleen (SP), Testes (TT) of the male (M) and female (F) bat.

To examine the evolutionary relationship among tissues, we performed hierarchical clustering of the gene expression profiles (Figure 1.3B). The brain, which has a different developmental pathway compared to the other organs, was classified as an outgroup. The spleen, lymph node, and bone marrow are all organs of the immune system and, as expected, clustered near each other. The peripheral blood contains some of the same cell types as the immune organs, thus, clustered near these tissues. Lastly, the gonads and kidney, which develop from the intermediate mesoderm, were grouped as neighbors in the tree. These results suggest that our transcriptome captured sufficient heterogeneity of gene expression to distinguish individual tissues while preserving their developmental relation-

ships.

## Gene ontology analysis

We further assessed biological validity of our transcriptome assembly through gene ontology (GO) analysis of tissue-specific expression profiles. We compared expression profile of each tissue with the average expression in the whole dataset, and identified the top 200 most differentially expressed genes based on a generalized linear modeling framework. Using this list, we examined the enriched GO biological process (BP) terms. Figure 1.4 shows the top 10 GO BP terms from the bone marrow, spleen, lymph nodes, and peripheral blood mononuclear cells (PBMCs). (For other tissues, see Appendix 3.5). Terms enriched for each tissue are consistent with their expected physiological functions.

## Identification of immune-related transcripts

*R. aegyptiacus* is a natural reservoir host for MARV, allowing for virus replication and dissemination with little to no pathological consequences (Amman et al. 2015; Middleton et al. 2007; Paweska et al. 2012, 2015; Swanepoel et al. 1996; Williamson et al. 1998, 2000). One important aspect of reservoir host biology is how their immune response compares to that of animal species that experiences severe disease, such as humans. Because of this, we examined the transcriptome for the presence of immune-related genes. We associated the R. aegyptiacus gene set with GO terms based on the human-specific gene ontology annotation. This resulted in 14,781 genes that mapped to 14,817 GO terms. We used CateGOrizer (Hu, Bao, and Reecy 2008) and applied the immune class GOSlim terms to

| Bone Marrow | Spleen |
|---|---|
| 1. (GO:0002376) immune system process | 1. (GO:0019752) carboxylic acid metabolic process |
| 2. (GO:0009611) response to wounding | 2. (GO:0043436) oxoacid metabolic process |
| 3. (GO:0050832) defense response to fungus | 3. (GO:0006082) organic acid metabolic process |
| 4. (GO:0007596) blood coagulation | 4. (GO:0055114) oxidation-reduction process |
| 5. (GO:0050817) coagulation | 5. (GO:0009063) cellular amino acid catabolic process |
| 6. (GO:0007599) hemostasis | 6. (GO:0044282) small molecule catabolic process |
| 7. (GO:0050878) regulation of body fluid levels | 7. (GO:0016054) organic acid catabolic process |
| 8. (GO:0042060) wound healing | 8. (GO:0046395) carboxylic acid catabolic process |
| 9. (GO:0007155) cell adhesion | 9. (GO:0044281) small molecule metabolic process |
| 10. (GO:0022610) biological adhesion | 10. (GO:0006520) cellular amino acid metabolic process |
| **Lymph node** | **PBMC** |
| 1. (GO:0046649) lymphocyte activation | 1. (GO:0007155) cell adhesion |
| 2. (GO:0045321) leukocyte activation | 2. (GO:0022610) biological adhesion |
| 3. (GO:0001775) cell activation | 3. (GO:0006935) chemotaxis |
| 4. (GO:0031295) T cell costimulation | 4. (GO:0042330) taxis |
| 5. (GO:0031294) lymphocyte costimulation | 5. (GO:0030198) extracellular matrix organization |
| 6. (GO:0002376) immune system process | 6. (GO:0043062) extracellular structure organization |
| 7. (GO:0002682) regulation of immune system process | 7. (GO:0006928) cellular component movement |
| 8. (GO:0002694) regulation of leukocyte activation | 8. (GO:0009605) response to external stimulus |
| 9. (GO:0042110) T cell activation | 9. (GO:0040011) locomotion |
| 10. (GO:0050865) regulation of cell activation | 10. (GO:0050896) response to stimulus |

Figure 1.4: **Top Ten Enriched Gene Ontology Biological Process Terms for bone marrow, spleen, lymph node, and PBMC**. In each panel, the terms are listed in descending order of significance of enrichment. These tissues, in particular are associated with different aspects of the immune system and these associations are observed within the GO BP terms identified.

identify immune-related genes from this set. Similar to previous studies in *P. alecto* and *A. jamaicensis*, we found that out of 14,817 GO terms, approximately 2.75% were associated with immune response (Papenfuss et al. 2012; Shaw et al. 2012). Amongst the most represented GO terms were cytokine production, lymphocyte activation, T cell activation, regulation of apoptosis, and regulation of lymphocyte activation (Figure 1.5).

We next searched for specific genes related to various aspects of the immune response in other mammals, primarily mice and humans. We started by evaluating the annotation of the transcriptome for the presence of anti-viral genes. A multitude of pattern recognition receptors were identified including toll-like receptors (TLRs) 1-9, RIG-I, MDA5, and

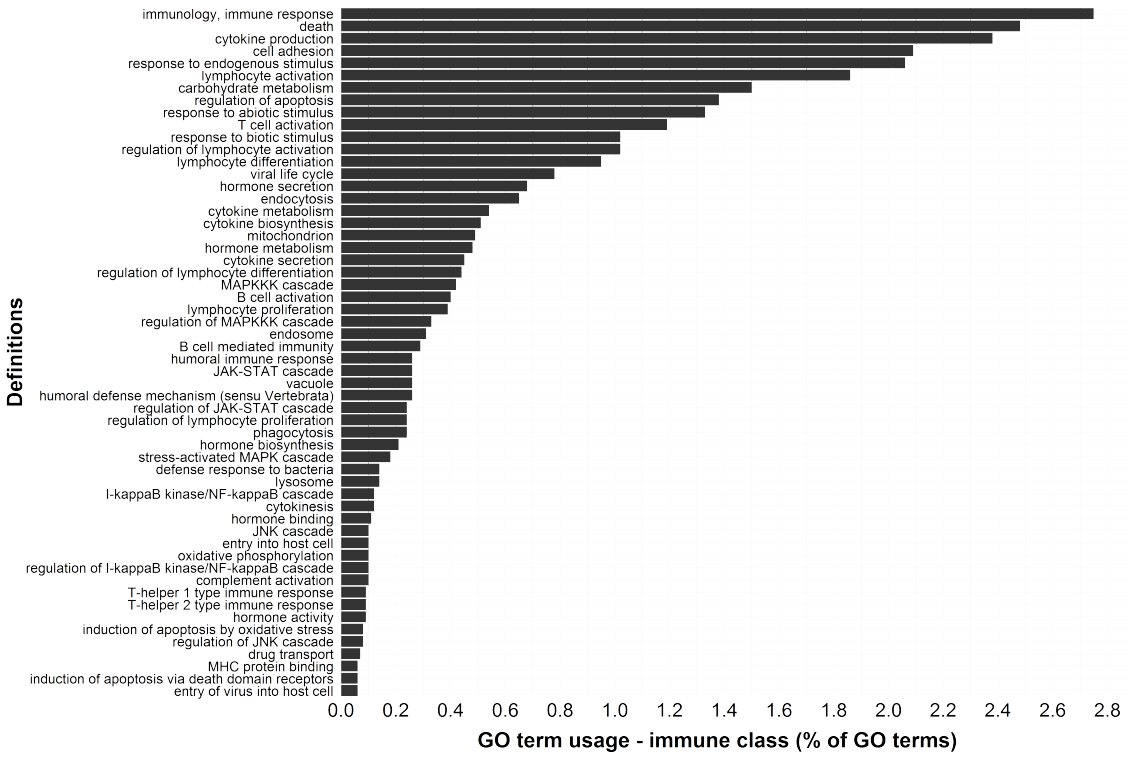


Figure 1.5: **Distribution of immune genes within the *R. aegyptiacus* transcriptome at the GO Slim level using CateGOrizer**. Genes annotated in the transcriptome were assessed for association with the immune response by analyzing them with CateGOrizer using the immune class GO Slim terms. The frequency shown is the percent of immune class GO slim terms associated with that particular pathway out of all the GO terms that were identified.

LGP2 along with the important scaffold and signaling molecules Myd88 and MAVS. A variety of antiviral molecules were also found, including Mx1 and Mx2, PKR, STING, IRF3, IRF5, IRF7, members of the IFIT and IFITM families, and ISG15. We also looked for the presence of type I, II, and III interferons (IFN). We were able to identify IFN*gamma*, IFN*gamma*2, and IFN*alpha*. Transcripts corresponding to the IFN receptor subunits IFNAR1 and IFNAR2 were also identified. IFN*alpha* and IFN*beta* have been previously characterized by cloning from stimulated cells (Omatsu et al. 2008). To eliminate the possibility that IFNB was not identified because of an impaired assembly, we aligned reads

prior to assembly to the IFNB sequence from *P. alecto* (*NCBI Eukaryotic genomes annotations*) (Figure 1.6).

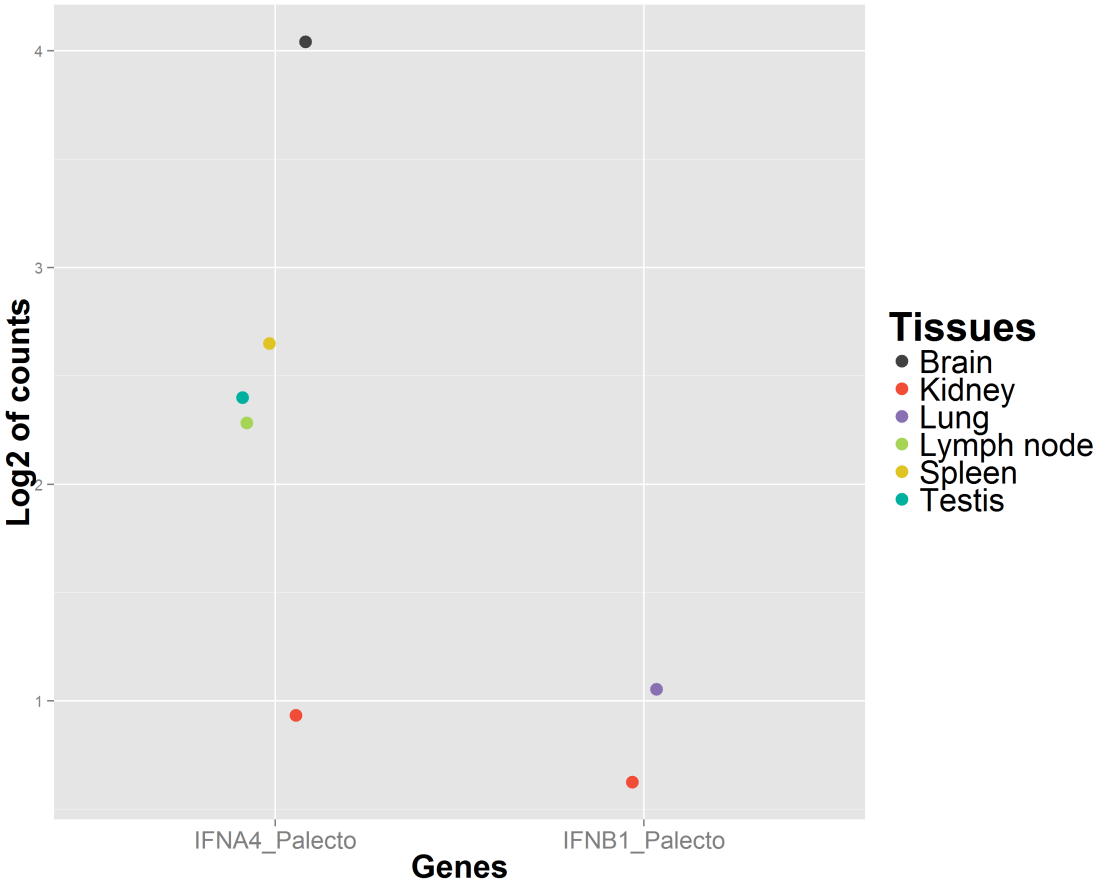


Figure 1.6: **Alignment of *R. aegyptiacus* reads to *P. alecto* transcripts**. The pre-processed reads are aligned to the interferon and immunoglobulin transcripts of *P. alecto* obtained from *NCBI Eukaryotic genomes annotations* and Papenfuss et al. 2012. The two sequences(IFNA4_Palecto and IFNB1_Palecto) used are XM_006918336.1 and XM_006918337.1, respectively.

Only 2 reads from *R. aegyptiacus* were detected which provided insufficient coverage to construct the transcript. These data suggest that IFNB expression in healthy tissues of *R. aegyptiacus* is low. This is consistent with other mammals in which IFNB is primarily expressed after exposure to a stimulus.

We then searched the transcriptome for genes associated with innate immune cells. We

found the genes CD14 and CD11c, which are useful for phenotyping macrophages and dendritic cells, as well as CD80 and CD86, which are useful for evaluating the activation status of these cells. Genes associated with natural killer (NK) cells, however, were much less obvious. We were able to identify the co-receptor CD56, but not CD16. Genes encoding for molecules in the killer cell lectin-like receptor (KLR) family, including NKG2A and NKG2D, were also not found. In other bat transcriptomes, coverage of NK cell-related genes was sparser than that of other mammals (Papenfuss et al. 2012; Shaw et al. 2012). These data suggest that expression of NK cell-related genes are either not present or present at low levels in *R. aegyptiacus* or that bats may contain a different NK cell receptor repertoire than other species.

Next, we examined the repertoire of genes associated with adaptive the immune response. We identified a variety of genes associated with T cell identification, activation, inhibition, and differentiation including CD3$\varepsilon$, CD4, CD8a, CD25, CD69, CCR7, PD-1, CTLA4, GATA3, foxp3, and Tbet. Interestingly, we were able to identify genes for the TCR$\alpha$ and TCR$\beta$ chains, but were unable to find transcripts for the TCR$\delta$ and TCR$\gamma$ chains. The transcriptome annotation for *P. alecto* included these genes, but they were at low levels (Papenfuss et al. 2012). This supports the notion that $\alpha\beta$ T cells are the predominant T cell subset in bats. We also looked at genes associated with B cells and were able to find CD19, CD20, CD27, as well as transcripts that were similar to the immunoglobulin heavy chains A, E, G, and M and the immunoglobulin light chains $\kappa$ and $\lambda$. Future analysis of the *R. aegyptiacus* genome would be useful in evaluating the immunoglobulin gene repertoire.

Finally, we studied the cytokine and chemokine repertoire, important for shaping both

innate and adaptive immune responses. We found a variety of transcripts corresponding to a wide array of both pro-inflammatory and anti-inflammatory cytokines. These included IL-2, IL-4, IL-5, IL-6, IL-12a, IL-12b, IL-17a, IL-23, IL-10, TGF$\beta$, TNF, IFN$\gamma$, IL-1$\beta$, CCL2, CCL5, and CXCL10. Altogether, the reference transcriptome generated for *R. aegyptiacus* provides an excellent foundation for investigating reservoir host immunology in bats.

## Novel Transcripts

There were 2,806,154 unannotated contigs from the nonredundant contig set (Figure 1.2B). Of those, 71.6% (2,008,503 contigs) did not have an ORF suggesting the majority of these contigs may be noncoding transcripts. To determine if the unannotated contigs were real or artifacts from the assembly, we used BLAST to align this set of contigs to the *P. alecto* genome and found that 96% (2,706,432 contigs) were aligned. To evaluate the possibility of an incomplete or impaired assembly, we grouped the aligned contigs into a total of 1,012,664 clusters based on the presence of overlapping sequences. This reduction suggests that multiple isoform expression patterns between different tissues may have affected our assembly or that our short read assembly may have been incomplete. Nonetheless, the number of unannotated contigs that aligned to the *P. alecto* genome suggests that many of these contigs, either coding or noncoding, may be novel transcripts shared within the order *Pteropodinae*. Future studies evaluating the conservation and possible functions of these sequences will be useful to determine the importance of these genetic elements. To validate novel contigs in *R. aegyptiacus* that appeared to be coding we utilized PCR. Primers

were designed to produce amplicons for eight highly expressed, unannotated contigs that contained ORFs longer than 400 bp. Using RNA isolated from the spleen, we were able to produce amplicons of the expected size from at least one bat (Figure 1.7 and Appendix 1).

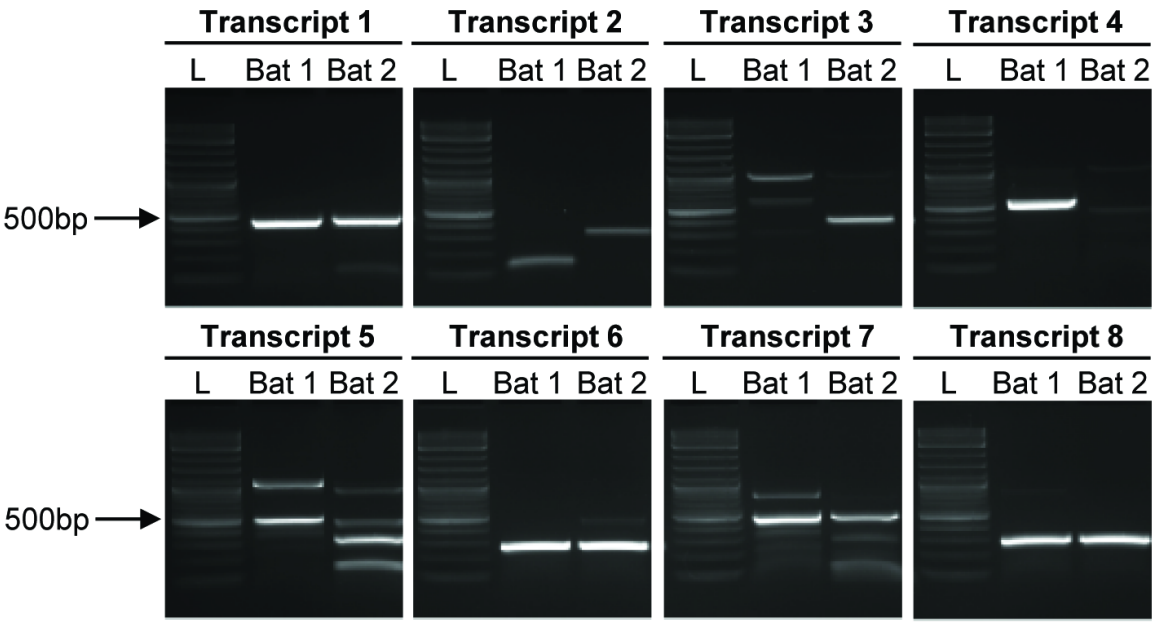


Figure 1.7: **Unannotated, novel transcripts from *R. aegyptiacus* were validated of by RT-PCR**. RNA from the spleen of both bats was reverse transcribed to make cDNA. The cDNA was amplified using primers specific for one of 8 novel transcripts that were unannotated in the assembly, but contained a complete ORF larger than 400 nucleotides. The expected product sizes were: transcript 1, 457bp; transcript 2, 450bp; transcript 3, 419bp; transcript 4, 548bp; transcript 5, 469bp, transcript 6, 277bp; transcript 7, 507 bp; and transcript 8, 301bp.

The sequences of these amplicons were found to match the expected sequence from the assembled ORF of the unannotated contig. These contigs also showed high sequence similarity to the *P. alecto* genome. In particular, six of the 8 validated transcripts showed sequence similarity higher than 75% at a query coverage greater than 64. The other two validated transcripts had a query coverage of 23 with 78.36% identity (transcript 1 in Figure 1.7) and a query coverage of 7 with 91.27% identity (transcript 2 in Figure 1.7) (Appendix 3.5); therefore, we hypothesize that these transcripts might be specific to *R. aegyp-*

*tiacus*. Further investigation is needed to fully understand the characteristics and biological functions associated with the proteins these contigs encode.

## 1.3 Conclusion

In this section, we presented the comprehensively annotated of transcriptome of *R.aegyptiacus* and assessed its quality and biological validity. This transcriptome will be an important resource to study bat immunology. In particular, it will facilitate the process of investigating differences in host responses between asymptomatic reservoir host species and species that exhibit severe pathology. It will also pave the way for the development of novel therapeutics and prevention approaches against emerging zoonotic virus outbreaks.

## 1.4 Material and Methods

### Sample preparation

Tissues and blood were collected from one male and one female adult *R. aegyptiacus* bats that were bred and housed at the colony established at the Center for Disease Control and Prevention, Atlanta, GA, USA (Amman et al 2015). Approximately 100mg of the following tissues were collected and homogenized in 1mL of Trizol LS (Invitrogen, Carlsbad, CA): liver(bat id:BAT7, BAT17), lung(BAT05, BAT15), heart(BAT03, BAT13), kidney (BAT04, BAT14), brain(BAT02, BAT12), axillary lymph nodes (bilateral, pooled) (BAT06, BAT16), spleen(BAT10, BAT19), bone marrow(BAT01, BAT11), and gonad(BAT08, BAT20). PBMCs(BAT08, BAT18) were isolated from the blood and stored in

Trizol LS as well.

RNA was extracted using the PureLink RNA Mini kit (Invitrogen, Carlsbad, CA). cDNA was synthesized using the TruSeq Stranded Total RNA Sample Prep Kit (Illumina, San Deigo, CA) according to the manufacturer's protocol. The libraries were evaluated for quality using the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA). After quantification by real-time PCR with the KAPA qPCR Kit (Kapa Biosystems, Woburn, MA), libraries were diluted to 10nM. Cluster amplification was performed on the Illumina cBot and libraries were sequenced on the Illumina HiSeq 2500. Eight of the female bat libraries were single-end, while the remaining tissues from the female bat and all tissues from the male bat were paired-end. All of the libraries sequenced were 125 bp in length. The average library depth was 66M reads (minimum 16M and maximum 98M).

## Ethics Statement

All experimental procedures were conducted with approval from the Centers for Disease Control and Prevention (CDC, Atlanta, GA, USA) Institutional Animal Care and Use Committee, and in strict accordance with the Guide for the Care and Use of Laboratory Animals (Committee for the Update of the Guide for the Care and Use of Laboratory Animals 2011). The CDC is an Association for Assessment and Accreditation of Laboratory Animal Care International fully accredited research facility. No human patient-derived clinical materials were used in these studies.

## *De novo* transcriptome assembly

We first examined the quality of the reads using FastQC v0.11.3 (*FastQC*). We also preprocessed the reads to remove the adapter sequence using cutadapt v1.5 (Martin 2011). We removed "AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC" from the forward strand and "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCG-GTGGTCGCCGTATCATT" from the reverse strand. We performed strand-specific de novo transcriptome assembly using Trinity r20140413p1 (Grabherr et al. 2011) with the parameters: "–normalize_reads" and "–SS_lib_type FR", along with its default parameters for all of our samples.

## Homology based annotation of the transcriptome

For annotation of contigs and clustering them into a gene model, we used Multiple Species Annotation pipeline, an nucleotide-based annotation approach that is more efficient and faster than BLASTX (Lee et al. 2014). To make a BLAST (Altschul et al. 1997) database for bats, we started with the complete "Nucleotide collection" (nt) database. We exported all accession numbers of the bat sequences at NCBI and made a subset database from nt using "blastdb_aliastool -db nt -dbtype nucl -gilist bats.sequence.gi.txt -title Bats -out Bats". Using the same type of query, we also created a database for primates including humans due to their extraordinarily well-annotated transcriptomes, which will maximize the power of our annotation pipeline. We then used BLAST to iteratively align the contigs to the bat db, the primate db, and finally nt using a subtractive approach: what did not align to the bat db was aligned to the primate db, and what did not align to the primate db was

aligned to nt.

## Sensitivity of *R.aegyptiacus* transcriptome

To assess the coverage of our transcriptome, we downloaded the *M. davidii*, *P. alecto*, and *M. brandtii* transcriptomes from NCBI Eukaryotic genomes annotations (*NCBI Eukaryotic genomes annotations*). We generated a BLAST index out of union of all contigs from our samples, and aligned the three bat contigs to our BLAST databases. We chose the alignment with 70% of sequence identity with maximum evalue of 1e-4.

## Nonredundant transcriptome assembly

To generate a nonredundant set of contigs, we iteratively merged individual assemblies using the the methods similar to the (Robertson et al. 2010) employed to merge the kmers. Using CD-HIT-EST v4.6 (Li and Godzik 2006) with sequence identity threshold of 0.99, we merged the first two pairs of contig sets (of sample $i$ and sample $i+1$) upto the final sample $n$. After each iteration, we merged the resulting merged contig sets using a similar approach until only one contig set remained.

## Canonical Coding Transcript Set

For the expression profiling, we generated a reference transcriptome consisting of transcripts each representing a gene model according to the following method: We first used TransDecoder(r20140413p1) (*TransDecoder*) to find the ORF of all transcripts. Then, based on the MSA pipeline, we chose a transcript with gene symbols and the longest ORF

in each gene cluster to capture the most information for downstream expression analysis. We did not consider the contigs mapped to nt database in this manuscript because obtaining feature files for all sequences in as a requirement of MSA pipeline was computationally impractical and a majority of the gene symbols (24,118) are captured in the bat and primate databases.

## Gene expression and Gene Ontology analysis

After a canonical transcript set was obtained, we used this as a transcriptome reference for expression analysis. We mapped the preprocessed reads to this reference using RSEM v1.2.19 (Li and Dewey 2011) and obtained a gene-to-count matrix. We removed the transcripts with expression variance equal to zero or with low expression (count $<= 10$). For MDS plot, we used the spearman correlaton as a distance measure and "cmdscale" from the "stats" package in R (R Core Team 2015). To explore the biological processes in each gene expression profile, we employed a one-to-all sample comparison using the EdgeR generalized linear model framework (McCarthy et al. 2012; Robinson, McCarthy, and Smyth 2010b) . For each tissue, we compared individual gene expression within the tissue versus the average expression of all other tissues. With each tissue having differently ranked gene lists, we then selected top 200 genes and ran gene ontology analysis using topGO (Alexa and Rahnenfuhrer 2010) with human-specific gene ontology annotation (Ashburner et al. 2000).

## Analysis of unannotated transcripts and Identification of novel transcripts and validation

We used BLAST (Altschul et al. 1997) to align unannotated contigs to the genome of *P. alecto* with the evalue of 1e-4 and query coverage of 40%. To cluster the aligned contigs into groups, we used bedtools (Quinlan and Hall 2010) setting the distance threshold parameter at 0. For transcripts that did not align with any similarity to bat, primate, or nt BLAST databases, we applied a series of filters to select for the coding transcripts to be validated. We used the following criteria: an ORF that was complete with both a start and stop codon, an ORF that was at least 400bp in size, and a transcript that was expressed (a read count $> 0$). We further selected for the novel transcripts with usuable primers using primer-BLAST (Ye et al. 2012). Using these criteria, the number of novel transcripts was narrowed down to a total of 8. The primers and expected amplicon size are listed in Appendix 1.

For validation, RNA was extracted from the spleen tissue of both the male and female bats using Trizol LS (Invitrogen, Carlsbad, CA). cDNA was synthesized from 2.5 $\mu$g of RNA using the Superscript III First-strand Synthesis SuperMix (Invitrogen, Carlsbad, CA). Amplicons for each of the primer sets were generated using Phusion HotStart Flex DNA polymerase (New England BioLabs, Ipswitch, MA) and run on a 1.5% agarose gel for visualization. The correct size amplicon was gel extracted, quantified, and Sanger sequenced on the Applied Biosystems 3730x1 DNA Analyzer.

# Chapter 2

## *Multilevel Analysis of Gene Expression*

Although the price of the high-throughput genomic and transcriptomic experiments have dropped significantly over the years, the number of libraries used in RNA-seq samples remain relatively small because of its exploratory nature, budgetary constraints, and ethical issues involving animal samples. Typical the number of biological replicates obtained for individual groups in RNA-seq experiments is three (Conesa et al. 2016), and many techniques have been developed to address this so-called "small n high p problem" such as edgeR or DESeq. However, as the experiment becomes more complex with increasing number of factors, which are interacting among themselves, it calls for a more sophisticated technique.

In this chapter, I will describe the motivation of statistical framework titled Multilevel Analysis of Gene Expression (MAGE) which collaborators and I developed to address the aforementioned challenges, in particular, the RNA-seq study involving the serially sacrificed rhesuses infected with Marburgvirus specified in Chapter 3. We also provide the theoretical justification and validation via posterior predictive checks for MAGE.

## 2.1 Background

### Traditional approach to tackling Gene Expression Analysis

Despite the ease of ability to quantify the global gene expression with the help of microarray or RNA-seq, it comes with another set of challenges. The data generated from high-throughput sequencing technology is stochastic and therefore robust statistical method is needed to estimate the underlying parameters of interest.

Normalization is one issue when different experimental units have different exposures, and one must adjust them for the fair comparison. However, the more important issue is the small number of replicates within each group. Even though it was possible to measure thousands of genes, only a handful number of samples, usually three (Conesa et al. 2016), could be used for an experiment because of a fixed budget and ethical issues. This results in lack of power and susceptibility to erroneous signals.

The numerous statistical tools that have been developed to analyze the high-throughput gene expression data both for microarray and RNA-seq (Auer and Doerge 2010). For microarray, the unit of measurement is the intensity whereas for RNA-seq gene expression are measured in the form of read counts. In a microarray, the gene expression value ranges reside in real numbers, but with RNA-seq, the read counts are in the non-negative integers. The majority of gene expression analysis tools for RNA-seq such as EdgeR, DESeq2, limma-Voom, and Sleuth uses employs a generalized linear modeling framework (Degexp et al. 2010; Law et al. 2014; Love, Huber, and Anders 2014; Pimentel et al. 2017; Robinson, McCarthy, and Smyth 2010a) with underlying likelihood model of Poisson or Negative Binomial with a factor with $K$ levels in which $k$-ith element corresponds to distinct biologi-

cal states such as normal, cancer, drugged, etc. In general, the parameters of interest are the means of K groups for a particular gene. To compare the means, the $log2(\frac{mu_2}{mu_1})$ is computed and its significance is used to determine if it's worth further exploration for the downstream analysis such as GSEA (Subramanian et al. 2005).

The challenge associated with these algorithms is to estimate the parameters of the model with the small number of experimental units. The core algorithm for DESeq (Degexp et al. 2010) or EdgeR (Robinson, McCarthy, and Smyth 2010a) is designed to accurately estimate the dispersion parameter in a negative binomial assuming that genes with the similar expression levels have more or less similar dispersion parameter. Voom (Law et al. 2014) and Sleuth (Pimentel et al. 2017) first log transform the count data so that the reads are in the shape of the normal distribution and then regularize variance via shrinkage approach roughly similar to DESeq.

All these tools are usually built for the case when the factor of interest is one. As the cost of RNA-seq drops, investigators started to perform experiments with more complex designs. When one has a multiple covariates such as time and tissue, the traditional linear model with the small sample size falls short. Another weakness of these classic tools is that they provide the estimated parameters in the form of point estimates. If one tries to explore the transformed parameters, the uncertainty might not be adequately accounted for.

Another variable of concern in the gene expression analysis is time. Since the biological process is a dynamic system, there is an immense interest in understanding the relationship between time and gene expression. For example, one may be interested in how a set of genes that change upon infection of a particular virus. When there is a dimension of time in one's gene expression data, however, the model becomes more complex to capture the

variance embedded in the temporal information (Bar-Joseph, Gitter, and Simon 2012; Spies and Ciaudo 2015). The classic tools mentioned in the previous section are optimized for the static gene expression analysis. There have been approaches to tackle time course RNA-seq data using approaches such as splines (Storey et al. 2005), Bayesian approach (Aryee et al. 2009), Gaussian Processes (Bar-Joseph et al. 2003; Heinonen et al. 2014), or mixed effect model (Sun et al. 2016). However, the majority of these tools have not been specifically developed for the RNA-seq data. Most of the time course data were done using microarray technologies (Aryee et al. 2009; Storey et al. 2005) or qPCR (Heinonen et al. 2014) and only for the case involving two groups (Sun et al. 2016). Therefore, the need for the flexible tool for analyzing the RNA-seq experiment with the complex experimental design is warranted.

## 2.2 Multilevel Analysis of Gene Expression

To address the challenges of analyzing an RNA-seq experiment with a complex experimental design involving many parameters, we developed the novel high-throughput gene expression analysis pipeline titled Multilevel Analysis of Gene Expression(MAGE). MAGE attempts to tackle gene expression analysis problem using a fully Bayesian approach and multilevel(hierarchical) modeling technique. This approach is useful in incorporating prior information, analyzing the quantity of interest to address complex statistical questions, and capturing the uncertainty inherent in the noisy system. With MAGE, one can take the set of raw RNA-seq read inputs and transform them into the full Bayesian posterior samples of the parameters of interest, with which one can perform useful statistical inference on

biological questions at hand. It combined the best practices of gene expression analysis, Bayesian statistics, and multilevel(hierarchical) modeling. Thanks to the advent of probabilistic programming languages such as Stan (Carpenter et al. 2016), it is also easy to improve and extend the underlying model. This has an additional beneficial effect in that researchers do not have to learn new and different tools to try multiple different models. Instead, they can use the metrics such as Leave-one-out cross-validation (LOO) and the widely applicable information criterion (WAIC) to compare and evaluate separate models to choose the best model (Vehtari, Gelman, and Gabry 2017). In Figure 2.1, the schema of MAGE pipeline is shown.

## Motivation of each component of MAGE

### Gene Expression Analysis

MAGE follows the typical framework of RNA-seq data analysis (Degexp et al. 2010; Robinson, McCarthy, and Smyth 2010a in that it models the count data with the generalized linear models (GLM) (Nelder and Baker 1972), in particular, negative binomial distributions. The benefit of this approach–as opposed to log transformation of counts, the approach taken by Law et al. 2014; Pimentel et al. 2017–is it has excellent theoretical framework for modeling count data (Nelder and Baker 1972) while accounting for the biological variability via overdispersion parameter. Furthermore, GLM provides you with the interpretability of estimated coefficients. For example, in the one factor model with two levels, the coefficient of a level corresponds to the log fold change with respect to the

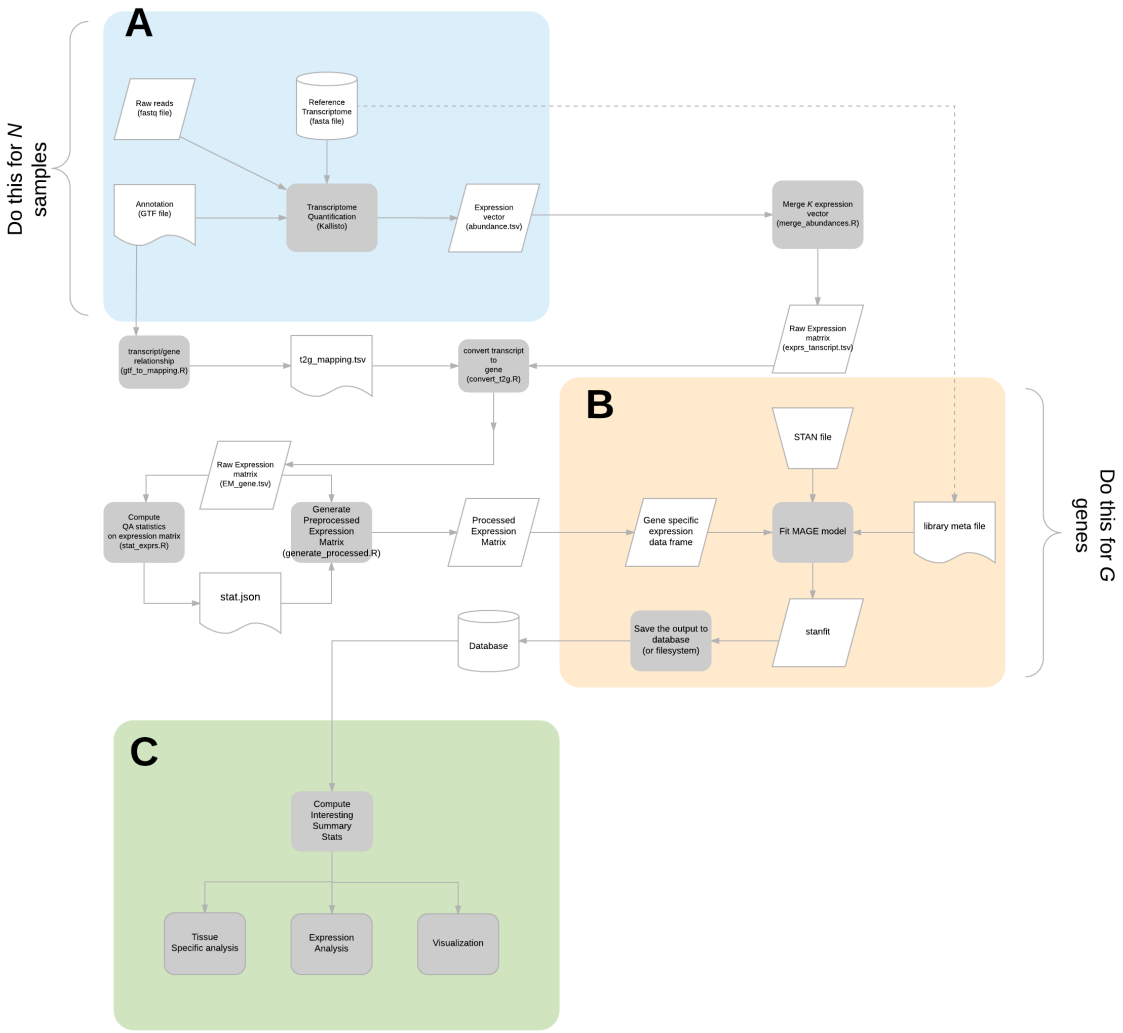


Figure 2.1: **MAGE Pipeline Schema**. A) Gene expression quantification B) Probabilistic Model Fitting C) Inference. An arrow represents dependency relationship. A parallelogram icon represent the output file from the previous step. a grey-colored round rectangle represents a particular process, implemented in a script. A document icon represents a reference file obtained either externally or from another file

reference group[1]. This ease of interpretability is important for classification purposes for the downstream analysis such as GSEA or pathway analysis.

### Bayesian Statistics

The Bayesian statistical framework comes with many benefits such as incorporating prior knowledge, regularizing parameters, having interpretability and extensibility (Gelman and Hill 2007; Gelman et al. 2014). In frequentist approach, parameters are considered as fixed constant values; in contrast, in Bayesian, parameters are viewed as random variables that have probabilistic nature. By adding prior information for parameters, one can incorporate known information that might not be present in the data or have parameters share information among one another in a robust manner. In Bayesian statistics, the first class object is the posterior probability of parameters given the data. With a posterior distribution, rather than a point estimate, one can transform the parameters however one wants to address the questions at hand statistically and compute expected values to answer interesting questions. Furthermore, incorporation of prior information becomes almost necessary when one tries to estimate a lot of parameters as a result of complex structure with not enough data to prevent under- or over-fitting (Gelman and Rubin 1995). Our goal is the accurate estimation of parameters of interest with the small number of replicates–a typical scenario in the high throughput gene expression data analysis–and without prior information, the estimates become unreliable.

[1] If $\mu_x = exp(b_x)$ and $\mu_y = exp(b_x + b_{x-y})$, then $b_{x-y} = log(\mu_y/\mu_x)$.

### Multilevel modeling

When there is multiple group information associated with data, as is the case with the data with complex experimental design, the number of parameters to estimate increase. This is because the system is usually modeled with the set of indicator variables whose cardinality increases as the number of levels increase. With more than one factor, estimation of parameters become unstable without imposing additional structure (Gelman et al. 2014). MAGE attempts to address this issue by employing multilevel modeling–also known as hierarchical modeling– where information from covariates are shared to stabilize the inference (Gelman and Hill 2007; Gelman et al. 2014). With multilevel modeling, there are usually fewer parameters are necessary to fit and the particular characteristics of the groups are learned from the data (McElreath 2015). Additional benefits of multilevel modeling include, but not limited to, its ability handle imbalance data, which is prevalent in the real world, stable and accurate estimation of parameters, and accounting for the variations between and within groups (ibid.). In addition, multilevel modeling can be viewed in the framework of the bias and variance trade-off. Ignoring the group structure leads to highly bias estimate whereas estimating parameters for individual group naïvely can lead to high variance. The former is called "complete pooling" approach whereas the latter "no-pooling"(Gelman et al. 2014). Multilevel modeling targets for something in between in a theoretically sound manner.

We attempt to bring this idea to the development of MAGE, combining Bayesian statistics and multilevel modeling to the high-throughput gene expression analysis. The advantage of MAGE is that it is general, flexible to effectively analyze the high-throughput

sequencing experiments with the added interpretability delivered by the accurate estimation of relevant parameters.

## Data description

As shown in Figure 2.1, the input data of the first step of the MAGE pipeline is the fastq files of $N$ libraries (Figure 2.1A). Each library $n$ can be first quantified by using RNA-seq gene expression quantification program such as Kallisto. The result of this operation is the vector of length $G$ where each element corresponds to the expression level of the gene $g$ for the library $n$ in the form of count. $N$ number of vectors can be combined into a $G \times N$ expression matrix. There is a separate table that contains the meta information for individual $s$ libraries such as library size, tissue, or time labels. The expression matrix and the meta table are the input for the MAGE model described in the following section.

## Model specification

MAGE has a following generative model specification [2] (Figure 2.2). In a typical high throughput gene expression analysis, we have $G$ genes, each of which are represented as a vector of length $N$ which is the number of libraries. We fit $G$ separate models for individual genes $g$. The generative model is that for each of the sample $n$ in each gene $g$, we draw $y_{g[n]}$ reads from Negative Binomial(NB) distribution with the mean $\mu_{g[n]}$ and the inverse dispersion parameter $\phi_g$.

[2]The initial version of MAGE was built for the study described in the Chapter 3, but the model can be modified easily for different studies with different experimental designs.

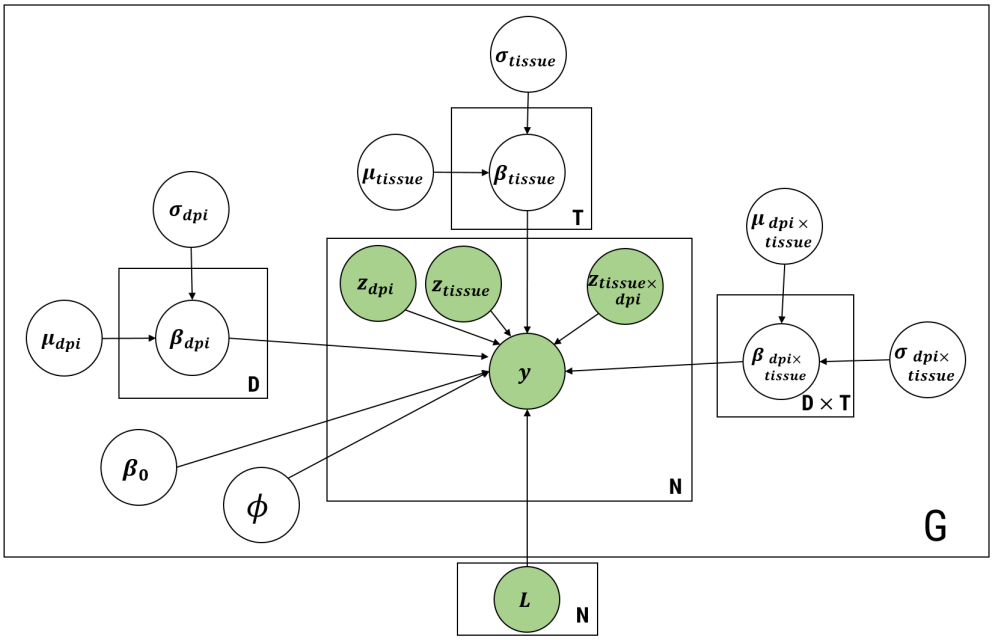


Figure 2.2: **Graphical Model of probablistic model used in MAGE**. Standard plate notation is employed. A empty node represents a random variable and a colored node represents the observed variable or meta information. An arrow between nodes represents a conditional relationship in which the target node is conditinally dependent on a source node. A plate represents a group of variables with its cardinality shown in the bottom right corner. There are $N$ libraries, $G$ genes, $D$ Day Post Infection (DPI) groups, $T$ tissues, and $D$ *x* $T$ interaction terms. $L$ represents a (effective) library size for $n$th library. For a particular library $n$ and gene $g$ , there are $y$ a read, $\phi$ a dispersion parameter, $b_0$ a intercept, and other betas the group specific parameters.

$$y_{g[n]} \sim NB(\mu_{g[n]}, \phi_g)$$

We choose NB to model because 1) $y_{g[n]}$ is realized as count data, 2) can account for overdispersion induced by the biological variability, and 3) a popular choice of probability distribution to model RNA-seq count data in many packages (Love, Huber, and Anders 2014; Nueda, Tarazona, and Conesa 2014; Robinson, McCarthy, and Smyth 2010a). The dispersion parameter can be either a plug-in estimate(via a package such as edgeR) or estimated from the data as a parameter in the MAGE model. The particular parameterization

used is as follows:

$$NB(y\,|\,\mu,\phi) = \binom{y+\phi-1}{y}\left(\frac{\mu}{\mu+\phi}\right)^{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}.$$

which has the following mean and variance specifications.

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \text{Var}[Y] = \mu + \frac{\mu^2}{\phi}.$$

We assume that the mean $\mu_{g[n]}$ has a following structure that the support of the mean within the non-negative range:

$$\mu_{g[n]} = L_{s[n]} exp(\theta_{g[n]})$$

Where $L_{s[n]}$ is the depth of the library $s$ for a $n$th sample and $\theta_g$ a linear sum of predictors. In other words, we use the expontential function as a link function with a offset term $log(L_{s[n]})$, which is equivalent to a model-based normalization approach employed in edgeR (Robinson, McCarthy, and Smyth 2010a). One can use the raw library size or TMM-normalized effective library size (Oshlack, Robinson, and Young 2010). We set $\theta_{g[n]}$ as follows.

$$\theta_{g[n]} = \beta_0 + \beta_{z_{\text{dpi[n]}}} + \beta_{z_{\text{tissue[n]}}} + \beta_{z_{\text{dpi x tissue[n]}}}$$

Here, we denote dpi[n] as one of $D$ Day Post Infection (DPI) labels, tissue[n] as one of $T$ tissue labels, and dpi x tissue[n] as one of $DxT$ DPI and tissue labels for $n$th sample. The model assumes that there is a baseline effect $\beta_0$(intercept) and the rest of the terms are

the deflections based on the membership of a sample to a particular group, which can be obtained from the meta table. In the model, there are two main effects--dpi and tissue-- and the interaction between the two factors. The interaction term is necessary because there is the substantive reason to believe that the different tissues will have different expression profiles across days. For individual coefficients other than intercepts, we put the hierarchical structure on the prior which has the effect of learning them adapatively from the data (McElreath 2015), which is one of the novel contributions to the existing approaches (Love, Huber, and Anders 2014; Robinson, McCarthy, and Smyth 2010a).

$$\beta_0 \sim N(0, 10)$$

$$\beta_{\text{dpi[n]}} \sim N(\mu_{\text{dpi}}, \sigma_{\text{dpi}})$$

$$\beta_{\text{tissue[n]}} \sim N(\mu_{\text{tissue}}, \sigma_{\text{tissue}})$$

$$\beta_{\text{dpixtissue[n]}} \sim N(\mu_{\text{dpixtissue}}, \sigma_{\text{dpixtissue}})$$

Here we use a weakly informative prior for the intercept; For standard deviation[3], we chose the value 10, which is a relatively large number in a log scale. We did this to accommodate the dynamic ranges of various gene expressions and different library sizes. With the library of a million reads $10^6 * exp(-10)$ corresponds to 45 read counts. Regardless of the interpretation, the choice of a particular number in the weakly informative prior has the little influence on estimation of parameters (hence the adverb "weakly") (Gelman et

[3]In Stan, the normal distribution is parameterized with mean and standard deviation, as opposed to mean and *variance*

al. 2014; Kruschke 2013). For the non-intercept terms, we put the multilevel structure to learn the parameters from the data. For hyper-parameters, we used the following weakly informative priors for regularizing effects (Gelman et al. 2014).

$$\mu_{\text{dpi}}, \mu_{\text{tissue}}, \mu_{\text{dpixtissue}} \sim N(0,5)$$

$$\sigma_{\text{dpi}}, \sigma_{\text{tissue}}, \sigma_{\text{dpixtissue}} \sim halfCauchy(0,2)$$

We chose 5 for standard deviation to stay conservative in effects of the factors. We chose the half-cauchy distribution to model all the scale parameters, which can be viewed as a t-distribution with one degree of freedom, for its flexibility of handling not only the near origin but other parameter spaces to capture outliers and is a recommended choice for default prior for variances (Gelman 2006; Polson and Scott 2012).

## Computation

To implement and fit the model to the data, we used Stan, a probabilistic programing language which compiles a given model to C++ and uses an Hamiltonian Monte Carlo (HMC) engine to draw posterior samples from the given model (Carpenter et al. 2016). In Stan, we used nonstandard parameterization to speed up the sampling process. The stan code is attached in the Appendix 3.5). We find that 1000 from 4 chains are sufficient to achieve MCMC convergence for our datasets; however this could be different for different datasets.

## 2.3 MCMC Convergence Diagnostics

Since we obtain the posterior distributions from the iterative simulations of Hamiltonian Monte Carlo sampling via Stan, we must assess whether MCMC sampling converged to a target joint distribution. The standard way of doing this is the visual diagnostics via trace plots. The central idea is to run multiple chains of MCMC and see if they mix well, which is a sign of the well covered posterior distribution of a given parameter. If the mixing does not occur, it is a sign of pathological cases. However, to do it for thousands of genes is not scalable and intractable; Alternative heuristics that can be used to diagnose MCMC convergence is to compute and monitor the potential scale reduction statistics called Rhat (Gelman and Rubin 1992)–the ratio of within chain variance to between-chain variance –for each parameter using samples from multiple chains. If the convergence has occurred, this statistics should be close to 1. The rule of thumb is that each of the model parameter should reach Rhat below 1.02 (Brooks and Gelman 1998).

The most important set of parameters of interest is the fold change. In Figure 2.3, we see that the majority of Rhat values computed are below the recommended value of 1.02.

## 2.4 Posterior Predictive Checks

The gold standard way of validating your model is to test the model against the unseen future data. This way we know that our model has learned the general data generating process. Usually, this is done via form of train-test data separation, leave one out cross validation, or k-fold cross validation (Friedman, Hastie, and Tibshirani 2001). However, this is not feasible in our case we only have a few data points and it is very costly to generate

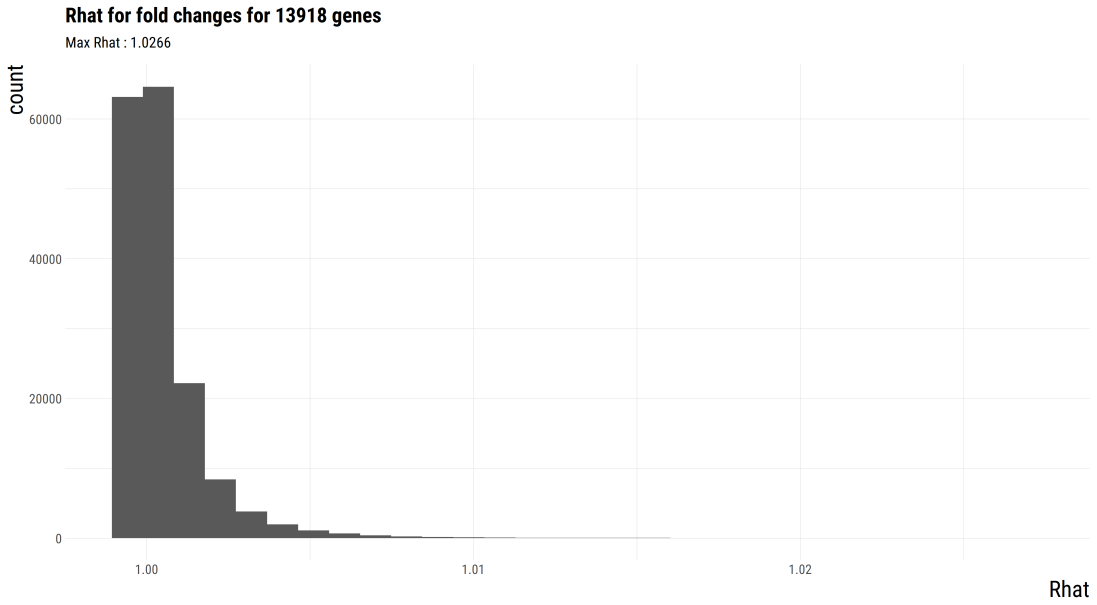


Figure 2.3: **MCMC convergence diagnostics**. Rhats for 167,016 fold changes(12 fold changes for each of 13918 genes) are below the recommended convergence threshold of 1.02, which suggests the MCMC convergence.

the data. Another approach to testing your model is to do simulation and assess type 1 and 2 error, which is a common way to validate your model. However, in our case, it's not that simple because we do not know the underlying data-generative process.

Posterior Predictive Checks is a way to validate one's model (Gelman et al. 2014). The idea is to compare the simulated samples from a proposed model to observed data to see if the posterior distribution well approximates a target distribution. If the simulated samples are drastically different from the data, it suggests that the current model needs an improvement.

In Figure 2.4, we plot the qqplot between predictive distribution and the empirical count for DDX58, NPC1, and group of genes. Predictive distributions of read counts are obtained by simulating the read count from the parameters being estimated (See "generated quan-

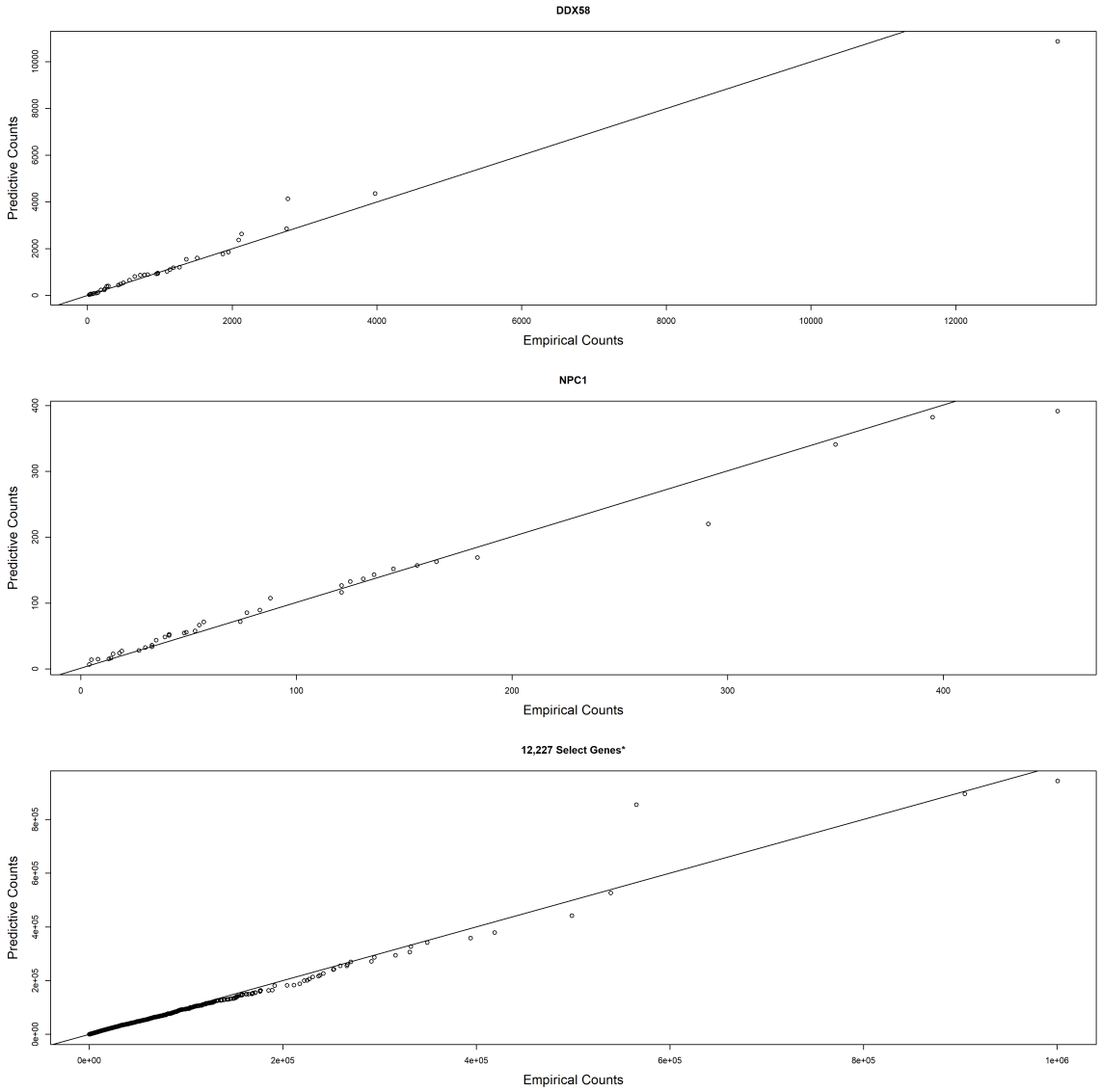


Figure 2.4: **Positive Predictive Checks for DDX58, NPC1, and Group of Select Genes**. In each panel, qqplot is plotted between empirical counts and predictive counts sampled from posterior distribution. The top, middle, and bottom panel shows the qqplot for DDX58, NPC1, and group of selected genes. Select genes are the genes in which 50% are samples have expression greater than 0 read count

tities" in Stan code 3.5). We observe that the simulated data generated from our model are more or less similar to the actual data for genes that have low expression (NPC1), high expression (DDX58), and average genes, which have more than 50% non-zero read count). The prediction of MAGE overlaps well with the actual data, which suggest that the model has accurately captured the structure well and be trusted for the inference in the downstream analysis.

## 2.5 Benchmark

To assess the validity and performance of the Bayesian approach taken by MAGE pipeline, the synthetic expression data with two experimental conditions was generated consisting of genes with known parameters selected from following:

- Mean expression level $E$ ranging between 10 and 1000 with the step size of 5
- Dispersion $D$ ranging between 0.5 and 2 with the step size of 0.1
- Expression state $S$ of a gene: "no change", "up regulated"(2 fold change), "down regulated"(0.5 fold change) with respect to individual gene's corresponding baseline mean expression value.

Furthermore, different experimental settings were generated based on the combination of following parameters:

- Proportion of differentially expressed genes $P$: 1%, 10%, or 50%
- Number of biological replicates per condition $N$: 3, 10, and 50

The code to simulate the data is in Appendix 3.5. The idea is to test the performance of MAGE in three different settings: easy, middle, difficult where $N$=50, 10, and 3, respectively using one factor model in both MAGE and EdgeR, a classic and popular gene expression analysis package to compare the performances (Robinson, McCarthy, and Smyth 2010a). EdgeR is shown to have a high sensitivity and a specificity compared to other gene expression analysis packages (Rapaport et al. 2013).

With EdgeR, the default GLM workflow was used, and for MAGE, the same hyperparameters and likelihoods as the previous section were used. The code used in the benchmark is in Appendix 3.5.

The important parameter of interest in applications that motivated the development of MAGE is a fold change (FC) between conditions because it can be used to perform differential gene expression analysis and dimensionality reduction (See the Section 2.6 and also Chapter 3). Therefore, the mean absolute error (MAE) with respect to the known FC is chosen to be the metric in the benchmark:

$$\text{MAE} = \frac{\sum_{g=1}^{G} \left| FC_{g,\text{real}} - FC_{g,estimated} \right|}{n}$$

where $G$ is the total number of genes. In EdgeR, the MLE estimate of FC was used and for MAGE, posterior median FC was used. MAE is computed for various expressions settings defined above. We computed each MAE using the three different simulations to add confidence intervals.

Output of the benchmark is shown in Figure 2.5. While both methods struggled when the number of biological replicates($N$) was low, MAGE had lower MAE in all different

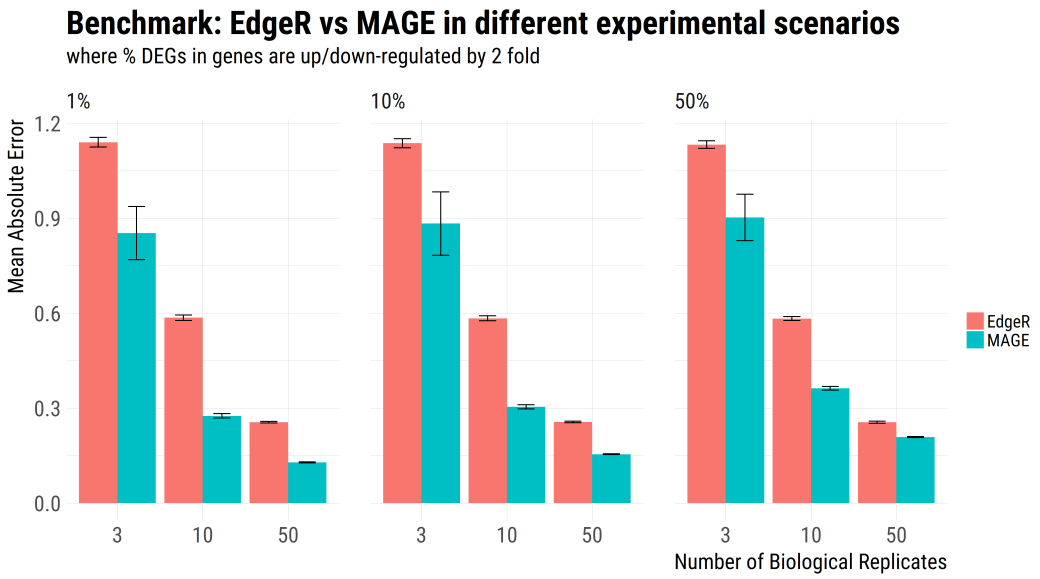


Figure 2.5: **Benchmark of MAGE Performance Using Synthetic Data Compared With EdgeR**.

possible configurations of $N$. The fact that MAGE consistently outperformed EdgeR in each setting in this simple benchmark of one factor model provides a confidence on the accuracy and robustness of the Bayesian approach employed by MAGE.

## 2.6 Applications

With the full posterior distributions, transforming the parameter and obtaining the Monte Carlo estimate of the expected value is straightforward:

$$E_{Pr(q|D)}[t(q)] = \int t(q)Pr(q|D)dq \approx \frac{1}{S}\sum_{s=1}^{S} t(q_s)$$

Where $q$ is a parameter of interest in the model, $t$ is a function, $S$ is the number of MCMC draws. That is, the expected value of any transformed parameter is a function of

the posterior distribution, and can be easily and accurately estimated with MCMC samples (Betancourt and Girolami 2015; Carpenter et al. 2016).

Armed with this knowledge, one can do all the classic tasks such as computing the fold change. For example, log2 fold change between dpi3 and dpi1 in Tracheobronchial Lymph Node (TbLN) can be computed with following:

$$P(log_2(FC^{TbLN}_{dpi3})|D) \approx \frac{1}{S}\sum_{s}^{S} \frac{((\beta^{dpi3}_{(s)} + \beta^{TbLN}_{(s)} + \beta^{dpi3xTbLN}_{(s)}) - (\beta^{dpi1}_{(s)} + \beta^{TbLN}_{(s)} + \beta^{dpi1xTbLN}_{(s)}))}{log(2)}$$

where $D$ represents the data, $s$ the $s$th sample in the posterior. To compute the log2FC in a tissue between two dpis, first the group specific parameters(such as dpi1/tbln and dpi3/tbln) are computed by combining the corresponding main effects and interactions. Then, their difference is computed which denotes the log fold change and the coefficients are computed in the log scale. The change of base is applied to make it in the log2FC scale. The visualization of the log2FC for three issues and various dpis with respect to corresponding dpi1 is shown in Figure 2.6.

One can compute the high density interval (HDI) to capture the region that covers $\alpha$ % of the distribution (Kruschke 2014):

$$HDI_\alpha(\theta) = (L,U), L,U \in R \text{ and } L < U s.t.$$

$$argmin_{(U-L)} Pr(L \leq \theta \leq U) = \alpha$$

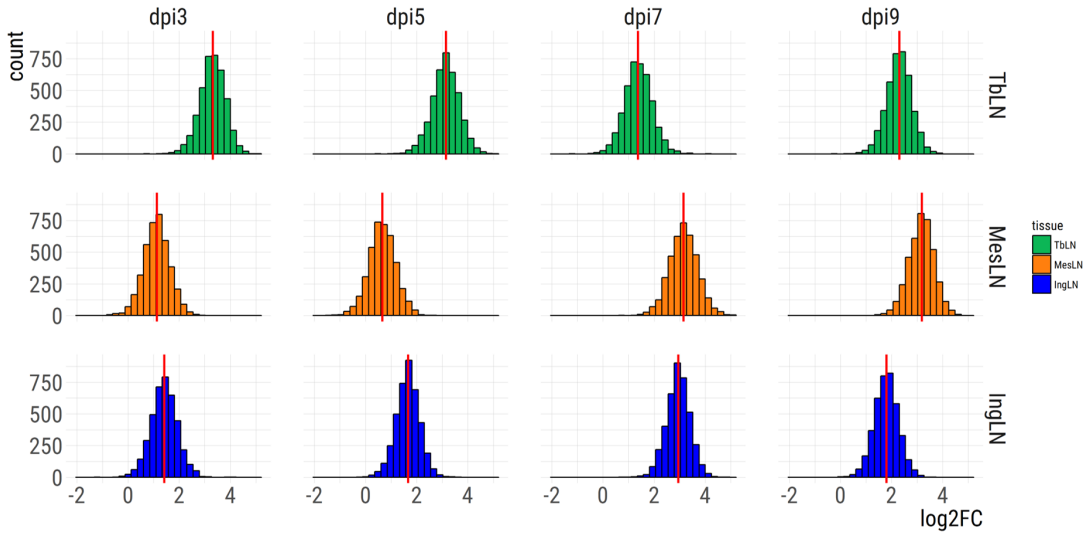


Figure 2.6: **Marginal Posterior log2 Fold Change for DDX58** .
Red line indicates the mean within each group

HDI has a better interpretability than the frequentist confidence interval in that the former denotes a most likely region for a given parameter whereas the latter has a slightly different meaning; $\alpha\%$ confidence interval represents the percentage of the theoretical confidence intervals that would contain a true parameter. In Chapter 2, we used HDI to perform the differential gene expression analysis.

## 2.7 Bayesian Differential Gene Expression Analysis

With the full joint posterior distribution, one can perform differential gene expression analysis in the framework of Bayesian decision theory. In general, there are two ways to make a decision in a Bayesian framework: Bayes Factor(BF)-based and Parameter Estimation(PE)-based methods (Kruschke and Liddell 2017). BF-based method constructs two separate models–null and alternative–where the former has a degenerate prior on a point mass(i.e. 0) and the latter has a diffuse prior over parameter values. With these

models, one constructs a BF statistics as following:

$$BF_{\text{null}} = \frac{D|M_{\text{null}}}{D|m_{\text{alt}}} = \frac{p(M_{\text{null}}|D)}{p(M_{\text{alt}}|D)} \Big/ \frac{p(M_{\text{null}})}{p(M_{\text{alt}})}$$

The goal is to choose the model explains the data best between the two competing models using this metric. Commonly, if $BF_{\text{null}}$ is greater than a threshold such as 10, the null model is accepted; On the contrary, if $BF_{\text{null}}$ is less than $\frac{1}{10}$, the alternative model is chosen instead. However, the usefulness of BF-method is shown to be limited (Gelman et al. 2014; Kruschke and Liddell 2017). For example, when the null prior probability is low or the models that are continuous rather than discrete, BF-based method might lead to the bias.

Another approach of making a decision is PE-method guided by HDI and the Region Of Practical Equivalence (ROPE) around the null value, the interval of values considered to be virtually null to account for the inherent uncertainty of the parameter values (Kruschke and Liddell 2017; Kruschke 2014). The procedure is following: if $\alpha\% - \text{HDI}$ is completely outside of ROPE, the null value is rejected and if $\alpha\% - \text{HDI}$ is completely inside the ROPE, the null value is accepted. The benefit of the PE-based approach is that it's intuitive, simple to perform, and general. Kruschke 2014 et al. extensively reviews the superiority of ROPE over BF in Kruschke and Liddell 2017; Kruschke 2014. To apply the PE-method to identify differentially expressed genes (DEG), we can obtain the posterior distribution of log2FC as mentioned in the previous section, choose the size of HDI, the ROPE-radius and the null value, and check if the given gene meets the criteria to be considered as differentially expressed. For example, we can construct the following algorithm to determine if a gene is

DEG:

1. lower bound of $HDI_\alpha(log2FC)$ is greater than $\gamma$, OR
2. upper bound of $HDI_\alpha(log2FC)$ is less than -$\gamma$

where $\alpha$ corresponds to the size of HDI(i.e., 90%) and $\gamma$ corresponds to the radius of the ROPE around 0. The Chapter 3 shows an application of this DEG algorithm.

## 2.8 Limitations

Currently, the limitations of the MAGE pipeline is that fitting all genes is computationally intensive. Although fitting individual genes with 4000 MCMC draws using Stan takes less than 30 seconds, the challenge is that it must be done so for about 20,000 genes. We addressed that challenge by fitting models in parallel using the Sun Grid Engine, which might not be available to general users. The current solution for a general user is using the clusters from Amazon Web Service(AWS) or a computer with multiple cores.

The parallelization sheds light on another limitation of MAGE: genes are assumed to be independent of each other. By making this assumption, we might have missed the additional parameter information that could have learned from combining all genes into one model. However, the benefit of making independence assumption for genes outweighs the cost. Performing MCMC algorithm to estimate parameters from a fully joint model that includes all genes will make the computation prohibitively slow. An open avenue of research is to develop an algorithm that combines all genes into one model while maintaining efficient fitting process.

## 2.9 Conclusions

In this chapter, we developed Multilevel Analysis of Gene Expression to accurately estimate the parameters of the gene expression model involving time course RNA-seq data using multilevel modeling and Bayesian approach. We showed the motivation and theoretical justification for MAGE and the validness of the proposed model via posterior predictive checks and feasibility of fitting thousands of simultaneous models by providing the various MCMC diagnostics. To our knowledge, MAGE is the first framework that applied the full Bayesian approach to gene expression data analysis. With the powerful and flexible multilevel modeling, the posterior draws for each gene in a particular biological system are provided, and with them, one can conduct a powerful and biologically interesting inference. The majority of the tools currently available are optimized for a simple one factor model, and more powerful tool is needed when the the number of factors increase while the experiment units stay the same. We were able to implement MAGE thanks to the development of probabilistic programming language such as Stan that provides a general framework of modeling and parallel computational resources that are becoming increasingly commonplace. The model is easy to extend and modify depending on the data set at hand. The software is available at https://github.com/RabadanLab/mage

Chapter 3

# *Transcriptomic evaluation of the host response to aerosolized Marburg virus Angola exposure in the lymph nodes of Rhesus macaques*

## 3.1 Background

Marburg virus (MARV) is a single strand, negative-sense RNA virus that belongs to the order *Mononegavirales*, the family *Filoviridae*, the genus *Marburgvirus*. Its functions are performed by the seven proteins encoded in its 19 kb long-genome--nucleoprotein (NP), VP35, VP40, glycoprotein (GP), VP30, VP24, and the polymerase (L).

Despite its simple composition, MARV is considered as one of the deadliest pathogens along with its close relative Ebola virus (EBOV) for its high infectivity and virulence (Brauburger et al. 2012; Mohamadzadeh, Chen, and Schmaljohn 2007; Slenczka and Klenk 2007). Classified as a biosafety level-4 (BSL-4) pathogen, MARV is capable of causing severe hemorrhagic fever disease in humans and Non-Human Primates (NHP) with symptoms including but not limited to fever, rash, malaise, diarrhea, vomiting, excessive bleeding, severe liver damage, and coagulation complication (Brauburger et al. 2012). With no clinically approved antivirals available currently, MARV poses a significant threat of pan-

demic and bioterrorism (Brauburger et al. 2012; Feldmann and Kiley 1999; Kiley et al. 1982; Kuhn et al. 2010; Nakayama and Saijo 2013).

MARV is first discovered and isolated in 1967 in the former Yugoslavia and the city of Marburg in Germany--hence its name--when laboratory workers became sick while handling the African Green Monkeys (*Chlorocebus aethiops*) imported from Uganda (Siegert 1972). Since then, it has caused a number of sporadic outbreaks (Gear et al. 1975; Johnson et al. 1996; Nikiforov et al. 1993; Smith et al. 1982) between 1970s and 1990s, followed by two major outbreaks. The first major MARV outbreak occurred in Durba, Democratic Republic of Congo (DRC) in 1998-2000 (Bausch et al. 2006), and the second major outbreak was in Uige Province in northern Angola, West Africa in 2004-2005 (Towner et al. 2006). In each case, the case fatality rate reached 83% (128 reported number of death) and 90% (227 reported number of death), respectively (Brauburger et al. 2012; *Chronology of Marburg Hemorrhagic Fever Outbreaks* 2014) (Figure 3.1). The former outbreak was a case where multiple strains of marburgvirus were introduced whereas the latter outbreak was caused by a single strain. These outbreaks collectively demonstrated highly infectious, pathogenic, and easily transmissible nature of MARV.

How is MARV so virulent in human and non-human primates? In the past, MARV and EBOV have been shown to trigger potent dysregulation of both innate and adaptive immune responses in both nonhuman primates and humans (Bosio et al. 2003; Brauburger et al. 2012; Mohamadzadeh, Chen, and Schmaljohn 2007). Initially, MARV targets sentinel cells such as dendritic, monocytes, macrophages in lymph nodes, liver, and spleen. In the late stage, MARV is rapidly replicated and disseminated to other organs (Brauburger et al. 2012).

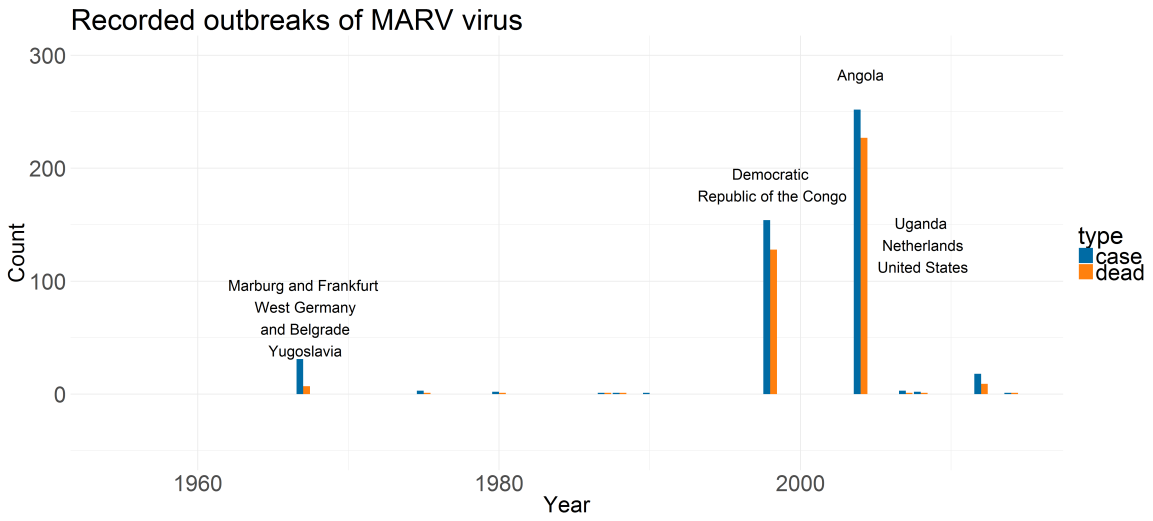


Figure 3.1: **History of Marburg Virus Outbreak**. MARV was introduced in 1972. There have been sporadic outbreaks between the 70s and 90s. There were two major outbreaks of MARV in late 90s and early 2000. A blue bar represents the number of exposed cases and the red bars the number of fatal cases.

However, the exact mechanism of MARV pathogenesis in animal hosts remains elusive. For example, *Rousettus aegyptiacus*, the recently identified host of MARV (Towner et al. 2007, 2009), is asymptomatic to virus; in contrast, to human and NHP such as *Rhesus macaques*, the pathology is severe.

There are features of MARV that render its research difficult. MARV is fatal to human and NHP if left untreated, and the time to death after the infection can range from 1 to 2 weeks (Geisbert et al. 2007; Mohamadzadeh, Chen, and Schmaljohn 2007). Furthermore, mouse models are not useful in MARV research as they do not fully recapitulate MARV disease phenotype of human. *Rhesus macaques* is an NHP model for MARV infection (Nakayama and Saijo 2013), but they are expensive to maintain and process, especially with the ban of import of Indian origin research. Nonetheless, the transcriptomic characterization of the natural host and spillover hosts such as NHPs or humans would contribute to our understanding of factors involved in the pathogenesis and host-virus interaction.

The present study explores the transcriptomic response in lymph nodes of *Rhesus macaques* to aerosolized Marburg virus Angola exposure. We chose lymph nodes for their importance in the immune response to infection and virus dissemination.

Cell lines and primary cells in a peripheral blood mononuclear cell (PBMC) have been used to investigate gene expression of hosts infected with MARV (Connor et al. 2015; Lin et al. 2015). However, the transcriptomic view of Marburg virus infection on the tissue samples of animal hosts will better characterize the host response to MARV at the site of infection and the secondary tissues. This will open the door for better understanding of immunology and expedite the development of effective treatments.

To characterize the immune responses to MARV in lymph nodes of rhesus macaques on a molecular level, 15 rhesus macaques were sequentially sacrificed via aerosol exposure to Marburg Virus Angola. Over the course of 9 days, and 3 types of lymph nodes (Tracheobronchial, Mesenteric, and Inguinal) were extracted from each sample and sequenced for the gene expression analysis.

With the novel pipeline MAGE discussed in Chapter 2, we obtained the posterior distribution of parameters relevant to expressions of 13,918 genes, with which we performed several biological inferences including differential gene expression analysis, pathway analysis, unsupervised expression archetype identification, and tissue-specific differential gene identification.

## 3.2 Results and Discussion

### Data, Study Design, and Analysis Strategy

In Figure 3.2, the analysis pipeline is shown along with the data description and study design. Briefly, to characterize the global gene expression pattern and individual gene expression changes, 15 rhesus macaques have been simultaneously infected with MARV via aerosol route. They were grouped into a group of three and each group has been sacrificed serially over 9-day post-infection (DPI) to Marburg virus, with 2 days apart from each group. Three lymph node tissues (Tracheobronchial (TbLN), Mesenteric (MesLN), and Inguinal (IngLN)) have been extracted from each rhesus macaque, and RNA-seq data were generated.

With the RNA-seq data, the quality for individual libraries were checked first. Total of 6 samples was dropped due to a low number of reads (See 3.5). Figure 3.3 shows the pattern of sample quality. After selecting the suitable transcriptome reference (see the next section (3.2)), the quantification of transcriptome was performed. The gene counts for each sample have been piped into the MAGE pipeline (Chapter 2) to estimate the posterior distribution of relevant parameters including dispersion, log2 Fold Change (log2FC), and baseline expression. Once the joint posterior distribution was obtained in the form of MCMC samples, the various biological analysis were performed and discussed in the Section 3.3.

### Reference Selection

One of the most important decisions for the RNA-seq analysis is a choice of a reference sequence against which reads are to be mapped (Conesa et al. 2016) since it determines

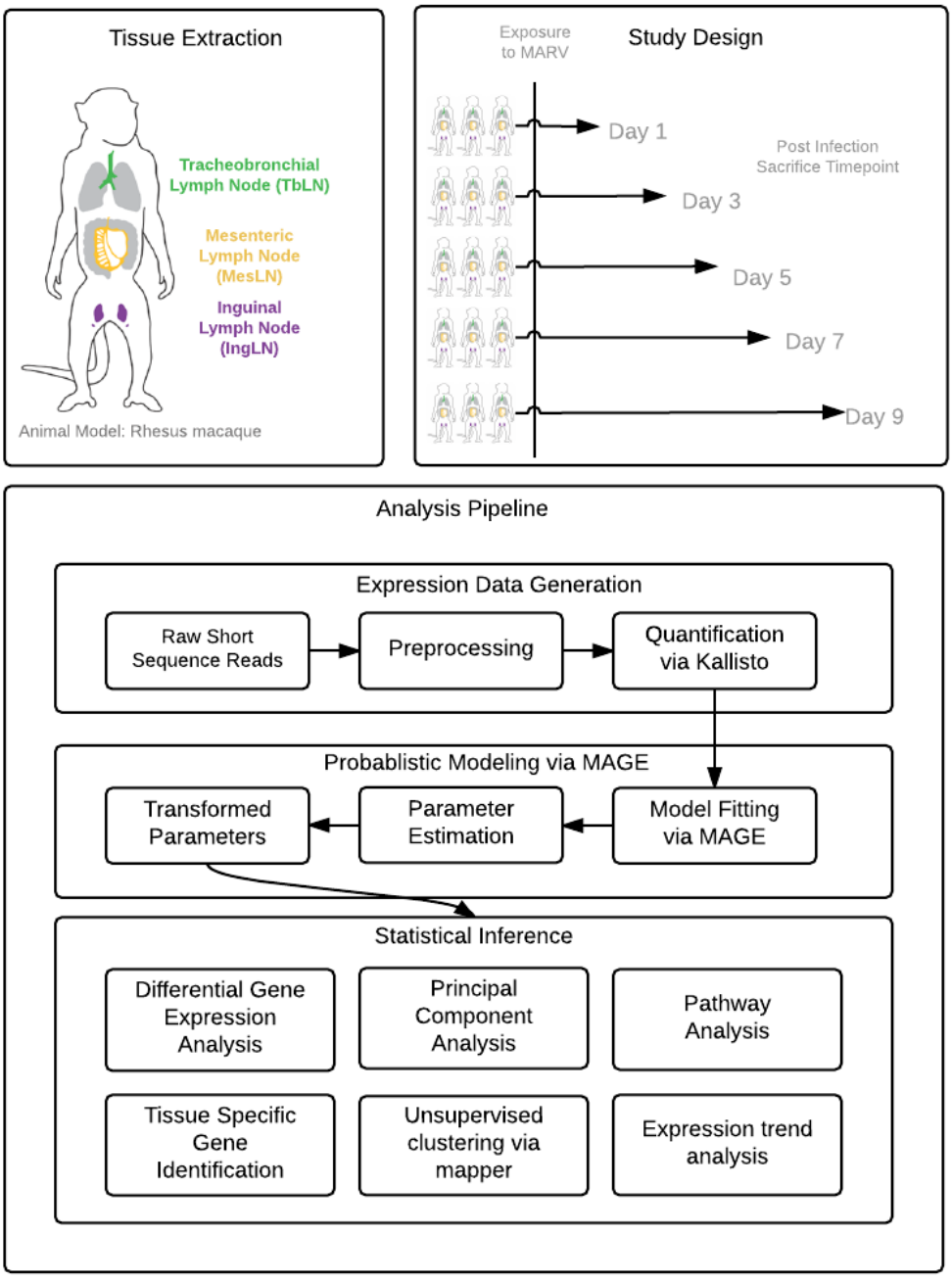


Figure 3.2: **Data, Study Design, and Analysis Strategy**. A) **Tissue extraction**: three lymph nodes--tracheobrochial, mesenteric, and inguinal-- have been extracted from an individual monkey; B) **Study Design**: 5 groups of three rhesus macaques have been serially sacrificed post exposure to MARV infection via aerosol route. Each group have been sacrificed at day post infection (DPI)1, 3, 5, 7, and 9 and tissues were extracted, processed, and sequenced; C) **Analysis Pipeline**: Analysis comprise three major steps 1) gene expression data generation 2) probabilistic modeling via MAGE, and 3) statistical inference using the posterior draws. Multiple sub-steps comprise each of the major step.

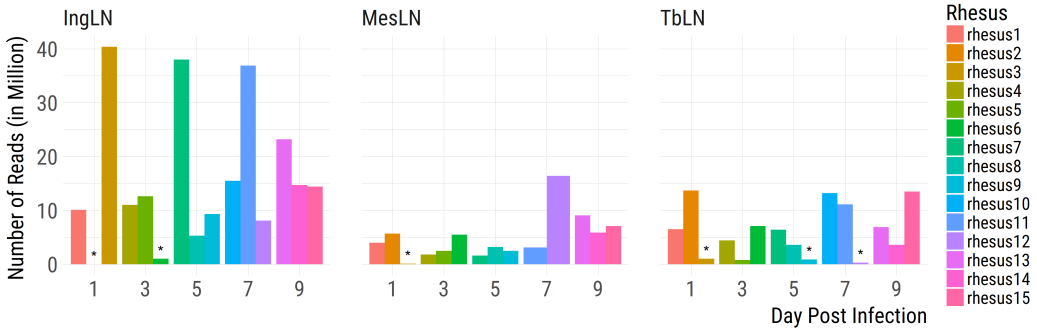


Figure 3.3: **Number of reads for individual library RNA-seq libraries**. The number of reads are plotted for individual libraries in dpi and tissue groups. Color represents unique rhesus samples. Starred are the libraries that are excluded due to low number of reads.

the feature space on which the analysis is performed. Multiple reference genome and transcriptome resources are available for rhesus macaques (Aken et al. 2016; Gibbs et al. 2007; O'Leary et al. 2015; Peng et al. 2014; Zimin et al. 2014). For the present study, our desiderata were to use the reference transcriptome with the high sensitivity (that the majority of sample reads are captured in the reference sequences) and high specificity (that there are less extraneous sequences in the reference transcriptome) as well as good annotations. To this end, we compared the three reference resources: 1) Refseq transcriptome based on RheMac2 genome (O'Leary et al. 2015), 2) Ensembl transcriptome based on RheMac2 (Aken et al. 2016), and 3) MacaM transcriptome (Zimin et al. 2014). After checking the percentages of reads mapped and percentage of reference sequences with no reads mapped (Figure 3.4), we confirmed that MacaM has both high sensitivity and specificity. We used MacaM for the downstream analysis discussed in the next section.

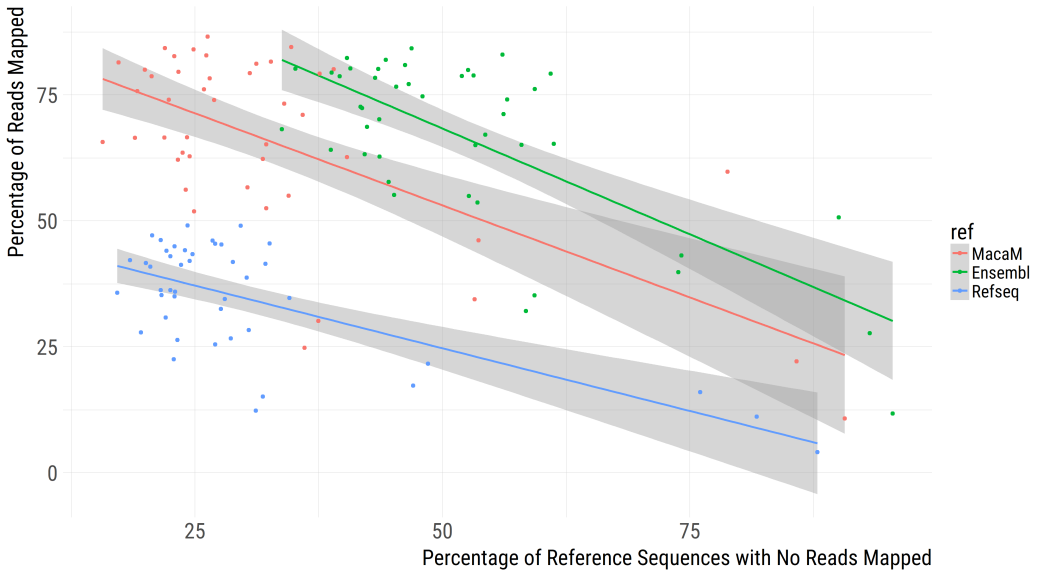


Figure 3.4: **Choosing the reference transcriptome for quantification**. Each dot represents a unique RNA-seq library. Different colors represent quantification using different annotation resources. Lines were fitted using linear model. Ideally, a good annotation has both high percentage of reads mapped, and low percentage of reference sequences with no reads mapped.

## 3.3 Statistical Inference

### Posterior Viral Load Estimated via RNA-seq

To characterize the viral transcriptome in the rhesuses infected with MARV, we quantified the viral reads and obtained posterior log2FC and baseline expression (See Method 3.5). Since we did not have the day 0 (control) samples, to compute the fold change, we used day 1 as a reference sample. This is justified by the fact that viral load is non-existence (Figure 3.5B). Considering the day 1 post infection as a reference, Posterior Log2 Fold Change was computed for subsequent DPI time points(day 3, 5, 7, and 9). In Figure 3.5A, we observe the high correlation between the viral load and posterior median log2FC with respect to dpi1 within a tissue. Over the days, however, the viral load increases dramatically

by the order of magnitude, with TbLN spikes up the quickest with 20 median log2FC at DPI3 (Figure 3.5C). This showed that the hosts were unable to antagonize the rapid viral replication of MARV.

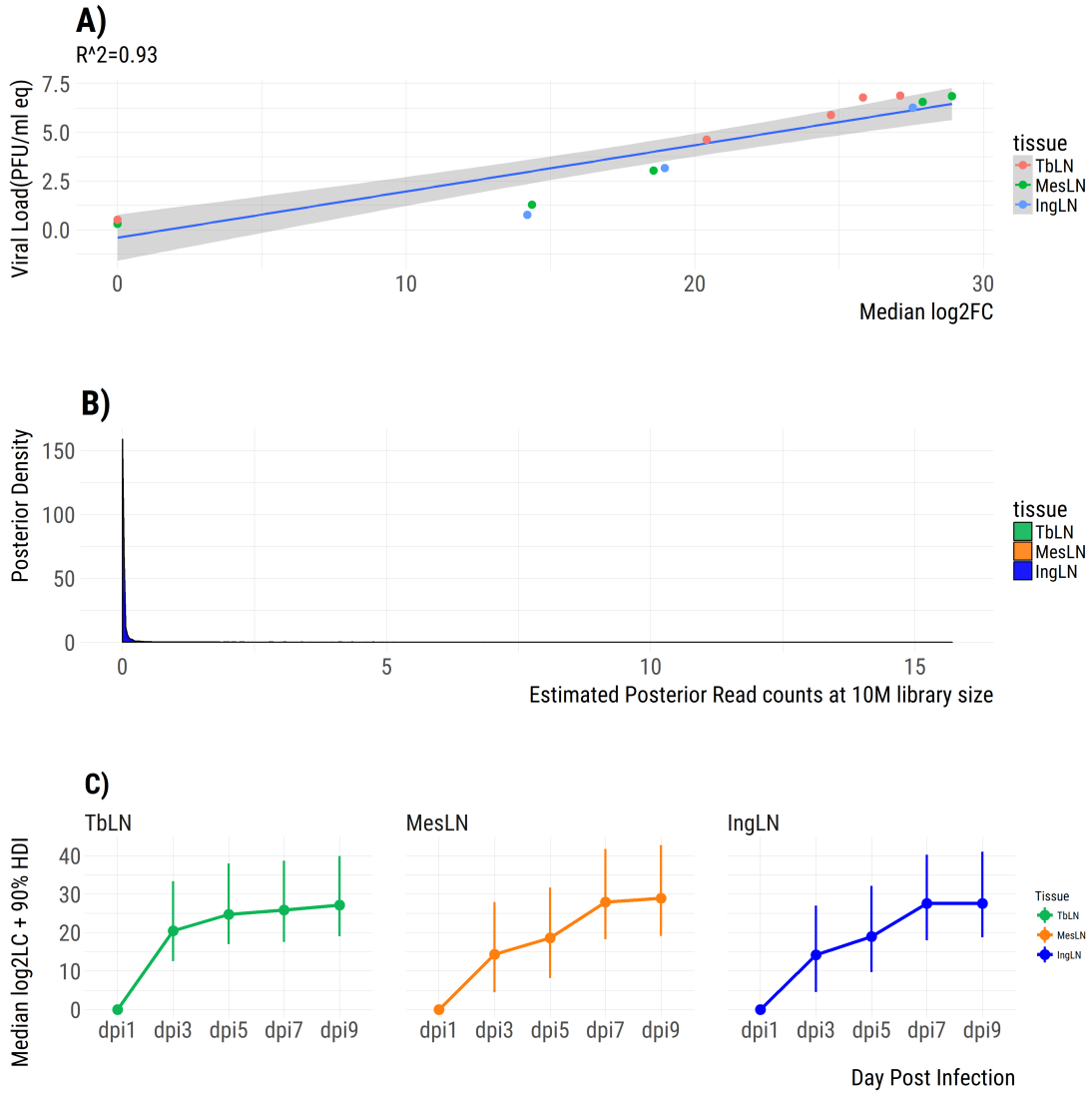


Figure 3.5: **Posterior Viral load Estimated via RNA-seq and MAGE. A) Correlation Between posterior median viral Fold Change and Viral Load** posterior median log2 Fold change can be used to predict the viral load measured independently ; **B) Base Expression of MARV** Base expression of MARV is near 0; **C) Posterior Log2 Fold Change of MARV** As the DPI increases, the viral load spikes up in all three tissues

## Global Expression Pattern of MARV infection in rhesus

The quantification of viral genes illustrated that the viral growth rapidly has increased. We were then interested in determining the host response to this MARV growth. To this end, we processed 13,918 individual host genes using the MAGE pipeline (Chapter 2). To view the gene expression pattern of MARV infection in rhesus macaques globally, we obtained the posterior median log2FC for individual genes and performed the principal component analysis on the gene to fold change matrix (Figure 3.6). We used log2FC rather than TPM or count as features due to the baseline expression differences among tissues (See Appendix 3.5). The first two dimensions of eigenspace collectively explained 53.1% variances in the expression profile represented by posterior log2FC. We first noted that the U-shaped topology of the projection of samples on the eigenspace is consistent with the ordering of day post infection. Interestingly, the PC1 axis seems to represent the genes involved in the early and late infection signatures whereas the PC2 axis represents the transcriptional signatures that are active the mid-infection stage. We also noted that the samples were clustered by DPI indicating that the activity of the distinct transcriptome profile in the time course of MARV infection.

## Differential Gene Expression Analysis

Next, we were interested in the biological processes occurring in each DPI in each tissue. To identify the set of differentially expressed genes (DEG) over time in each DPI, we compared each DPI group to DPI1 in each tissue to obtain the log2FC. Since we have the posterior distribution of log2FC, we defined that gene expression for a gene is differ-

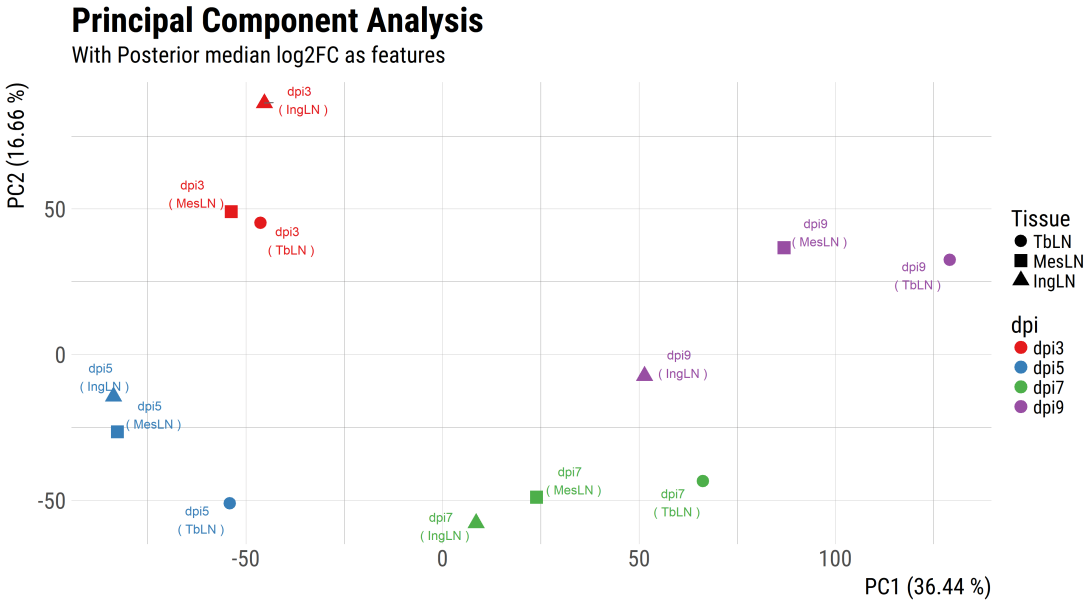


Figure 3.6: **Principal Component Analysis using Median Log2 Fold Change**. Points represent the individual dpi groups; shapes represent tissues.

ential if the two parameters--absolute log2FC and high density interval (HDI)-- satisfy the following conditions:

1. upper bound of $HDI_90(log2FC)$ has to be greater than 1.5, OR
2. lower bound of $HDI_90(log2FC)$ has to be less than -1.5

We chose these thresholds to capture the genes with strong effect sizes (the read count approximately 2.8 times higher/lower than that in dpi 1) and statistical confidence (90% of the probability mass is above/below the threshold). We tested the grid of parameters (Figure 3.7) and the relative amount of DEGs was more or less consistent among different parameter sets.

Overall, the DPI increase results in the number of genes.

Interestingly, there are more up-regulated genes than down-regulated genes in MARV

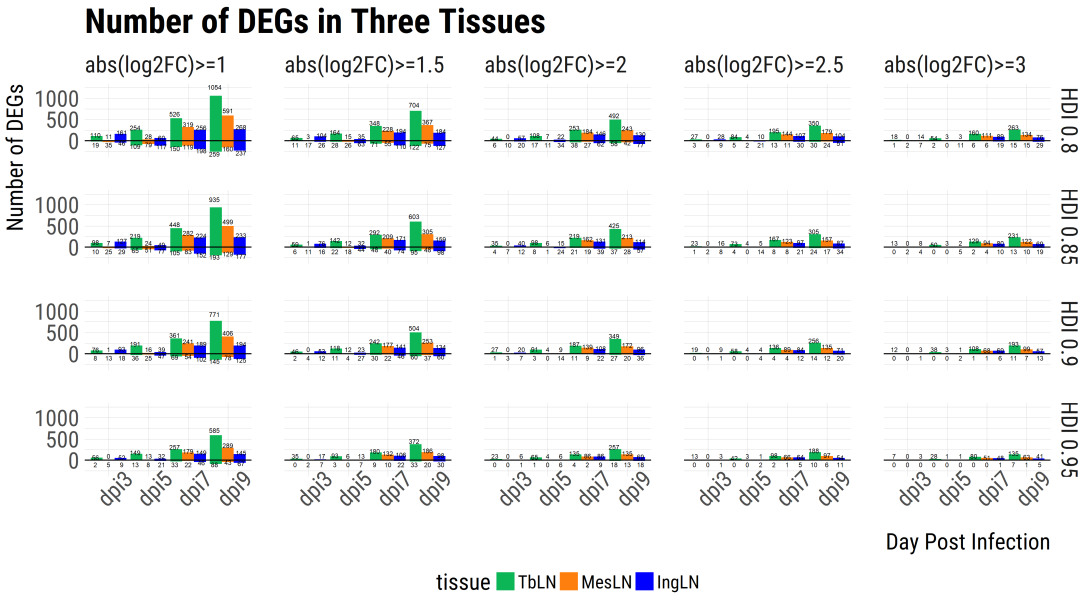


Figure 3.7: **Tuning the parameters for Differentially Expressed Gene Analysis**. The bar above the horizontal line corresponds to the number of up-regulated genes where as the bar below corresponds to the number of down-regulated genes.

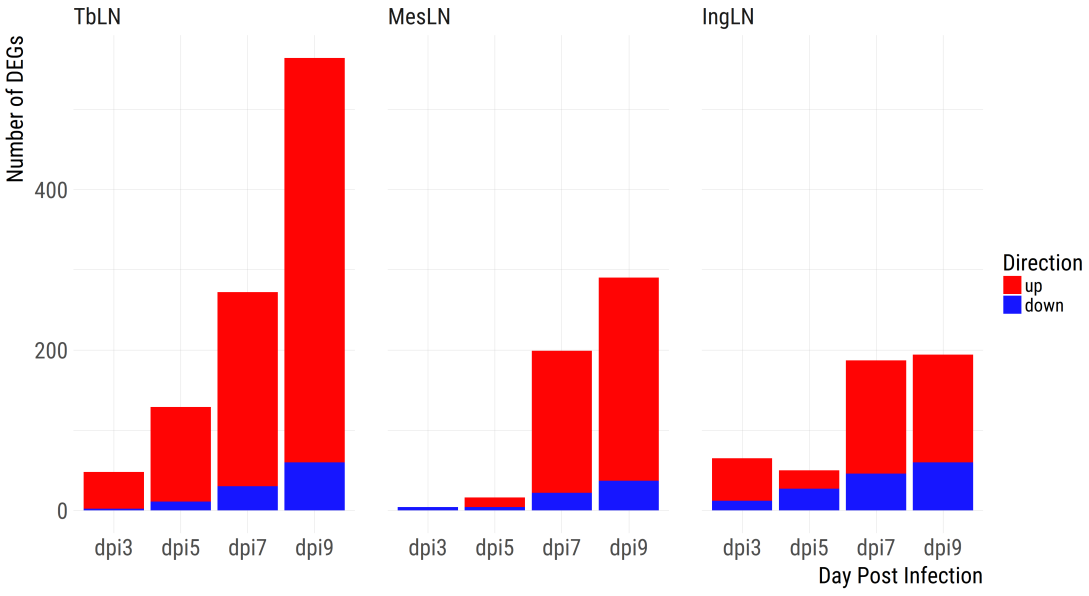


Figure 3.8: **Differentially Expressed Gene at HDI 90% and abs(log2FC) $>$ 1.5**.

infected samples. This maybe due to the fact that immune systems are actively engaged upon infection which requires the production of signaling molecules such as cytokines and chemokines and immune effector genes such as interferon stimulated genes. In fact, many genes that are differentially expressed are immune related. Furthermore, the number of DEGs is the highest in TbLN followed by MesLN, and IngLN (Figure 3.8). This is also consistent with the fact that TbLN serves as the draining lymph node and primary site of immune response activation. Indeed, the number of differentially expressed genes are correlated with the viral load (Pearson correlation of 0.73, p-value =0.008).

## Pathway Analysis

We now were interested in combining individual gene expression information and inferring biological pathways that might be enriched. To perform pathway analysis, we uploaded gene expression tables for each lymph node at each time after infection into the Ingenuity Pathway Analysis (IPA) program. The data was filtered using the same parameters as we used previously to identify differentially expressed genes ($abs(log2FC) > 1.5$ , HDI0.9). We then ran core analyses on all 12 data sets using DEG genes. This yielded information for each lymph node at each time point. We focused on the canonical pathways enriched in each tissue with a p-value ¡ 0.05 and found that 27 pathways are enriched in all 3 lymph nodes during the course of infection (Figure 3.9) and that 27 pathways are enriched only in the TbLN during infection (Figure 3.10).

The top 5 pathways that are enriched amongst all lymph nodes evaluated relate to antiviral host responses and immune cell activation and migration. Similar to the patterns

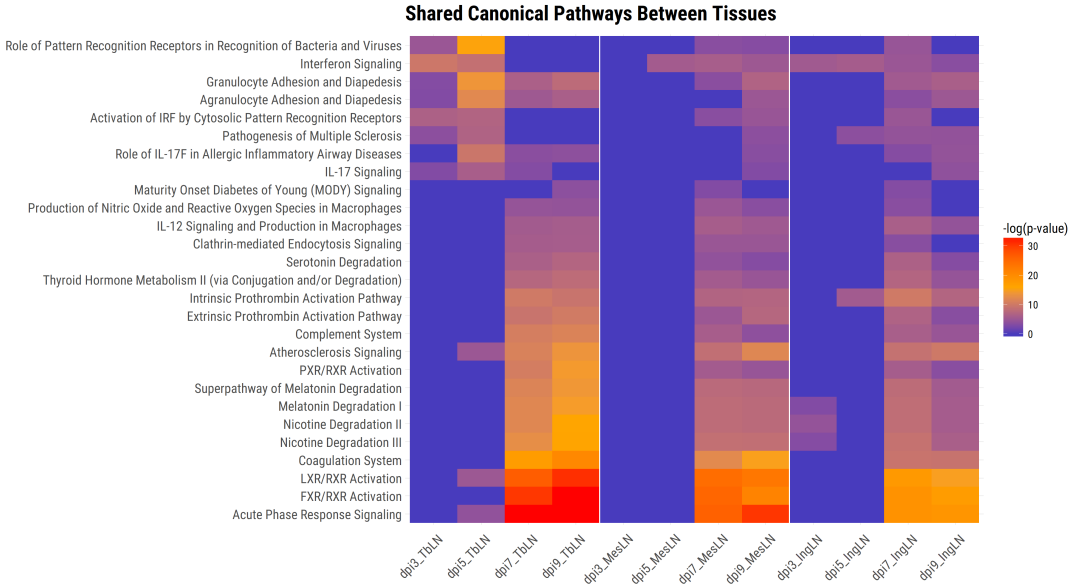


Figure 3.9: **Canonical Pathways Shared among Tissues**. p-value indicates the overlap p-value between the gene set and our data set computed using Fisher's Exact Test

observed in the number of DEGs in each tissue, the immune pathways are enriched first in the TbLN followed by the MesLN and IngLN which corresponds to the kinetics of MARV replication which occurs first in the TbLN followed by the other LNs. With TbLN being the draining lymph node, we expect to see an increase in immune cell migration and activation in this tissue shortly after infection.

Later during infection, we observe enrichment of pathways that correspond to the acute phase response, macrophage function, and the complement system. We also see enrichment of coagulation-related pathways which agrees with previous data that MARV infection results in late stage coagulopathies as well as the appearance of fibrin deposits and thrombi in tissue sections of the TbLN. Furthermore, we also see enrichment for a variety of metabolic pathways which may be a result of virus infection and the disruption of normal cellular function and/or a results of tissue damage induced during MARV infection.

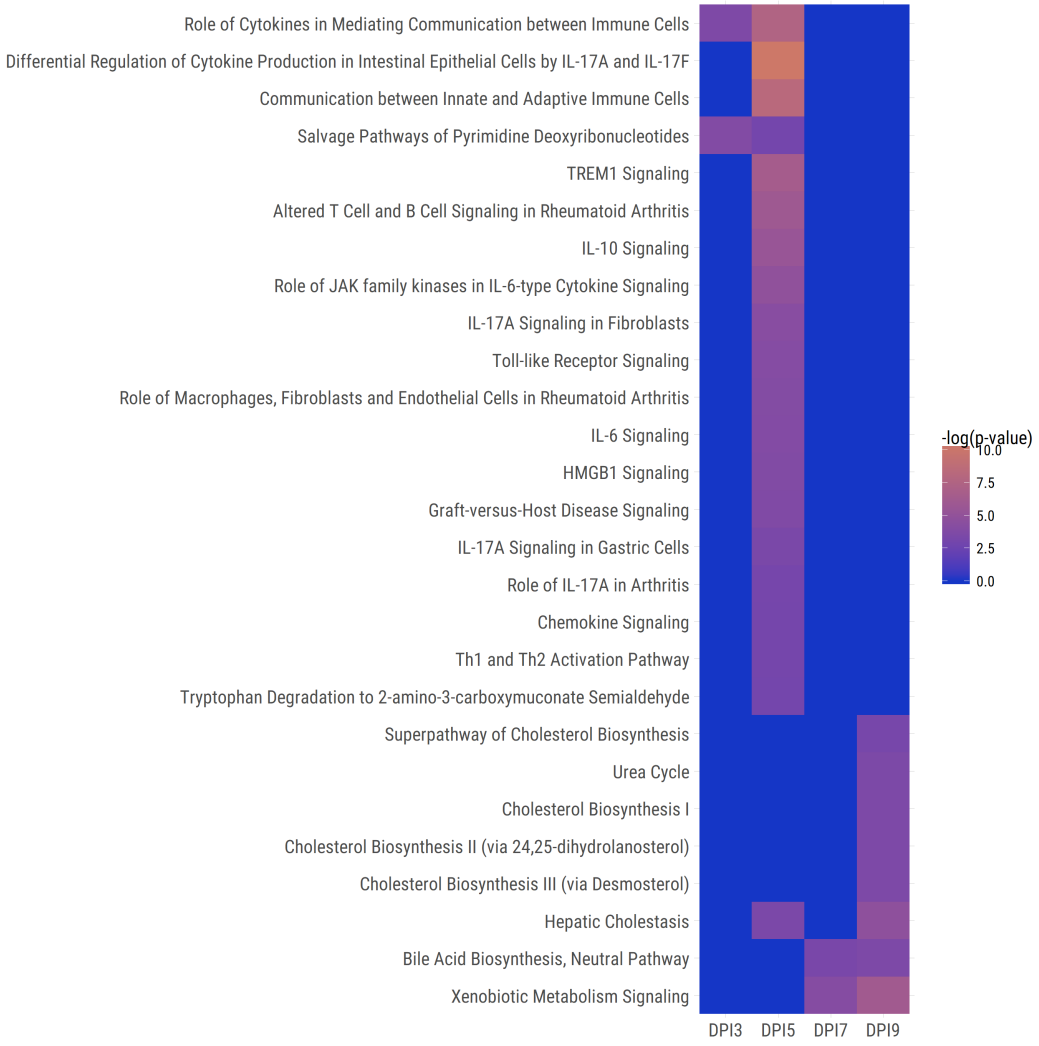


Figure 3.10: **Canonical Pathways in TbLN only**.

The pathways that are enriched only in the TbLN are again primarily associated with the immune response. Rather than being only antiviral and macrophage-specific responses, we now see enrichment of pathways that deal with regulating and coordinating innate and adaptive response including those pathways that focus more on development and skewing of T and B cell responses. Again, this further supports the TbLN as the draining and priming lymph node for mounting and coordinating the immune response to aerosolized MARV infection. These pathways are most highly enriched 5 days after infection, after which point the enrichment scores(-log(p-value)) decline. This correlates with what is observed during histological analysis of the TbLN where tissue architecture, in particular germinal center formation is maintained through 5 days after infection, but becomes abolished by 7 and 9 days after infection correlating with an increase in lymphocytolysis and apoptosis as determined by TUNEL staining (data not shown). Interestingly, lymph node tissue architecture is mostly maintained in the MesLN and IngLN through 9 days after infection with little to no apoptotic cells. Previous studies on MARV pathogenesis have shown that a decrease in lymphocytes and dysregulated immune response occur later during infection and contribute to disease pathogenesis (Geisbert et al. 2008, 2009). These data suggest that part of this immune destruction is occurring in the draining lymph node.

## Unsupervised Clustering to Uncover Expression Archetype

In the previous section, we analyzed the gene expression profiles by mapping the DEGs to the knowledge base provided by IPA. This is useful for the identification of activity or inhibition of the known pathways at a particular time point. In addition to known pathways,

we were also interested in characterizing the archetypical expression patterns existing in the expression profile as a whole.

Bar-Joseph et al. illustrated archetypical trends common in genes in biological systems in Bar-Joseph, Gitter, and Simon 2012. The patterns they described include “Sustained” pattern in which an expression level of a gene continually increases or decreases as a function of time or“Impulse” pattern in which there is a spike of an expression followed by the rapid decrease.

Since we have a full posterior distribution of the system, we can transform it to identify the trends that exist in our expression profile. To this end, we transformed the posterior log2FC into the rank distribution with 1 being the lowest fold change and 5 being the highest fold change. In effect, we generated the posterior distribution for the 120-category nominal variable which level corresponds to the possible rank of a log2FC, which we refer as “waveform”. With this, we can identify, for example, the set of genes with the upward trend(“1,2,3,4,5”) and those with the downward trend(“5,4,3,2,1”).

We looked at the most frequent waveform patterns in all genes (Figure 3.11). Interestingly, we observed that the distribution of waveforms is not uniform; there are certain waveforms that are more frequent than the others. For example, the top patterns comprise the waveforms that have ranked the dpi7 and dpi9 as rank 4 and rank 5, respectively, indicating that majority of genes are up-regulated in the later days.

Next, since each gene has 120 possible waveforms in each of three tissues, we computed the entropy for those probabilities for individual genes and compared the distribution of entropies among tissues (Figure 3.12).

Given that the upper bound of the entropy for the given probability vector is approxi-

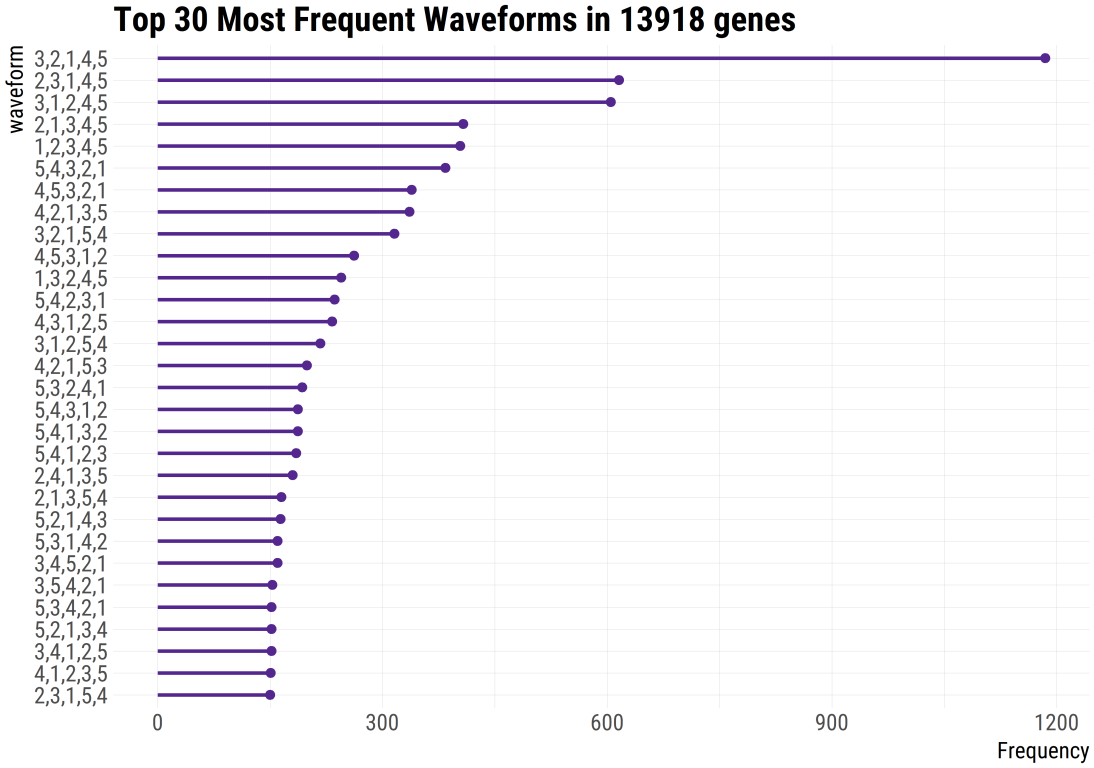


Figure 3.11: **Top 30 Most Frequent Waveforms in 13918 Genes**. y axis represents the the sequential rank of log2FC and x-axis represents frequency

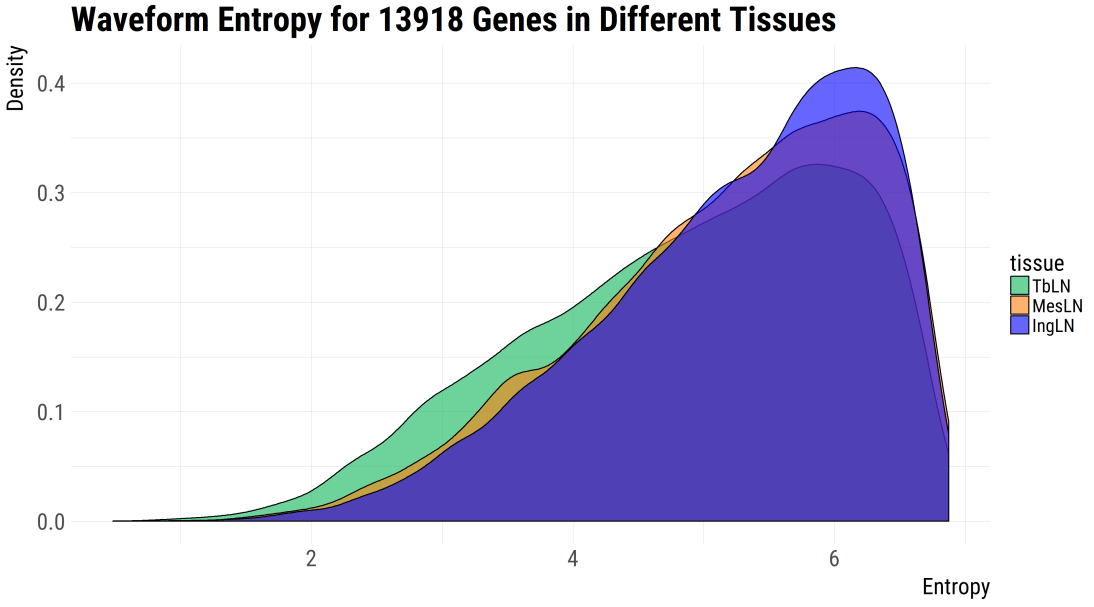


Figure 3.12: **Waveform Entropy for 13918 Genes in Different Tissues**.

mately 6.91(Log2(120), the majority of the genes are near 6, which suggests that the majority of genes do not have a skewed pattern. However, we see that the distribution of TbLN is lightly skewed left followed by MesLN indicating that genes in TbLN and MesLN have more pronounced patterns than those in IngLN on a global scale. This is consistent with the fact the physical locations of these tissues and the route of infection.

To uncover the number and the type of archetypical expression using both direction and magnitude of expression in an unsupervised fashion, we used the Mapper algorithm which is based on topological data analysis (TDA) techniques (Carlsson 2009; Singh, Mémoli, and Carlsson 2007). Recently, The effectiveness of TDA has been demonstrated in many applications in biology for its ability to reduce the dimensionality while preserving the topological features of original multidimensional space (Cámara 2017; Rizvi et al. 2017; Singh, Mémoli, and Carlsson 2007). Mapper, in particular, is an algorithm that takes the point cloud in the high dimensional space and attempts to represent it in the low dimension by reducing them into simplicial complexes. Representing individual genes as a vector of posterior median Log2 fold changes at DPI 3, 5, 7, 9, and considering only the genes that have at least one differential log2FC (as specified in the Section 3.3) in any DPIs, we obtain a $788x12$ matrix, which can be viewed as a point cloud with 788 points in the 12-dimensional space. After applying the z-score transformation on each dimension to prevent the late days log2FC biasing the inference, we performed the mapper algorithm using Ayasdi (Carlsson 2009; Lum et al. 2013 . The output of the Mapper is a network in which a node corresponds to a set of observations (in our case, genes) and the edge corresponds to the existence of a shared element between nodes. Elements in each node are determined by the nearness of observations in the original higher dimensional space.

We chose the resolution of 30 and gain of 3 in the parameters for Mapper for the downstream analysis (See Method 3.5) to create the network shown in Figure 3.13. In this network, there are 6 clusters detected via the community detection algorithm (See Method 3.5). The next panel shows the archetypical gene expression within each cluster for each tissue. Interestingly, cluster 1 captures the continually down-regulated genes, whereas cluster 4 captures gradually increasing genes. Cluster 2 and cluster 5 consists of genes that spike up after DPI 5, but the former has the higher magnitude. These sets of genes might be interesting to further explore since these are the ones that have higher expression in the late stage of infection. Cluster 3 seems to contain genes that go up in day 3 and down in dpi5 but goes back up in dpi7 and dpi9 whereas cluster 6 seem to have different patterns among tissues.

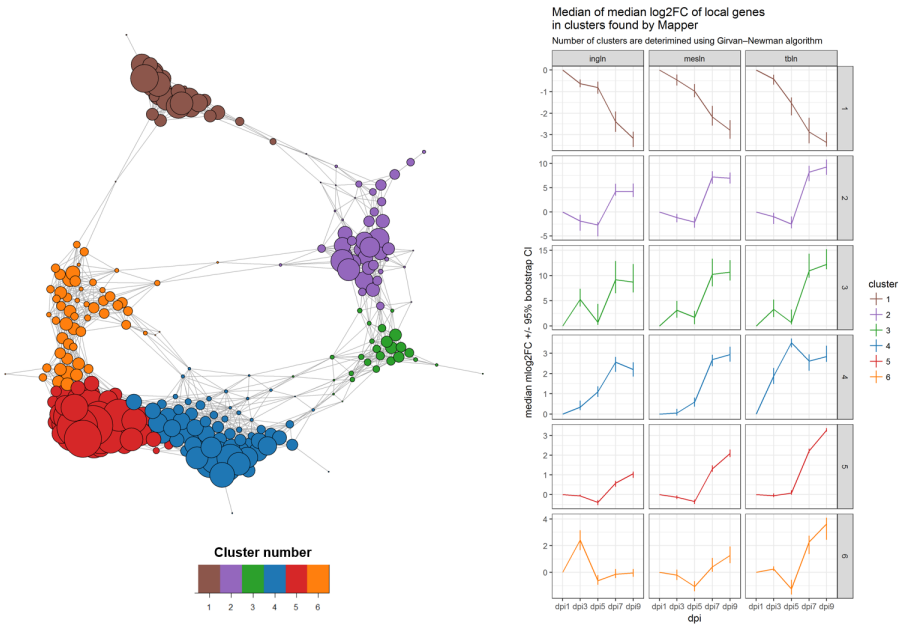


Figure 3.13: **Gene Clusters**.

We further examined the genes exclusively present in each cluster, as shown in Table 3.1, and performed the gene ontology enrichment analysis to capture potential biologi-

| Cluster | Genes |
|---|---|
| 1 | ADCK5, CCDC42B, CD33, CD40LG, CLEC3B, COLEC10, FAM92B, FOLR2, GPIHBP1, GRIA2, GRID2, ITGAD, KCNK12, LAT, LRRC26, LTB, OIP5, OR4K15, PEBP4, PLA2G2D, PVALB, SAMD3, SGCA, SMR3B, SYNDIG1, TNFRSF25, TPSAB1, EDN3, FAM180B, IGSF10, KCNJ16, LACRT, LGALS12, LRRC38, MYOZ2, NCCRP1, VPREB1, C2orf74, CDH4, CELF4, DACT3, PER2, POFUT1, POLI, RAB27B, SERTAD4, TPPP3, TTYH1, ZFP42, ACMSD, CHD3, **CXCR2**, EBI3, EGFLAM, FAIM3, FLT3LG, GRIA1, TGFBI, C2CD4C, CD1B, DLEU7, DNAH2, HS3ST2, INSM1, KCNJ5, MEPE, MIXL1, PDZRN4, SLC30A3, SPRR2D, STK31, WNT10B, ZCCHC12, CHRNA1, HOXC10, MGAT5B, MYO18B, PTPRZ1, THEG, AMICA1, ARHGAP8, C1QA, CCL21, CD37, GCSAM, KCNIP1, LAMP3, LEF1, LTA, MEOX2, MS4A1, NREP, PAQR6, PDCD4, PIK3IP1, RGS13, SCD5, STAG3, ZNF581, BIRC7, CD1E, GDF10, GRIN2A, KCNG3, OTOA, NTM, C1QTNF9, LY86, DBP, FAM81A, TVP23A |
| 2 | ABCC12, ACSM4, ARID3C, CLEC2L, CNTN2, GALNTL5, KLK6, KLK9, LRRC3B, NPY, SEZ6, SPATA21, TMPRSS6, TRIM50, WISP3, ACSM1, AHSP, ASGR1, HPX, OXT, SERPIND1, RIMBP3B, AGXT2L1, AMBP, CYP17A1, F11, GAS2, GLYAT, HABP2, HGD, HNF4A, HPD, PAQR9, PTPRD, SERPINC1, SLC13A1, SLC38A3, SLC38A4, SLCO4A1, F9, FGG, METTL7B, AFP, ALDOB, APOA2, BHMT, CPS1, FGB, HAO2, LBP, PLA2G2A, PON1, PRHOXNB, SAA4, SLC1A1, UGT2A3, ABP1, AHSG, ANGPTL3, C8A, CREB3L3, HAO1, HNF4G, HRG, IGFBP1, MTTP, PAH, SERPINA4, SERPINA7, SLC10A1, SLC22A1, ST8SIA3, TAT, TTR, UGT1A9, DEFA1, FA2H, OVOL1, TRIM10, CYP2C18, **SAA2**, **CCL23**, TDO2, ANPEP, NSUN7, SLC4A1, AIM1L, LRRC31, MYH4, ABCB11, C8B, CFHR5, CRP, CYP2C8, F13B, FETUB, GYS2, LECT2, MBL2, MIA2, SAA2-SAA4, SERPINA6, SLC17A4, SLC30A10, SMLR1, UGT1A1, UGT1A6, UGT2B17, UGT3A1, ALB, CPN2, PLA2G12B, PLG, SLC17A1, UGT1A3, UNC93A, UPP2 |
| 3 | AQP9, C19orf59, C4BPB, CSF3, FGL1, GBA3, GPR77, HAL, ORM1, TMPRSS2, AADAC, AZGP1, C9, CYP2A6, FABP1, FOXA1, HAMP, PRR15L, RAB17, RGS7BP, SLC17A3, SLC2A2, UGT2B4, UGT2B7, AMDHD1, C4BPA, CYP2C9, SLCO1A2, FGA, GC, KNG1, SFTPC, TM4SF4, RBP2, ADH6, AFM, APCS, C10orf90, APOB, CPB2 |
| 4 | ADORA2B, APOBEC3A, **CCL1**, **CXCL10**, CXCL11, FOSL1, GPR84, HRASLS2, IFI27, **IL1RN**, KCTD14, MYO7B, OASL, PTX3, RSAD2, S100P, TM4SF19, C15orf48, **CCL3L3**, CCL4, CEACAM6, CXCL2, **CXCL9**, **CXCR1**, DHRS9, DUSP5, FCN1, HCAR2, IFNG, INHBA, KIAA0895, MT1X, MT2A, S100A8, SOD2, SRXN1, SYTL3, TFPI2, THBS1, TIMP1, UNC5B, WARS, CMPK2, IFI44L, **IFI6**, IFIT1B, IFIT3, **ISG15**, AKR1B10, CYP4F3, GDF15, GNLY, LEAP2, NQO1, ZC3H12C, CP, CXADR, **CXCL6**, DESI1, EMP2, FFAR2, HSD11B1, OPN1LW, OSGIN1, SCG2, TF, C8G, CTH, FPR3, GZMA, **IL6**, S100A9, **TNFAIP6**, ALOX5AP, BST2, C1orf162, C3AR1, CASP4, **CCR1**, CLEC4A, CLEC7A, CSF3R, DYNLT1, FCGR2B, FOSL2, GBP1, GBP2, GBP3, GBP6, GLRX, IDO1, IL18BP, **IL1B**, KYNU, NBN, OSM, PI4K2B, PLA2G4C, PMAIP1, SAMD9L, SERPINB1, SOCS1, TLR2, TLR4, TNFSF10, TXN, UBE2L6, F2, HFE2, BCL2L14, DEGS2, APOL2, CCRL2, **DDX58**, DNAJB1, GCH1, HERC5, HSPA1A, IFIH1, IFIT5, **IRF7**, ISG20, MNDA, MX2, NT5C3, OAS2, ODF3B, TMEM140, TYMP, **USP18**, ADRA1A, SERPINE1, ATF3, BATF2, CASP5, CCL11, **CCL2**, DDX60, EIF2AK2, NMI, SOCS3, XAF1, CCL15, CHAC1, FAM167B, HPDL, IFI44, MT1M, TMEM255A, CCNA1, **IFIT2**, **MX1**, OAS1, IFNA2, **IL27**, **CCL7**, IL22, LCN2 |
| 5 | AOX1, CFHR2, CLMN, CYP1B1, DMD, DNAJC24, ENPP1, ENTPD5, FAM134B, FST, IFNAR1, IRAK4, KIAA1217, KRBA2, LGR4, LIMD1, MACF1, MAL2, MANSC1, MED18, MET, MSMO1, NR1I2, PARD6G, PCYOX1, PDP2, PHACTR4, PIGW, PLOD2, PPAPDC1B, PRKAR2A, PROX1, PRR11, PTPN14, RAB2B, SERPINA5, SGTB, SLC16A1, SLC25A15, SPAG9, SRR, STRADB, TAF8, UGGT1, APOF, C17orf75, F11R, FAM151B, FAM184B, FMO5, FZD4, GNE, LOXL4, LZTS1, PROS1, PTPN3, SERPINE2, TSKU, ABI3BP, AKR1C3, ATP11C, BCAP29, BIVM, BNIP3, CFL2, CLCN5, CLDN1, CLU, CYP51A1, EIF5A2, ELL2, ENPP4, ERRFI1, FAM20B, FAM216A, GINM1, IAPP, ID2, IGF2R, IL6ST, ITGAV, ITGB1, LMAN1, LRRC58, LYVE1, MAPK6, MRPL30, PGRMC1, PLA2G12A, PON2, PSMD12, SCARB2, SDC2, SEMA3A, SEMA6B, SLC25A13, SMIM13, TMEM245, TMTC3, UBXN2A, UTP23, ZDHHC20, DCBLD2, FN1, LRP1, NMT2, NT5E, PPAP2B, PXMP4, RORA, SGMS2, SNTB2, TMEM97, TXNDC5, PNP, SERPING1, CALU, CLOCK, FAM8A1, FLNB, GXYLT1, KIF1B, LIFR, LPGAT1, MAVS, RND3, SEL1L, TMEM192, TMEM33, TMEM56, AAED1, AK4, ASS1, DPY19L4, INSIG1, MGST1, PLIN2, PYGL, RDH11, SC5DL, SLC31A1, SLC35A3, TFPI, ACSL3, ATP1B1, HILPDA, RMDN2, DST, TAF13, ZFP1, ARL5B, SLC11A2, TMEM81, CCNE1, FAM111B, ABCD3, SOWAHC, SLC2A13 |
| 6 | AGER, ALDH1B1, BCAM, BGN, CALHM1, CLDN18, CLIC5, COBL, COX4I2, CYP4B1, DSP, HPGD, IL13, INMT, ITIH3, KCNS3, LMO3, MAP2, RDH16, RHPN1, S100A14, TNNC1, SUSD5, TBX4, CAMK2N1, EPAS1, FOXF1, HLF, HOPX, ID1, NEBL, PRKCE, SIK1, A1CF, ANXA3, CAV2, HPN, LPPR1, NR1H4, PPP1R14C, SLC7A2, SLCO1B1, SULT1C4, TEAD4, TSPAN6, VNN1, PRSS12, SERPINA10, BHMT2, BMP7, C6, KIF21A, MYH14, SEC14L4, TTPA, CA14, CCL16, CDKL1, CYP2A13, CYP2J2, FAM180A, G6PC, GRAMD1C, HFE, LRG1, SHROOM2, AGTR1, GGH, GPLD1, AKR1C1, DHCR24, FZD5, LPHN2, NAV2, N4BP3, XDH, C5, RBP4, FCN3, MPV17L, CCDC68, RELN, DSG2, CLUL1, SSUH2, TCF21, WDR49, ANXA8, KRT7 |

Table 3.1: **Genes in Clusters Induced By Mapper**. Red colored genes are the genes reported to be differentially expressed in peripheral blood mononuclear cells of rhesus infected with MARV by Connor et al.(Connor et al. 2015)

cal processes that might explain the data.

In Table 3.2, the top 5 gene ontology terms are shown for each cluster.

The cluster 1 contains the continually down-regulated genes, and the notable genes are CD40LG, LTA, and CXCR2 (**fig:cluster1˙genes**). For example, CD40LG and LTA encode for proteins which are the members of TNF superfamily and primarily produced in lymphocytes of many types. The rapid decrease in their expressions may be associated with lymphocyte apoptosis, lymphocyte migration, or independent down-regulation mechanism during infection. Further studies are required to pinpoint the exact cause.

Interestingly, the majority of GO terms corresponding to the cluster 4 corresponds to immune pathways. Indeed, we capture the important immune genes such as IL6, IFNG, IL1B, and CCL2. The GO terms with strong significance show up in cluster 2 and 6, which

| Cluster | GO.ID | Term | Annotated | Significant | Expected | classicFisher |
|---|---|---|---|---|---|---|
| 1 | GO:0006813 | potassium ion transport | 10 | 6 | 1.35 | 0.00069 |
| 1 | GO:0071804 | cellular potassium ion transport | 10 | 6 | 1.35 | 0.00069 |
| 1 | GO:0071805 | potassium ion transmembrane transport | 10 | 6 | 1.35 | 0.00069 |
| 1 | GO:0007268 | synaptic transmission | 28 | 10 | 3.77 | 0.00201 |
| 1 | GO:0099536 | synaptic signaling | 28 | 10 | 3.77 | 0.00201 |
| 2 | GO:0019752 | carboxylic acid metabolic process | 102 | 31 | 15.16 | 9.90E-06 |
| 2 | GO:0006082 | organic acid metabolic process | 111 | 32 | 16.5 | 2.40E-05 |
| 2 | GO:0043436 | oxoacid metabolic process | 107 | 31 | 15.91 | 3.10E-05 |
| 2 | GO:0006063 | uronic acid metabolic process | 10 | 7 | 1.49 | 0.00011 |
| 2 | GO:0019585 | glucuronate metabolic process | 10 | 7 | 1.49 | 0.00011 |
| 3 | GO:0008202 | steroid metabolic process | 51 | 8 | 2.65 | 0.003 |
| 3 | GO:0010955 | negative regulation of protein processin... | 10 | 3 | 0.52 | 0.012 |
| 3 | GO:1903318 | negative regulation of protein maturatio... | 10 | 3 | 0.52 | 0.012 |
| 3 | GO:0009896 | positive regulation of catabolic process | 19 | 4 | 0.99 | 0.014 |
| 3 | GO:0072376 | protein activation cascade | 29 | 5 | 1.5 | 0.014 |
| 4 | GO:0006952 | defense response | 182 | 87 | 37.52 | 3.10E-23 |
| 4 | GO:0002376 | immune system process | 205 | 90 | 42.27 | 8.90E-21 |
| 4 | GO:0006955 | immune response | 157 | 76 | 32.37 | 8.20E-20 |
| 4 | GO:0043207 | response to external biotic stimulus | 99 | 57 | 20.41 | 8.40E-19 |
| 4 | GO:0051707 | response to other organism | 99 | 57 | 20.41 | 8.40E-19 |
| 5 | GO:0010256 | endomembrane system organization | 19 | 11 | 3.68 | 0.00018 |
| 5 | GO:0034613 | cellular protein localization | 49 | 19 | 9.48 | 0.00083 |
| 5 | GO:0070727 | cellular macromolecule localization | 49 | 19 | 9.48 | 0.00083 |
| 5 | GO:0044249 | cellular biosynthetic process | 201 | 54 | 38.9 | 0.0013 |
| 5 | GO:0048667 | cell morphogenesis involved in neuron di... | 26 | 12 | 5.03 | 0.00141 |
| 6 | GO:0072358 | cardiovascular system development | 79 | 22 | 9.2 | 2.00E-05 |
| 6 | GO:0072359 | circulatory system development | 79 | 22 | 9.2 | 2.00E-05 |
| 6 | GO:0048568 | embryonic organ development | 21 | 10 | 2.44 | 3.30E-05 |
| 6 | GO:0009888 | tissue development | 98 | 24 | 11.41 | 8.30E-05 |
| 6 | GO:0014706 | striated muscle tissue development | 23 | 10 | 2.68 | 8.70E-05 |

Table 3.2: **Top 5 Enriched Gene Ontology Terms in Each Clusters**.

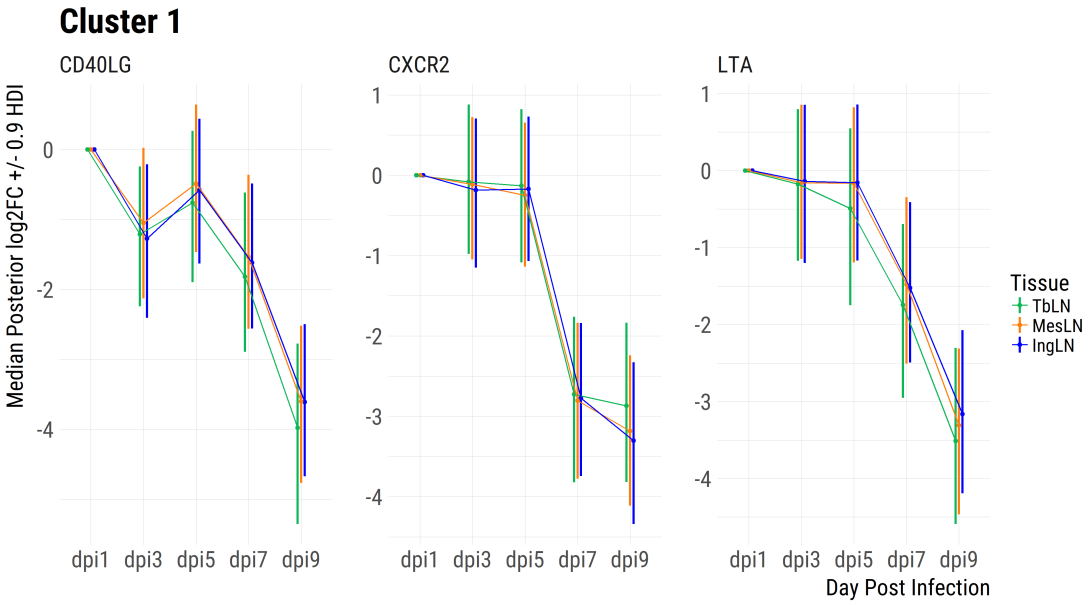


Figure 3.14: **Genes in cluster 1**.

are related to metabolic pathways and developmental pathways. It is important to keep in mind that the genes that are annotated with a particular GO term might have drastically expression pattern(i.e., genes that are anti-correlated), and thus this analysis may not fully uncover the biological meaning behind our archetypical gene sets. Further studies are necessary to study individual clusters or a set of genes in each cluster to identify potentially novel biological markers, receptors, and immune effector genes.

## Identification of Tissue-specific Genes

We wanted to identify the set of genes that have a unique pattern in a particular tissue compared to other tissues. This is in part motivated by the fact that phenotypically TbLN is drastically different the other two lymph nodes. To that end, we first defined the Day-wise Tissue Specificity Score (DTSS) for each gene $g$ at a particular dpi $d$. In particular, DTSS is defined as the posterior probability with the following criteria:

1. Effect Size Difference: $|log_2FC_{g,d}(t) - log_2FC_{g,d}(x)| > \kappa$
2. Uniqueness: $-\zeta < log_2FC_{g,d}(x) < \zeta$

Where $t$ is a target tissue and $x$ is the tissue that is being compared. In other words, log2FC of tissue $t$ at dpi d has to be greater than log2FC of tissue $x$ at dpi $d$ by $\kappa$ to ensure high effect size difference, and log2FC of tissue $x$ (the one that is compared against) should have a small magnitude $\zeta$ to ensure uniqueness for $d$ and no changes for $x$.

For illustration, we look at the DTSS at dpi 9 for TbLN against IngLN computed with posterior samples for DDX58 and TMEM97 in Figure 3.15. We chose these genes to illustrate the two cases, the one which would yield the low DTSS and the other the high

DTSS. We set $\kappa$ and $\zeta$ to be 1. With DDX58, we can see that at dpi9, it is not preferentially expressed in TbLN compared to IngLN because it has low effect size difference and low uniqueness, resulting in low DTSS (3.75%). In contrast, with TMEM97, the effect size difference is high and the expression is unique in TbLN, resulting in high DTSS (92.83%).

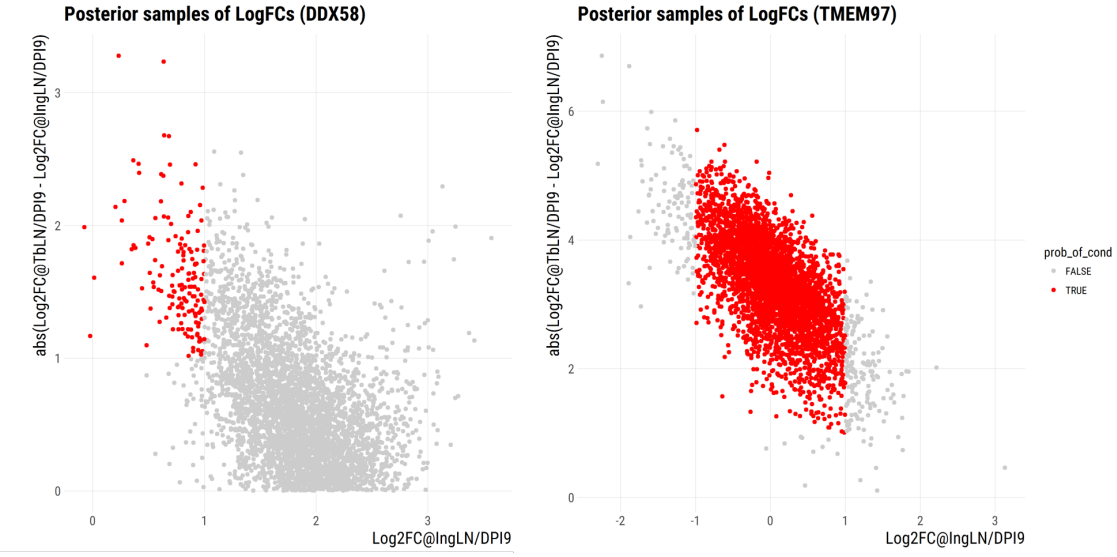


Figure 3.15: **Day-wise Tissue Specific Score for DPI9 and TbLN**.

Since we have the joint posterior distribution, we can compute the three sets of joint tissue specific score (JTSS) scores for global (G-JTSS), early (E-JTSS), and late (L-JTSS) stage of infection, which are defined as follows:

$$\text{G-JTSS} = Pr\left(\bigwedge_{d\in D, x\in X, x\neq t} |log_2FC_{g,d}(t) - log_2FC_{g,d}(x)| > \kappa \wedge |log_2FC_{g,d}(x)| < \zeta\right)$$

$$\text{E-JTSS} = Pr\left(\bigwedge_{d\in \{3,5\}, x\in X, x\neq t} |log_2FC_{g,d}(t) - log_2FC_{g,d}(x)| > \kappa \wedge |log_2FC_{g,d}(x)| < \zeta\right)$$

$$\text{L-JTSS} = Pr\left(\bigwedge_{d\in \{7,9\}, x\in X, x\neq t} |log_2FC_{g,d}(t) - log_2FC_{g,d}(x)| > \kappa \wedge |log_2FC_{g,d}(x)| < \zeta\right)$$

In other words, JTSSs are derived from DTSS contributing to individual days and individual tissues that are being compared. In Figure 3.16), we show the top 10 G-JTSS, E-JTSS, and L-JTSS for TbLN, MesLN, and IngLN.

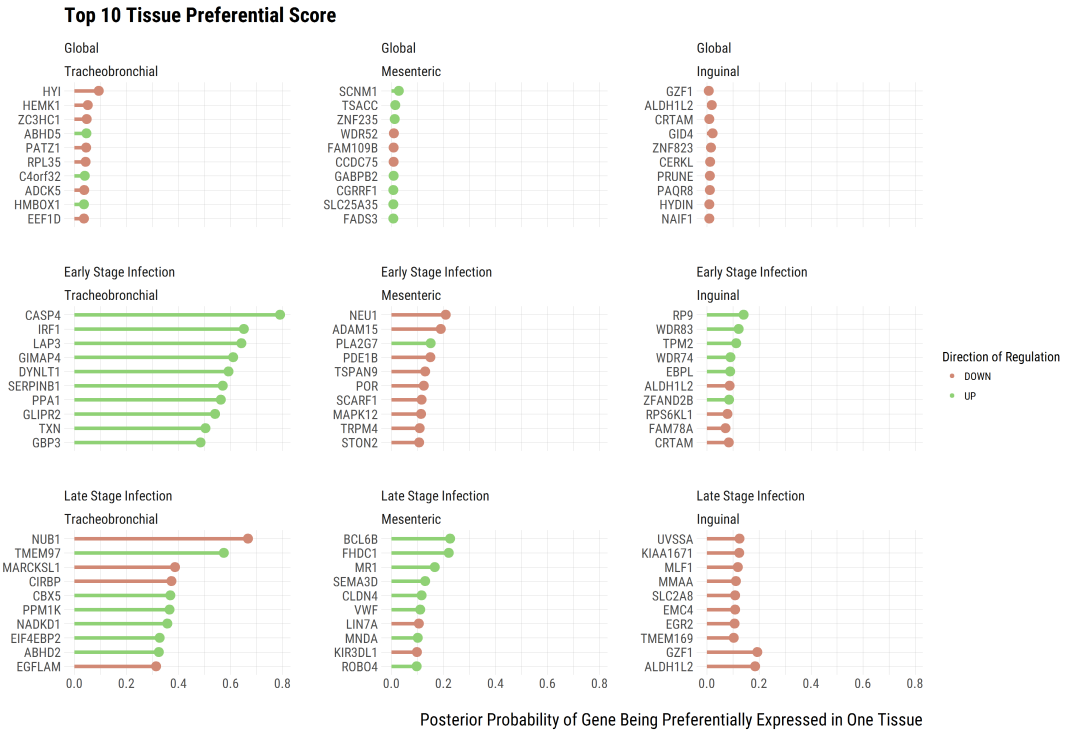


Figure 3.16: **Joint Tissue Specific Score**.

We noted that there are no genes that have preferential tissue expression throughout infection, indicated by the low posterior probability. When we look at E-JTSS and L-JTSS, we see that TbLN has several genes that have a relatively high posterior probability for tissue specificity. For example, CASP4, IRF1, and DYNLT1 are positively regulated in the early stage of infection in TbLN only. Interestingly, in MesLN,the majority of the top genes are down-regulated in the early stage of infection, but the posterior probabilities are low compared to others. In the last stage of infection, there are two genes that have the high posterior probability of L-JTSS, NUB1 and TMEM97. In particular, TbLN specific

late gene TMEM97 is interesting in that it is known to be a negative regulator of NPC1 (Ebrahimi-Fakhari et al. 2016), which is a protein required for filovirus entry (White and Schornberg 2012). Why would MARV let a protein down-regulate the proteins required for entry? There are a couple of potential explanations for this observation. One hypothesis that MARV actively down-regulate its own host protein for some unknown purposes. It has been reported that some virus is known to negatively regulate its host receptor; for instance, Hepatitis C Virus is known to down-regulates the host receptor CD81 required for entry (Ke and Chen 2013); the reason for phenomenon still remains unclear. Another hypothesis is the up-regulation of TMEM97 may be counterbalanced by post-transcriptional or translational modifications for degradation. Further studies are necessary to investigate this seemingly paradoxical event.

## Interactive MARV Infection Gene Expression Viewer

To facilitate the spread of information and generating hypotheses, we developed the interactive web application using R Shiny (Chang et al. 2017). With this tool, a user can view the genes in 40 different immune pathways (Figure 3.17) and how their log2FC change over time.

This is useful for obtaining a holistic view of pathway members gene expression profile changes upon MARV infection.

If users are interested in knowing the baseline expression and the log2FC for individual genes, they can do it at “Baseline & Posterior log2FC” in the web application 3.18.

This is a great tool to generate hypothesis and validate the output from pathway

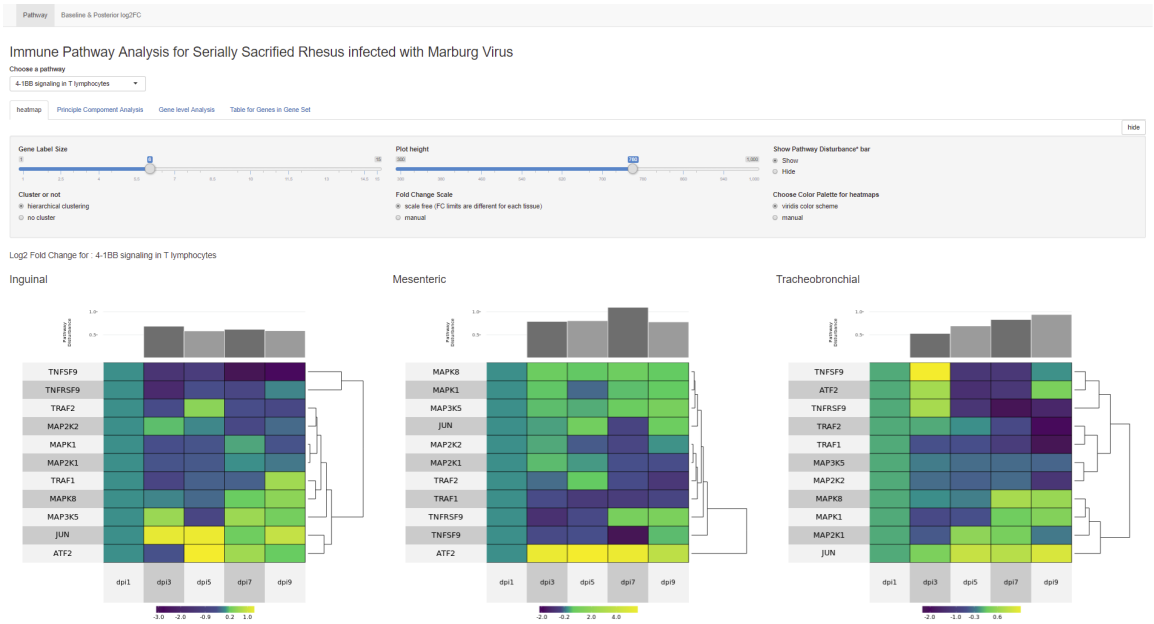


Figure 3.17: **Interactive Gene Expression Viewer - Immune pathway**. An end user can explore the gene expressions of the set of genes associated with a particular immune pathways.


Figure 3.18: **Interactive Gene Expression Viwer - Gene Explorer**. A base expression and log2FC for a particular gene can be queried

analysis. For example, the pathway analysis from Figure 3.9 indicated "Interferon Signaling" related genes are enriched. We can see from Figure 3.18 the pathway member OASL is indeed up-regulated over DPI. What is nice about seeing log2FC over time is that we can view the kinetics are different among tissues; there is a delay of OASL transcription between TbLN and IngLN. The expression viewer can be accessed at `https://rabadan.c2b2.columbia.edu/marburg`.

## 3.4 Conclusion

In this chapter, we explored the global characterization of transcriptional profile changes of three lymph nodes of rhesus monkeys upon infection of MARV. We demonstrated the multiple biological findings powered with the posterior samples provided with MAGE pipeline discussed in Chapter 2. First, we showed that the viral load can be accurately predicted with MAGE posterior and showed that the amount of viral RNA goes up over time, indicating active viral replication in the samples. The MARV replication is shown to be the fastest in TbLN, potentially because it is a draining lymph node. We then looked at the global expression pattern via PCA; infection patterns are distinguished in each dpi group, and infection pattern nicely divides into early and late stages. For differential gene expression analysis, DEG goes up as DPI goes up, but it's the highest in TbLN. Pathway analysis reveals that innate systems are activated in all three lymph nodes, but with slight of delay in MesLN and IngLN compared to TbLN. Among the differentially expressed genes, there are six gene expression archetype groups, which we identified via Mapper algorithm. Lastly, we identified a set of early and late genes that are preferentially expressed in TbLN.

The biological findings and the visualization tool provided in this study will contribute to the better understanding of host responses to MARV infection.

## 3.5 Material and Methods

### Animals, virus, and aerosol exposure

The samples used for these analyses were previously collected in a study by Lin et al (Lin et al. 2015). Research was conducted under IACUC approved protocol in compliance with the Animal Welfare Ace, PHS Policy, and other deferral statues and regulations relating to animals and experiments involving animals. The facility where this research was conducted is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International and adheres to principles stated in the Guide for the Care and Use of Laboratory Animals. Briefly, 15 Rhesus macaques were obtained from World Wide Primates (Miami, FL) and randomly assigned to sacrifice groups (N = 3; days 1, 3, 5, 7, and 9 post-exposure). NHPs were acclimated to the BSL-4 laboratory prior to the beginning of the study. At days -8 and -7, NHPs received a physical examination and baseline parameters for hematology, serum chemistry, coagulation, and cytokine expression were determined. On day 0, animals were exposed to a small particle aerosol dose of 1,000 PFU of Marburg virus Angola (Marburg virus H. sapiens-tc/ANG/2005/Angola-1379c; GenBank: DQ447653.1) using the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) head-only automated bioaerosol exposure system (ABES-II). Starting virus concentrations and exposure dose were confirmed by plaque assay (ibid.).

## Sample collection

Following euthanasia, a complete necropsy was performed on each animal in the BSL-4 laboratory. Blood collected at euthanasia and sera and PBMCs were isolated. Tissues were collected and a portion was fixed in 10% neutral buffered formalin for a minimum of 21 days for pathological examination and another portion was snap frozen and homogenized in Trizol LS for virus genome quantification and RNA sequencing. For this study, we focused on the tracheobronchial, mesenteric, and inguinal lymph nodes. These samples were removed from the BSL-4 laboratory upon sufficient inactivation in Trizol LS.

## RNA sequencing

RNA was isolated from the lymph node samples using the PureLink RNA Mini kit (Thermo Fisher Scientific). Libraries were generated on the Sciclone G3 liquid handling robot (Perkin Elmer) using the TruSeq RNA Library Prep Kit v2 (Illumina). Library quality was evaluated suing the DNA 1K Chip on the LabChipGX (Perkin Elmer) and quantified by qPCR using the KAPA Complete qPCR kit for Illumina libraries (Kapa Biosystems). Libraries were diluted to 10nM and cluster generation was performed on the Illumina cBot. Libraries were sequenced on the HiSeq 2500 using a paired end 1x100bp, single index format.

## RNA-seq analysis

Sequence quality filtering and adapter trimming was performed using TrimGalore v0.3.7 (*Trim Galore!*) in paired-end mode with a quality phred score cutoff of

20, a maximum trimming error rate of 0.1, and a minimum required adapter overlap of 1 (-f fastq -e 0.1 -q 20 -O 1). The adapter sequences AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC and AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT were used to trim adapters. After trimming, each RNA-seq sample was checked for quality using fastQC v0.10.1 (*FastQC*) and multiqc (v0.7) (Ewels et al. 2016) to aggregate the results. For gene expression quantification, we used Kallisto v0.43.0 (Bray et al. 2015) to map the preprocessed reads to the transcriptome reference sequences generated from the *Macaca mulatta* genome annotation version the MacaM Rhesus Genome v7 (Zimin et al. 2014). The output from Kallisto generated the estimated count and TPM (transcripts per million) for each transcript. To transform the transcript level quantification to gene level expression, we chose to use the isoform with the highest mean TPM across all samples to be the representative isoform for gene level quantification. Finally, we excluded any sample that had less than 500000 mapped reads. A total of 6 samples were excluded from further analysis.

## Viral Transcriptome Quantification

For viral transcriptome quantification, we mapped the RNA reads to MARV genome (Lake Victoria marburgvirus strain Uganda 371Bat2007, complete genome (GenkBank: FJ750958.1)) using Kallisto v0.43.0 (Bray et al. 2015). We summed the reads mapped to individual genes in each sample to obtain a pooled estimate for viral load. We used the resulting output for the MAGE pipeline and obtained the posterior for log2FC and baseline expression.

## Global Gene Expression Analysis

Global gene expression analysis was performed by Principle component analysis(PCA) using "prcomp" function in R. PCA was performed using the posterior Log2 fold change for 13,918 genes as the input with scaling and centering.

## Pathway Analysis

Ingenuity Pathway Analysis is used (*Ingenuity Pathway Analysis, QIAGEN Inc.*) for the canonical pathway analysis. For the input of the pathway analysis, we used the differentially expressed genes selected using 90%high density interval and the ROPE radius of 1.5. We only considerd the pathways that has the p-value less than 0.05.

## Unsupervised Clustering via Mapper

Since Mapper requires to provide the two tunable parameters, *resolution* and *gain*, we performed the grid search on these parameters (Figure 3) in 3.5 and chose the parameter set that maximizes the Dunn index (Dunn 1973) of the clustering in the resulting network $G$, which is defined as following:

$$DI(G) = \frac{min_{i \leq j \leq m} D(c_i, c_j)}{max_{1 \leq k \leq m} T(c_k)}$$

Where $D$ corresponds to a distance function (max distance between members in two clusters $c_i$ and $c_j$), and $m$ the number of clusters, and $T$ corresponds to the intra distance function. The clustering algorithm for a network, we used the community detection algorithm based on edge betweenness (Girvan and Newman 2002). We used the "cluster-edge-

betweenness" function from igraph in R (Csardi and Nepusz 2006). There were 8 singleton nodes that did not have with any edge with other nodes. These singleton nodes contain 9 genes, which are BMP5, C19orf77, DNAJA4, FKBP4, IFNA4, KCTD19, KLKB1, OTX1, and TMIGD1, and they are removed from the downstream analysis.

# *Conclusion*

In my thesis, I have discussed the three important components of the transcriptome analysis and how they contributed to our understanding of Marburg virus infection and immune responses in animal hosts. In the first of my thesis, I have presented the comprehensive *de novo* reference transcriptome of *Rousettus aegyptiacus*. The challenge I addressed and contribution I made is to come up with the reference transcriptome without genome sequences available. There I used an iterative approach to combine the individual assemblies and performed homology-based annotation to obtain the transcript level annotation of previously uncharacterized *Rousettus aegyptiacus* transcriptome. I tested the transcriptome against several metrics to show its validity and made biological remarks about the transcriptome of *Rousettus aegyptiacus*. In the next chapter of my thesis, I delved into the development of the statistical machinery combining the best practices of gene expression analysis, Bayesian statistics, and multilevel modeling. I constructed the pipeline in which the raw sequence input get transformed into the posterior MCMC samples. The model implemented exploited the multilevel modeling structure of the data, resulting in regularized estimates of parameters, with which users can make practical and useful inference for biological questions. I showed that the model is valid, scalable, and useful for the RNAseq experiments with the complex experimental designs. In the last part of my thesis,

I employed MAGE pipeline to characterize the immune responses of the three lymph node tissues of *Macaca mulata* infected MARV over the course of 9 days. The global expression pattern was the separation between early and late stages. Furthermore, I identified a set of genes which are differentially expressed at each DPI compared to day 1 and inferred the implicated pathways. I also identified the genes which are preferentially expressed in TbLN and not in other Lymph nodes. Lastly, I clustered the genes into six groups in an unsupervised manner using a topological method to capture the distinctive trends present in the expression profile of lymph nodes of rhesuses infected with MARV. The pipeline and references I developed and the findings I made with these tools will contribute to the growing body of Marburg virus research and helping researchers in the field generate more interesting biological questions and puzzles to tackle this deadly virus.

# *Bibliography*


Adams, Mark D et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project". In: *Science* 252.5013, pp. 1651–1656.

Aken, Bronwen L et al. (2016). "The Ensembl gene annotation system". In: *Database* 2016.

Alexa, Adrian and Jorg Rahnenfuhrer (2010). *topGO: topGO: Enrichment analysis for Gene Ontology*. R package version 2.18.0.

Altschul, Stephen F et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic acids research* 25.17, pp. 3389–3402.

Amman, Brian R et al. (2012). "Seasonal pulses of Marburg virus circulation in juvenile Rousettus aegyptiacus bats coincide with periods of increased risk of human infection". In: *PLoS pathogens* 8.10, e1002877.

Amman, Brian R et al. (2015). "Oral shedding of Marburg virus in experimentally infected Egyptian fruit bats (Rousettus Aegyptiacus)". In: *Journal of wildlife diseases* 51.1, pp. 113–124.

Aryee, Martin J et al. (2009). "An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation)". In: *BMC bioinformatics* 10.1, p. 409.

Ashburner, Michael et al. (2000). "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1, pp. 25–29.

Auer, Paul L and RW Doerge (2010). "Statistical design and analysis of RNA sequencing data". In: *Genetics* 185.2, pp. 405–416.

Bar-Joseph, Ziv, Anthony Gitter, and Itamar Simon (2012). "Studying and modelling dynamic biological processes using time-series gene expression data". In: *Nature reviews. Genetics* 13.8, p. 552.

Bar-Joseph, Ziv et al. (2003). "Continuous representations of time-series gene expression data". In: *Journal of Computational Biology* 10.3-4, pp. 341–356.

Barczak, Amy K et al. (2012). "RNA signatures allow rapid identification of pathogens and antibiotic susceptibilities". In: *Proceedings of the National Academy of Sciences* 109.16, pp. 6217–6222.

Bausch, Daniel G et al. (2006). "Marburg hemorrhagic fever associated with multiple genetic lineages of virus". In: *New England Journal of Medicine* 355.9, pp. 909–919.

Bean, Andrew GD et al. (2013). "Studying immunity to zoonotic diseases in the natural host [mdash] keeping it real". In: *Nature Reviews Immunology* 13.12, pp. 851–861.

Betancourt, Michael and Mark Girolami (2015). "Hamiltonian Monte Carlo for hierarchical models". In: *Current trends in Bayesian methodology with applications* 79, p. 30.

Bolstad, Benjamin M et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics* 19.2, pp. 185–193.

Bosio, Catharine M et al. (2003). "Ebola and Marburg viruses replicate in monocyte-derived dendritic cells without inducing the production of cytokines and full maturation". In: *The Journal of infectious diseases* 188.11, pp. 1630–1638.

Brauburger, Kristina et al. (2012). "Forty-five years of Marburg virus research". In: *Viruses* 4.10, pp. 1878–1927.

Brawand, David et al. (2011). "The evolution of gene expression levels in mammalian organs". In: *Nature* 478.7369, pp. 343–348.

Bray, Nicolas et al. (2015). "Near-optimal RNA-Seq quantification". In: *arXiv preprint arXiv:1505.02710*.

Brazma, Alvis et al. (2001). "Minimum information about a microarray experiment (MIAME)toward standards for microarray data". In: *Nature genetics* 29.4, pp. 365–371.

Brooks, Stephen P and Andrew Gelman (1998). "General methods for monitoring convergence of iterative simulations". In: *Journal of computational and graphical statistics* 7.4, pp. 434–455.

Calisher, Charles H et al. (2006). "Bats: important reservoir hosts of emerging viruses". In: *Clinical microbiology reviews* 19.3, pp. 531–545.

Cámara, Pablo G (2017). "Topological methods for genomics: Present and future directions". In: *Current Opinion in Systems Biology* 1, pp. 95–101.

Carlsson, Gunnar (2009). "Topology and data". In: *Bulletin of the American Mathematical Society* 46.2, pp. 255–308.

Carpenter, Bob et al. (2016). "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 20, pp. 1–37.

Chang, Winston et al. (2017). *shiny: Web Application Framework for R*. R package version 1.0.3. URL: https://CRAN.R-project.org/package=shiny.

*Chronology of Marburg Hemorrhagic Fever Outbreaks* (2014). https://www.cdc.gov/vhf/marburg/resources/outbreak-table.html. [Online; accessed 2017-08-21].

Chua, KB et al. (2000). "Nipah virus: a recently emergent deadly paramyxovirus". In: *Science* 288.5470, pp. 1432–1435.

Conesa, Ana et al. (2016). "A survey of best practices for RNA-seq data analysis". In: *Genome biology* 17.1, p. 13.

Connor, John H et al. (2015). "Transcriptional profiling of the immune response to Marburg virus infection". In: *Journal of virology* 89.19, pp. 9865–9874.

Csardi, Gabor and Tamas Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems, p. 1695. URL: http://igraph.org.

Degexp, Degexp et al. (2010). "DEGseq". In: 2.

Dobin, Alexander et al. (2013). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21.

Duggan, David J et al. (1999). "Expression profiling using cDNA microarrays." In: *Nature genetics* 21.

Dunn, Joseph C (1973). "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters". In:

Ebrahimi-Fakhari, Darius et al. (2016). "Reduction of TMEM97 increases NPC1 protein levels and restores cholesterol trafficking in Niemann-pick type C1 disease cells". In: *Human molecular genetics* 25.16, pp. 3588–3599.

Ewels, Philip et al. (2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19, pp. 3047–3048.

*FastQC*. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Feldmann, H and MP Kiley (1999). "Classification, structure, and replication of filoviruses." In: *Current topics in microbiology and immunology* 235, p. 1.

Feldmann, Heinz and Thomas W Geisbert (2011). "Ebola haemorrhagic fever". In: *The Lancet* 377.9768, pp. 849–862.

Field, Hume, Brad McCall, and Janine Barrett (1999). "Australian bat lyssavirus infection in a captive juvenile black flying fox." In: *Emerging infectious diseases* 5.3, p. 438.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

Gear, Js S et al. (1975). "Outbreake of Marburg virus disease in Johannesburg." In: *Br Med J* 4.5995, pp. 489–493.

Geisbert, Thomas W et al. (2007). "Marburg virus Angola infection of rhesus macaques: pathogenesis and treatment with recombinant nematode anticoagulant protein c2". In: *The Journal of infectious diseases* 196.Supplement_2, S372–S381.

Geisbert, Thomas W et al. (2008). "Vesicular stomatitis virus-based vaccines protect nonhuman primates against aerosol challenge with Ebola and Marburg viruses". In: *Vaccine* 26.52, pp. 6894–6900.

Geisbert, Thomas W et al. (2009). "Single-injection vaccine protects nonhuman primates against infection with marburg virus and three species of ebola virus". In: *Journal of virology* 83.14, pp. 7296–7304.

Gelman, Andrew and Jennifer Hill (2007). *Data analysis using regression and multilevel-hierarchical models*. Vol. 1. Cambridge University Press New York, NY, USA.

Gelman, Andrew and Donald B Rubin (1992). "Inference from iterative simulation using multiple sequences". In: *Statistical science*, pp. 457–472.

— (1995). "Avoiding model selection in Bayesian social research". In: *Sociological methodology* 25, pp. 165–173.

Gelman, Andrew et al. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". In: *Bayesian analysis* 1.3, pp. 515–534.

Gelman, Andrew et al. (2014). *Bayesian data analysis*. Vol. 2. CRC press Boca Raton, FL.

Gibbs, Richard A et al. (2007). "Evolutionary and biomedical insights from the rhesus macaque genome". In: *science* 316.5822, pp. 222–234.

Girvan, Michelle and Mark EJ Newman (2002). "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12, pp. 7821–7826.

Grabherr, Manfred G et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome". In: *Nature biotechnology* 29.7, pp. 644–652.

Groot, Raoul J de et al. (2013). "Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group". In: *Journal of virology* 87.14, pp. 7790–7792.

Heinonen, Markus et al. (2014). "Detecting time periods of differential gene expression using Gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction". In: *Bioinformatics* 31.5, pp. 728–735.

Hu, Zhi-Liang, Jie Bao, and James M Reecy (2008). "CateGOrizer: a web-based program to batch analyze gene ontology classification categories". In: *Online Journal of Bioinformatics* 9.2, pp. 108–112.

*Ingenuity Pathway Analysis, QIAGEN Inc.* URL: `https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis`.

Johnson, ED et al. (1996). "Characterization of a new Marburg virus isolated from a 1987 fatal case in Kenya". In: *Archives of Virology Supplement*, pp. 101–114.

Jones, Megan EB et al. (2015). "Experimental Inoculation of Egyptian Rousette Bats (Rousettus aegyptiacus) with Viruses of the Ebolavirus and Marburgvirus Genera". In: *Viruses* 7.7, pp. 3420–3442.

Ke, Po-Yuan and Steve S-L Chen (2013). "Active RNA replication of hepatitis C virus downregulates CD81 expression". In: *PloS one* 8.1, e54866.

Kiley, MP et al. (1982). "Filoviridae: a taxonomic home for Marburg and Ebola viruses?" In: *Intervirology* 18.1-2, pp. 24–32.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg (2015). "HISAT: a fast spliced aligner with low memory requirements". In: *Nature methods* 12.4, pp. 357–360.

Kim, Daehwan et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome biology* 14.4, R36.

Krämer, Andreas et al. (2013). "Causal analysis approaches in ingenuity pathway analysis". In: *Bioinformatics* 30.4, pp. 523–530.

Kruschke, John K (2013). "Bayesian estimation supersedes the t test." In: *Journal of Experimental Psychology: General* 142.2, p. 573.

Kruschke, John K and Torrin M Liddell (2017). "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective". In: *Psychonomic Bulletin & Review*, pp. 1–29.

Kruschke, John (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kuhn, Jens H et al. (2010). "Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations". In: *Archives of virology* 155.12, pp. 2083–2103.

Law, Charity W et al. (2014). "voom : precision weights unlock linear model analysis tools for RNA-seq read counts". In: pp. 1–17.

Lee, Albert et al. (2014). "Transcriptome reconstruction and annotation of cynomolgus and African green monkey". In: *BMC genomics* 15.1, p. 846.

Leroy, Eric M et al. (2005). "Fruit bats as reservoirs of Ebola virus". In: *Nature* 438.7068, pp. 575–576.

Li, Bo and Colin N Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC bioinformatics* 12.1, p. 323.

Li, Weizhong and Adam Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". In: *Bioinformatics* 22.13, pp. 1658–1659.

Li, Wendong et al. (2005). "Bats are natural reservoirs of SARS-like coronaviruses". In: *Science* 310.5748, pp. 676–679.

Liang, Peng, Arthur B Pardee, et al. (1992). "Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction". In: *Science* 257.5072, pp. 967–971.

Lin, Kenny L et al. (2015). "Temporal characterization of Marburg virus Angola infection following aerosol challenge in rhesus macaques". In: *Journal of virology* 89.19, pp. 9875–9885.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: pp. 1–21. DOI: 10.1186/s13059-014-0550-8.

Lum, PY et al. (2013). "Extracting insights from the shape of complex data using topology". In: *Scientific reports* 3, p. 1236.

Martin, Marcel (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet. journal* 17.1, pp–10.

McCarthy et al. (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic Acids Research* 40.10, pp. –9.

McElreath, Richard (2015). *Statistical Rethinking. A Bayesian Course with Examples in R and Stan*. Chapman and HallCRC.

Middleton, DJ et al. (2007). "Experimental Nipah virus infection in pteropid bats (Pteropus poliocephalus)". In: *Journal of comparative pathology* 136.4, pp. 266–272.

Mohamadzadeh, Mansour, Lieping Chen, and Alan L Schmaljohn (2007). "How Ebola and Marburg viruses battle the immune system". In: *Nature reviews. Immunology* 7.7, p. 556.

Moratelli, Ricardo and Charles H Calisher (2015). "Bats and zoonotic viruses: can we confidently link bats with emerging deadly viruses?" In: *Memorias do Instituto Oswaldo Cruz* 110.1, pp. 1–22.

Morens, David M, Gregory K Folkers, and Anthony S Fauci (2004). "The challenge of emerging and re-emerging infectious diseases". In: *Nature* 430.6996, p. 242.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature methods* 5.7, pp. 621–628.

*NCBI Eukaryotic genomes annotations*. URL: `http://www.ncbi.nlm.nih.gov/genome/annotation\_euk/all/`.

Nakayama, Eri and Masayuki Saijo (2013). "Animal models for Ebola and Marburg virus infections". In: *Frontiers in microbiology* 4.

Nelder, John Ashworth and R Jacob Baker (1972). *Generalized linear models*. Wiley Online Library.

Nikiforov, VV et al. (1993). "A case of a laboratory infection with Marburg fever". In: *Zhurnal mikrobiologii, epidemiologii, i immunobiologii* 3, pp. 104–106.

Nueda, María José, Sonia Tarazona, and Ana Conesa (2014). "Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series". In: *Bioinformatics* 30.18, pp. 2598–2602.

O'Leary, Nuala A et al. (2015). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic acids research* 44.D1, pp. D733–D745.

Ogawa, Hirohito et al. (2015). "Seroepidemiological prevalence of multiple species of filoviruses in fruit bats (Eidolon helvum) migrating in Africa". In: *Journal of Infectious Diseases*, jiv063.

Olival, Kevin J et al. (2013). "Ebola virus antibodies in fruit bats, Bangladesh". In: *Emerging infectious diseases* 19.2, p. 270.

Omatsu, Tsutomu et al. (2008). "Induction and sequencing of Rousette bat interferon $\alpha$ and $\beta$ genes". In: *Veterinary immunology and immunopathology* 124.1, pp. 169–176.

Oshlack, Alicia, Mark D Robinson, and Matthew D Young (2010). "From RNA-seq reads to differential expression results". In: *Genome biology* 11.12, p. 220.

Pan, Wei (2002). "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments". In: *Bioinformatics* 18.4, pp. 546–554.

Papenfuss, Anthony T et al. (2012). "The immune gene repertoire of an important viral reservoir, the Australian black flying fox". In: *BMC genomics* 13.1, p. 261.

Paweska, Janusz T et al. (2012). "Virological and serological findings in Rousettus aegyptiacus experimentally inoculated with vero cells-adapted hogan strain of Marburg virus". In: *PloS one* 7.9, e45479.

Paweska, Janusz T et al. (2015). "Lack of Marburg Virus Transmission From Experimentally Infected to Susceptible In-Contact Egyptian Fruit Bats". In: *Journal of Infectious Diseases*, jiv132.

Peng, Xinxia et al. (2014). "Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR)". In: *Nucleic acids research* 43.D1, pp. D737–D742.

Pimentel, Harold et al. (2017). "Differential analysis of RNA-seq incorporating quantification uncertainty". In: *Nature Publishing Group* 14.7. ISSN: 1548-7091. DOI: 10.1038/nmeth.4324. URL: http://dx.doi.org/10.1038/nmeth.4324.

Polson, Nicholas G, James G Scott, et al. (2012). "On the half-Cauchy prior for a global scale parameter". In: *Bayesian Analysis* 7.4, pp. 887–902.

Pourrut, Xavier et al. (2009). "Large serological survey showing cocirculation of Ebola and Marburg viruses in Gabonese bat populations, and a high seroprevalence of both viruses in Rousettus aegyptiacus". In: *BMC infectious diseases* 9.1, p. 159.

Quinlan, Aaron R and Ira M Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Rapaport, Franck et al. (2013). "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data". In: *Genome biology* 14.9, p. 3158.

Rizvi, Abbas H et al. (2017). "Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development". In: *Nature Biotechnology* 35.6, pp. 551–560.

Robertson, Gordon et al. (2010). "De novo assembly and analysis of RNA-seq data". In: *Nature methods* 7.11, pp. 909–912.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010a). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140.

— (2010b). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26, pp. –1.

Saéz, Almudena Marí et al. (2015). "Investigating the zoonotic origin of the West African Ebola epidemic". In: *EMBO molecular medicine* 7.1, pp. 17–23.

Schena, Mark et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *SCIENCE-NEW YORK THEN WASHINGTON-*, pp. 467–467.

Seim, Inge et al. (2013). "Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii". In: *Nature communications* 4.

Shaw, Timothy I et al. (2012). "Transcriptome sequencing and annotation for the Jamaican fruit bat (Artibeus jamaicensis)". In: *PloS one* 7.11, e48472.

Siegert, Rudolf (1972). "Marburg virus". In: *Canine Distemper Virus*. Springer, pp. 97–153.

Singh, Gurjeet, Facundo Mémoli, and Gunnar E Carlsson (2007). "Topological methods for the analysis of high dimensional data sets and 3d object recognition." In: *SPBG*, pp. 91–100.

Slenczka, Werner and Hans Dieter Klenk (2007). "Forty years of Marburg virus". In: *The Journal of infectious diseases* 196.Supplement_2, S131–S135.

Smith, DH et al. (1982). "Marburg-virus disease in Kenya". In: *The Lancet* 319.8276, pp. 816–820.

Smith, Ina et al. (2011). "Identifying Hendra virus diversity in pteropid bats". In: *PLoS One* 6.9, e25275.

Smyth, Gordon K (2005). "Limma: linear models for microarray data". In: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, pp. 397–420.

Spies, Daniel and Constance Ciaudo (2015). "Dynamics in transcriptomics: advancements in RNA-seq time course and downstream analysis". In: *Computational and structural biotechnology journal* 13, pp. 469–477.

Storey, John D et al. (2005). "Significance analysis of time course microarray experiments". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.36, pp. 12837–12842.

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550.

Sun, Xiaoxiao et al. (2016). "Statistical inference for time course RNA-Seq data using a negative binomial mixed-effect model". In: *BMC bioinformatics* 17.1, p. 324.

Swanepoel, Robert et al. (1996). "Experimental inoculation of plants and animals with Ebola virus." In: *Emerging infectious diseases* 2.4, p. 321.

Swanepoel, Robert et al. (2007). "Studies of reservoir hosts for Marburg virus". In: *Emerging infectious diseases* 13.12, p. 1847.

Teeling, Emma C et al. (2002). "Microbat paraphyly and the convergent evolution of a key innovation in Old World rhinolophoid microbats". In: *Proceedings of the National Academy of Sciences* 99.3, pp. 1431–1436.

Towner, Jonathan S et al. (2006). "Marburgvirus genomics and association with a large hemorrhagic fever outbreak in Angola". In: *Journal of virology* 80.13, pp. 6497–6516.

Towner, Jonathan S et al. (2007). “Marburg virus infection detected in a common African bat”. In: *PLoS One* 2.8, e764.

Towner, Jonathan S et al. (2009). “Isolation of genetically diverse Marburg viruses from Egyptian fruit bats”. In: *PLoS pathogens* 5.7, e1000536.

*TransDecoder*. URL: `https://transdecoder.github.io/`.

*Trim Galore!* URL: `https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`.

Van’t Veer, Laura J et al. (2002). “Gene expression profiling predicts clinical outcome of breast cancer”. In: *nature* 415.6871, pp. 530–536.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and Computing* 27.5, pp. 1413–1432.

Velculescu, Victor E et al. (1995). “Serial analysis of gene expression”. In: *Science* 270.5235, p. 484.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature reviews genetics* 10.1, pp. 57–63.

Watson, Andrew et al. (1998). “Technology for microarray analysis of gene expression”. In: *Current opinion in biotechnology* 9.6, pp. 609–614.

White, Judith M and Kathryn L Schornberg (2012). “A new player in the puzzle of filovirus entry”. In: *Nature Reviews Microbiology* 10.5, pp. 317–322.

Wikipedia (2004). *Cell Theory*. [Online; accessed 2017-10-18]. URL: `Celltheory`.

Williamson, MM et al. (1998). “Transmission studies of Hendra virus (equine morbillivirus) in fruit bats, horses and cats”. In: *Australian Veterinary Journal* 76.12, pp. 813–818.

Williamson, MM et al. (2000). “Experimental hendra virus infectionin pregnant guinea-pigs and fruit Bats (Pteropus poliocephalus)”. In: *Journal of Comparative Pathology* 122.2, pp. 201–207.

Wilson, Don E and DeeAnn M Reeder (2005). *Mammal species of the world: a taxonomic and geographic reference*. Baltimore: Johns Hopkins University Press.

Ye, Jian et al. (2012). “Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction”. In: *BMC bioinformatics* 13.1, p. 134.

Ye, Shui Qing et al. (2002). “Microarray, SAGE and their applications to cardiovascular diseases”. In: *Cell research* 12.2, pp. 105–115.

Young, Peter L et al. (1996). “Serologic evidence for the presence in Pteropus bats of a paramyxovirus related to equine morbillivirus.” In: *Emerging infectious diseases* 2.3, p. 239.

Yu, Guangchuang and Qing-Yu He (2016). “ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization”. In: *Molecular BioSystems* 12.2, pp. 477–479.

Zhang, Guojie et al. (2013). “Comparative analysis of bat genomes provides insight into the evolution of flight and immunity”. In: *Science* 339.6118, pp. 456–460.

Zimin, Aleksey V et al. (2014). “A new rhesus macaque assembly and annotation for next-generation sequencing analyses”. In: *Biol Direct* 9, p. 20.

# *Appendix*

## Chapter 1 Supplementary Table 1 - Tissue Specific Gene Ontology Terms

| Tissue | GO.ID | Term | Annotated | Significant | Expected | P-Value |
|---|---|---|---|---|---|---|
| BM | GO:0002376 | immune system process | 2007.00 | 53.00 | 22.44 | 0.00 |
| BM | GO:0009611 | response to wounding | 883.00 | 30.00 | 9.87 | 0.00 |
| BM | GO:0050832 | defense response to fungus | 15.00 | 5.00 | 0.17 | 0.00 |
| BM | GO:0007596 | blood coagulation | 490.00 | 19.00 | 5.48 | 0.00 |
| BM | GO:0050817 | coagulation | 492.00 | 19.00 | 5.50 | 0.00 |
| BM | GO:0007599 | hemostasis | 495.00 | 19.00 | 5.54 | 0.00 |
| BM | GO:0050878 | regulation of body fluid levels | 608.00 | 21.00 | 6.80 | 0.00 |
| BM | GO:0042060 | wound healing | 627.00 | 21.00 | 7.01 | 0.00 |
| BM | GO:0007155 | cell adhesion | 959.00 | 27.00 | 10.72 | 0.00 |
| BM | GO:0022610 | biological adhesion | 963.00 | 27.00 | 10.77 | 0.00 |
| BR | GO:0007268 | synaptic transmission | 712.00 | 49.00 | 8.02 | 0.00 |
| BR | GO:0007267 | cell-cell signaling | 1122.00 | 54.00 | 12.63 | 0.00 |
| BR | GO:0050804 | regulation of synaptic transmission | 224.00 | 22.00 | 2.52 | 0.00 |
| BR | GO:0007399 | nervous system development | 1811.00 | 58.00 | 20.39 | 0.00 |
| BR | GO:0006836 | neurotransmitter transport | 180.00 | 19.00 | 2.03 | 0.00 |
| BR | GO:0001505 | regulation of neurotransmitter levels | 181.00 | 19.00 | 2.04 | 0.00 |
| BR | GO:0048489 | synaptic vesicle transport | 119.00 | 16.00 | 1.34 | 0.00 |
| BR | GO:0097480 | establishment of synaptic vesicle locali... | 119.00 | 16.00 | 1.34 | 0.00 |
| BR | GO:0097479 | synaptic vesicle localization | 121.00 | 16.00 | 1.36 | 0.00 |
| BR | GO:0031175 | neuron projection development | 772.00 | 35.00 | 8.69 | 0.00 |
| HT | GO:0003012 | muscle system process | 322.00 | 39.00 | 3.62 | 0.00 |
| HT | GO:0006936 | muscle contraction | 275.00 | 36.00 | 3.10 | 0.00 |
| HT | GO:0060047 | heart contraction | 165.00 | 30.00 | 1.86 | 0.00 |
| HT | GO:0003015 | heart process | 166.00 | 30.00 | 1.87 | 0.00 |
| HT | GO:0061061 | muscle structure development | 496.00 | 42.00 | 5.58 | 0.00 |
| HT | GO:0048738 | cardiac muscle tissue development | 155.00 | 27.00 | 1.74 | 0.00 |
| HT | GO:0008016 | regulation of heart contraction | 139.00 | 26.00 | 1.56 | 0.00 |
| HT | GO:0060537 | muscle tissue development | 319.00 | 34.00 | 3.59 | 0.00 |
| HT | GO:0014706 | striated muscle tissue development | 307.00 | 33.00 | 3.46 | 0.00 |
| HT | GO:0055002 | striated muscle cell development | 150.00 | 25.00 | 1.69 | 0.00 |
| KY | GO:0055085 | transmembrane transport | 1125.00 | 47.00 | 13.34 | 0.00 |
| KY | GO:0006811 | ion transport | 1282.00 | 50.00 | 15.20 | 0.00 |
| KY | GO:0006820 | anion transport | 465.00 | 27.00 | 5.51 | 0.00 |
| KY | GO:0007588 | excretion | 58.00 | 10.00 | 0.69 | 0.00 |
| KY | GO:0001822 | kidney development | 236.00 | 17.00 | 2.80 | 0.00 |
| KY | GO:0006814 | sodium ion transport | 182.00 | 15.00 | 2.16 | 0.00 |
| KY | GO:0034220 | ion transmembrane transport | 767.00 | 30.00 | 9.09 | 0.00 |
| KY | GO:0015711 | organic anion transport | 353.00 | 20.00 | 4.18 | 0.00 |
| KY | GO:0072001 | renal system development | 251.00 | 17.00 | 2.98 | 0.00 |
| KY | GO:0001655 | urogenital system development | 286.00 | 18.00 | 3.39 | 0.00 |
| LG | GO:0045333 | cellular respiration | 150.00 | 11.00 | 1.69 | 0.00 |
| LG | GO:0022904 | respiratory electron transport chain | 99.00 | 9.00 | 1.11 | 0.00 |
| LG | GO:0022900 | electron transport chain | 100.00 | 9.00 | 1.13 | 0.00 |
| LG | GO:0001944 | vasculature development | 528.00 | 20.00 | 5.94 | 0.00 |
| LG | GO:0072001 | renal system development | 251.00 | 13.00 | 2.83 | 0.00 |
| LG | GO:0006120 | mitochondrial electron transport, NADH t... | 42.00 | 6.00 | 0.47 | 0.00 |
| LG | GO:0048514 | blood vessel morphogenesis | 449.00 | 17.00 | 5.05 | 0.00 |
| LG | GO:0001822 | kidney development | 236.00 | 12.00 | 2.66 | 0.00 |
| LG | GO:0001568 | blood vessel development | 510.00 | 18.00 | 5.74 | 0.00 |
| LG | GO:0042773 | ATP synthesis coupled electron transport | 50.00 | 6.00 | 0.56 | 0.00 |
| LN | GO:0046649 | lymphocyte activation | 491.00 | 27.00 | 5.23 | 0.00 |
| LN | GO:0045321 | leukocyte activation | 581.00 | 29.00 | 6.19 | 0.00 |
| LN | GO:0001775 | cell activation | 800.00 | 31.00 | 8.53 | 0.00 |
| LN | GO:0031295 | T cell costimulation | 58.00 | 10.00 | 0.62 | 0.00 |
| LN | GO:0031294 | lymphocyte costimulation | 59.00 | 10.00 | 0.63 | 0.00 |
| LN | GO:0002376 | immune system process | 2007.00 | 50.00 | 21.40 | 0.00 |
| LN | GO:0002682 | regulation of immune system process | 1037.00 | 34.00 | 11.06 | 0.00 |
| LN | GO:0002694 | regulation of leukocyte activation | 337.00 | 19.00 | 3.59 | 0.00 |
| LN | GO:0042110 | T cell activation | 352.00 | 19.00 | 3.75 | 0.00 |
| LN | GO:0050865 | regulation of cell activation | 367.00 | 19.00 | 3.91 | 0.00 |
| LV | GO:1901605 | alpha-amino acid metabolic process | 200.00 | 25.00 | 2.07 | 0.00 |
| LV | GO:0019752 | carboxylic acid metabolic process | 871.00 | 41.00 | 9.03 | 0.00 |
| LV | GO:0043436 | oxoacid metabolic process | 977.00 | 43.00 | 10.12 | 0.00 |
| LV | GO:0006082 | organic acid metabolic process | 994.00 | 43.00 | 10.30 | 0.00 |
| LV | GO:0006520 | cellular amino acid metabolic process | 429.00 | 29.00 | 4.45 | 0.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| LV | GO:0009063 | cellular amino acid catabolic process | 105.00 | 17.00 | 1.09 | 0.00 |
| LV | GO:1901606 | alpha-amino acid catabolic process | 82.00 | 15.00 | 0.85 | 0.00 |
| LV | GO:0044282 | small molecule catabolic process | 242.00 | 21.00 | 2.51 | 0.00 |
| LV | GO:0055114 | oxidation-reduction process | 878.00 | 37.00 | 9.10 | 0.00 |
| LV | GO:0016054 | organic acid catabolic process | 191.00 | 19.00 | 1.98 | 0.00 |
| OV | GO:0007018 | microtubule-based movement | 187.00 | 13.00 | 1.59 | 0.00 |
| OV | GO:0022414 | reproductive process | 976.00 | 28.00 | 8.30 | 0.00 |
| OV | GO:0008406 | gonad development | 172.00 | 12.00 | 1.46 | 0.00 |
| OV | GO:0008584 | male gonad development | 108.00 | 10.00 | 0.92 | 0.00 |
| OV | GO:0046546 | development of primary male sexual chara... | 108.00 | 10.00 | 0.92 | 0.00 |
| OV | GO:0045137 | development of primary sexual characteri... | 177.00 | 12.00 | 1.50 | 0.00 |
| OV | GO:0008585 | female gonad development | 85.00 | 9.00 | 0.72 | 0.00 |
| OV | GO:0046545 | development of primary female sexual cha... | 89.00 | 9.00 | 0.76 | 0.00 |
| OV | GO:0044702 | single organism reproductive process | 878.00 | 25.00 | 7.46 | 0.00 |
| OV | GO:0044703 | multi-organism reproductive process | 707.00 | 22.00 | 6.01 | 0.00 |
| PB | GO:0007155 | cell adhesion | 959.00 | 48.00 | 11.58 | 0.00 |
| PB | GO:0022610 | biological adhesion | 963.00 | 48.00 | 11.63 | 0.00 |
| PB | GO:0006935 | chemotaxis | 616.00 | 37.00 | 7.44 | 0.00 |
| PB | GO:0042330 | taxis | 616.00 | 37.00 | 7.44 | 0.00 |
| PB | GO:0030198 | extracellular matrix organization | 354.00 | 25.00 | 4.28 | 0.00 |
| PB | GO:0043062 | extracellular structure organization | 355.00 | 25.00 | 4.29 | 0.00 |
| PB | GO:0006928 | cellular component movement | 1562.00 | 52.00 | 18.87 | 0.00 |
| PB | GO:0009605 | response to external stimulus | 1782.00 | 56.00 | 21.52 | 0.00 |
| PB | GO:0040011 | locomotion | 1412.00 | 49.00 | 17.05 | 0.00 |
| PB | GO:0050896 | response to stimulus | 6462.00 | 120.00 | 78.05 | 0.00 |
| SP | GO:0019752 | carboxylic acid metabolic process | 871.00 | 43.00 | 10.91 | 0.00 |
| SP | GO:0043436 | oxoacid metabolic process | 977.00 | 43.00 | 12.24 | 0.00 |
| SP | GO:0006082 | organic acid metabolic process | 994.00 | 43.00 | 12.45 | 0.00 |
| SP | GO:0055114 | oxidation-reduction process | 878.00 | 40.00 | 11.00 | 0.00 |
| SP | GO:0009063 | cellular amino acid catabolic process | 105.00 | 15.00 | 1.32 | 0.00 |
| SP | GO:0044282 | small molecule catabolic process | 242.00 | 20.00 | 3.03 | 0.00 |
| SP | GO:0016054 | organic acid catabolic process | 191.00 | 18.00 | 2.39 | 0.00 |
| SP | GO:0046395 | carboxylic acid catabolic process | 191.00 | 18.00 | 2.39 | 0.00 |
| SP | GO:0044281 | small molecule metabolic process | 2563.00 | 67.00 | 32.10 | 0.00 |
| SP | GO:0006520 | cellular amino acid metabolic process | 429.00 | 23.00 | 5.37 | 0.00 |
| TT | GO:0046661 | male sex differentiation | 133.00 | 13.00 | 1.24 | 0.00 |
| TT | GO:0022414 | reproductive process | 976.00 | 31.00 | 9.10 | 0.00 |
| TT | GO:0007548 | sex differentiation | 227.00 | 15.00 | 2.12 | 0.00 |
| TT | GO:0008584 | male gonad development | 108.00 | 11.00 | 1.01 | 0.00 |
| TT | GO:0046546 | development of primary male sexual chara... | 108.00 | 11.00 | 1.01 | 0.00 |
| TT | GO:0045137 | development of primary sexual characteri... | 177.00 | 13.00 | 1.65 | 0.00 |
| TT | GO:0032501 | multicellular organismal process | 5406.00 | 81.00 | 50.38 | 0.00 |
| TT | GO:0044702 | single organism reproductive process | 878.00 | 27.00 | 8.18 | 0.00 |
| TT | GO:0007283 | spermatogenesis | 360.00 | 17.00 | 3.35 | 0.00 |
| TT | GO:0048232 | male gamete generation | 361.00 | 17.00 | 3.36 | 0.00 |

Table 1: Enriched Biological Processes in Tissues

| Tissue | GO.ID | Term | Annotated | Significant | Expected | P-Value |
|---|---|---|---|---|---|---|
| BM | GO:0050786 | RAGE receptor binding | 10.00 | 4.00 | 0.11 | 0.00 |
| BM | GO:0001047 | core promoter binding | 117.00 | 8.00 | 1.29 | 0.00 |
| BM | GO:0000979 | RNA polymerase II core promoter sequence... | 44.00 | 5.00 | 0.49 | 0.00 |
| BM | GO:0001046 | core promoter sequence-specific DNA bind... | 75.00 | 6.00 | 0.83 | 0.00 |
| BM | GO:0005518 | collagen binding | 57.00 | 5.00 | 0.63 | 0.00 |
| BM | GO:0042802 | identical protein binding | 1010.00 | 23.00 | 11.15 | 0.00 |
| BM | GO:0042803 | protein homodimerization activity | 587.00 | 16.00 | 6.48 | 0.00 |
| BM | GO:0008301 | DNA binding, bending | 17.00 | 3.00 | 0.19 | 0.00 |
| BM | GO:0050840 | extracellular matrix binding | 44.00 | 4.00 | 0.49 | 0.00 |
| BM | GO:0005504 | fatty acid binding | 23.00 | 3.00 | 0.25 | 0.00 |
| BR | GO:0005215 | transporter activity | 1081.00 | 31.00 | 11.21 | 0.00 |
| BR | GO:0005216 | ion channel activity | 342.00 | 15.00 | 3.55 | 0.00 |
| BR | GO:0022838 | substrate-specific channel activity | 353.00 | 15.00 | 3.66 | 0.00 |
| BR | GO:0046873 | metal ion transmembrane transporter acti... | 363.00 | 15.00 | 3.76 | 0.00 |
| BR | GO:0015267 | channel activity | 370.00 | 15.00 | 3.84 | 0.00 |
| BR | GO:0022803 | passive transmembrane transporter activi... | 370.00 | 15.00 | 3.84 | 0.00 |
| BR | GO:0022836 | gated channel activity | 281.00 | 13.00 | 2.91 | 0.00 |
| BR | GO:0022843 | voltage-gated cation channel activity | 133.00 | 9.00 | 1.38 | 0.00 |
| BR | GO:0022890 | inorganic cation transmembrane transport... | 435.00 | 16.00 | 4.51 | 0.00 |
| BR | GO:0015075 | ion transmembrane transporter activity | 708.00 | 21.00 | 7.34 | 0.00 |
| HT | GO:0008092 | cytoskeletal protein binding | 742.00 | 36.00 | 7.97 | 0.00 |
| HT | GO:0008307 | structural constituent of muscle | 41.00 | 11.00 | 0.44 | 0.00 |
| HT | GO:0003779 | actin binding | 366.00 | 24.00 | 3.93 | 0.00 |
| HT | GO:0044325 | ion channel binding | 76.00 | 10.00 | 0.82 | 0.00 |
| HT | GO:0042805 | actinin binding | 26.00 | 7.00 | 0.28 | 0.00 |
| HT | GO:0005198 | structural molecule activity | 520.00 | 22.00 | 5.59 | 0.00 |
| HT | GO:0051393 | alpha-actinin binding | 22.00 | 6.00 | 0.24 | 0.00 |
| HT | GO:0031432 | titin binding | 12.00 | 5.00 | 0.13 | 0.00 |
| HT | GO:0032036 | myosin heavy chain binding | 10.00 | 3.00 | 0.11 | 0.00 |
| HT | GO:0048037 | cofactor binding | 241.00 | 10.00 | 2.59 | 0.00 |
| KY | GO:0015291 | secondary active transmembrane transport... | 180.00 | 24.00 | 2.16 | 0.00 |
| KY | GO:0022857 | transmembrane transporter activity | 821.00 | 43.00 | 9.86 | 0.00 |
| KY | GO:0022891 | substrate-specific transmembrane transpo... | 756.00 | 41.00 | 9.08 | 0.00 |
| KY | GO:0022804 | active transmembrane transporter activit... | 278.00 | 26.00 | 3.34 | 0.00 |
| KY | GO:0015075 | ion transmembrane transporter activity | 708.00 | 39.00 | 8.51 | 0.00 |
| KY | GO:0005215 | transporter activity | 1081.00 | 47.00 | 12.99 | 0.00 |
| KY | GO:0008509 | anion transmembrane transporter activity | 246.00 | 23.00 | 2.96 | 0.00 |
| KY | GO:0022892 | substrate-specific transporter activity | 894.00 | 41.00 | 10.74 | 0.00 |
| KY | GO:0015293 | symporter activity | 124.00 | 16.00 | 1.49 | 0.00 |
| KY | GO:0008514 | organic anion transmembrane transporter ... | 147.00 | 17.00 | 1.77 | 0.00 |
| LG | GO:0003954 | NADH dehydrogenase activity | 38.00 | 7.00 | 0.42 | 0.00 |
| LG | GO:0008137 | NADH dehydrogenase (ubiquinone) activity | 38.00 | 7.00 | 0.42 | 0.00 |
| LG | GO:0050136 | NADH dehydrogenase (quinone) activity | 38.00 | 7.00 | 0.42 | 0.00 |
| LG | GO:0016651 | oxidoreductase activity, acting on NAD(P... | 81.00 | 9.00 | 0.89 | 0.00 |
| LG | GO:0016655 | oxidoreductase activity, acting on NAD(P... | 49.00 | 7.00 | 0.54 | 0.00 |
| LG | GO:0022857 | transmembrane transporter activity | 821.00 | 22.00 | 9.01 | 0.00 |
| LG | GO:0005215 | transporter activity | 1081.00 | 25.00 | 11.86 | 0.00 |
| LG | GO:0022892 | substrate-specific transporter activity | 894.00 | 21.00 | 9.81 | 0.00 |
| LG | GO:0005178 | integrin binding | 101.00 | 6.00 | 1.11 | 0.00 |
| LG | GO:0051183 | vitamin transporter activity | 18.00 | 3.00 | 0.20 | 0.00 |
| LN | GO:0004872 | receptor activity | 958.00 | 23.00 | 9.44 | 0.00 |
| LN | GO:0005102 | receptor binding | 1140.00 | 25.00 | 11.23 | 0.00 |
| LN | GO:0016614 | oxidoreductase activity, acting on CH-OH... | 98.00 | 6.00 | 0.97 | 0.00 |
| LN | GO:0004888 | transmembrane signaling receptor activit... | 670.00 | 16.00 | 6.60 | 0.00 |
| LN | GO:0003823 | antigen binding | 48.00 | 4.00 | 0.47 | 0.00 |
| LN | GO:0038023 | signaling receptor activity | 762.00 | 17.00 | 7.51 | 0.00 |
| LN | GO:0050839 | cell adhesion molecule binding | 167.00 | 7.00 | 1.64 | 0.00 |
| LN | GO:0005178 | integrin binding | 101.00 | 5.00 | 0.99 | 0.00 |
| LN | GO:0008009 | chemokine activity | 32.00 | 3.00 | 0.32 | 0.00 |
| LN | GO:0005283 | sodium:amino acid symporter activity | 10.00 | 2.00 | 0.10 | 0.00 |
| LV | GO:0016491 | oxidoreductase activity | 575.00 | 24.00 | 5.75 | 0.00 |
| LV | GO:0005319 | lipid transporter activity | 81.00 | 9.00 | 0.81 | 0.00 |
| LV | GO:0008236 | serine-type peptidase activity | 141.00 | 11.00 | 1.41 | 0.00 |
| LV | GO:0016645 | oxidoreductase activity, acting on the C... | 26.00 | 6.00 | 0.26 | 0.00 |
| LV | GO:0017171 | serine hydrolase activity | 143.00 | 11.00 | 1.43 | 0.00 |
| LV | GO:0004252 | serine-type endopeptidase activity | 121.00 | 10.00 | 1.21 | 0.00 |
| LV | GO:0005215 | transporter activity | 1081.00 | 28.00 | 10.81 | 0.00 |
| LV | GO:0048037 | cofactor binding | 241.00 | 12.00 | 2.41 | 0.00 |
| LV | GO:0003824 | catalytic activity | 4858.00 | 74.00 | 48.57 | 0.00 |
| LV | GO:0019842 | vitamin binding | 72.00 | 7.00 | 0.72 | 0.00 |
| OV | GO:0003777 | microtubule motor activity | 75.00 | 11.00 | 0.66 | 0.00 |
| OV | GO:0003774 | motor activity | 124.00 | 11.00 | 1.09 | 0.00 |
| OV | GO:0005509 | calcium ion binding | 571.00 | 14.00 | 5.03 | 0.00 |
| OV | GO:0016887 | ATPase activity | 367.00 | 10.00 | 3.23 | 0.00 |
| OV | GO:0005201 | extracellular matrix structural constitu... | 64.00 | 4.00 | 0.56 | 0.00 |
| OV | GO:0004867 | serine-type endopeptidase inhibitor acti... | 72.00 | 4.00 | 0.63 | 0.00 |
| OV | GO:0005198 | structural molecule activity | 520.00 | 11.00 | 4.58 | 0.01 |
| OV | GO:0005160 | transforming growth factor beta receptor... | 44.00 | 3.00 | 0.39 | 0.01 |
| OV | GO:0043565 | sequence-specific DNA binding | 681.00 | 13.00 | 6.00 | 0.01 |
| OV | GO:0017111 | nucleoside-triphosphatase activity | 698.00 | 12.00 | 6.15 | 0.02 |
| PB | GO:0004872 | receptor activity | 958.00 | 35.00 | 11.51 | 0.00 |
| PB | GO:0038023 | signaling receptor activity | 762.00 | 29.00 | 9.15 | 0.00 |
| PB | GO:0004888 | transmembrane signaling receptor activit... | 670.00 | 26.00 | 8.05 | 0.00 |
| PB | GO:0005201 | extracellular matrix structural constitu... | 64.00 | 8.00 | 0.77 | 0.00 |
| PB | GO:0005509 | calcium ion binding | 571.00 | 20.00 | 6.86 | 0.00 |
| PB | GO:0004871 | signal transducer activity | 1076.00 | 29.00 | 12.93 | 0.00 |
| PB | GO:0060089 | molecular transducer activity | 1076.00 | 29.00 | 12.93 | 0.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PB | GO:0003779 | actin binding | 366.00 | 15.00 | 4.40 | 0.00 |
| PB | GO:0019838 | growth factor binding | 107.00 | 8.00 | 1.29 | 0.00 |
| PB | GO:0005518 | collagen binding | 57.00 | 6.00 | 0.68 | 0.00 |
| SP | GO:0016491 | oxidoreductase activity | 575.00 | 30.00 | 6.78 | 0.00 |
| SP | GO:0003824 | catalytic activity | 4858.00 | 94.00 | 57.27 | 0.00 |
| SP | GO:0016655 | oxidoreductase activity, acting on NAD(P... | 49.00 | 9.00 | 0.58 | 0.00 |
| SP | GO:0003954 | NADH dehydrogenase activity | 38.00 | 8.00 | 0.45 | 0.00 |
| SP | GO:0008137 | NADH dehydrogenase (ubiquinone) activity | 38.00 | 8.00 | 0.45 | 0.00 |
| SP | GO:0050136 | NADH dehydrogenase (quinone) activity | 38.00 | 8.00 | 0.45 | 0.00 |
| SP | GO:0016651 | oxidoreductase activity, acting on NAD(P... | 81.00 | 10.00 | 0.95 | 0.00 |
| SP | GO:0048037 | cofactor binding | 241.00 | 15.00 | 2.84 | 0.00 |
| SP | GO:0031406 | carboxylic acid binding | 172.00 | 11.00 | 2.03 | 0.00 |
| SP | GO:0043177 | organic acid binding | 173.00 | 11.00 | 2.04 | 0.00 |
| TT | GO:0046983 | protein dimerization activity | 902.00 | 20.00 | 8.28 | 0.00 |
| TT | GO:0004871 | signal transducer activity | 1076.00 | 21.00 | 9.88 | 0.00 |
| TT | GO:0060089 | molecular transducer activity | 1076.00 | 21.00 | 9.88 | 0.00 |
| TT | GO:0042803 | protein homodimerization activity | 587.00 | 14.00 | 5.39 | 0.00 |
| TT | GO:0008528 | G-protein coupled peptide receptor activ... | 97.00 | 5.00 | 0.89 | 0.00 |
| TT | GO:0001653 | peptide receptor activity | 99.00 | 5.00 | 0.91 | 0.00 |
| TT | GO:0004872 | receptor activity | 958.00 | 17.00 | 8.79 | 0.01 |
| TT | GO:0003705 | RNA polymerase II distal enhancer sequen... | 81.00 | 4.00 | 0.74 | 0.01 |
| TT | GO:0010181 | FMN binding | 14.00 | 2.00 | 0.13 | 0.01 |
| TT | GO:0090079 | translation regulator activity, nucleic ... | 14.00 | 2.00 | 0.13 | 0.01 |

Table 2: Enriched Molecular Functions in Tissues

| Tissue | GO.ID | Term | Annotated | Significant | Expected | P-Value |
|---|---|---|---|---|---|---|
| BM | GO:0071944 | cell periphery | 3877.00 | 70.00 | 43.51 | 4.1E-6 |
| BM | GO:0031226 | intrinsic component of plasma membrane | 1234.00 | 31.00 | 13.85 | 1.5E-5 |
| BM | GO:0005886 | plasma membrane | 3800.00 | 67.00 | 42.64 | 1.9E-5 |
| BM | GO:0005887 | integral component of plasma membrane | 1193.00 | 30.00 | 13.39 | 2.0E-5 |
| BM | GO:0005576 | extracellular region | 3559.00 | 63.00 | 39.94 | 3.6E-5 |
| BM | GO:0044459 | plasma membrane part | 2030.00 | 42.00 | 22.78 | 4.1E-5 |
| BM | GO:0044421 | extracellular region part | 3122.00 | 56.00 | 35.03 | 9.0E-5 |
| BM | GO:0032587 | ruffle membrane | 75.00 | 6.00 | 0.84 | 1.9E-4 |
| BM | GO:0043230 | extracellular organelle | 2385.00 | 44.00 | 26.76 | 3.8E-4 |
| BM | GO:0065010 | extracellular membrane-bounded organelle | 2385.00 | 44.00 | 26.76 | 3.8E-4 |
| BR | GO:0045202 | synapse | 537.00 | 38.00 | 6.18 | 6.6E-20 |
| BR | GO:0097458 | neuron part | 896.00 | 45.00 | 10.31 | 1.0E-17 |
| BR | GO:0044456 | synapse part | 396.00 | 31.00 | 4.56 | 1.6E-17 |
| BR | GO:0043005 | neuron projection | 739.00 | 40.00 | 8.50 | 7.9E-17 |
| BR | GO:0030424 | axon | 300.00 | 22.00 | 3.45 | 4.2E-12 |
| BR | GO:0008021 | synaptic vesicle | 116.00 | 14.00 | 1.33 | 5.9E-11 |
| BR | GO:0042995 | cell projection | 1436.00 | 45.00 | 16.52 | 2.1E-10 |
| BR | GO:0005886 | plasma membrane | 3800.00 | 80.00 | 43.72 | 8.7E-10 |
| BR | GO:0071944 | cell periphery | 3877.00 | 80.00 | 44.61 | 2.4E-9 |
| BR | GO:0048786 | presynaptic active zone | 24.00 | 7.00 | 0.28 | 6.8E-9 |
| HT | GO:0030016 | myofibril | 189.00 | 44.00 | 2.11 | < 1e-30 |
| HT | GO:0030017 | sarcomere | 170.00 | 42.00 | 1.90 | < 1e-30 |
| HT | GO:0043292 | contractile fiber | 200.00 | 44.00 | 2.23 | < 1e-30 |
| HT | GO:0044449 | contractile fiber part | 185.00 | 42.00 | 2.06 | < 1e-30 |
| HT | GO:0031674 | I band | 115.00 | 30.00 | 1.28 | < 1e-30 |
| HT | GO:0030018 | Z disc | 103.00 | 28.00 | 1.15 | < 1e-30 |
| HT | GO:0015629 | actin cytoskeleton | 407.00 | 27.00 | 4.54 | 6.2E-14 |
| HT | GO:0016528 | sarcoplasm | 62.00 | 12.00 | 0.69 | 3.2E-12 |
| HT | GO:0031672 | A band | 30.00 | 9.00 | 0.33 | 2.5E-11 |
| HT | GO:0036379 | myofilament | 24.00 | 8.00 | 0.27 | 1.2E-10 |
| KY | GO:0098590 | plasma membrane region | 395.00 | 35.00 | 4.69 | 4.9E-21 |
| KY | GO:0016324 | apical plasma membrane | 229.00 | 25.00 | 2.72 | 2.3E-17 |
| KY | GO:0045177 | apical part of cell | 303.00 | 27.00 | 3.59 | 2.1E-16 |
| KY | GO:0043230 | extracellular organelle | 2385.00 | 72.00 | 28.29 | 1.3E-15 |
| KY | GO:0065010 | extracellular membrane-bounded organelle | 2385.00 | 72.00 | 28.29 | 1.3E-15 |
| KY | GO:0070062 | extracellular vesicular exosome | 2385.00 | 72.00 | 28.29 | 1.3E-15 |
| KY | GO:0005903 | brush border | 65.00 | 14.00 | 0.77 | 2.3E-14 |
| KY | GO:0044421 | extracellular region part | 3122.00 | 78.00 | 37.03 | 1.6E-12 |
| KY | GO:0016323 | basolateral plasma membrane | 175.00 | 18.00 | 2.08 | 2.5E-12 |
| KY | GO:0005576 | extracellular region | 3559.00 | 83.00 | 42.21 | 7.8E-12 |
| LG | GO:0005576 | extracellular region | 3559.00 | 78.00 | 40.19 | 7.5E-11 |
| LG | GO:0044421 | extracellular region part | 3122.00 | 71.00 | 35.26 | 2.1E-10 |
| LG | GO:0031226 | intrinsic component of plasma membrane | 1234.00 | 36.00 | 13.94 | 8.3E-8 |
| LG | GO:0043230 | extracellular organelle | 2385.00 | 54.00 | 26.93 | 1.3E-7 |
| LG | GO:0065010 | extracellular membrane-bounded organelle | 2385.00 | 54.00 | 26.93 | 1.3E-7 |
| LG | GO:0070062 | extracellular vesicular exosome | 2385.00 | 54.00 | 26.93 | 1.3E-7 |
| LG | GO:0031988 | membrane-bounded vesicle | 3025.00 | 63.00 | 34.16 | 1.4E-7 |
| LG | GO:0031982 | vesicle | 3122.00 | 64.00 | 35.26 | 1.9E-7 |
| LG | GO:0005747 | mitochondrial respiratory chain complex ... | 41.00 | 7.00 | 0.46 | 3.3E-7 |
| LG | GO:0030964 | NADH dehydrogenase complex | 41.00 | 7.00 | 0.46 | 3.3E-7 |
| LN | GO:0009897 | external side of plasma membrane | 199.00 | 15.00 | 2.01 | 1.5E-9 |
| LN | GO:0071944 | cell periphery | 3877.00 | 71.00 | 39.10 | 9.8E-9 |
| LN | GO:0098552 | side of membrane | 271.00 | 15.00 | 2.73 | 9.5E-8 |
| LN | GO:0044459 | plasma membrane part | 2030.00 | 45.00 | 20.47 | 1.1E-7 |
| LN | GO:0005886 | plasma membrane | 3800.00 | 67.00 | 38.32 | 1.9E-7 |
| LN | GO:0005576 | extracellular region | 3559.00 | 64.00 | 35.89 | 2.2E-7 |
| LN | GO:0001772 | immunological synapse | 29.00 | 5.00 | 0.29 | 9.5E-6 |
| LN | GO:0044421 | extracellular region part | 3122.00 | 54.00 | 31.49 | 1.2E-5 |
| LN | GO:0044425 | membrane part | 5144.00 | 76.00 | 51.88 | 2.5E-5 |
| LN | GO:0009986 | cell surface | 604.00 | 18.00 | 6.09 | 3.6E-5 |
| LV | GO:0005576 | extracellular region | 3559.00 | 73.00 | 36.15 | 2.8E-11 |
| LV | GO:0044421 | extracellular region part | 3122.00 | 65.00 | 31.71 | 4.8E-10 |
| LV | GO:0043230 | extracellular organelle | 2385.00 | 55.00 | 24.22 | 5.9E-10 |
| LV | GO:0065010 | extracellular membrane-bounded organelle | 2385.00 | 55.00 | 24.22 | 5.9E-10 |
| LV | GO:0070062 | extracellular vesicular exosome | 2385.00 | 55.00 | 24.22 | 5.9E-10 |
| LV | GO:0031988 | membrane-bounded vesicle | 3025.00 | 57.00 | 30.72 | 4.4E-7 |
| LV | GO:0031982 | vesicle | 3122.00 | 57.00 | 31.71 | 1.3E-6 |
| LV | GO:0005615 | extracellular space | 972.00 | 26.00 | 9.87 | 4.6E-6 |
| LV | GO:0016323 | basolateral plasma membrane | 175.00 | 10.00 | 1.78 | 1.2E-5 |
| LV | GO:0072562 | blood microparticle | 92.00 | 7.00 | 0.93 | 4.1E-5 |
| OV | GO:0005929 | cilium | 360.00 | 28.00 | 3.14 | 3.9E-19 |
| OV | GO:0030286 | dynein complex | 39.00 | 11.00 | 0.34 | 2E-14 |
| OV | GO:0005930 | axoneme | 66.00 | 12.00 | 0.58 | 3.8E-13 |
| OV | GO:0097014 | ciliary cytoplasm | 66.00 | 12.00 | 0.58 | 3.8E-13 |
| OV | GO:0032838 | cell projection cytoplasm | 82.00 | 12.00 | 0.72 | 5.7E-12 |
| OV | GO:0042995 | cell projection | 1436.00 | 40.00 | 12.54 | 1.1E-11 |
| OV | GO:0005858 | axonemal dynein complex | 12.00 | 6.00 | 0.10 | 3.4E-10 |
| OV | GO:0044447 | axoneme part | 23.00 | 7.00 | 0.20 | 7.1E-10 |
| OV | GO:0005868 | cytoplasmic dynein complex | 26.00 | 7.00 | 0.23 | 1.9E-9 |
| OV | GO:0044441 | ciliary part | 238.00 | 14.00 | 2.08 | 2E-8 |
| PB | GO:0031012 | extracellular matrix | 367.00 | 26.00 | 4.35 | 1.9E-13 |
| PB | GO:0071944 | cell periphery | 3877.00 | 90.00 | 45.98 | 4.7E-13 |
| PB | GO:0005578 | proteinaceous extracellular matrix | 306.00 | 23.00 | 3.63 | 1.6E-12 |
| PB | GO:0005886 | plasma membrane | 3800.00 | 85.00 | 45.07 | 3.5E-11 |
| PB | GO:0044420 | extracellular matrix part | 119.00 | 14.00 | 1.41 | 1.2E-10 |
| PB | GO:0009897 | external side of plasma membrane | 199.00 | 16.00 | 2.36 | 1.8E-9 |
| PB | GO:0009986 | cell surface | 604.00 | 27.00 | 7.16 | 2.4E-9 |
| PB | GO:0044459 | plasma membrane part | 2030.00 | 54.00 | 24.08 | 3E-9 |
| PB | GO:0098552 | side of membrane | 271.00 | 18.00 | 3.21 | 3.4E-9 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PB | GO:0005576 | extracellular region | 3559.00 | 75.00 | 42.21 | 2.4E-8 |
| SP | GO:0005739 | mitochondrion | 1419.00 | 57.00 | 17.44 | 1.2E-16 |
| SP | GO:0005576 | extracellular region | 3559.00 | 92.00 | 43.73 | 2.9E-15 |
| SP | GO:0044421 | extracellular region part | 3122.00 | 82.00 | 38.36 | 1.6E-13 |
| SP | GO:0044429 | mitochondrial part | 771.00 | 37.00 | 9.47 | 5.4E-13 |
| SP | GO:1990204 | oxidoreductase complex | 78.00 | 13.00 | 0.96 | 9.9E-12 |
| SP | GO:0005759 | mitochondrial matrix | 314.00 | 22.00 | 3.86 | 3.9E-11 |
| SP | GO:0043230 | extracellular organelle | 2385.00 | 62.00 | 29.30 | 1.3E-9 |
| SP | GO:0065010 | extracellular membrane-bounded organelle | 2385.00 | 62.00 | 29.30 | 1.3E-9 |
| SP | GO:0070062 | extracellular vesicular exosome | 2385.00 | 62.00 | 29.30 | 1.3E-9 |
| SP | GO:0005740 | mitochondrial envelope | 586.00 | 27.00 | 7.20 | 2.7E-9 |
| TT | GO:0000795 | synaptonemal complex | 31.00 | 4.00 | 0.29 | 1.9E-4 |
| TT | GO:0071944 | cell periphery | 3877.00 | 54.00 | 36.35 | 5.9E-4 |
| TT | GO:0005886 | plasma membrane | 3800.00 | 51.00 | 35.62 | 2.2E-3 |
| TT | GO:0000794 | condensed nuclear chromosome | 74.00 | 4.00 | 0.69 | 5.1E-3 |
| TT | GO:0043186 | P granule | 13.00 | 2.00 | 0.12 | 6.3E-3 |
| TT | GO:0045495 | pole plasm | 13.00 | 2.00 | 0.12 | 6.3E-3 |
| TT | GO:0060293 | germ plasm | 13.00 | 2.00 | 0.12 | 6.3E-3 |
| TT | GO:0009986 | cell surface | 604.00 | 12.00 | 5.66 | 1.1E-2 |
| TT | GO:0030027 | lamellipodium | 150.00 | 5.00 | 1.41 | 1.3E-2 |
| TT | GO:0032420 | stereocilium | 25.00 | 2.00 | 0.23 | 2.273E-2 |

Table 3: Enriched Cellular Comparments in Tissues

# Chapter 1 Supplementary Table 2

red means low median expressions, so we decided not to test it
green means identitical ORFs discovered

| transcript | tid | first | firstV | medianEi | tissue | cmpt | type | tlen | clen | olen | c2t_ratio | orientati | cds_seq | t_seq | orf_seq | no_blast | vscore | Forward | Reverse | amplicon | fstart | fstop | rstart | rstop |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Transcript 1 | BAT16_c303122_g1_i2_V2 | BAT18_M | 3216 | 474 | PB | c303122_g | complete | 6493 | 1329 | 443 | 0.204682 | 561-1889( | ATGGACG | GCGGGGC | MDEEPAI | 1 | 97.01925 | ATGAGA | GCTGGG | 457 | 707 | 726 | 1163 | 1144 |
| Transcript 2 | BAT14_c319832_g1_i2_V2 | BAT05_F_ | 375 | 182 | LG | c319832_g | complete | 2970 | 1623 | 541 | 0.546465 | 941-2563( | ATGCGGA | AGTCCGC | MRTLRAH | 1 | 99.45657 | CACAGA | GGGGCC | 450 | 919 | 938 | 1368 | 1350 |
| Transcript 3 | BAT14_c306764_g1_i2_V2 | BAT16_M | 246 | 104 | LN | c306764_g | complete | 2187 | 1356 | 452 | 0.620027 | 104-1459( | ATGTCTC | AAAAAAA | MSPPLLQ | 1 | 64.48285 | CAATTTI | AGACAG | 419 | 440 | 459 | 858 | 839 |
| Transcript 4 | BAT04_c186391_g1_i1_V2 | BAT14_M | 685 | 87.5 | KY | c186391_g | complete | 3587 | 1215 | 405 | 0.338723 | 765-1979( | ATGAGAC | GGCCACG | MRPAASP | 1 | 29.63828 | CCCATT | TTGCCG | 548 | 348 | 362 | 890 | 871 |
| Transcript 5 | BAT20_c424841_g1_i1_V2 | BAT14_M | 221.6 | 53.73 | KY | c424841_g | complete | 3121 | 1572 | 524 | 0.503685 | 573-2144( | ATGCGCA | CACGACG | MRTEAGT | 1 | 27.06298 | AACCTC | CCGACA | 469 | 870 | 889 | 1338 | 1319 |
| Transcript 6 | BAT05_c336348_g1_i2_V2 | BAT05_F_ | 215.79 | 52.63 | LG | c336348_g | complete | 4514 | 1398 | 466 | 0.309705 | 1280-267 | ATGTGGA | CCCCGCC | MWSTGA( | 1 | 16.29968 | TCTTTC | AATGCA | 277 | 308 | 327 | 584 | 565 |
| Transcript 7 | BAT02_c232985_g2_i1_V2 | BAT02_F_ | 380.31 | 4.79 | BR | c232985_g | complete | 2080 | 1296 | 432 | 0.623077 | 589-1884( | ATGTTCA | CTGGGCA | MFKLPELR | 1 | 2.984538 | ACCAAT | GCCACA | 507 | 370 | 389 | 876 | 857 |
| Transcript 8 | BAT12_c420427_g1_i1_V2 | BAT12_M | 450.11 | 4.24 | BR | c420427_g | complete | 2397 | 1296 | 432 | 0.540676 | 258-1553( | ATGTTCA | GAGCGGC | MFKLPELR | 1 | 2.292466 | ACGCCT | TCCTCC | 301 | 363 | 382 | 663 | 644 |

Figure 1: **Novel Transcript Information**. Various Information on 8 novel coding transcripts are provided including average expression value, transcript length, CDS length, ORF length, transcript sequence, cds sequence, ORF sequence, primers used, and expected amplicon sizes

## Chapter 1 Supplementary Table 3

| ID | qseqid | sseqid | pident | length | mismatch | gapopen | qstart | qend | sstart | send | evalue | bitscore | qcovs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.00 | BAT16_c303122_g1_i2_V2 | NW_006429611.1 | 78.36 | 1594.00 | 246.00 | 60.00 | 88.00 | 1599.00 | 1139854.00 | 1138278.00 | 0.00 | 941.00 | 23.00 |
| 2.00 | BAT14_c319832_g1_i2_V2 | NW_006441953.1 | 80.45 | 3121.00 | 412.00 | 132.00 | 2.00 | 2970.00 | 32358573.00 | 32361647.00 | 0.00 | 2202.00 | 99.00 |
| 3.00 | BAT14_c306764_g1_i2_V2 | NW_006432945.1 | 76.64 | 1066.00 | 128.00 | 61.00 | 1205.00 | 2165.00 | 17465840.00 | 17466889.00 | 0.00 | 477.00 | 78.00 |
| 3.00 | BAT14_c306764_g1_i2_V2 | NW_006432945.1 | 84.20 | 772.00 | 99.00 | 17.00 | 452.00 | 1215.00 | 17465022.00 | 17465778.00 | 0.00 | 728.00 | 78.00 |
| 4.00 | BAT04_c186391_g1_i1_V2 | NW_006492169.1 | 91.27 | 252.00 | 19.00 | 3.00 | 1.00 | 249.00 | 618.00 | 367.00 | 0.00 | 340.00 | 7.00 |
| 5.00 | BAT20_c424841_g1_i1_V2 | NW_006436283.1 | 82.51 | 1189.00 | 133.00 | 33.00 | 1060.00 | 2202.00 | 724224.00 | 725383.00 | 0.00 | 974.00 | 80.00 |
| 5.00 | BAT20_c424841_g1_i1_V2 | NW_006436283.1 | 83.49 | 757.00 | 99.00 | 19.00 | 309.00 | 1049.00 | 723422.00 | 724168.00 | 0.00 | 682.00 | 80.00 |
| 5.00 | BAT20_c424841_g1_i1_V2 | NW_006436283.1 | 92.36 | 602.00 | 42.00 | 3.00 | 2403.00 | 3001.00 | 725580.00 | 726180.00 | 0.00 | 854.00 | 80.00 |
| 6.00 | BAT05_c336348_g1_i2_V2 | NW_006436283.1 | 82.61 | 1190.00 | 133.00 | 32.00 | 1723.00 | 2868.00 | 724224.00 | 725383.00 | 0.00 | 983.00 | 89.00 |
| 6.00 | BAT05_c336348_g1_i2_V2 | NW_006436283.1 | 84.15 | 757.00 | 94.00 | 20.00 | 972.00 | 1712.00 | 723422.00 | 724168.00 | 0.00 | 710.00 | 89.00 |
| 6.00 | BAT05_c336348_g1_i2_V2 | NW_006436283.1 | 87.21 | 1188.00 | 121.00 | 16.00 | 3069.00 | 4231.00 | 725580.00 | 726761.00 | 0.00 | 1323.00 | 89.00 |
| 6.00 | BAT05_c336348_g1_i2_V2 | NW_006436283.1 | 88.10 | 966.00 | 97.00 | 15.00 | 20.00 | 969.00 | 722417.00 | 723380.00 | 0.00 | 1131.00 | 89.00 |
| 7.00 | BAT02_c232985_g2_i1_V2 | NW_006436291.1 | 83.60 | 744.00 | 81.00 | 21.00 | 1221.00 | 1942.00 | 85385.00 | 84661.00 | 0.00 | 660.00 | 64.00 |
| 7.00 | BAT02_c232985_g2_i1_V2 | NW_006436291.1 | 86.01 | 629.00 | 78.00 | 9.00 | 604.00 | 1223.00 | 86042.00 | 85415.00 | 0.00 | 665.00 | 64.00 |
| 8.00 | BAT12_c420427_g1_i1_V2 | NW_006436291.1 | 78.87 | 530.00 | 57.00 | 24.00 | 1630.00 | 2109.00 | 84602.00 | 84078.00 | 0.00 | 307.00 | 84.00 |
| 8.00 | BAT12_c420427_g1_i1_V2 | NW_006436291.1 | 83.74 | 744.00 | 80.00 | 21.00 | 890.00 | 1611.00 | 85385.00 | 84661.00 | 0.00 | 665.00 | 84.00 |
| 8.00 | BAT12_c420427_g1_i1_V2 | NW_006436291.1 | 86.17 | 629.00 | 77.00 | 9.00 | 273.00 | 892.00 | 86042.00 | 85415.00 | 0.00 | 671.00 | 84.00 |
| 8.00 | BAT12_c420427_g1_i1_V2 | NW_006436291.1 | 88.66 | 194.00 | 21.00 | 1.00 | 2104.00 | 2297.00 | 84044.00 | 83852.00 | 0.00 | 235.00 | 84.00 |

Table 4: BLAST results of validated novel transcripts

## Code Examples

Below is the stan code that was used in MAGE.

```
data {
  // N's
  int<lower=1> N;          // Number of observations
  int<lower=1> N_dpi;      // Number of day post infection type
  int<lower=1> N_tissue;   // Number of tissue types
  int<lower=1> N_beta;     // Number of tissue x dpi interaction terms

  // Predictors
  int obs_to_dpi_index[N];
  int obs_to_tissue_index[N];
  int obs_to_interaction_index[N];

  // Response variable and offset term
  int y[N];            // counts
  vector[N] offset_term;
}

parameters {
  real mu_intercept;

  real beta_tissue_raw[N_tissue];
  real beta_dpi_raw[N_dpi];
  real beta_interaction_raw[N_dpi*N_tissue];

  // Precision(inverse dispersion) parameter
  real<lower=0> phi;

  // hyperprior
  real mu_dpi;
  real<lower=0> sigma_dpi;

  //tissue
  real mu_tissue;
  real<lower=0> sigma_tissue;

  //interaction
  real mu_interaction;
  real<lower=0> sigma_interaction;
}
transformed parameters {
  real beta_dpi[N_dpi];
  real beta_tissue[N_tissue];
  real beta_interaction[N_dpi*N_tissue];
    // Combine all
  for( n in 1:N ) {
      beta_dpi[obs_to_dpi_index[n]] = 5*mu_dpi + sigma_dpi*beta_dpi_raw[
    obs_to_dpi_index[n]];
      beta_tissue[obs_to_tissue_index[n]] = 5*mu_tissue + sigma_tissue*
    beta_tissue_raw[obs_to_tissue_index[n]] ;
```

```
      beta_interaction[obs_to_interaction_index[n]] = 5*mu_interaction +
     sigma_interaction*beta_interaction_raw[obs_to_interaction_index[n
    ]];

  }

}
model {
  vector[N] yhat;
  mu_intercept ~ normal(0, 10);
  //hyper prior for dpi
  mu_dpi ~ normal(0, 1);
  sigma_dpi ~ cauchy(0, 2);

  //hyerprior for tissue
  mu_tissue ~ normal(0, 1);
  sigma_tissue ~ cauchy(0, 2);

  //hyerprior for tissue x dpi interaction
  mu_interaction ~ normal(0, 1);
  sigma_interaction ~ cauchy(0, 2);

  //prior for inverse dispersion
  phi ~ cauchy(0, 2);

  beta_tissue_raw ~ normal(0, 1);
  beta_dpi_raw ~ normal(0, 1);
  beta_interaction_raw ~ normal(0, 1);

  for( n in 1:N ) {
    yhat[n] =  mu_intercept +
                    beta_dpi[obs_to_dpi_index[n]] +
                    beta_tissue[obs_to_tissue_index[n]] +
                    beta_interaction[obs_to_interaction_index[n]] +
                    log(offset_term[n]);
  }
  // log linear negative binomial regression
  y ~ neg_binomial_2_log(yhat, phi);
}
generated quantities {
  vector[N] yhat;
  vector[N] log_lik;

  real dpi1_ingln;
  real dpi3_ingln;
  real dpi5_ingln;
  real dpi7_ingln;
  real dpi9_ingln;
  real dpi1_mesln;
  real dpi3_mesln;
  real dpi5_mesln;
  real dpi7_mesln;
  real dpi9_mesln;
  real dpi1_tbln;
```

```
  real dpi3_tbln;
  real dpi5_tbln;
  real dpi7_tbln;
  real dpi9_tbln;
  real fc_dpi3_ingln;
  real fc_dpi5_ingln;
  real fc_dpi7_ingln;
  real fc_dpi9_ingln;
  real fc_dpi3_mesln;
  real fc_dpi5_mesln;
  real fc_dpi7_mesln;
  real fc_dpi9_mesln;
  real fc_dpi3_tbln ;
  real fc_dpi5_tbln ;
  real fc_dpi7_tbln ;
  real fc_dpi9_tbln ;

  dpi1_tbln  = beta_dpi[1] + beta_tissue[3] + beta_interaction[3];
  dpi3_tbln  = beta_dpi[2] + beta_tissue[3] + beta_interaction[6];
  dpi5_tbln  = beta_dpi[3] + beta_tissue[3] + beta_interaction[9];
  dpi7_tbln  = beta_dpi[4] + beta_tissue[3] + beta_interaction[12];
  dpi9_tbln  = beta_dpi[5] + beta_tissue[3] + beta_interaction[15];

  dpi1_mesln = beta_dpi[1] + beta_tissue[2] + beta_interaction[2];
  dpi3_mesln = beta_dpi[2] + beta_tissue[2] + beta_interaction[5];
  dpi5_mesln = beta_dpi[3] + beta_tissue[2] + beta_interaction[8];
  dpi7_mesln = beta_dpi[4] + beta_tissue[2] + beta_interaction[11];
  dpi9_mesln = beta_dpi[5] + beta_tissue[2] + beta_interaction[14];

  dpi1_ingln = beta_dpi[1] + beta_tissue[1] + beta_interaction[1];
  dpi3_ingln = beta_dpi[2] + beta_tissue[1] + beta_interaction[4];
  dpi5_ingln = beta_dpi[3] + beta_tissue[1] + beta_interaction[7];
  dpi7_ingln = beta_dpi[4] + beta_tissue[1] + beta_interaction[10];
  dpi9_ingln = beta_dpi[5] + beta_tissue[1] + beta_interaction[13];

  // fold change
  fc_dpi3_tbln = (dpi3_tbln - dpi1_tbln)/log(2);
  fc_dpi5_tbln = (dpi5_tbln - dpi1_tbln)/log(2);
  fc_dpi7_tbln = (dpi7_tbln - dpi1_tbln)/log(2);
  fc_dpi9_tbln = (dpi9_tbln - dpi1_tbln)/log(2);

  fc_dpi3_mesln = (dpi3_mesln - dpi1_mesln)/log(2);
  fc_dpi5_mesln = (dpi5_mesln - dpi1_mesln)/log(2);
  fc_dpi7_mesln = (dpi7_mesln - dpi1_mesln)/log(2);
  fc_dpi9_mesln = (dpi9_mesln - dpi1_mesln)/log(2);

  fc_dpi3_ingln = (dpi3_ingln - dpi1_ingln)/log(2);
  fc_dpi5_ingln = (dpi5_ingln - dpi1_ingln)/log(2);
  fc_dpi7_ingln = (dpi7_ingln - dpi1_ingln)/log(2);
  fc_dpi9_ingln = (dpi9_ingln - dpi1_ingln)/log(2);

  for (n in 1:N) {
    yhat[n] =  mu_intercept +
                    beta_dpi[obs_to_dpi_index[n]] +
```

```
                        beta_tissue[obs_to_tissue_index[n]] +
                        beta_interaction[obs_to_interaction_index[n]] +
                        log(offset_term[n]);
    // preferred Stan syntax as of version 2.10.0
    log_lik[n] = neg_binomial_2_log_lpmf(y[n] |  yhat[n] , phi);
  }

}
```

## Code to Generate Synthetic Data

```
library(tidyverse)

rgampois_proposed <- function(n, mu, precision) {
  prob <- precision / (precision + mu)
  #  this definition is directly copied off of the documentation of
    rnbionom :
  # "...An alternative parametrization (often used in ecology) is
  # by the mean mu (see above), and size, the dispersion parameter,
  # where prob = size/(size+mu). The variance is mu + mu^2/size in this
    parametrization...."

  rnbinom(n = n, size = precision,  prob = prob)
}


#' Simulate the RNA_seq samples for a gene in two conditions
#'
#' @param n Number of replicates to make
#' @param exprs Mean Expressions
#' @param fc Fold change
#' @param dispers Dispersion
#'
#' @return
#' @export
#'
#' @examples
simulate <- function(n, exprs, fc, dispers) {
  s1 <-
    rgampois_proposed(n = n,
                      mu = exprs,
                      precision = 1 / dispers)

  # multiply fold change for the second group
  s2 <-
    rgampois_proposed(n = n,
                      mu = exprs * fc,
                      precision = 1 / dispers)

  list(rep = seq(1, n),
       s1 = s1,
```

```
        s2 = s2)
}

#' This function determines what % of the data is DEG
adjust_meta  <- function(meta, PERC = 0.1) {
  split_df <-  base::split(meta, meta$fc)
  name_split_df <- names(split_df)
  to_sample <- setdiff(name_split_df, "1")
  res <- list()
  for (sample_name in to_sample) {
    sampled_df <-
      dplyr::sample_n(split_df[[sample_name]], base::NROW(split_df[["1"
    ]]) * PERC /
                        length(to_sample))
    res[[sample_name]] <-  sampled_df
  }

  dplyr::bind_rows(split_df[["1"]], res)

}

# Simulate the RNA-seq data
# Parameters ----
exprs <-
  seq(10, 1000, by = 5)
dispers <- seq(0.5, 2, by = 0.1) # dispersion
fc <- c(1, 2, 0.5)  # fold change


# Simulate the RNA-seq data
simulate_with_n <- function(N_REP, meta, PREFIX) {
  data_sim_tmp <- meta %>%
    mutate(sample = pmap(list(n = N_REP, exprs, fc, dispers), simulate))

  data_sim <- data_sim_tmp %>%
    mutate(sample = map(sample, as_tibble)) %>%
    unnest() %>%
    gather(sample, data, 5:6) %>%
    mutate(target_id = paste0("r", exprs, "_fc", fc, "_d", dispers)) %>%
    mutate(id = paste0(sample, "_", rep))

  DEPTH <- FALSE
  if (DEPTH) {
    sim_meta <- data_sim %>%
      distinct(id) %>%
      mutate(size_factor = rnorm(
        n = NROW(.),
        mean = 1,
        sd = 0.1
      )) %>%
      mutate(depth = 1e7 * size_factor)

    data_sim <- data_sim %>%
      mutate(data = data / 1e7) %>%
```

```
      left_join(sim_meta, by = "id") %>%
      mutate(data = as.integer(data * depth))
  }

  # Generate wide matrix
  data_sim_w <- data_sim %>%
    select(target_id, id, data) %>%
    spread(id, data)

  cname <- colnames(data_sim_w)[-1]

  # Save wide expression data ----
  write_tsv(data_sim_w,
            paste0(PREFIX, N_REP, ".tsv"))

  invisible(list(data_sim_w, meta, meta_sample))
}

simulate_all <- function(meta, SET, PERC, NREP) {
  #SET <- 1
  #PERC <- 0.1
  meta_adj <- adjust_meta(meta, PERC = PERC)
  set.seed(as.integer(20 * SET))
  data <-
    simulate_with_n(
      NREP,
      meta_adj,
      PREFIX = paste0(
        "tests/benchmark_data/S",
        SET,
        "_P",
        PERC,
        "_N"
      )
    )

}

# Construct multiple scenarios ----
meta <- as_tibble(expand.grid(
  exprs = exprs,
  fc = fc,
  dispers = dispers
))

SETS <- seq(1, 3)
PERCS <- c(0.01, 0.1, 0.5)
NREPS <- c(3, 10, 50)
sim_param_table <- expand.grid(sets = SETS,
            percs = PERCS,
            nreps = NREPS) %>%
  as_tibble
```

## Workflow used for Benchmark

```r
df_to_edg <- function(df) {
// converts data frame to expression matrix with gene names in rownames and
  exp_mat <- df_to_expression_matrix(df)

  # instantiation
  dge <- edgeR::DGEList(counts = exp_mat)

  # normalization
  dge <- edgeR::calcNormFactors(dge, method = "TMM")

  # estimation
  dge <- edgeR::estimateGLMCommonDisp(dge)
  dge <- edgeR::estimateGLMTrendedDisp(dge)
  dge <- edgeR::estimateGLMTagwiseDisp(dge)

  dge
}

dge_edger <- df_to_edg(df)
mod <- model.matrix( ~ sid, data = meta_sample)
dge_fit <- edgeR::glmFit(dge_edger, mod)
dge_lrt <- edgeR::glmLRT(dge_fit, coef = "sids2")
res_edger <- dge_lrt$table
```

```
// Stan model used in benchmark
data {
  // N's
  int<lower=1> N;          // Number of observations
  int<lower=1> N_factor;      // Number of day post infection type

  // Covariate and reponse
  int idx_b[N];

  int y[N];            // counts
  vector[N] depth;
}

parameters {
  real alpha; // intercept
  real beta_raw[N_factor]; // random effect
  real<lower=0> phi; // 1/dispersion

  real mu_beta; // hyperprior mean
  real<lower=0> sigma_beta;    // hyperpror std

}
transformed parameters {
  real beta[N_factor];

  for( n in 1:N ) {
      beta[idx_b[n]] = 5*mu_beta + sigma_beta*beta_raw[idx_b[n]];
```

```
  }

}
model {
  vector[N] yhat;

  // Hyperparms
  mu_beta ~ normal(0, 1);
  sigma_beta ~ cauchy(0, 2);

  alpha ~ normal(0, 10);
  beta_raw ~ normal(0, 1);

  //prior for inverse dispersion
  phi ~ cauchy(0, 2);


  for( n in 1:N ) {
    yhat[n] =  alpha +
                    beta[idx_b[n]] +
                    log(depth[n]);
  }
  // log linear negative binomial regression
  y ~ neg_binomial_2_log(yhat, phi);
}
generated quantities {
  vector[N] yhat;
  vector[N] log_lik;

  for (n in 1:N) {
    yhat[n] =  alpha +
                    beta[idx_b[n]] +
                    log(depth[n]);
    // preferred Stan syntax as of version 2.10.0
    log_lik[n] = neg_binomial_2_log_lpmf(y[n] |  yhat[n] , phi);
  }

}
```

## PCA using TPM

In Figure 2, PCA based on sample-wise TPM as features is performed. This demonstrates that using the TPM or raw counts in PCA may not clearly show the global pattern that was shown with log2FC due to the heterogeneity that exists among tissues at the baseline.

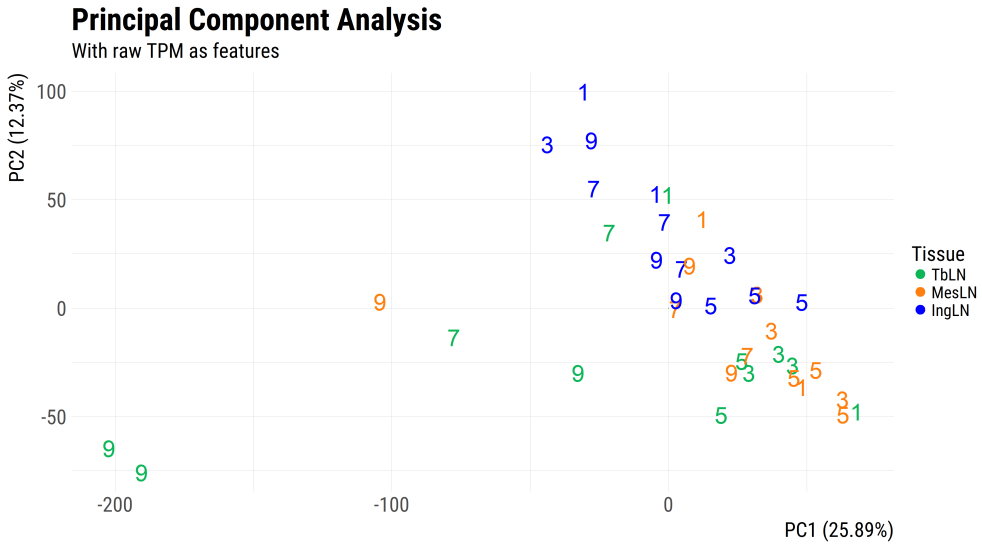


Figure 2: **Principal Component Analysis using raw TPM**.

## Mapper Parameter Selection

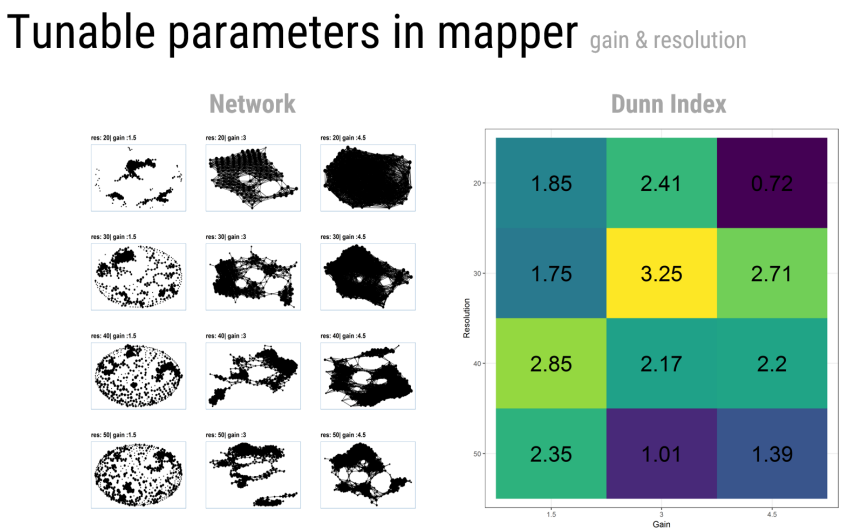


Figure 3: **Tuning Mapper Parameters**. Multiple networks can be generated based on the two parameters, resolution and gain. The first panel shows the grid of networks generated based on the set of specific parameters. The second panel shows the corresponding dunn index for a network.