PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Identification of integrated proteomics and transcriptomics signature of alcohol-associated liver disease using machine learning

Stanislav Listopad[1¤]*, Christophe Magnan[1], Le Z. Day[2], Aliya Asghar[3], Andrew Stolz[4], John A. Tayek[5], Zhang-Xu Liu[4], Jon M. Jacobs[2], Timothy R. Morgan[3], Trina M. Norden-Krichmar[1,6]*

1 Department of Computer Science, University of California, Irvine, California, United States of America, **2** Biological Sciences Division and Environmental and Molecular Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, United States of America, **3** Medical and Research Services, VA Long Beach Healthcare System, Long Beach, California, United States of America, **4** Division of Gastrointestinal & Liver Diseases, Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America, **5** Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Department of Internal Medicine, David Geffen School of Medicine, University of California Los Angeles, Torrance, California, United States of America, **6** Department of Epidemiology and Biostatistics, University of California, Irvine, California, United States of America

¤ Current address: Department of Neuroscience, Scripps Research, La Jolla, California, United States of America
* slistopa@uci.edu (SL); tnordenk@uci.edu (TMN-K)

Citation: Listopad S, Magnan C, Day LZ, Asghar A, Stolz A, Tayek JA, et al. (2024) Identification of integrated proteomics and transcriptomics signature of alcohol-associated liver disease using machine learning. PLOS Digit Health 3(2): e0000447. https://doi.org/10.1371/journal.pdig.0000447

Editor: Nicole Yee-Key Li-Jessen, McGill University, CANADA

Received: September 8, 2023

Accepted: January 9, 2024

Published: February 9, 2024

Data Availability Statement: The human RNA raw sequencing data in this study requires deposit into the Database of Genotypes and Phenotypes (dbGAP) of the National Center for Biotechnology Information (United States National Library of Medicine) with controlled access. The data will be available through dbGaP (https://www.ncbi.nlm.nih.gov/gap/) under accession number: phs003112.v1.p1. The public RNA data used for validation in this study is available in the GEO database under accession number GSE142530

Abstract

Distinguishing between alcohol-associated hepatitis (AH) and alcohol-associated cirrhosis (AC) remains a diagnostic challenge. In this study, we used machine learning with transcriptomics and proteomics data from liver tissue and peripheral mononuclear blood cells (PBMCs) to classify patients with alcohol-associated liver disease. The conditions in the study were AH, AC, and healthy controls. We processed 98 PBMC RNAseq samples, 55 PBMC proteomic samples, 48 liver RNAseq samples, and 53 liver proteomic samples. First, we built separate classification and feature selection pipelines for transcriptomics and proteomics data. The liver tissue models were validated in independent liver tissue datasets. Next, we built integrated gene and protein expression models that allowed us to identify combined gene-protein biomarker panels. For liver tissue, we attained 90% nested-cross validation accuracy in our dataset and 82% accuracy in the independent validation dataset using transcriptomic data. We attained 100% nested-cross validation accuracy in our dataset and 61% accuracy in the independent validation dataset using proteomic data. For PBMCs, we attained 83% and 89% accuracy with transcriptomic and proteomic data, respectively. The integration of the two data types resulted in improved classification accuracy for PBMCs, but not liver tissue. We also identified the following gene-protein matches within the gene-protein biomarker panels: *CLEC4M-CLC4M*, *GSTA1-GSTA2* for liver tissue and *SELENBP1-SBP1* for PBMCs. In this study, machine learning models had high classification accuracy for both transcriptomics and proteomics data, across liver tissue and PBMCs. The integration of transcriptomics and proteomics into a multi-omics model yielded

(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE142530). Proteomic data can be found in the MassIVE repository under accession number MSV000089168.

Funding: Funding for this study was provided to the researchers in the Southern California Alcoholic Hepatitis Consortium (SCAHC) by the National Institute on Alcohol Abuse and Alcoholism (NIAAA, https://www.niaaa.nih.gov/) award numbers: U01AA021838 (TMNK), U01AA021886 (TRM), U01AA021884 (TRM), U01AA021918 (JMJ), and U01AA021857 (ZXL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

improvement in classification accuracy for the PBMC data. The set of integrated gene-protein biomarkers for PBMCs show promise toward developing a liquid biopsy for alcohol-associated liver disease.

Author summary

Alcohol-associated cirrhosis and alcohol-associated hepatitis can be difficult to classify clinically. Previously, we established that these two diseases can be differentiated using RNA sequencing gene expression data collected from either liver tissue biopsies or from peripheral blood mononuclear cells (PBMCs), which are extracted from blood samples. In the current study, we investigated whether using protein expression data, in addition to gene expression data, would improve our machine learning models' ability to distinguish between the two alcohol-associated liver diseases and enable identification of gene and protein biomarkers. We found that our models accurately classified alcohol-associated liver diseases with each data type. We were also able to identify promising tissue and blood-based diagnostic gene and protein biomarkers. Additionally, we have demonstrated that challenges present in analyzing small sample size, high dimensional genomic data can be addressed through careful application of appropriate software, bioinformatics, and machine learning methods. By applying these computational approaches to this liver disease genomics data set, we have identified blood-based diagnostic biomarkers of liver disease that will potentially contribute to the development of highly accurate blood tests that will replace invasive liver biopsies.

Introduction

In this study, we focused on alcohol-associated hepatitis (AH) and alcohol-associated cirrhosis (AC) because these are deadly liver conditions with similar clinical presentation. In 2019 there were 23,780 deaths from alcohol-associated cirrhosis (AC) in United States [1]. This is more than triple the number of deaths from alcohol-associated cirrhosis in 1999. The patients with alcohol-associated liver disease (ALD) account for 18% of liver transplants [2]. However, attaining a liver transplant as an ALD patient is difficult, since donor livers are scarce and there are concerns about allocation to individuals with alcohol addiction [2]. Typically, a 6-month abstinence from alcohol is required to be a candidate for liver transplant [2]. Many of ALD patients have alcohol-associated hepatitis (AH) a condition which carries mortality as high as 50% at 3 months [3]. For the severe AH patients, the 6-month abstinence requirement can be tantamount to a death sentence [2]. When carefully selected, ALD patients can benefit from liver transplantation [4,5,6,7]. Currently, establishing AH diagnosis can require liver biopsy, typically done using a transjugular route [3]. Liver biopsy has several limitations, such as procedural risk of internal bleeding, high cost, and patient dissatisfaction. Thus, development of a non-invasive test that can reliably distinguish between AH and AC would be beneficial. Currently, there are a large number of imaging and blood tests for diagnosis of liver cirrhosis [8]. However, liver biopsy remains the current standard for diagnosis [9]. Further improvement in accuracy of non-invasive tests is necessary to reduce the need for liver biopsy [10].

In a previous study, we established that gene expression biomarkers from liver tissue and peripheral mononuclear blood cells (PBMCs) can be used with a multiclass machine learning approach to successfully distinguish between multiple liver diseases [11]. In the present study,

in addition to transcriptomic data, we also obtained proteomic data for participants from the same cohort [12]. Addition of proteomic data presented new opportunities, but it also further increased the ratio of feature size to sample size. This made overfitting a greater challenge than when we only used the gene expression data. First, we compared how well gene and protein biomarkers could be used to classify these conditions separately. Then we examined whether further improvement in classification accuracy could be obtained by combining transcriptomic and proteomic data. As part of the classification process, we have identified the most effective gene and protein biomarkers of alcohol-associated liver disease. We also examined the degree of concordance between top differentially expressed proteins and genes for the three conditions. The gene and protein biomarkers identified in this study, with further validation, could be used to develop new highly accurate blood tests to distinguish between various types of ALD.

Materials and methods

Study population

This study was primarily conducted using biospecimens collected from participants enrolled by the Southern California Alcoholic Hepatitis Consortium (SCAHC). The protocol was approved by the IRB, and informed written consent was obtained from all participants. The liver tissue from participants with AC and healthy controls were obtained from the liver tissue cell distribution system (LTCDS) at University of Minnesota. The study population demographics for liver tissue and PBMC samples for transcriptomic and proteomic analyses can be found in Tables 1 and 2.

The biospecimens consisted of 98 PBMC RNAseq samples, 55 PBMC proteomic samples, 48 liver tissue RNAseq samples, and 53 liver tissue proteomic samples. The liver diseases

Table 1. Study population demographics (liver) for proteomic and RNAseq analysis.

	Liver tissue samples (proteomics)			Liver tissue samples (transcriptomics)		
	AH	**CT**	**AC**	**AH**	**CT**	**AC**
	n = 33	**n = 10**	**n = 10**	**n = 32**	**n = 8**	**n = 8**
Age: mean ± std	42.7 ± 11.4	56 ± 8.6	51.9 ± 13.1	43.3 ± 11.3	55.4 ± 4.3	54.2 ± 6.9*
MELD: mean ± std	25.2 ± 5.7	NA	32 ± 6.1*	25.1 ± 5.7	NA	NA
Maddrey's DF: mean	53.3 ± 22.2	NA	NA	52.3 ± 22.1	NA	NA
BMI: mean ± std	29 ± 5.3	NA	25.6 ± 8.4*	29.4 ± 5.9	NA	NA
Gender: N (percent)						
Female	3(9.1%)	0(0.0%)	0(0.0%)	3(9.4%)	0(0.0%)	0(0.0%)
Male	30(90.9%)	10(100%)	9(90%)	29(90.6%)	7(87.5%)	5(62.5%)
Ethnicity: N (percent)						
Hispanic	25(75.8%)	NA	0(0.0%)	25 (78.1%)	NA	0 (0.0%)
NHW	5(15.1%)	NA	5(50%)	5 (15.6%)	NA	4 (50.0%)
Black	2(6.1%)	NA	0(0.0%)	1 (3.1%)	NA	0 (0.0%)
Other	1(3.0%)	NA	0(0.0%)	1 (3.1%)	NA	0 (0.0%)
Source	SCAHC	LTCDS	LTCDS	SCAHC	LTCDS	LTCDS

Abbreviations: AC, alcohol-associated cirrhosis; AH, alcohol-associated hepatitis; CT, healthy controls; MELD, model for end-stage liver disease; NHW, non-Hispanic White; NA, not available; SCAHC, Southern California Alcoholic Hepatitis Consortium.

*Missing MELD scores for 7 proteomic AC samples, BMI for 8 proteomic AC samples, and age for 3 transcriptomic AC samples.

https://doi.org/10.1371/journal.pdig.0000447.t001

Table 2. Study population demographics (PBMCs) for proteomic and RNAseq analysis.

	PBMC samples (proteomics)			PBMC samples (transcriptomics)		
	AH	**CT**	**AC**	**AH**	**CT**	**AC**
	n = 20	**n = 22**	**n = 13**	**n = 38**	**n = 20**	**n = 40**
Age: mean ± std	48.7 ± 11.6	34.8 ± 15.1	54.2 ± 11.2	47.3 ± 11.5	35.9 ± 15.6	54.5 ± 9.7
MELD: mean ± std	24.5 ± 3.6	7.5 ± 2.5	13.6 ± 6.7	25 ± 3.8	7.3 ± 2.6	13.4 ± 5.8
Maddrey's DF: mean	49.3 ± 17.3	2.5 ± 7.8	22.1 ± 23.3	52.6 ± 20.7	2.4 ± 8.1	21.1 ± 19.1
BMI: mean ± std	29.6 ± 5.5	27.1 ± 4	30 ± 4.8	30 ± 6.2	27 ± 3.5	30.4 ± 5.1
Gender: N (percent)						
Female	1(5%)	10(45.4%)	0(0.0%)	1 (2.6%)	8 (40.0%)	0 (0.0%)
Male	19(95%)	12(54.6%)	13(100%)	37 (97.4%)	12 (60.0%)	40(100.0%)
Ethnicity: N (percent)						
Hispanic	12(60%)	12(54.5%)	10(76.9%)	25 (65.8%)	8 (40.0%)	25 (62.5%)
NHW	5(25%)	0(0.0%)	2(15.4%)	10 (26.3%)	0 (0.0%)	13 (32.5%)
Black	2(10%)	1(4.5%)	0(0.0%)	2 (5.3%)	2 (10.0%)	1 (2.5%)
Other	1(5%)	12(54.5%)	1(7.7%)	1 (2.6%)	10 (50.0%)	1 (2.5%)
Source	SCAHC	SCAHC	SCAHC	SCAHC	SCAHC	SCAHC

*The ethnicity and sex percentages may not add up to 100% due to missing data.

Abbreviations: AC, alcohol-associated cirrhosis; AH, alcohol-associated hepatitis; CT, healthy controls; LTCDS, Liver Tissue Cell Distribution System; MELD, model for end-stage liver disease; NHW, non-Hispanic White; NA, not available; SCAHC, Southern California Alcoholic Hepatitis Consortium.

https://doi.org/10.1371/journal.pdig.0000447.t002

represented were encoded with two letter symbols as follows: alcohol-associated hepatitis (AH) and alcohol-associated cirrhosis (AC). Most of the AC participants within the SCAHC study were expected to be in-patients with decompensated cirrhosis. Best efforts were made during recruitment of the AH and AC groups within SCAHC study to match based on age, gender, and ethnicity. Severity-based matching was not possible due to small sample size. One of the main reasons for small sample size in our study and in publicly available data sets, is difficulty in recruiting patients with AH. AH has a low incidence rate of an estimated 4.5 hospitalizations per 100,000 person per year [13]. Additional information about the inclusion and exclusion criteria, sample collection, sample processing, and preliminary data processing can be found in S1 Text.

Partitioning samples into datasets

Because some proteomic and transcriptomic samples came from the same participants, while others did not, we implemented a strategy to partition and balance the samples in the datasets into matched and unmatched sets. Table 3 summarizes the degree of matching between proteomic and transcriptomic samples in liver tissue and PBMC. For several algorithms in the pipeline, some of the unmatched subsets were too small. Therefore, we moved some matched samples into unmatched sample categories, and we will refer to these new categories as "balanced matched" and "balanced unmatched" subsets. We divided our data into the following dataset categories described below.

Full datasets. These datasets are composed of all available samples for the given tissue and genomic datatype: PBMC 3-Way Full proteomics, PBMC 3-Way Full RNAseq, Liver 3-Way Full proteomics, and Liver 3-Way Full RNAseq.

Table 3. The degree of matching between proteomic and transcriptomic samples for PBMC and liver tissue. The numbers in parenthesis denote the number of samples that were moved from matched category into matched balanced and unmatched balanced categories.

	PBMC (proteomics)			PBMC (transcriptomics)		
	AH	CT	AC	AH	CT	AC
Full	20	22	13	38	20	40
Matched	18	19	13	18	19	13
Unmatched	2	3	0	20	1	27
Matched Balanced	9(-9)	12(-7)	6(-7)	9(-9)	12(-7)	6(-7)
Unmatched Balanced	11(+9)	10(+7)	7(+7)	29(+9)	8(+7)	34(+7)
	Liver (proteomics)			Liver (transcriptomics)		
	AH	CT	AC	AH	CT	AC
Full	33	10	10	32	8	8
Matched	29	3	5	29	3	5
Unmatched	4	7	5	3	5	3
Matched Balanced	24(-5)	3	3(-2)	24(-5)	3	3(-2)
Unmatched Balanced	9(+5)	7	7(+2)	8(+5)	5	5(+2)

https://doi.org/10.1371/journal.pdig.0000447.t003

Unmatched balanced datasets. These datasets consist of a mixture of matched and unmatched samples: PBMC 3-Way Unmatched Balanced proteomics, PBMC 3-Way Unmatched Balanced RNAseq, Liver 3-Way Unmatched Balanced proteomics, and Liver 3-Way Unmatched Balanced RNAseq.

Matched balanced datasets. These datasets consist of only matched samples, such that for each RNAseq sample there is also a proteomic sample obtained from the same individual: PBMC 3-Way Matched Balanced proteomics, PBMC 3-Way Matched Balanced RNAseq, Liver 3-Way Matched Balanced proteomics, and Liver 3-Way Matched Balanced RNAseq.

Matched balanced integrated datasets. These datasets were formed by merging the proteomic and RNAseq data from Matched Balanced datasets: PBMC 3-Way Matched Balanced Integrated and Liver 3-Way Matched Balanced Integrated.

Validation dataset

We validated our proteomic liver tissue machine learning (ML) models using data obtained from MassIVE repository (accession number MSV000089168) [12]. This dataset contained liver tissue proteomic data from participants with AH (n = 6) and healthy controls (n = 12). Notably, the healthy controls came from two different sources, 7 from University of Louisville and 5 from John Hopkins University. Publicly available proteomic data from PBMCs was not available for the conditions in our study, and therefore, only the liver tissue datasets were validated using independent data. Information regarding the RNAseq liver tissue validation dataset can be found in our previous publication [11].

RNAseq Classification and Feature Selection Pipeline

The detailed methods used to classify RNAseq counts and identify best genes are described in [11]. Briefly, the classification was performed using nested cross-validation with feature selection. Features were selected using either differential expression software or information gain algorithm. Additionally, outlier features were removed prior to feature selection. Domain expertise was incorporated into the pipeline via enrichment analysis. For each dataset, multiple pipeline configurations were executed, resulting in multiple, promising, candidate gene sets. For each dataset, we then selected a single best gene set that maximized classification performance and in-silico biological relevancy (attained via enrichment analysis), while minimizing

the gene set size. The methods used throughout were focused on minimizing the possibility of overfitting. Note that for any given pipeline configuration, there is a resultant set of genes (candidate gene set). Subsequently, when referring to candidate or best gene sets, we are also referring to the pipeline configurations that resulted in those gene sets.

Proteomic Classification and Feature Selection Pipeline

Methods used to classify proteomic counts and identify best proteins were similar to the methods used for analysis of RNAseq data with the following exceptions.

Feature sizes. The feature sizes for proteomic data were largely based on our findings when dealing with RNAseq data. Due to smaller number of proteomic samples the maximum number of features used was reduced from 500 to 200. The following feature sizes were selected: 15, 25, 35, 50, 60, 70, 80, 90, 100, 150, and 200.

Imputation. Unlike the RNAseq data, the proteomic data contained missing values. We used median and replacement with zero imputation strategies to address this. Median imputation replaces missing values using the median along each column (feature, in this case protein). Zero imputation replaces all missing values with zeros.

Imputed values were used for proteins that were missing data for a small number of samples. The following imputation thresholds were used 0%, 5%, and 10%. That is, values for a given protein were only imputed if less than the threshold % of total samples were missing data. Threshold of 0% means no imputation took place and all proteins with missing values were removed.

Differential expression feature selection. Cuffdiff [14] was used for the differential expression analysis of the RNAseq data, while we used INFERNORDN to perform differential expression analysis with proteomic counts [15]. Proteins were filtered by q-value ≤ 0.05. Afterward, any proteins that had too much missing data (above imputation threshold) were removed.

In silico biological validation and best protein set selection. Enrichr [16], which was used for RNAseq data analysis, was replaced with AGOTOOL [17] for enrichment analysis of proteins. When selecting the best protein set, an identical algorithm was used for both transcriptomic and proteomic data, with one exception. That is, for proteomic data, protein sets produced by configurations with the least imputation were preferred for selection.

Analysis outline

The analysis pipeline was divided into the 3 stages, which are shown in Fig 1.

Stage 1 (No Integration). In the first stage, we used machine learning approaches with nested cross-validation to separately classify the Full datasets (Liver 3-Way RNAseq Full, Liver 3-Way Proteomics Full, PBMC 3-Way RNAseq Full, and PBMC 3-Way Proteomics Full). This enabled us to identify the best genes and proteins, independently of each other, for both sample types using our RNAseq and proteomic pipelines. Refer to Fig 2 for the classification performance for Stage 1.

Stage 2 (Integration).

Part A:

We performed the same type of analyses as in Stage 1, i.e. nested cross-validation, to classify the Liver 3-Way Unmatched Balanced and PBMC 3-Way Unmatched Balanced gene and protein datasets. Each pipeline configuration produced a unique candidate gene/protein set. We noted several best performing candidate gene and protein sets for later use in parts B and C.

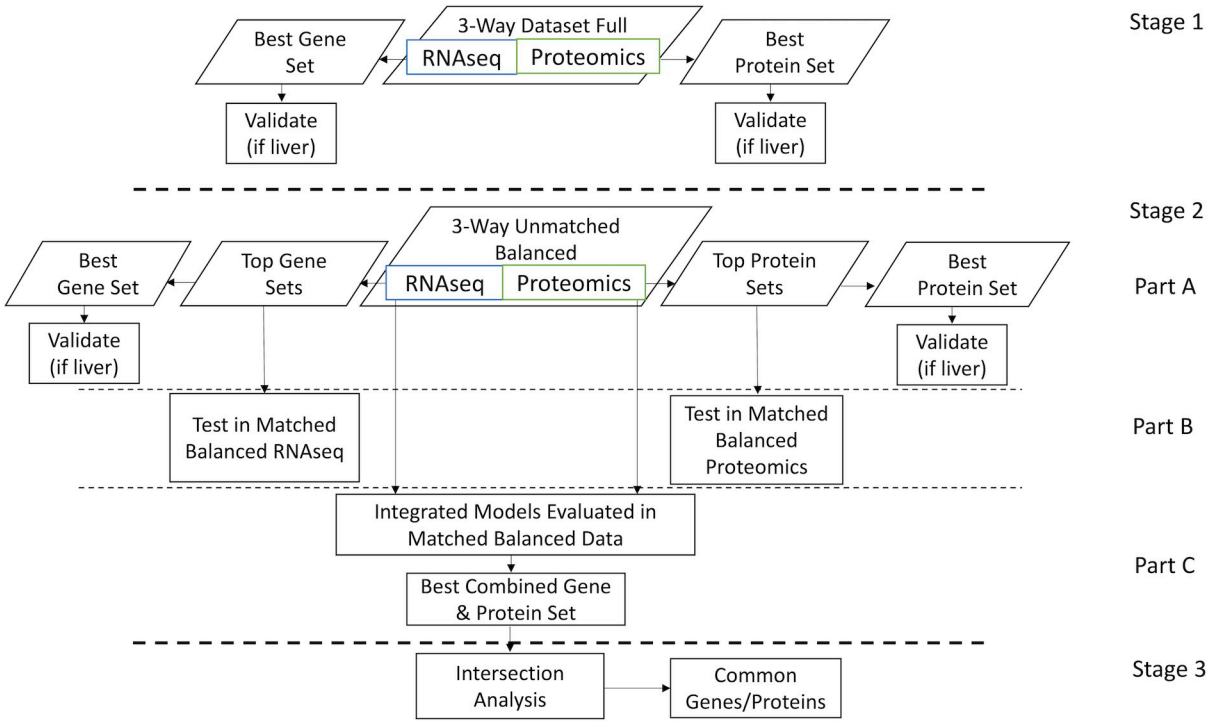

Fig 1. Flowchart of the 3 stages of the analysis. Stage 1: Separate analyses of full RNAseq and proteomics datasets (Liver 3-Way RNAseq Full, Liver 3-Way Proteomics Full, PBMC 3-Way RNAseq Full, and PBMC 3-Way Proteomics Full). To simplify the flowchart, we are only showing one representative dataset, which we will refer to as "3-Way Full Datasets". Stage 2: Training ML models in unmatched balanced data with subsequent testing and integration in matched balanced data. Part A: Identification of top transcriptomic and proteomic pipeline configurations along with their corresponding gene and protein sets for unmatched balanced datasets. Part B: Evaluation of top performing models with their corresponding gene and protein sets from part A in matched balanced data. Part C: Integration of paired sets of the top performing gene and proteomics models with their corresponding gene and protein sets, in matched balanced data. Stage 3: Intersection analysis of the combined best gene-protein sets for liver samples and for PBMCs.

https://doi.org/10.1371/journal.pdig.0000447.g001

Part B:

We trained classifiers, corresponding to the best performing RNAseq and proteomic ML pipeline configurations from part A, on the entirety of unmatched balanced data. The resulting ML models were then tested in matched balanced data. This would serve as a reference, to which we could later compare the integrated model, as shown in Fig 3.

Part C:

Pairings of the best performing RNAseq and proteomic ML models for each sample type from part B (using their corresponding gene/protein sets) were integrated and evaluated in matched balanced data using cross-validation (Table AA in S1 Text for models tested for the liver samples, and Table AD in S1 Text for the PBMC models tested). The integration was performed by supplying the output prediction probabilities from each pair of RNAseq and proteomic models as input into an integrated model. The pair of candidate gene and candidate protein sets that attained the best classification accuracy was reported as the best combined gene and protein panel. The performance of integrated model in matched balanced data was compared to the performance of separate (RNAseq and proteomic) models in matched balanced data (from part B) as shown in Fig 3.

Stage 3 (Intersection). In the third stage, we examined which genes and proteins matched within the best gene and protein panel. That is, we can consider a protein and a gene that codes for it, as a match.

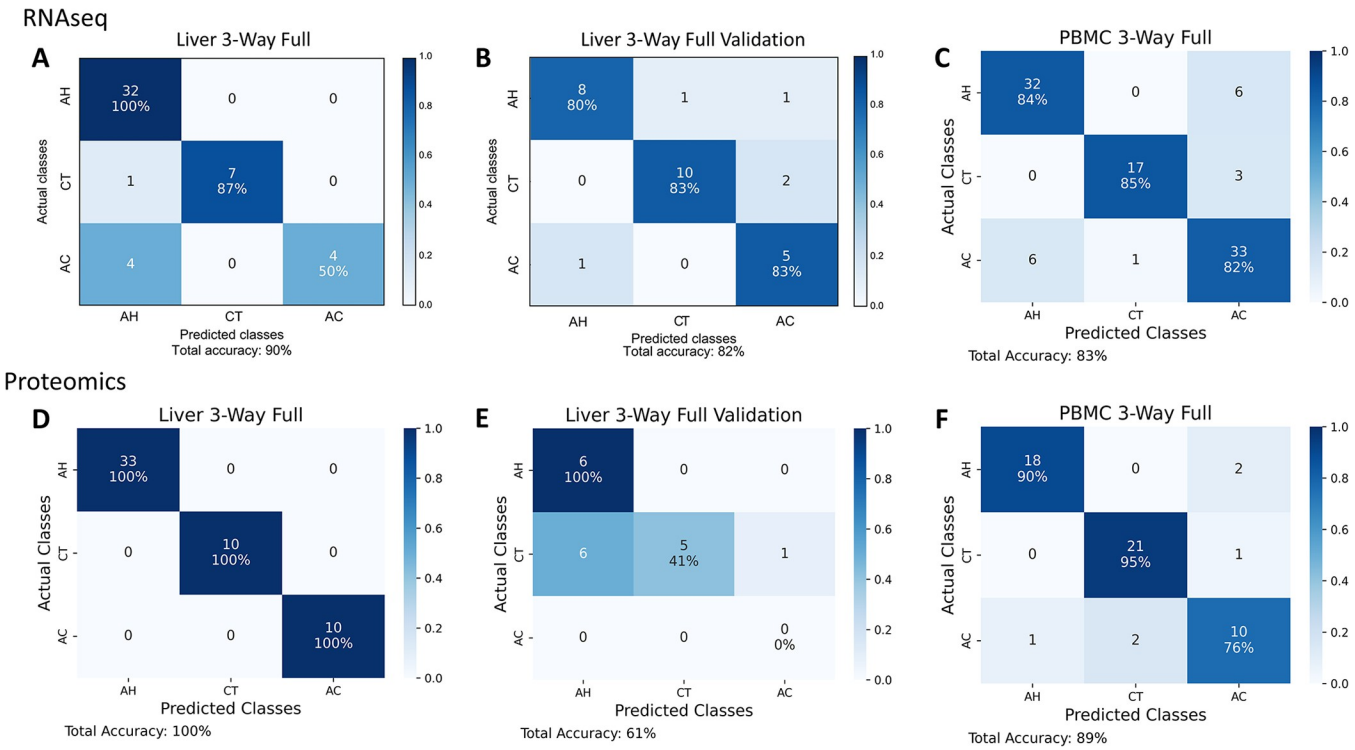

Fig 2. Confusion matrices corresponding to the best gene and protein sets of the full datasets and the liver tissue validation datasets. The Liver 3-way Full best gene and protein sets contained 33 genes and 27 proteins, respectively. The PBMC 3-Way Full best gene and protein sets contained 16 genes and 28 proteins, respectively. (A) Confusion matrix for classification of Liver 3-Way Full RNAseq dataset using best gene set identified by filter feature selection. The diagonal contains the number and percentage of the correctly predicted samples. (B) Confusion matrix for classification of AH, AC, and healthy control (CT) samples within independent validation RNAseq dataset. (C) Confusion matrix for classification of PBMC 3-Way Full RNAseq dataset using best gene set identified by filter feature selection. (D) Confusion matrix for classification of Liver 3-Way Full proteomic dataset using best protein set identified by filter feature selection. (E) Confusion matrix for classification of AH, AC, and CT samples within independent validation proteomic dataset. (F) Confusion matrix for classification of PBMC 3-Way Full proteomic dataset using best protein set identified by filter feature selection.

https://doi.org/10.1371/journal.pdig.0000447.g002

Validation in independent liver tissue data

All liver tissue ML models (RNAseq and proteomic) were validated in independent liver tissue validation data. Briefly, the ML model that performed best during nested cross-validation was trained on entirety of our liver tissue data. This trained classifier was then evaluated in independent liver tissue validation data. The methods for independent validation were identical for both RNAseq and proteomic datatypes. The further description of these methods can be found in our previous publication [11] methods.

Machine learning classifiers

The classifiers used in the individual analysis of the transcriptomic and proteomic data were: k nearest neighbors (kNN), logistic regression (LR), and support vector machine (SVM). For the integrated transcriptomic and proteomic analysis, we used only logistic regression and linear kernel SVM classifiers, due to ease of interpretation. Within the integrated model, the models that directly utilized the RNAseq and proteomic counts were either LR or linear kernel SVM. The classifier that used the prediction probabilities supplied via the RNAseq and proteomic models was LR with default hyperparameters. The LR model has been shown to be well suited for small sample size proteomic data previously [18]. Both LR and SVM classifiers were regularized.

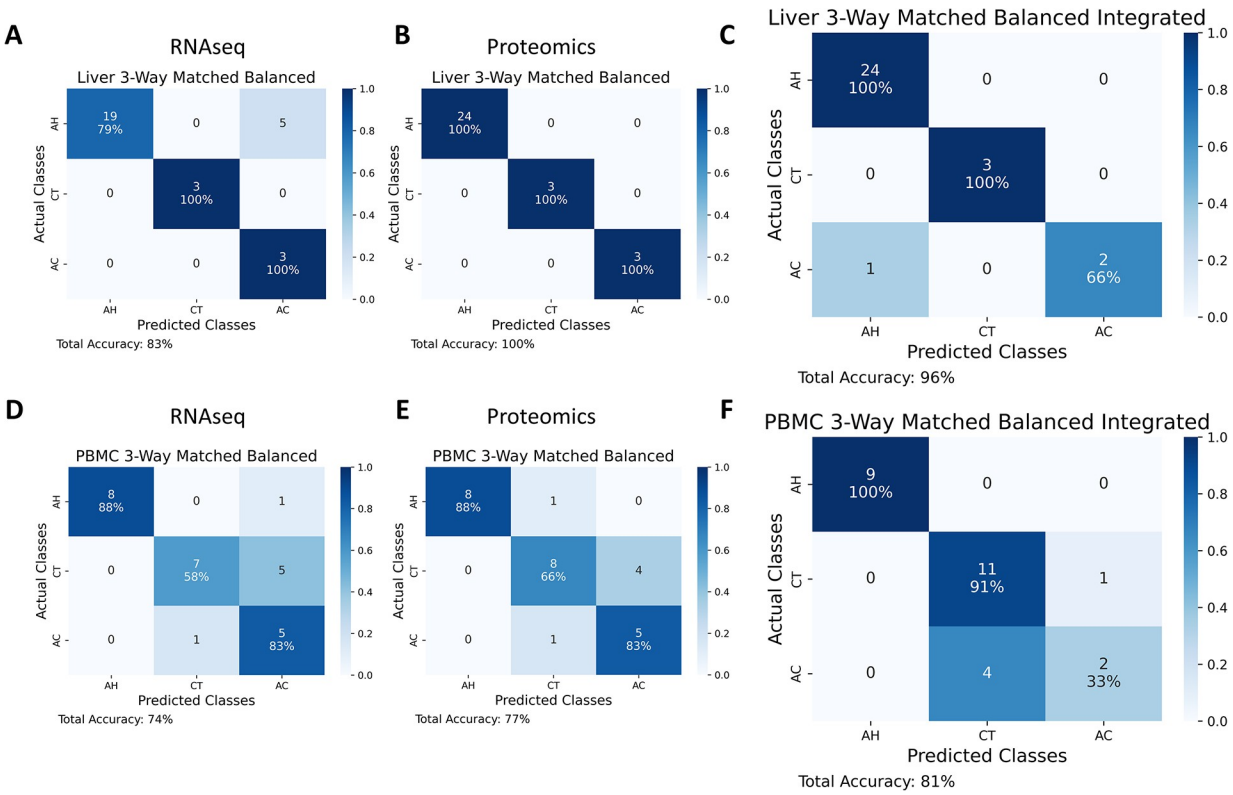

Fig 3. Confusion matrices corresponding to the best gene and protein sets in the matched balanced data set tested separately, and tested with the integrated gene/protein set. Confusion matrices corresponding to the best gene and protein sets (59 genes and 19 proteins, respectively) evaluated within Liver 3-Way Matched Balanced data and within PBMC 3-Way Matched Balanced data (16 genes and 33 proteins, respectively). (A) Confusion matrix for classification of Liver 3-Way Matched Balanced RNAseq dataset using best gene set identified by filter feature selection. (B) Confusion matrix for classification of Liver 3-Way Matched Balanced proteomic dataset using best protein set identified by filter feature selection. (C) Confusion matrix for classification of Liver 3-Way Matched Balanced dataset using a combination of best gene and protein sets. (D) Confusion matrix for classification of PBMC 3-Way Matched Balanced RNAseq dataset using best gene set identified by filter feature selection. (E) Confusion matrix for classification of PBMC 3-Way Matched Balanced proteomic dataset using best protein set identified by filter feature selection. (F) Confusion matrix for classification of PBMC 3-Way Matched Balanced dataset using a combination of best gene and protein sets.

https://doi.org/10.1371/journal.pdig.0000447.g003

Feature importance

The combined gene-protein panels for integrated Liver 3-Way and integrated PBMC 3-Way datasets were evaluated for feature importance. Feature importance was evaluated separately for genes and proteins due to the nature of machine learning architecture. The feature importance was evaluated using trained model coefficients. Visualizations of feature importance for integrated Liver 3-Way and integrated PBMC 3-Way datasets can be found in S1 Text.

Summary of computational methods

Table 4 contains a summary of the computational methods used in the final configurations of the ML models for the RNAseq and Proteomics datasets. Further details can be found in S1 Text.

Results

Classification of Liver 3-Way Full (AH vs Healthy vs AC)

The gene and protein sets produced via various methods were compared according to classification performance and biological validation scores in order to select the best gene and protein

Table 4. Summary of methods used with transcriptomic and proteomic data types.

Data Type	Feature Selection	Feature Sizes	Imputation	ML Classifiers	In-silico Biological Validation
Transcriptomic	Filter (DE, IG)	10, 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500	None	LR, kNN, SVM	Enrichr
Proteomic	Filter (DE)	15, 25, 35, 50, 60, 70, 80, 90, 100, 150, 200	Median and Zero (Thresholds: 0%, 5%, and 10%)	LR, kNN, SVM	AGOTOOL

https://doi.org/10.1371/journal.pdig.0000447.t004

sets. The best gene set contained 33 genes, attained 90% accuracy in main data and 82% accuracy in validation data (Fig 2A and 2B). The best protein set contained 27 proteins, and attained 100% accuracy in main data and 61% accuracy in validation data (Fig 2D and 2E). RNAseq and proteomic data proved similarly effective at classifying our Liver 3-Way samples. However, the best gene set derived from RNAseq data achieved better performance in RNAseq validation data than the best protein set derived from proteomic data achieved in proteomic validation data. The heatmaps of RNAseq and proteomic counts can be found in Figures A-H in S1 Text. The enriched pathways, tissues, and diseases for best gene and protein sets can be found in the Tables E and H in S1 Text. The best gene and protein sets for each dataset are shown in Table 5.

Table 5. Best genes and proteins for each dataset. For the integrated datasets, the matching genes and proteins are bolded.

Dataset	Genes	Proteins
Liver 3-Way Full	*AKR1B10*, *C15orf52*, *CFTR*, *CREB3L3*, *CXCL6*, *CYP2A7*, *CYP2B6*, *DBNDD1*, *EEF1A2*, *EPS8L1*, *FAM198A*, *FCGR3B*, *FCN3*, *FITM1*, *GPC3*, *GPNMB*, *HAMP*, *HAO2*, *IGSF9*, *KRT23*, *LCN2*, *LYZ*, *MMP7*, *MT1G*, *PLA2G2A*, *PPP1R1A*, *RGS1*, *S100A8*, *SCTR*, *STAG3*, *TMEM132A*, *TREM2*, *VCAN*.	*ACBP*, *ADH1A*, *ADH1B*, *ADH4*, *ADH6*, *ALBU*, *ARF3*, *CD34*, *CO1A2*, *CP1A2*, *CP3A4*, *CP3A7*, *CRP*, *DDTL*, *ERI3*, *FABPL*, *GSTA1*, *GSTA2*, *GSTM4*, *H2B1C*, *K2C79*, *K2C80*, *LDH6A*, *MFAP4*, *PAL4C*, *SAA1*, *UDB17*.
PBMC 3-Way Full	*ETS2*, *FLVCR2*, *FPR1*, *GRB10*, *IMPA2*, *ITGAM*, *ITGB2*, *LILRA5*, *MYO7A*, *PTGR1*, *RAB31*, *RNASE2*, *SERPINB1*, *SLC36A1*, *ST14*, *TLR4*.	*APOA1*, *BLVRB*, *CATS*, *CSRP1*, *EST1*, *FIBA*, *FIBB*, *FIBG*, *G6B*, *GP1BB*, *GPIX*, *HBD*, *ILK*, *ITA2B*, *ITB3*, *LTBP1*, *MYL9*, *PMGE*, *RAP1A*, *RSU1*, *SDPR*, *SEP11*, *SRC*, *TBA4A*, *TOR4A*, *TSP1*, *URP2*, *VINC*.
Liver 3-Way Matched Balanced Integrated	*ACKR1*, *AKR1B10*, *BBOX1*, *C15orf52*, *CFTR*, ***CLEC4M***, *CREB3L3*, *CSF3R*, *CXCL1*, *CXCL6*, *DCDC2*, *DHODH*, *DHRS2*, *F3*, *FABP4*, *FAM118A*, *FCGR3B*, *FCN3*, *GADD45B*, *GADD45G*, *GPC3*, ***GSTA2***, *HAMP*, *HAO2*, *ID4*, *IGSF9*, *IL7R*, *KRT23*, *LBP*, *LCN2*, *LRG1*, *MARCO*, *MMP7*, *MT1A*, *MT1G*, *MT1H*, *MT1M*, *MT1X*, *MUC13*, *MUC6*, *NRTN*, *PAPLN*, *PID1*, *PLA2G2A*, *PLCB1*, *PPP1R1A*, *S100A12*, *S100A8*, *S100A9*, *SLC13A5*, *SLC22A1*, *SOCS1*, *SPINK1*, *STAG3*, *STMN2*, *TREM2*, *TRIB3*, *VSIG2*, *VTCN1*.	*ACBP*, *ADH1A*, *ADH1B*, *ADH4*, *ADH6*, *ALBU*, *ASSY*, *CD34*, ***CLC4M***, *CO1A2*, *CP1A2 CRP*, *CYB5*, *ERI3*, ***GSTA1***, *HBAZ*, *LDH6A*, *SAA1*, *UDB17*.
PBMC 3-Way Matched Balanced Integrated	*AHSP*, *ALAS2*, *CA1*, *CD177*, *CDK10*, *EHMT1*, *HBD*, *HBM*, *IFI27*, *IL1R2*, *MECP2*, *MMP8*, *MMP9*, ***SELENBP1***, *SLC4A1*, *TANGO2*.	*ACTN1*, *ALBU*, *CCL5*, *CXCL7*, *FHL1*, *FIBA*, *FIBB*, *FIBG*, *FRIL*, *FSTL1*, *GP1BB*, *ILK*, *ITA2B*, *ITB1*, *ITB3*, *LIMS1*, *LYSC*, *MYL9*, *PP14A*, *RAP1A*, *RS4Y1*, ***SBP1***, *SDPR*, *TBA4A*, *TBA8*, *TBB1*, *TPM2*, *TRML1*, *TSN15*, *TSP1*, *URP2*, *VINC*, *VTDB*.

https://doi.org/10.1371/journal.pdig.0000447.t005

Classification of PBMC 3-Way Full (AH vs Healthy vs AC)

The best gene set contained 16 genes and attained 83% accuracy in main data (Fig 2C). The best protein set contained 28 proteins and attained 89% accuracy in main data (Fig 2F). RNA-seq and proteomic data proved equally effective at classifying our PBMC 3-Way samples. The heatmaps of RNAseq and proteomic counts can be found in Figures I-L in S1 Text. The enriched pathways, tissues, and diseases for best gene and protein sets can be found in the Tables K and N in S1 Text. The best gene and protein sets for each dataset are shown in Table 5.

Classification of Liver 3-Way Matched Balanced (AH vs Healthy vs AC)

Integration of genes and proteins. The best gene set and protein set derived from Liver 3-Way Unmatched Balanced datasets were evaluated in Liver 3-Way Matched Balanced datasets separately and in combination. Using the best gene set of 59 genes we attained 83% classification accuracy within matched balanced RNAseq data (Fig 3A). Using the best protein set of 19 proteins we attained 100% classification accuracy within matched balanced proteomic data (Fig 3B). Using a combination of best gene and protein sets, we attained 96% accuracy in matched balanced integrated data (Fig 3C). Additionally, we generated a one-vs-rest micro-averaged receiver operating characteristic (ROC) curve for the integrated Liver 3-Way model, which resulted in AUC of 1.0 (Figure AE in S1 Text). The constituent transcriptomic (59 genes) and proteomic (19 proteins) models resulted in AUCs of 0.94 and 1.0 respectively (Figures AF and AG in S1 Text).

Intersection. Additionally, we examined which biomarkers were shared between the best gene and protein sets of the integrated model with liver tissue. The *CLEC4M-CLC4M*, *GSTA1-GSTA2* were found in common. The *CLEC4M-CLC4M* was a direct match, while the *GSTA1* (protein) was a familial match with *GSTA2* (gene). If the genes and proteins had been selected randomly from among significantly differentially expressed genes and proteins, an expected 0.12 would be shared. Calculation of expected value can be found in S1 Text. Therefore, we have identified more biomarkers in common than expected. Best gene and protein sets were commonly enriched for several different inflammation pathways. The best protein set was more strongly enriched for metabolism pathways than the best gene set (Tables Q and T in S1 Text).

Classification of PBMC 3-Way Matched Balanced (AH vs Healthy vs AC)

Integration of genes and proteins. The best gene and protein sets derived from PBMC 3-Way Unmatched Balanced datasets were evaluated in PBMC 3-Way Matched Balanced datasets separately and in combination. Using the best gene set of 16 genes we attained 74% classification accuracy within matched balanced RNAseq data (Fig 3D). Using the best protein set of 33 proteins we attained 77% classification accuracy within matched balanced proteomic data (Fig 3E). Using a combination of best gene and protein sets, we attained 81% accuracy in matched balanced integrated data (Fig 3F). We also generated a one-vs-rest micro-averaged ROC curve for the integrated PBMC 3-Way model, which resulted in AUC of 0.96 (Figure AK in S1 Text). The constituent transcriptomic (16 genes) and proteomic (33 proteins) models resulted in identical AUCs of 0.89 (Figures AL and AM in S1 Text).

Intersection. With the integrated model for PBMCs, the *SELENBP1-SBP1* gene-protein was found in common between the best gene and protein sets. For a random selection from the significantly differentially expressed genes and proteins, we calculated that an expected 0.05 would be shared. Thus, more biomarkers were found to be shared than expected. The best

gene and protein sets for PBMCs were mainly enriched for several different inflammation and cancer related pathways (Tables W and Z in S1 Text).

Discussion

In this study, we used machine learning approaches with transcriptomics and proteomics data from liver tissue and PBMCs to effectively classify samples from participants with alcohol-associated hepatitis (AH), alcohol-associated cirrhosis (AC), and healthy controls. Liver tissue models outperformed PBMC models by a small margin in our data. Both transcriptomic and proteomic liver tissue ML models generalized relatively well in the independent validation data. Overall, the transcriptomic and proteomic models performed similarly well in each sample type.

The integration of proteomic and transcriptomic data did not increase classification accuracy with liver tissue, mainly because the classification accuracy was already high in both data types separately. For PBMCs, on the other hand, the integration improved classification accuracy slightly. While the performance of PBMC biomarkers is less than that of liver tissue biomarkers for classification of ALDs, the integration of multiple -omics data types could help close the gap in the future. To our knowledge, this is the first study in which a combined PBMC gene-protein expression biomarker panel has been identified for distinguishing AH, AC, and healthy controls.

Of special interest are the gene-protein matches present in the combined gene-protein sets identified for Liver 3-Way and PBMC 3-Way Matched Balanced Integrated datasets. All the matched liver tissue genes have been established as relevant biomarkers of liver disease in prior literature. *CLEC4M* has been identified as prognostic liver tissue biomarker of hepatocellular carcinoma [19]. *GSTA1* and *GSTA2* have been previously identified as biomarkers of liver injury (including ethanol injury) and hepatocellular carcinoma respectively [20,21]. Less is known about the role of the matched PBMC genes in liver disease. Differential expression of *SELENBP1* in PBMCs of hepatocellular carcinoma patients has been established previously [22]. The differential expressions of these biomarkers in both transcriptomic and proteomic data increases our confidence in their significance.

The gene-protein panels for Liver 3-Way and PBMC 3-Way integrated datasets were examined using enrichment analysis. The genes and proteins were examined separately. For Liver 3-Way the proteins were overwhelmingly enriched for metabolic pathways, including ethanol metabolism (Table AB in S1 Text). Notably, many of the key liver proteins are alcohol dehydrogenases, some of which have been implicated in alcohol and liver disorders [23,24]. Other notable proteins include *CRP*, *SAA1*, *ALBU*. All of these have been previously established as diagnostic biomarkers of inflammatory liver diseases [25,26,27]. The genes were enriched for homeostasis, metabolism, and inflammatory pathways (Table AC in S1 Text). For PBMC 3-Way both the genes and proteins were enriched for blood processes, immune system functions, and cellular movement (Tables AE and AF in S1 Text). Some of the PBMC proteins have been previously connected to liver disease including *FSTL1*, *TSP1*, *CCL5*, and *TPM2* [28,29,30,31]. Overall, the identified genes and proteins are consistent with previous findings.

We have discussed the importance of using appropriate ML methods for analysis of small sample size RNAseq data [11] previously. Our recommendations for analysis of small sample size proteomic data are largely similar. In addition to the importance of filter feature selection we would like to highlight the importance of nested cross-validation (NCV) and performing feature selection within both inner and outer loops of NCV. The use of nested cross validation is necessary to separate model selection and evaluation if hyperparameter tuning is being done. Meanwhile, it is necessary to perform feature selection within nested cross validation to

avoid data leakage and the resulting bias [32]. The use of in-silico biological relevancy (via enrichment analysis) in our pipeline was also important as it decreased overfitting by favoring feature sets that corresponded to existing literature.

The liver tissue proteomics model's performance in independent validation data was lower than expected. The healthy control samples in independent validation proteomic dataset were collected from two different clinical sources. Most misclassified healthy controls were from one of the two sources. The heterogeneity in healthy samples may explain their unexpectedly poor classification performance. The PBMC models could not be independently validated due to lack of relevant public data. However, the methods used to derive the best biomarkers were identical in both tissues. The integrated models also could not be validated due to lack of appropriate publicly available genomic data in which both RNAseq and proteomics were available for the same individuals. A larger sample size and an independent integrated validation cohort are needed to further investigate these biomarkers.

Integrating two -omics datatypes further amplified the challenges we encountered in our earlier work [11]. The number of genes and proteins for each sample is much larger than the number of samples in our dataset. This makes data prone to overfitting, since a complex model can perfectly separate a small number of samples. Some of the other challenges were ensuring that the integrated model did not have a bias toward transcriptomic or proteomic features, performing feature selection with integrated gene and protein expression data, and addressing partial matching between our transcriptomic and proteomic samples (most were obtained from the same individuals, but some were not).

Overall, the integration of proteomic and transcriptomic data from liver tissue and PBMCs for ALD proved promising in two aspects. In the case of PBMCs in our study, combining transcriptomic and proteomic biomarkers was more effective than using either type of biomarkers alone for classification. Additionally, by examining both transcriptomic and proteomic data, we were able to identify gene-protein pairs that were significantly differentially expressed in both domains and were thus more likely to be relevant to the liver disease conditions in question. The possibility of using PBMCs to distinguish among alcohol-associated liver diseases is encouraging, and the relevant biomarkers warrant further examination.

Supporting information

S1 Text. Supplemental methods and supplemental results for this study.
(PDF)

Acknowledgments

The authors would like to thank and acknowledge that the participant recruitment and sample collection for the PBMCs and the AH liver tissue biopsies were performed by the SCAHC at the following locations: Long Beach Veterans Healthcare System (VALB), Long Beach, CA [Jessica Clare Gozum, Sheena Cruz, Hema Buddha, Yuxin Ouyang, Gregory Botwin, Lauren MacHarg, Monique French]; Harbor-UCLA Medical Center, Torrance, CA [Lavanya Cherukuri, Sajad Hamal, Wayne Fleischman, Divya Birudaraju]; University of Southern California (USC), Los Angeles, CA [Christy Rico, Susan Milstein, Carol Jones, John Donovan, Neil Kaplowitz]; VA Loma Linda, CA [Daniel Chen-Kang Chao]; and VA Albuquerque [Joseph Alcorn]. The authors would also like to thank and acknowledge the members of the UC Irvine Genomics High-Throughput Facility (GHTF) for their role in the RNA extraction and sequencing of the samples. The liver tissue from participants with AC and healthy control were obtained from the LTCDS at University of Minnesota. (https://med.umn.edu/pathology/

research/liver-tissue-system). Portions of this manuscript were submitted as a thesis in partial fulfillment of the requirements for the degree of Doctor of Philosophy (S.L.).

Author Contributions

Conceptualization: Stanislav Listopad, Timothy R. Morgan, Trina M. Norden-Krichmar.

Data curation: Stanislav Listopad, Christophe Magnan, Le Z. Day, Aliya Asghar, Zhang-Xu Liu, Jon M. Jacobs, Trina M. Norden-Krichmar.

Formal analysis: Stanislav Listopad, Christophe Magnan, Le Z. Day, Zhang-Xu Liu, Jon M. Jacobs, Trina M. Norden-Krichmar.

Funding acquisition: Zhang-Xu Liu, Jon M. Jacobs, Timothy R. Morgan, Trina M. Norden-Krichmar.

Investigation: Le Z. Day, Aliya Asghar, Andrew Stolz, John A. Tayek, Zhang-Xu Liu, Jon M. Jacobs, Timothy R. Morgan.

Methodology: Stanislav Listopad.

Project administration: Zhang-Xu Liu, Jon M. Jacobs, Timothy R. Morgan, Trina M. Norden-Krichmar.

Resources: Aliya Asghar, Andrew Stolz, John A. Tayek, Zhang-Xu Liu, Jon M. Jacobs, Timothy R. Morgan, Trina M. Norden-Krichmar.

Software: Stanislav Listopad, Christophe Magnan.

Supervision: Zhang-Xu Liu, Jon M. Jacobs, Timothy R. Morgan, Trina M. Norden-Krichmar.

Validation: Stanislav Listopad, Trina M. Norden-Krichmar.

Visualization: Stanislav Listopad.

Writing – original draft: Stanislav Listopad.

Writing – review & editing: Stanislav Listopad, Christophe Magnan, Le Z. Day, Aliya Asghar, Andrew Stolz, John A. Tayek, Zhang-Xu Liu, Jon M. Jacobs, Timothy R. Morgan, Trina M. Norden-Krichmar.

References

1. Termeie O, Fiedler L, Martinez L, Foster J, Perumareddi P, Levine RS, et al. Alarming Trends: mortality from alcoholic cirrhosis in the United States. The American Journal of Medicine. 2022 May 27; 135 (10):1263–1266. https://doi.org/10.1016/j.amjmed.2022.05.015 PMID: 35636480
2. Mellinger JL, Volk ML. Transplantation for alcohol-related liver disease: is it fair? Alcohol and Alcoholism. 2017 Dec 11; 53(2):173–177. https://doi.org/10.1093/alcalc/agx105 PMID: 29236944
3. Thursz M, Morgan TR. Treatment of severe alcoholic hepatitis. Gastroenterology. 2016 Mar 4; 150 (8):1823–1834. https://doi.org/10.1053/j.gastro.2016.02.074 PMID: 26948886
4. Mathurin P, Moreno C, Samuel D, Dumortier J, Salleron J, Durand F, et al. Early liver transplantation for severe alcoholic hepatitis. The New England Journal of Medicine. 2011 Nov 10; 365:1790–1800. https://doi.org/10.1056/NEJMoa1105703 PMID: 22070476
5. Im GY, Kim-Schluger L, Shenoy A, Schubert E, Goel A, Friedman SL, et al. Early liver transplantation for severe alcoholic hepatitis in the United States–a single-center experience. American Journal of Transplantation. 2015 Dec 28; 16(3):841–849. https://doi.org/10.1111/ajt.13586 PMID: 26710309
6. Lee BP, Chen P, Haugen C, Hernaez R, Gurakar A, Philosophe B, et al. Three-year results of a pilot program in early liver transplantation for severe alcoholic hepatitis. Annals of Surgery. 2017 Jan; 265 (1):20–29. https://doi.org/10.1097/SLA.0000000000001831 PMID: 27280501

7. Singal AK, Bashar H, Anand BS, Jampana SC, Singal V, Kuo Y. Outcomes after liver transplantation for alcoholic hepatitis are similar to alcoholic cirrhosis: exploratory analysis from the UNOS database. Hepatology. 2012 Mar 18; 55(5):1398–1405. https://doi.org/10.1002/hep.25544 PMID: 22213344

8. Soresi M, Giannitrapani L, Cervello M, Licata A, Montalto G. Non invasive tools for the diagnosis of liver cirrhosis. World Journal of Gastroenterology. 2014 Dec 28; 20(48):18131–18150. https://doi.org/10.3748/wjg.v20.i48.18131 PMID: 25561782

9. Berger D, Desai V, Janardhan S. Con: liver biopsy remains the gold standard to evaluate fibrosis in patients with nonalcoholic fatty liver disease. Clinical Liver Disease. 2019 Apr 30; 13(4):114–116. https://doi.org/10.1002/cld.740 PMID: 31061705

10. Lambrecht J, Verhulst S, Mannaerts I, Reynaert H, Grunsven LA. Prospects in non-invasive assessment of liver fibrosis: liquid biopsy as the future gold standard? Molecular Basis of Disease. 2018 Jan 9; 1864(4):1024–1036. https://doi.org/10.1016/j.bbadis.2018.01.009 PMID: 29329986

11. Listopad S, Magnan C, Asghar A, Stolz A, Tayek JA, Liu Z, et al. Differentiating between liver diseases by applying multiclass machine learning approaches to transcriptomics of liver tissue or blood based samples. JHEP Reports. 2022 Aug 18; 4(10). https://doi.org/10.1016/j.jhepr.2022.100560 PMID: 36119721

12. Hardesty J, Day L, Warner J, Warner D, Gritsenko M, Asghar A, et al. Hepatic protein and phosphoprotein signatures of alcohol-associated cirrhosis and hepatitis. The American Journal of Pathology. 2022 Apr 28; 192(7):1066–1082. https://doi.org/10.1016/j.ajpath.2022.04.004 PMID: 35490715

13. Mandrekar P, Bataller R, Tsukamoto H, Gao B. Alcoholic hepatitis: Translational approaches to develop targeted therapies. Hepatology. 2016 Apr 15; 64(4):1343–1355. https://doi.org/10.1002/hep.28530 PMID: 26940353

14. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012 Mar 1; 7(3):562–578. https://doi.org/10.1038/nprot.2012.016 PMID: 22383036

15. Polpitiya AD, Qian W, Jaitly N, Petyuk VA, Adkins JN, Camp DG, et al. DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics. 2008 May 3; 24(13):1556–8. https://doi.org/10.1093/bioinformatics/btn217 PMID: 18453552

16. Chen EY, Tan CM, Kou Y, Duan QN, Wang ZC, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. Bmc Bioinformatics 2013 Apr 15; 14. https://doi.org/10.1186/1471-2105-14-128 PMID: 23586463

17. Schölz C, Lyon D, Refsgaard JC, Jensen LJ, Choudhary C, Weinert BT. Avoiding abundance bias in the functional annotation of post-translationally modified proteins. Nat Methods. 2015 Nov; 12(11):1003–4. https://doi.org/10.1038/nmeth.3621 PMID: 26513550

18. Niu L, Thiele M, Geyer PE, Rasmussen DN, Webel HE, Santos A, et al. Noninvasive proteomic biomarkers for alcohol-related liver disease. Nature Medicine. 2022 Jun 2; 28(6):1277–1287. https://doi.org/10.1038/s41591-022-01850-y PMID: 35654907

19. Luo L, Chen L, Ke K, Zhao B, Wang L, Zhang C, et al. High expression levels of CLEC4M indicate poor prognosis in patients with hepatocellular carcinoma. Oncology Letters. 2020 Jan 13; 19(3):1711–1720. https://doi.org/10.3892/ol.2020.11294 PMID: 32194663

20. Ma X, Liu F, Li M, Li Z, Lin Y, Li R, et al. Expression of gluthathione S-transferase A1, a phase II drug-metabolizing enzyme in acute hepatic injury on mice. Experimental and Therapeutic Medicine. 2017 Aug 17; 14(4):3798–3804. https://doi.org/10.3892/etm.2017.4957 PMID: 29042982

21. Ng KT, Yeung OW, Lam YF, Liu J, Liu H, Pang L, et al. Gluthathione S-transferase A2 promotes hepatocellular carcinoma recurrence after liver transplantation through modulating reactive oxygen species metabolism. Cell Death Discovery. 2021 Jul 21; 7(1). https://doi.org/10.1038/s41420-021-00569-y PMID: 34290233

22. Han Z, Feng W, Hu R, Ge Q, Ma W, Zhang W, et al. RNA-seq profiling reveals PBMC RNA as potential biomarker for hepatocellular carcinoma. Scientific Reports. 2021 Sep 7; 11(1). https://doi.org/10.1038/s41598-021-96952-x PMID: 34493740

23. Liu X, Li T, Kong D, You H, Kong F, Tang R. Prognostic implications of alcohol dehydrogenases in hepatocellular carcinoma. BMC Cancer. 2020 Dec 7; 20(1). https://doi.org/10.1186/s12885-020-07689-1 PMID: 33287761

24. Ehlers CL, Liang T, Gizer IR. ADH and ALDH polymorphisms and alcohol dependence in Mexican and Native American. The American Journal of Drug and Alcohol Abuse. 2012 Sep; 38(5):389–394. https://doi.org/10.3109/00952990.2012.694526 PMID: 22931071

25. Vanbiervliet G, Breton FL, Rosenthal-Allieri M, Gelsi E, Marine-Barjoan E, Anty R, et al. Serum C-reactive protein: A non-invasive marker of alcoholic hepatitis. Scandinavian Journal of Gastroenterology. 2006 Dec; 41(12):1473–1479. https://doi.org/10.1080/00365520600842195 PMID: 17101579

26. Li D, Xie P, Zhao S, Zhao J, Yao Y, Zhao Y, et al. Hepatocytes derived increased SAA1 promotes intrahepatic platelet aggregation and aggravates liver inflammation in NAFLD. Biochemical and Biophysical Research Communications. 2021 Apr 1; 555:54–60. https://doi.org/10.1016/j.bbrc.2021.02.124 PMID: 33813276

27. Pares A, Deulofeu R, Cisneros L, Escorsell A, Salmeron JM, Caballeria J, et al. Albumin dialysis improves hepatic encephalopathy and decreases circulating phenolic aromatic amino acids in patients with alcoholic hepatis and severe liver failure. Critical Care. 2009 Jan 28; 13(1). https://doi.org/10.1186/cc7697 PMID: 19175915

28. Gu G, Xue H, Yang X, Nie Y, Qian X. Role of follistatin-like protein 1 in liver diseases. Experimental Biology and Medicine. 2022 Dec 19; 248(3):193–200. https://doi.org/10.1177/15353702221142604 PMID: 36533576

29. Li Y, Turpin CP, Wang S. Role of thrombospondin 1 in liver diseases. Hepatology Research. 2016 Aug 30; 47(2);186–193. https://doi.org/10.1111/hepr.12787 PMID: 27492250

30. Ambade A, Lowe P, Kodys K, Catalano D, Gyongyosi B, Cho Y, et al. Pharmacological inihibition of CCR2/5 signaling prevents and reverses alcohol-induced liver damage, steatosis, and inflammation in mice. Hepatology. 2019 Feb 12; 69(3);1105–1121. https://doi.org/10.1002/hep.30249 PMID: 30179264

31. Safaei A, Tavirani MR, Oskouei AA, Azodi MZ, Mohebbi SR, Nikzamir AR. Protein-protein interaction network analysis of cirrhosis liver disease. Gastroenterology and Hepatology From Bed to Bench. 2016; 9(2);114–23. PMID: 27099671

32. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights into Imaging. 2021 Nov 24; 12. https://doi.org/10.1186/s13244-021-01115-1 PMID: 34817740