Transcription Unit Position as a Predictor of Liver Disease State: A Machine Learning Approach

Team 3: Amogh Anand, Anokhi Haria, Yun-Shuo Chou, Xinyue Lu

Department of Biological Sciences, Carnegie Mellon University

Course number: 03-713

Dr. Joel McManus

05/04/2023

Table of Contents

Introduction

Chronic and excessive consumption of alcohol has devastating effects on the liver. Initially, it leads to hepatic steatosis or an accumulation of fat in the hepatocytes, commonly known as "fatty-liver." One of the central mechanisms proposed for hepatic steatosis is insulin resistance, which commonly occurs in obesity, metabolic syndrome, and also, alcoholism. At this stage, however, prompt control of the metabolic and lifestyle parameters can lead to a decrease in the fat accumulation. Continued alcohol consumption further adds insult to the injury and the individual hepatocytes undergo inflammation, apoptosis, necrosis… leading to Alcoholic hepatitis (AH). Despite these inflammatory processes, the overall hepatic function is largely preserved and rarely leads to presentation of overt disease. Sometimes, they may coincidentally show up as elevation in hepatic enzymes (Aspartate and Alanine aminotransferases) and can be correlated with a concurrent increase in the Gamma-glutamyl Transferase levels.

Further exposure causes fibrosis of the hepatic parenchyma leading to Alcoholic Liver Fibrosis, which eventually, irreversibly becomes Alcoholic Cirrhosis. In these stages, patients present clinically with decompensated liver disease and biochemical investigations coupled with ultrasonography are used in the diagnosis. The obvious risks involve fulminant hepatitis, hepatic failure, and the development of hepatocellular carcinoma (HCC) in susceptible, albeit a minority of this population. According to the World Health Organization (WHO)[1], alcohol is responsible for over 3 million deaths worldwide each year, and liver diseases caused by alcohol consumption are a significant contributor to this number.

Early detection, prompt interventions, and lifestyle modifications are crucial in the management of AH to prevent progression into eventual cirrhosis and other complications. However, identifying individuals who are at high risk of developing these diseases can be challenging. Traditional diagnostic methods for liver diseases may not always detect the disease in its early stages, which can delay treatment and lead to worse outcomes. For instance, liver function tests may not be abnormal until significant liver damage has already occurred. Moreover, liver biopsy, the gold standard for diagnosing liver disease, is invasive and carries the risk of complications. In addition, patients with early-stage liver disease may not show any symptoms or signs, which can make it difficult for clinicians to identify the disease. Therefore, there is a need for a more accurate and non-invasive diagnostic tool that can detect liver disease in its early stages.

Machine learning (ML) models can provide a powerful tool for predicting the likelihood of developing alcoholic hepatitis and cirrhosis based on genetic data. These models can analyze large amounts of data and identify patterns and relationships that may not be immediately apparent to clinicians. By incorporating a wide range of genetic markers, ML models can provide a more comprehensive assessment of an individual's risk for developing liver disease, which can help clinicians to identify high-risk patients at an earlier stage.

Alcoholic Hepatitis and Alternative Transcription

Severe alcoholic hepatitis (AH), has a high potential for rapid progression to cirrhosis (1). Studies have suggested that excessive alcohol consumption may promote hepatocyte dedifferentiation by increasing the proportion of proliferating or becoming proliferative hepatocytes in rats (2). The Hippo signaling pathway plays a crucial role in organ growth during fetal development and adult tissue regeneration by regulating the transcriptional regulators YAP and TAZ, which promote cell proliferation and epithelial-mesenchymal transitions (3).

In adult hepatocytes, the RNA-splicing factor ESRP2 controls the activity of key Hippo signaling components, Nf2 and Csnk1d, by retaining exons to generate longer adult mRNAs (4). This conversion from fetal splicing variants to adult splicing variants enables the suppression of YAP/TAZ by Hippo kinases with higher activity. However, chronic alcohol ingestion has been shown to suppress ESRP2 in adult hepatocytes, allowing relatively inactive fetal splicing variants of Hippo kinases to accumulate and reactivate YAP/TAZ (5,1). This adult to fetal reprogramming response to alcohol has also been observed in other models of alcohol-induced liver injury, including the National Institute on Alcohol Abuse and Alcoholism model of chronic plus acute alcohol binging, the Lieber DeCarli HF diet model of voluntary chronic alcohol diet consumption and binging, and the long-term HF/3,5-diethoxycarbonyl-1,4-dihydrocollidine diet plus alcohol feeding model (6-8).

The 3'UTR and 5'UTR regions of genes are important in regulating gene expression and can be used to identify transcriptional units, which are regions of the genome that are transcribed into RNA. In this report, we present a pipeline that utilizes machine learning algorithms to predict the likelihood of developing alcoholic hepatitis and cirrhosis based on genetic data, including data from the 3'UTR and 5'UTR regions. Our pipeline includes pre-processing genetic data, feature selection, model training and testing, and evaluation of model performance. By identifying individuals who are at high risk of developing alcoholic hepatitis and cirrhosis, clinicians can provide targeted interventions and preventive measures to reduce the risk of disease progression and improve patient outcomes.

The pipeline comprises two parts, the first of which involves the automated download and processing of data to generate .bed files. RNA-Seq reads are aligned to a reference genome using HISAT2, and transcriptional units are identified using the mountainClimber tool, which includes the analysis of the 3'UTR and 5'UTR regions. These files serve as input for PART II, where machine learning models are employed to predict the risk of alcoholic hepatitis and cirrhosis.

Early detection and intervention are critical for managing these diseases, and the pipeline offers a potent tool for this purpose.

Accessing Data

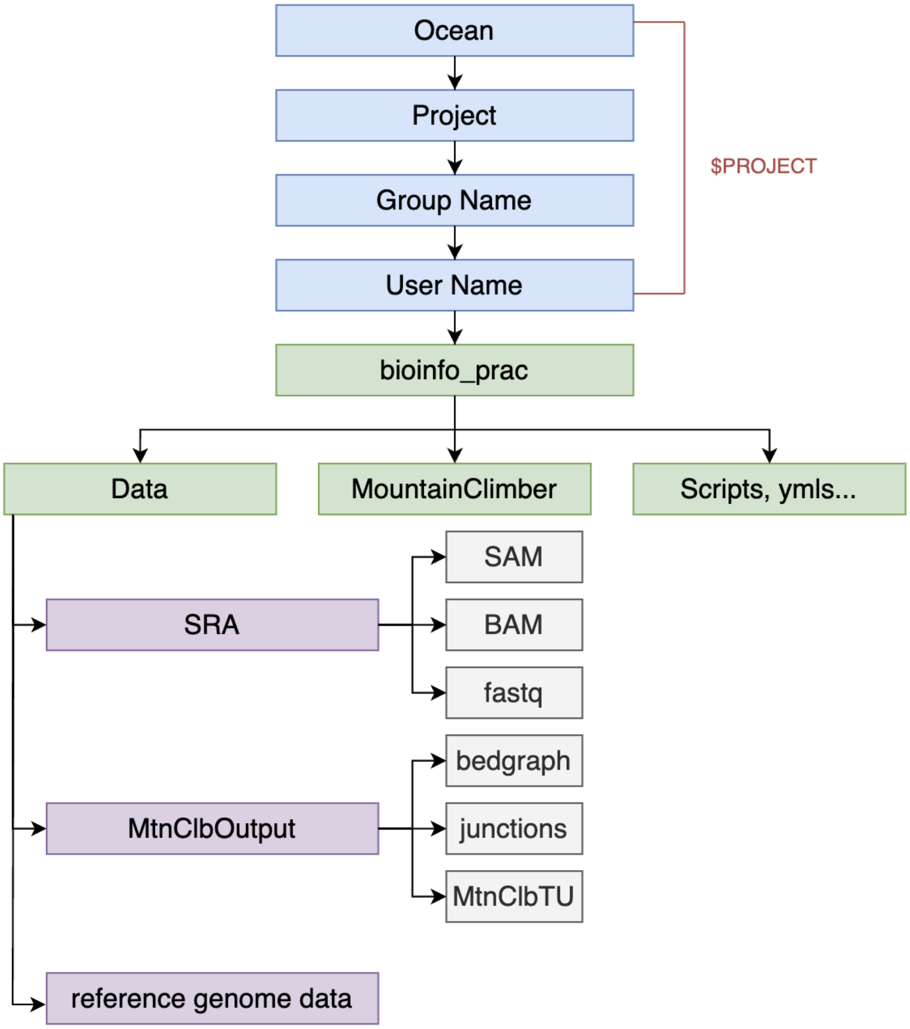


Figure 1: Structure of directories

When the pipeline is run according to the instructions outlined in the User Manual, there is a specific structure of directories created to store the data and output files. The data is stored in a directory named "data", which contains subdirectories for each step of the pipeline. For example, the "SRA" directory contains the SAM, BAM, and converted fastq files and the "mountainClimber" directory contains the output files generated by the mountainClimber tool. The pipeline scripts are stored in a separate directory, named "scripts". This structure of directories helps to keep the data organized and makes it easier to navigate and analyze the output files. When the user runs our pipeline this diagram will help them to navigate the directories more easily and locate specific files. It can also help to ensure that the pipeline is executed correctly and that the outputs are stored in the correct locations.

Raw Input Data:

The dataset, which was utilized in the research article titled "Integrated Multi-Omics Analysis Reveals Glucose Metabolic Reprogramming and Identifies a Novel Hexokinase in Alcoholic Hepatitis" by Massey et al[2]., is accessible on NCBI GEO. The pipeline for this dataset involves processing 27 files, which can be obtained by executing the SRA_download_fastq.sh shell script as described in the accompanying User Manual.

Human Genome:

Reference Genome:

The human genome is used as a reference for alignment of sequencing reads obtained from a sample under study. In this pipeline, the shell script, **SRA_process.sh**, is downloading the human reference genome GRCh38 and building an index using the HISAT2 software, which will be used to align sequencing reads to the reference genome.

hg38 Genome Size:

The human genome sizes are needed to obtain information about the size of each chromosome in the genome, which is necessary for the subsequent analysis of the RNA sequencing data. This information is necessary as an input for mountainClimber in order to generate bed files with information about transcription units.

Packages

Conda:

Conda is an open-source, cross-platform tool designed to manage packages, dependencies, and environments for various programming languages, including Python, R, Lua, Scala, and Java. It is particularly popular among data science teams that primarily use Python. While traditional Python users typically rely on pip for package management and venv for environment management, conda offers a comprehensive solution by efficiently handling both packages and working environments in one tool.

SRA toolkit:

The Sequence Read Archive (SRA) Toolkit is a collection of command-line tools and libraries for working with high-throughput sequencing data in the SRA format. It allows users to download SRA data, convert SRA files to other formats, and perform quality control and filtering of SRA data.

HiSAT2:

HiSAT2 is a fast and accurate alignment tool for mapping RNA sequencing reads to a reference genome. It uses a hierarchical indexing strategy to achieve fast alignment while maintaining high sensitivity and specificity. HiSAT2 can output the alignments in standard SAM/BAM format, which can be further processed by downstream analysis tools.

Samtools:

Samtools is a suite of tools for working with SAM/BAM files, which are standard formats for storing aligned sequencing reads. Samtools can be used to manipulate and filter SAM/BAM files, as well as to perform various operations such as sorting, merging, and indexing.

MountainClimber:

MountainClimberTU is a component of the MountainClimber tool (https://github.com/gxiaolab/mountainClimber) that is designed to identify transcription units (TUs) from RNA-Seq data. It employs a de novo approach to call TUs independently for each sample. The tool generates a set of TUs for each sample, which can then be used for downstream analysis such as identifying alternative transcription start sites (ATS) and alternative polyadenylation sites (APA).

sklearn:

Scikit-learn is a popular Python library for machine learning that provides various algorithms and tools for data preprocessing, feature engineering, model selection, and evaluation. It is built on top of NumPy, SciPy, and matplotlib and is designed to work seamlessly with these libraries.

Overview of Pipeline Implementation

This pipeline was implemented on Bridges2 supercomputing systems. All the scripts and files mentioned can be found in our Github repository.

This pipeline is divided into two parts, PART I includes downloading the data and processing it and PART II includes using the machine learning models to predict disease states.

For PART I of the pipeline, the first step is to download the raw data in a fastq format by running **SRA_download_fastq.sh.** This bash script automates the process of downloading and converting SRA sequencing data to FASTQ format using the **SRA toolkit**. The script creates a "fastq" directory in "data/SRA", loops over a range of SRA accessions, and checks if the corresponding FASTQ files already exist. If not, it runs the fastq-dump command to download and convert the SRA data into two compressed FASTQ files, saving them to the output directory. If the files exist, the script skips that accession and moves on. Then for the processing of the data, **SRA_process.sh** is run. This script downloads and unzips a **reference genome** (GRCh38). It then builds an index for the genome using **HISAT2**, a tool for aligning RNA-Seq reads to a reference genome. After building the index, the script uses the index to align RNA-Seq reads in FASTQ format (stored in "data/SRA") to the reference genome. Alignment is performed using HISAT2 and saves the output in SAM format (stored in "data/SRA". Next, each **SAM** file is **converted** to **BAM** format using **Samtools**, a popular suite of tools for manipulating BAM files. Finally the last step for PART I of this pipeline is to run **MC_process.sh.** The script downloads gene size information and clones the mountainClimber repository, creates output directories, and generates **junction reads** and **bedgraph files** for each BAM file. It then **sorts** and **trims** the generated files and runs the **mountainClimber** tool to identify transcriptional units. The outputs after running mountainClimber are **.bed** files which are then used as input for PART II of the pipeline and are analyzed using visualization tools such as **(Integrative Genomics Viewer) IGV**.

For PART II of the pipeline, we first used **gen_embedding.py** to extract features for further multi-class machine learning prediction. This script takes in BED files that contain transcriptional unit information and generates features for each chromosome. It first reads in the chromosome length from a file, then reads each BED file and extracts the transcriptional unit information for the specified chromosome, binning the data by a given resolution. It then performs principal component analysis (PCA) on the concatenated binned data from all samples to generate reduced-dimensional embeddings, and concatenates the embeddings for all chromosomes into a single feature matrix, which is saved as a numpy file. We then used **predict.py** which utilized the features generated in the previous step as input to perform multi-class classification. The script loads features and labels from the given file paths, initializes a KFold object with 5 splits, and trains a specified classification model like Logistic Regression with L1 penalty (LR(L1 penalty), Logistic Regression (LR), SVM, or Random Forest (RF) using cross-validation. The resulting accuracy score, F1 score, and confusion matrix are printed and plotted using the defined functions to enable further comparison.

Results

PART I

The bed files downloaded from running PART I of the pipeline were visualized using IGV, a powerful tool for visualizing and interpreting genomic data. The IGV analysis for the control sample showed that the TUs were well-aligned with the reference genome, indicating that the pipeline was successful in identifying and processing the relevant genetic data (Figure 3). In addition, the TUs were found to be evenly distributed across the genome, with no significant clustering observed.

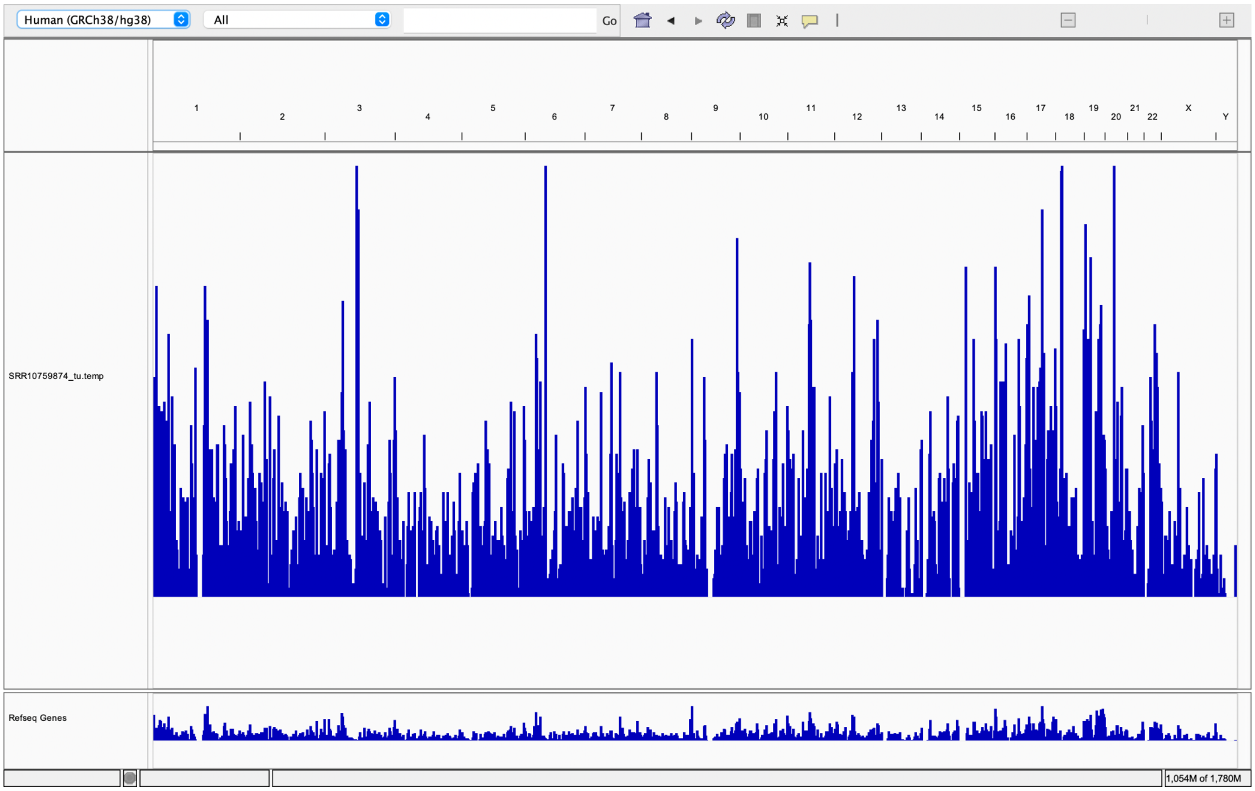


Figure 3: Visualization of Transcription Units obtained from the .bed file for a control sample (SRR10759874)

IGV analysis of a sample with severe alcoholic hepatitis (Figure 4) and cirrhosis (Figure 5) reveal an increased spike coverage in comparison to the reference genome. The spikes are higher and more frequent, indicating regions of the genome that have undergone structural variations or changes in copy number. In addition, these diseases are associated with significant changes in gene expression, and increased spike intensity in certain regions may indicate the upregulation of specific genes or changes in transcriptional activity. Overall, the increased spike intensity in these samples may be indicative of underlying genetic and epigenetic changes that are associated with disease development and progression.

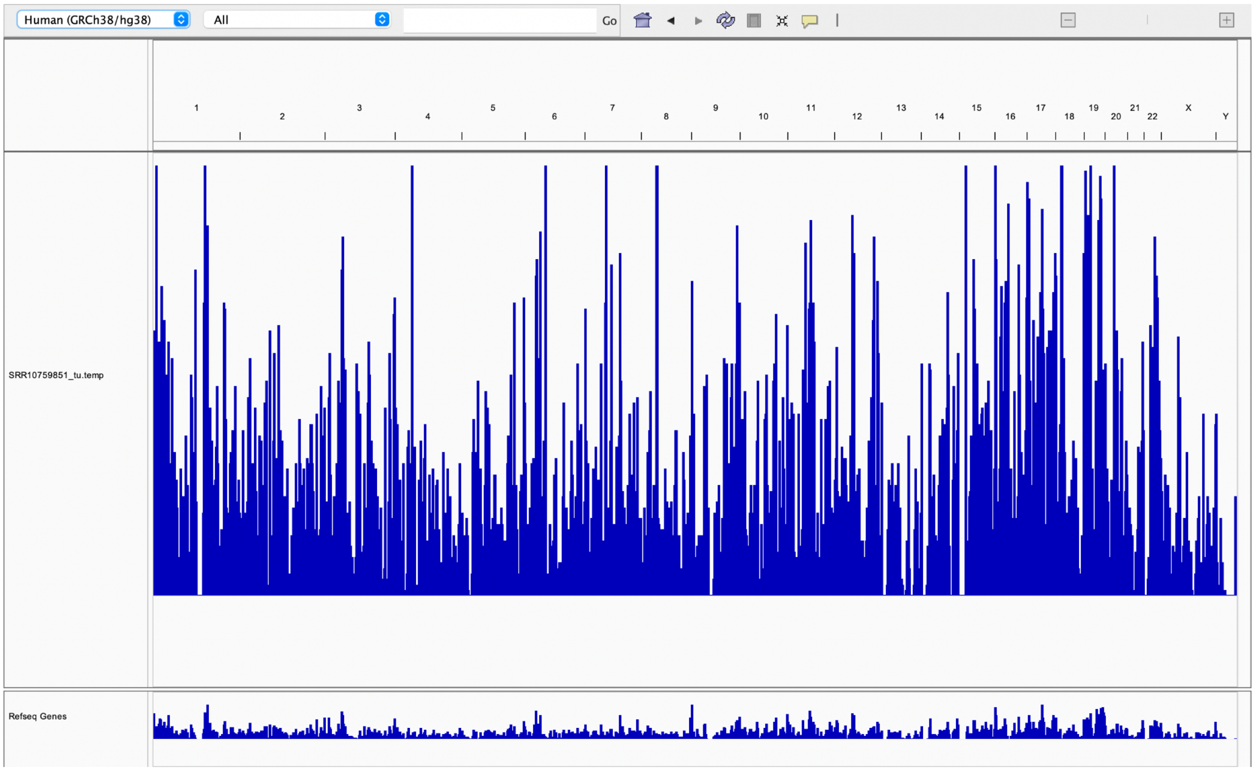


Figure 4: Visualization of Transcription Units obtained from the .bed file for a sample with severe AH (SRR10759851)

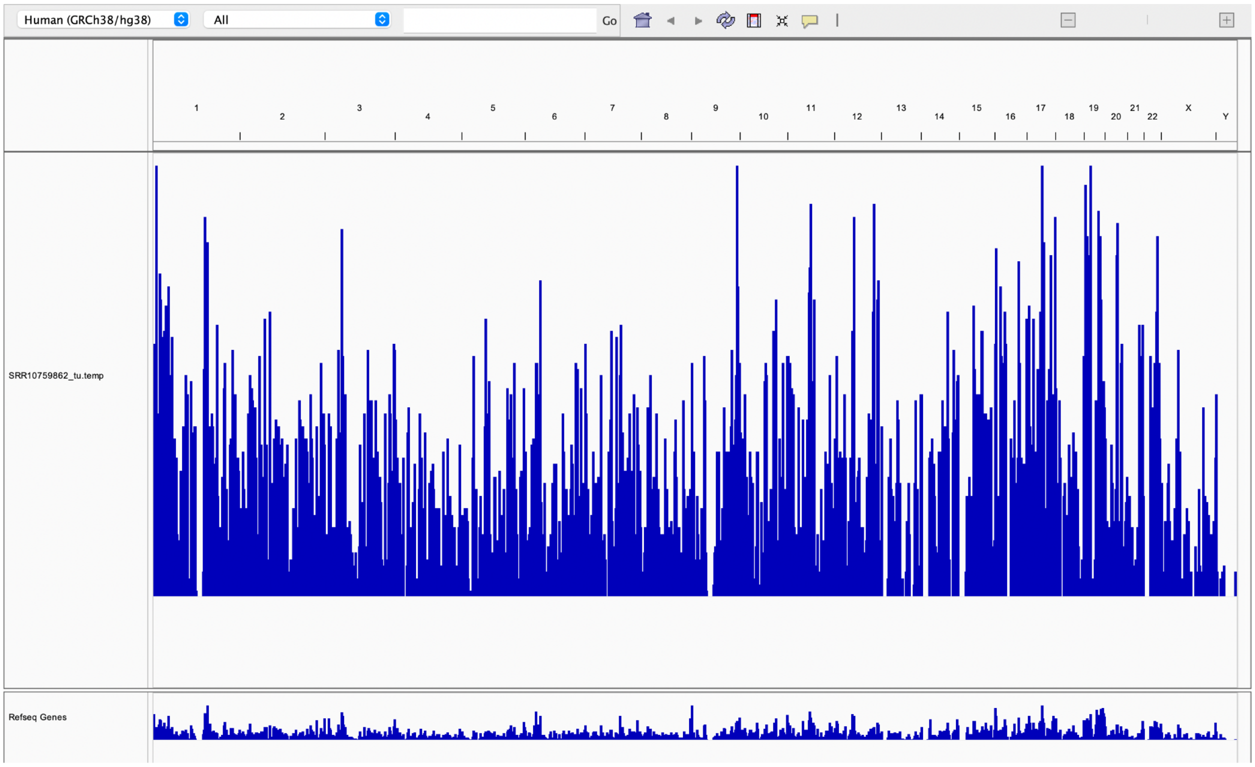


Figure 5: Visualization of Transcription Units obtained from the .bed file for a sample with cirrhosis (SRR10759862)

The results obtained from the analysis of the bed files using IGV revealed significant differences in the peak intensity between the control and the two diseased samples, particularly in the 15-21 chromosome region. The peaks in the AH and cirrhosis samples were more frequent and taller, indicating changes in structural variation or transcription activity. Further examination of chromosome 16, which harbors ESRP2, revealed a noticeable suppression of gene expression in the AH and cirrhosis samples compared to the control (Figure 6). In fact, the cirrhosis sample exhibited the greatest degree of suppression, which could be attributed to the disease's chronic nature. These findings suggest that structural variations or changes in transcriptional activity in the chromosome 16 region could be contributing to the suppression of ESRP2 expression in patients with alcoholic hepatitis and cirrhosis.

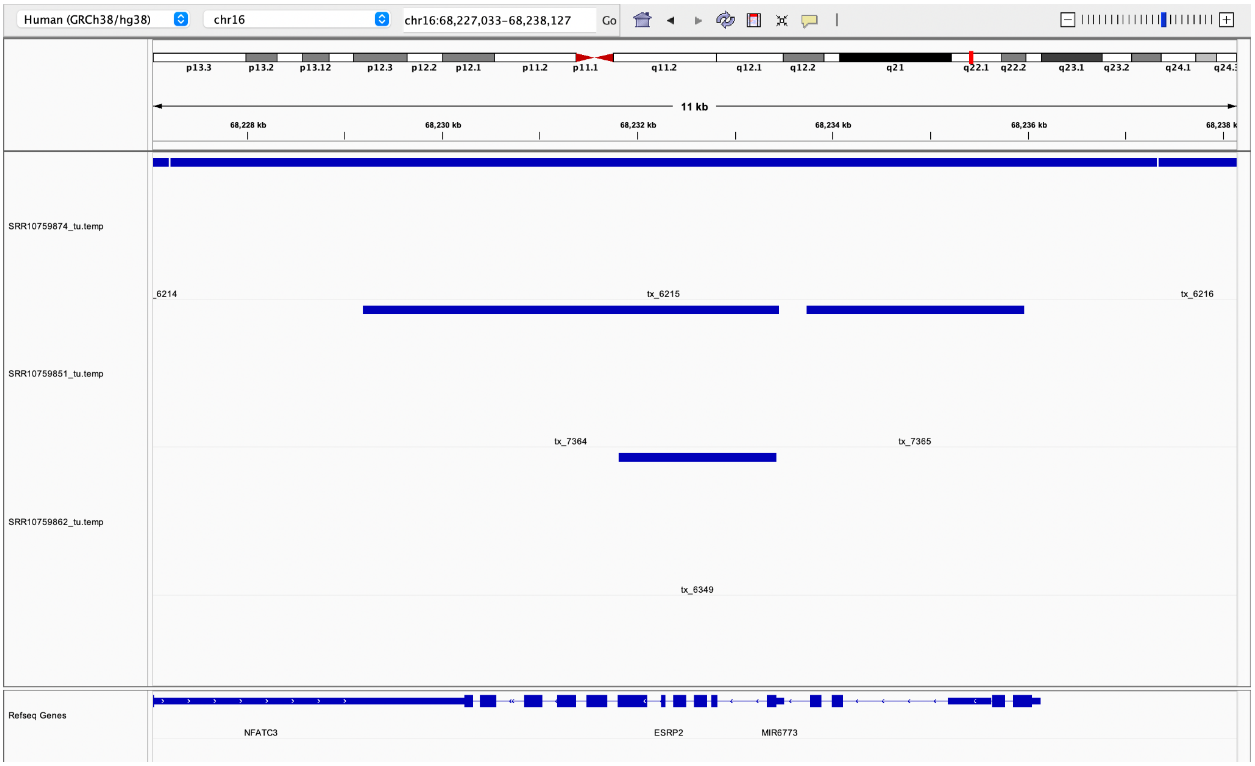


Figure 6: Visualization of gene expression for the ESRP2 gene on chromosome 16 for control, alcoholic hepatitis, and cirrhosis samples (top to bottom).

PART II

For the machine learning classification model, we enabled 4 different multi-class machine learning models in our pipeline. We performed the 5-fold cross validation and calculated the accuracy and f1-score for each model (Figure 7). We can observe that Random Forest has the best performance among all models. It has an accuracy of 0.75 and a f1-score of 0.721.

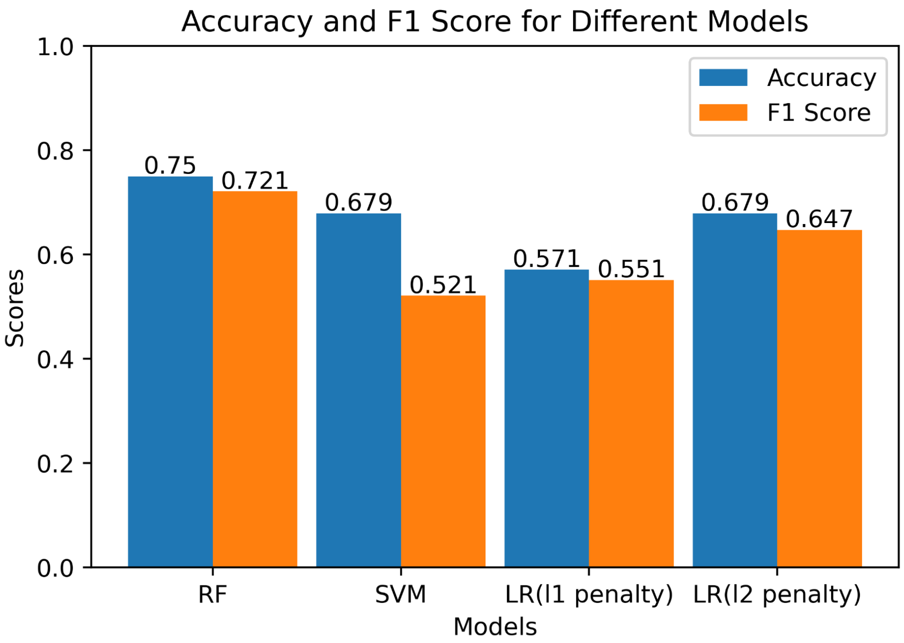
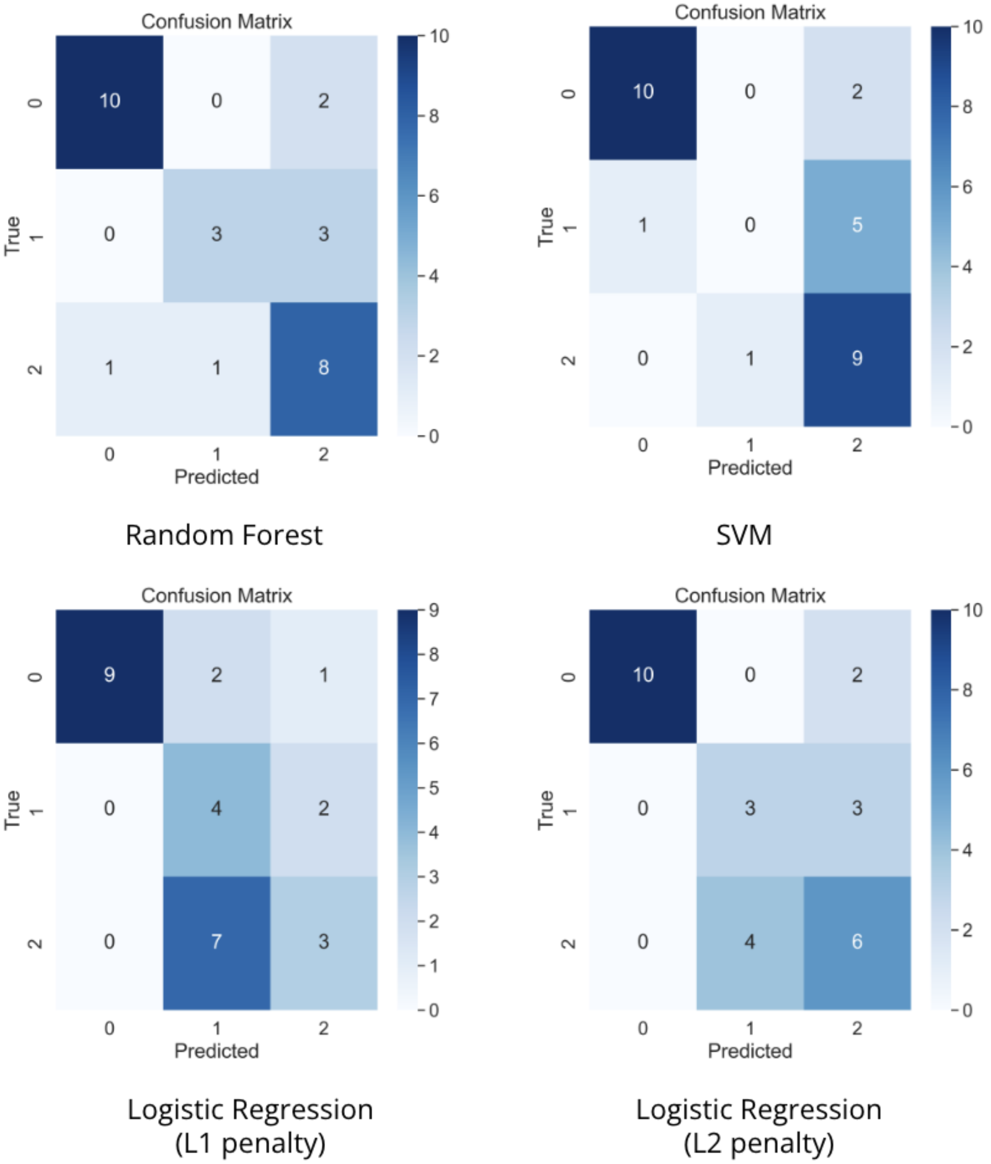


Figure 7: Overall performance comparison for different machine learning models

The confusion matrix was also generated for each classification model, as shown in Figure 8. The samples and labels used were control (12 samples, label 0), cirrhosis without AH (6 samples, label 1), and severe AH (10 samples, label 2), resulting in a 3 x 3 table since there were three different classes. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. The cells along the diagonal represent the instances that have the correct labels for the three different classes, while other cells represent the instances that are classified incorrectly. It was observed that all of the models performed very well in classifying control samples, with only a few of them being misclassified as diseased instances. However, the models performed worse when distinguishing between the two disease states. For instance, logistic regression with L1 penalty misclassified 7 samples with label 2 as label 1. Therefore, one possible direction for improvement is to enhance the feature generation method to achieve better classification performance.



Random Forest

SVM

Logistic Regression
(L1 penalty)

Logistic Regression
(L2 penalty)

Figure 8: Confusion matrices for different machine learning models

Discussion

In our study, we aimed to detect liver alcoholic hepatitis and cirrhosis by utilizing machine learning models that incorporate transcription start and end site information. Our approach consists of two main components: PART I, which focuses on data acquisition and processing, and PART II, which applies machine learning models for classification.

During PART I, we begin by downloading and converting SRA sequencing data into FASTQ format using a dedicated script. Subsequently, the data is aligned to a reference genome and transformed into BAM format with the utilization of Samtools. We then employ MountainClimber to identify transcriptional units, generating .bed files that serve as input for PART II.

In PART II, after obtaining the .bed files containing the transcriptional unit information, we perform dimensionality reduction to extract relevant features. These features are then fed into various machine learning classification models, including logistic regression, logistic regression with L1 penalty, random forest, and support vector machines (SVM), in order to predict the disease state more accurately.

The analysis of the .bed files obtained from the first part of our pipeline using Integrative Genomics Viewer (IGV) played a crucial role in interpreting the genomic data. As depicted in Figure 3, the analysis of the control sample demonstrated that the transcriptional units (TUs) were well-aligned with the reference genome. This finding suggests that our pipeline successfully identified and processed relevant genetic information. Furthermore, the distribution of TUs across the genome was observed to be even, with no significant clustering.

When analyzing samples from patients with severe alcoholic hepatitis and cirrhosis, as shown in Figures 4 and 5 respectively, we observed an increased spike coverage compared to the reference genome. The presence of higher and more frequent spikes may point to regions of the genome that have undergone structural variations or copy number changes, which are often linked to these diseases. Additionally, the increased spike intensity in specific regions could be indicative of alterations in gene expression or transcriptional activity.

IGV analysis of bed files showed significant differences in peak intensity between the control and the two diseased samples in the 15-21 chromosome region, with AH and cirrhosis samples displaying taller and more frequent peaks, suggesting changes in transcriptional activity or structural variation. Further examination of chromosome 16, where ESRP2 is located, showed suppressed gene expression in AH and cirrhosis samples compared to the control, with cirrhosis exhibiting the greatest degree of suppression. These findings suggest that structural variations or transcriptional changes in the 15-21 chromosome region may be contributing to ESRP2 suppression in alcoholic hepatitis and cirrhosis patients.

Despite the promising results of our study, there are several limitations that should be acknowledged, and potential avenues for future research can be explored to further refine our approach. One limitation of our study is the use of predefined hyperparameters for the HISAT2 aligner and MountainClimber tool, which may not be optimal for our specific application. The performance of these tools could be further improved by exploring different parameter configurations through a systematic hyperparameter tuning process. By comparing the results of various setups, we can identify the most suitable hyperparameters to enhance the accuracy and efficiency of the alignment and transcriptional unit identification steps, potentially leading to a more robust and reliable pipeline for the detection of liver alcoholic hepatitis and cirrhosis.

Additionally, the machine learning models employed in this study might benefit from the inclusion of more advanced algorithms, such as deep learning techniques, to improve the accuracy of disease state prediction. Another area for improvement lies in expanding the sample size and including diverse populations to ensure the generalizability of our findings. Furthermore, integrating other sources of omics data, such as transcriptomics, proteomics, and metabolomics, could provide a more comprehensive understanding of the underlying biological processes and improve the predictive power of our models. Lastly, validation of the identified genomic regions and their potential functional implications through experimental approaches would be essential to establish a more robust link between the observed patterns and the pathogenesis of liver alcoholic hepatitis and cirrhosis.

References

1. World Health Organization. (n.d.). Alcohol. World Health Organization. Retrieved May 2, 2023, from https://www.who.int/news-room/fact-sheets/detail/alcohol

2. Massey, V., Parrish, A., Argemi, J., Moreno, M., Mello, A., García-Rocha, M., Altamirano, J., Odena, G., Dubuquoy, L., Louvet, A., Martinez, C., Adrover, A., Affò, S., Morales-Ibanez, O., Sancho-Bru, P., Millán, C., Alvarado-Tapias, E., Morales-Arraez, D., Caballería, J., Mann, J., … Bataller, R. (2021). Integrated Multiomics Reveals Glucose Use Reprogramming and Identifies a Novel Hexokinase in Alcoholic Hepatitis. Gastroenterology, 160(5), 1725–1740.e2. https://doi.org/10.1053/j.gastro.2020.12.008