Introduction

Despite remarkable progress in patient management, liver cancer remains the fourth leading cause of cancer mortality worldwide [1]. Within the United States, an estimated 42,830 people were diagnosed with liver cancer in 2019 and an estimated 30,160 died from the disease, according to the American Cancer Society [2]. Sadly, it is one of the few types of cancer with increases in both incidence and mortality rates of about 3% per year in the United States [2]. The prevalent cancer is hepatocellular carcinoma (HCC) accounting for 70%–90% of all newly diagnosed liver cancers. Well supported risk factors include hepatitis B virus/hepatitis C virus (HBV/HCV) infection, nonalcoholic steatohepatitis, alcoholism, and smoking [2]. The 5-year survival rate of HCC varies widely across different populations, with an average rate of less than 32% [1]. It is a highly heterogeneous disease entity with a complex etiology, which makes prediction of disease prognosis and clinical outcomes very challenging [2]. This is further complicated by very limited effective therapeutic strategies. Thus a critical unmet medical need pivots around discovery of actionable diagnostic and prognostic markers predictive of clinical outcome are needed to guide therapeutic decisions.

To address this critical unmet medical need, we propose the application of a computational framework using Machine Learning to gene expression and somatic mutation data for molecular classification of liver cancer and predicting clinical outcomes. Our working hypothesis is that genomic alterations in the tumor transcriptome and genome would lead to measurable changes that affect therapeutic decision making and that application of ML would enable development of more accurate prediction models at the point of care. To test this hypothesis, we used publicly available data gene expression and mutation data on 369 patients diagnosed with HCC and 136 control samples from The Cancer Genome Atlas (TCGA). The developed methods were validated on an independent cohort.

Methods

Liver Hepatocellular Carcinoma (LIHC) and normal tissue RNA transcript data was downloaded through the publicly available The Cancer Genome Atlas (TCGA) database and processed using Python. The data was filtered so that any genes without expression values were removed and balanced so that there were equal tumor and normal sample counts (136 each). Normalization and feature selection were performed using the EdgeR package in R to identify differentially expressed genes between the tumor and control groups. The data was re-filtered to include only the differentially expressed genes identified from the previous step and was subsequently split 75:25 into training and testing sets. The data sets were trained and tested using a machine learning classifier. This step was repeated using different subsets of genes based off increasing p-value thresholds as determined by the differential expression analysis. The classifier was designed using the Weka Python package, and the outline of the final model can be visualized below (Figure 1). Different models were created, tuned, and tested against the data to determine the best model in terms of sensitivity and specificity of the classification task. A subset of significant genes were derived from the best performing model and used for pathway analysis. The final subset of genes were used to repeat the methodology on tumor data separated into two groups by outcome (alive or dead).

Results: Comparison by clinical outcome

Figure 2.b demonstrates the top 25 differentially expressed genes between normal and tumor groups. These genes are hypothesized to hold predictive value for the classification task. Figure 4.a demonstrates the top 25 differentially expressed genes between alive and dead groups. The genes highlighted in yellow are those which were also found in the list when using the genes yield the highest predictive value for the classification task. The genes used for this task were based off the genes yield the highest predictive value for the classification task.

Figure 4.b shows the classification accuracy when comparing clinical outcomes of disease (alive vs dead) using the different models. To test if the models were good predictors for tumorigenicity through machine learning methods. Pathway analysis of the final list reveals genes involved in historically defined anti-apoptotic and pro-proliferative pathways, as well as those involved in cytokine signaling, cellular senescence and differentiation, hepatic disease, and viral infection. Genes are known miRNAs involved with cancer, and 7 are protooncogenes in cancer. The usefulness of the same method is illustrated again by comparing clinical outcome of disease within the tumor groups the list of genes produced in this step are hypothesized to explicate the genes most predictive of poor prognosis of disease. Many of the same genes are significant predictors of cancer compared to normal tissue. The overall method demonstrates genes which are apt candidates for biomarker analysis. It is important to reiterate that cancer is an extremely heterogeneous disease, even within the same cancer type; the method doesn’t currently account for subclasses of Hepatocellular Carcinoma, but HCC tumorigenesis as a whole. Future improvements of the model should aim to cluster the single tumor group into appropriate subgroups. One way of achieving this goal is to include more robust predictive data such as methylation, histone modification, mutation, and copy number variant information to increase sensitivity between subgroups.

Conclusions

The list of differentially expressed genes were derived from raw transcriptome data and were demonstrated to be good predictors for tumorigenicity through machine learning methods. Pathway analysis of the final list reveals genes involved in historically defined anti-apoptotic and pro-proliferative pathways, as well as those involved in cytokine signaling, cellular senescence and differentiation, hepatic disease, and viral infection. Genes are known miRNAs involved with cancer, and 7 are protooncogenes in cancer. The usefulness of the same method is illustrated again by comparing clinical outcome of disease within the tumor groups the list of genes produced in this step are hypothesized to explicate the genes most predictive of poor prognosis of disease. Many of the same genes are significant predictors of cancer compared to normal tissue. The overall method demonstrates genes which are apt candidates for biomarker analysis. It is important to reiterate that cancer is an extremely heterogeneous disease, even within the same cancer type; the method doesn’t currently account for subclasses of Hepatocellular Carcinoma, but HCC tumorigenesis as a whole. Future improvements of the model should aim to cluster the single tumor group into appropriate subgroups. One way of achieving this goal is to include more robust predictive data such as methylation, histone modification, mutation, and copy number variant information to increase sensitivity between subgroups.

Regardless, this work serves as a formidable outline for any prospective pipeline.