May 13th, 2023

Final Paper

Understanding Cuticular Expansion in Ticks for Identification of New Molecular Targets

Jeshurun Ekow Biney

8001398281, BIOL 792

I. Overview

Ixodes scapularis is a type of tick that feeds on blood and is known to transmit various harmful pathogens such as bacteria, viruses, and protozoan parasites. The spirochete Borrelia burgdorferi, which causes Lyme disease, is primarily transmitted by this tick, resulting in over 400,000 cases annually in the US. The tick's ability to feed on blood for up to 10 days is a crucial factor in the successful transmission of pathogens. Therefore, identifying ways to manage this vector would help protect against multiple diseases. During the feeding period, female ticks increase in size by approximately 100-fold and undergo cuticle expansion through remodeling to accommodate the blood meal. However, the molecular mechanisms underlying this phenomenon are not well understood. Thus, the goal of this study is to use a multi-omics approach to better understand the blood-feeding process in I. scapularis. By analyzing the transcriptome of the tick epidermis at different stages of feeding, we aim to identify candidate genes and proteins that could be targeted for tick management.

To achieve this, we will employ various tools and packages from anaconda (bioconda), R, and our previous knowledge of Linux, bash scripting, python, version control, and data visualization to conduct transcriptomic analysis.

Figure 1.0 below sums up project.

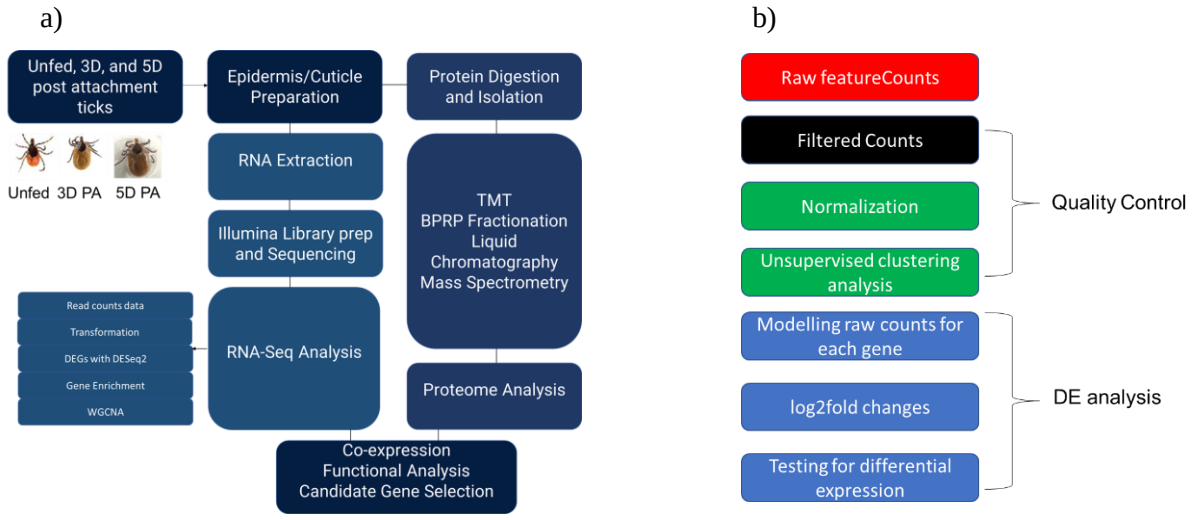


Fig 1.0 a) Process flow diagram showing the path to the transcriptomic b) Cuticle transcriptome analysis outlined in a step by step process

II. METHODS

The architecture plan described here is composed of four paradigms. The first step in the data analysis process involves the sequence quality control to filter the data at hand. This involves using fastQC and MultiQC interfaces. The second process concerns the alignment with the base pairs and a summary of the aligned pairs. This is followed by a differential expression analysis of the ticks at each feeding stage. The final step is the gene enrichment analysis dedicated to exploring significantly overexpressed and under expressed genes. Here is a summary of the flow process.

Sequence quality control

To ensure the quality of the raw fastq files, we will perform quality control using FastQC and aggregate the results with MultiQC. Next, we will filter sequence pairs based on nucleotide base quality and trim them using Trimmomatic. After trimming, we will assess the sequence quality of the reads using FastQC and summarize the results with MultiQC.

Sequence Alignment and Feature Detection/ Read Count

The trimmed read pairs will then be aligned to the Ixodes scapularis genome using the HISAT2 splice-aware read alignment tool. We will compress the aligned reads from SAM format to BAM format using SAMtools and use the featureCounts tool of the subread package to summarize the raw counts of read and read pairs that aligned with genes into a count matrix.

Differential Expression Analysis

To identify genes that are differentially expressed in the slow-stage fed tick and the rapid-stage fed tick epidermis compared to the unfed tick epidermis, we will use the R package DESeq2. We will visualize the differentially expressed genes using volcano plots and heatmaps in R.

Gene Enrichment Analysis

For gene enrichment analysis, we will use the GOEnrichment python package in Bioconda to perform GO enrichment analysis on significantly overexpressed and under-expressed genes. The enriched GO terms will be visualized using dot plots. Methods:

III. APPLICATION OF CLASS TECHNIQUES

Part I - Linux Refresh: The first step in the data analysis process involves refreshing the Linux command line interface. This involves using the terminal to navigate the file system, manipulate files and directories, and execute commands.

Part II - Version Control, Git, and GitHub: Version control is a critical step in managing and tracking changes made to code and data. Git allows multiple people to collaborate on a project simultaneously while keeping track of changes made to code and data. Github is a cloud-based platform that hosts Git repositories and allows for easy collaboration, code sharing, and version control.

Part III – R package: R is a powerful programming language widely used for data analysis. R allows for easy data exploration, analysis, and visualization. We will also use heatmaps in R

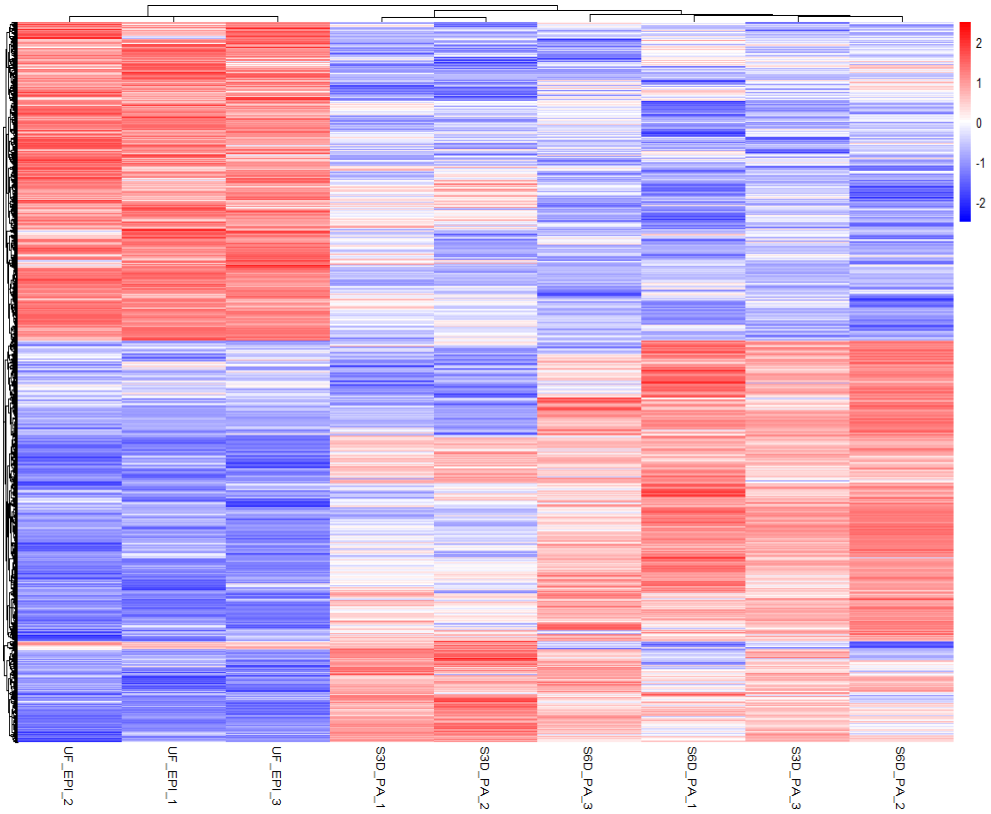
Part IV - Data Visualization: Data visualization is an important step in data analysis, as it allows us to explore patterns, relationships, and trends in data. We will learn how to use Matplotlib and Seaborn, two popular Python libraries used for creating static and interactive visualizations.

Part V - Data Cleaning, Relational Databases: Data cleaning involves preparing and cleaning the data before analysis. We will learn how to use Pandas to clean and preprocess data, and how to store and query data in a relational database using SQLite.

In conclusion, the data analysis pipeline for studying tick-borne pathogens involves multiple steps, including Linux command line basics, version control using Git and Github, data cleaning and preprocessing using Python and Pandas, data visualization using Matplotlib and Seaborn, machine learning using scikit-learn, and working with high-performance clusters for large-scale data analysis.

IV. Results

Heat Map of Differentially Expressed Genes



Differentially expressed genes in I. scapularis cuticle