UNIVERSITY OF CALIFORNIA,
IRVINE

Towards integrated genomics data analyses to facilitate identification of diagnostic biomarkers

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Stanislav Listopad

Dissertation Committee:
Assistant Professor Trina M. Norden-Krichmar, Chair
Professor Xiaohui Xie
Professor Ali Mortazavi

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLDEGMENTS

I would like to thank my committee chair, Assistant Professor Trina Norden-Krichmar, for sharing her excitement about and knowledge of bioinformatics with me. I also would like to thank her for countless hours spent brainstorming ideas with me, mentoring me, and otherwise guiding me.

I would like to thank my committee members, Professors Xiaohui Xie and Ali Mortazavi, who masterfully taught me a great deal about machine learning and genomics respectively.

I would like to thank Professor Jeffrey Krichmar and Chancellor's Professor Nikil Dutt for being remarkably supportive mentors during my formative years at University of California, Irvine.

I would like to thank Professor of Teaching Richard E. Pattis who inspired me tremendously both as an undergraduate and graduate student. Richard Pattis is an exemplar of organization and efficiency who guided generations of computer science students at UCI.

I would like to thank my parents, grandparents, sister, and many friends (especially Joseph, Jisoo, Jules, Vandy, Jimmy, and Jeff) who have shown incredible compassion and support to me as I advanced through the program.

I thank Springer Nature for permission to include Chapter 2 of my dissertation, which was originally published in *BMC Bioinformatics*, under a Creative Commons license. License: https://creativecommons.org/licenses/by/4.0/.

Listopad S and Norden-Krichmar TM. A-Lister: a tool for analysis of differentially expressed omics entities across multiple pairwise comparisons. BMC Bioinformatics. 2019;20. doi: https://doi.org/10.1186/s12859-019-3121-x

I thank Elsevier Inc for permission to include Chapter 3 of my dissertation, which was originally published in *JHEP Reports*, under a Creative Commons license. License: https://creativecommons.org/licenses/by/4.0/.

Listopad S, Magnan C, Asghar A, Stolz A, Tayek JA, Liu Z-X, Morgan TR, and Norden-Krichmar TM. Differentiating between liver diseases by applying multiclass machine learning approaches to transcriptomics of liver tissue or blood based samples. JHEP Reports. 2022;4(10). doi: https://doi.org/10.1016/j.jhepr.2022.100560

Funding for Chapters 2, 3, and 4 was provided by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) award numbers: U01AA021838(Norden-Krichmar), U01AA021886 (Morgan), U01AA021884 (Morgan), U01AA021918 (Jacobs), and U01AA021857 (Liu).

VITA

Stanislav Listopad

2016 B.S. in Computer Science and Engineering, University of California, Irvine

2022 Ph.D. in Computer Science, University of California, Irvine

FIELD OF STUDY

Computer Science with focus in Bioinformatics

PUBLICATIONS

Listopad S, Magnan C, Asghar A, Stolz A, Tayek JA, Liu Z-X, Morgan TR, and Norden-Krichmar TM. Differentiating between liver diseases by applying multiclass machine learning approaches to transcriptomics of liver tissue or blood based samples. JHEP Reports. 2022;4(10). doi: https://doi.org/10.1016/j.jhepr.2022.100560

Listopad S and Norden-Krichmar TM. A-Lister: a tool for analysis of differentially expressed omics entities across multiple pairwise comparisons. BMC Bioinformatics. 2019;20. doi: https://doi.org/10.1186/s12859-019-3121-x

Venkadesh S, Komendantov AO, **Listopad S**, Scott EO, Jong KD, Krichmar JL, and Ascoli GA. Evolving Simple Models of Diverse Intrinsic Dynamics in Hippocampal Neuron Types. Frontiers in Neuroinformatics. 2018;12(8). doi: https://doi.org/10.3389/fninf.2018.00008

ABSTRACT OF THE DISSERTATION

Towards integrated genomics data analyses to facilitate identification of diagnostic biomarkers

by

Stanislav Listopad

Doctor of Philosophy in Computer Science

University of California, Irvine, 2022

Assistant Professor Trina M. Norden-Krichmar, Chair

While the total amount of genomic data has rapidly increased over the past decade, most individual biomedical research studies are still limited to small numbers of participant samples due to the high costs of recruitment, sequencing, data storage, and data analysis. This results in many data sets with a low number of samples, but a very large number of features across multiple genomic data types. Appropriately handling the small sample size data sets and integrating multiple genomic data types is essential for identifying actionable diagnostic biomarkers. The overarching goal of my dissertation is to address some of these challenges using software engineering, bioinformatics, and machine learning methods. In this document, I will cover the three major projects of my dissertation. First, I will describe *A-Lister*, a software tool that I developed to filter, compare, and combine items across multiple differential expression files, to facilitate data integration and feature selection. Second, I implemented a multiclass machine learning approach to classify liver disease and identify gene expression biomarkers using a transcriptomics liver disease dataset. As part of this analysis, I have implemented a variety of bioinformatic pipelines, feature selection techniques, and machine learning classifiers to classify small sample size RNAseq data. Third, I created an integrated model using both transcriptomics and proteomics data to identify a combined gene and protein biomarker panel to

classify liver disease. The tools and methods developed in my dissertation are not specific to liver disease, but are intended for use with any small sample size genomics datasets to aid in biomarker discovery.

CHAPTER 1: INTRODUCTION

The amount of genomic data generated daily rivals that of YouTube and Twitter (1). This data is extremely heterogeneous in nature and is composed of many small datasets from individual studies, as well as a few large ones from meta-studies, clinical trials, and biobanks. Analyzing small sample size genomic data is difficult because it is inherently highly dimensional. Typical genetic, transcriptomic, proteomic, and other -omics datasets contain 10,000s of features. Additionally, it is often desirable to integrate multiple different -omics, further compounding the high dimensionality problem. One of the most common types of analysis with genomic data is biomarker discovery; a task that has proven to be highly challenging (2). The overarching goal of my dissertation is to address some of these challenges using software engineering, bioinformatics, and machine learning methods. It is composed of the following aims: 1) build a software tool that enables integration of differential expression data, 2) implement suitable machine learning methods and incorporate domain knowledge in order to identify diagnostic biomarkers within a small sample size multi-class transcriptomics dataset, and 3) integrate transcriptomics data with proteomics data in order to enhance understanding of the conditions under study.

For my first aim, I used a standard software process model to develop a bioinformatics tool called A-Lister (3). The purpose of the tool was to allow for convenient analysis of any differential expression dataset. The requirements gathering phase was performed by examining formats of differential transcriptomic, proteomic, and methylation expression files available in public genomic repositories. I then evaluated other requirements, such as data privacy, ease of use, core features, target runtime, etc. Data privacy proved an especially important issue, since some of the data expected to be analyzed through the tool was human genomic data. Human

genomic data is protected by the Health Insurance Portability and Accountability Act (HIPAA). Due to concerns about HIPAA compliance, I decided against implementing the tool as a web app, and instead developed it as a desktop application. In order to make the accessible to the largest user base, we assumed that the user would have minimal programming background. Therefore, we developed both a command line and a graphical user interface. The core features available consisted of filter and set operations that could be performed on pairwise comparisons within differential expression files. Additionally, it was developed to analyze multiple differential expression files at once in an integrated manner. The runtime for A-Lister was kept under five seconds, which was deemed reasonable. The tool was implemented in Python, since runtime was less of a concern than ease of development and maintenance. During implementation, test-driven development was used in order to minimize occurrence of bugs. An effort was made to minimize dependency on third party packages in order to diminish future maintenance costs. A-Lister code is available on GitHub, and the manuscript was published in BMC Bioinformatics (3). A-Lister was used within subsequent aims of my dissertation. Currently, A-Lister has also been cited by two publications (4, 5).

For my second aim, I developed a multiclass machine learning framework to classify disease conditions using small sample size transcriptomics data. I illustrated the functionality and utility of this aim with a liver disease transcriptomics dataset. The liver disease participants were enrolled through the Southern California Alcoholic Hepatitis Consortium (SCAHC). Both liver tissue and blood samples were taken from these participants. Additionally, the data set contained some liver tissue samples from the University of Minnesota Liver Tissue Cell Distribution System (LTCDS). The biospecimens consisted of 137 peripheral blood mononuclear cell (PBMC) samples and 67 liver tissue samples. The samples represented five distinct conditions:

alcohol-associated hepatitis (AH), alcohol-associated cirrhosis (AC), non-alcohol-associated fatty liver disease (NAFLD), chronic viral hepatitis C infection (HCV), and healthy controls (CT). Standard bioinformatics pipelines were used to process the raw sequencing data and attain gene expression counts. Differential expression analysis was performed on the gene expression counts. I then implemented and compared numerous different bioinformatic pipelines, feature transformation and selection methods, machine learning classifiers, performance metrics, outlier filtering strategies, and functional relevance scoring systems in context of small sample size transcriptomics data. The analyses were challenging due to feature size (10,000s) being magnitudes larger than sample size (100s) within the dataset. I was able to successfully classify the liver disease conditions, and identify the most effective gene expression biomarkers, culminating in the publication of this study in JHEP Reports (6).

For my third aim, I analyzed and integrated transcriptomic and proteomic data. Both types of genomic data were obtained through SCAHC and LTCDS. The goals of this analysis were to compare effectiveness of transcriptomic and proteomic biomarkers, evaluate utility of a combined gene and protein expression biomarker panel, and identify matching gene-protein pairs. The challenges present in the second aim were compounded, since our sample size stayed the same, while our feature size increased due to combination of both gene and protein expression features. I was able to successfully demonstrate the versatility of my classification and feature selection pipeline developed within the second aim by successfully applying it to proteomic data. This allowed us to compare the effectiveness of transcriptomic and proteomic biomarkers using similar methods. I implemented an ensemble model to perform integrated gene and protein expression analysis. The output of the ensemble model was a combined gene-protein

expression panel, which I then analyzed for gene-protein matches. The results of this analysis are currently in preparation for publication.

CHAPTER 2: A-Lister: a tool for analysis of differentially expressed omics entities across multiple pairwise comparisons

Published in BMC Bioinformatics, November 19th 2019:

Listopad S and Norden-Krichmar TM. A-Lister: a tool for analysis of differentially expressed omics entities across multiple pairwise comparisons. BMC Bioinformatics. 2019;20. doi: https://doi.org/10.1186/s12859-019-3121-x

Abstract

Background

Researchers commonly analyze lists of differentially expressed entities (DEEs), such as differentially expressed genes (DEGs), differentially expressed proteins (DEPs), and differentially methylated positions/regions (DMPs/DMRs), across multiple pairwise comparisons. Large biological studies can involve multiple conditions, tissues, and timepoints that result in dozens of pairwise comparisons. Manually filtering and comparing lists of DEEs across multiple pairwise comparisons, typically done by writing custom code, is a cumbersome task that can be streamlined and standardized.

Results

A-Lister is a lightweight command line and graphical user interface tool written in Python. It can be executed in a differential expression mode or generic name list mode. In differential expression mode, A-Lister accepts as input delimited text files that are output by differential expression tools such as DESeq2, edgeR, Cuffdiff, and limma. To allow for the most flexibility in input ID types, to avoid database installation requirements, and to allow for secure offline use, A-Lister does not validate or impose restrictions on entity ID names. Users can specify thresholds to filter the input file(s) by column(s) such as *p*-value, q-value, and fold

change. Additionally, users can filter the pairwise comparisons within the input files by fold change direction (sign). Queries composed of intersection, fuzzy intersection, difference, and union set operations can also be performed on any number of pairwise comparisons. Thus, the user can filter and compare any number of pairwise comparisons within a single A-Lister differential expression command.

In generic name list mode, A-Lister accepts delimited text files containing lists of names as input. Queries composed of intersection, fuzzy intersection, difference, and union set operations can then be performed across these lists of names.

Conclusions

A-Lister is a flexible tool that enables the user to rapidly narrow down large lists of DEEs to a small number of most significant entities. These entities can then be further analyzed using visualization, pathway analysis, and other bioinformatics tools.

Background

With the recent explosion of genomic data, researchers are storing, cleaning, processing, and analyzing increasingly large volumes of data (1). Differential expression studies account for a large portion of this data. Entire differential expression analysis pipelines have been built specifically for analyzing data generated in differential expression studies. These pipelines often end at the differential expression analysis step, the output of which is lists of differentially expressed entities (e.g. genes, proteins, etc.) (7). Each file represents all the entities that were differentially expressed between two conditions (e.g. control vs. drug A). In addition to the entity names themselves, fold changes, *p*-values, and many other categories of information are commonly listed within the differential expression files (8,9,10,11). Filtering and comparing files

of differentially expressed entities is a common task that is usually done by writing custom Perl, Python, or R scripts. This task can grow cumbersome when dealing with many pairwise comparisons. A-Lister addresses that concern by allowing the user to filter and compare any number of pairwise comparisons across any number of differential expression files within a single command. A-Lister accepts most common delimited (tab, comma, colon, semicolon, and space) text files containing differential expression data and is thus compatible with most differential expression tools.

A-Lister is intended for use within bioinformatics analysis pipelines between the differential expression (DE) analysis and the visualization/pathway analysis steps. A-Lister narrows down lists of differentially expressed entities produced by differential expression analysis tools. These entities can then be further analyzed using visualization, pathway, and other bioinformatics software.

Implementation

A-Lister is written in Python 3.7. A-Lister is freely available on GitHub at (12). The command line interface (CLI) version can be run in Windows, Mac, and Unix operating systems. The graphical user interface (GUI) version can be used to generate and launch A-Lister commands.

Workflow and output

An A-Lister command can be written and executed directly at a command line, or the command can be generated and executed through the GUI. All relevant input is supplied within a single A-Lister command. There are two commands available: diff-expression and name-list.

The diff-expression command is used to execute A-Lister in differential expression (DE) mode. The name-list command is used to execute A-Lister in generic name list mode.

Below is a description of how A-Lister executes diff-expression and name-list commands (Figure 2.1). Once the command is entered, A-Lister proceeds to validate it. If the command is valid, the program reads in the input files provided by the user. If specified by the user, the data within the input files may be filtered by any column. Furthermore, in DE mode the individual pairwise comparisons can be filtered by direction (sign of fold change). Set operations are then performed on the groups (name list mode) or pairwise comparisons (DE mode) as specified within the query. Once the query is executed, a delimited list of the resultant entity names and the count, is written into the result file. A system dump file is also output containing additional information regarding A-Lister's execution that can be helpful with debugging or validation. Additionally, in DE mode, the filtered copies of the original input files are output. These files are obtained by filtering the original input files by the result.

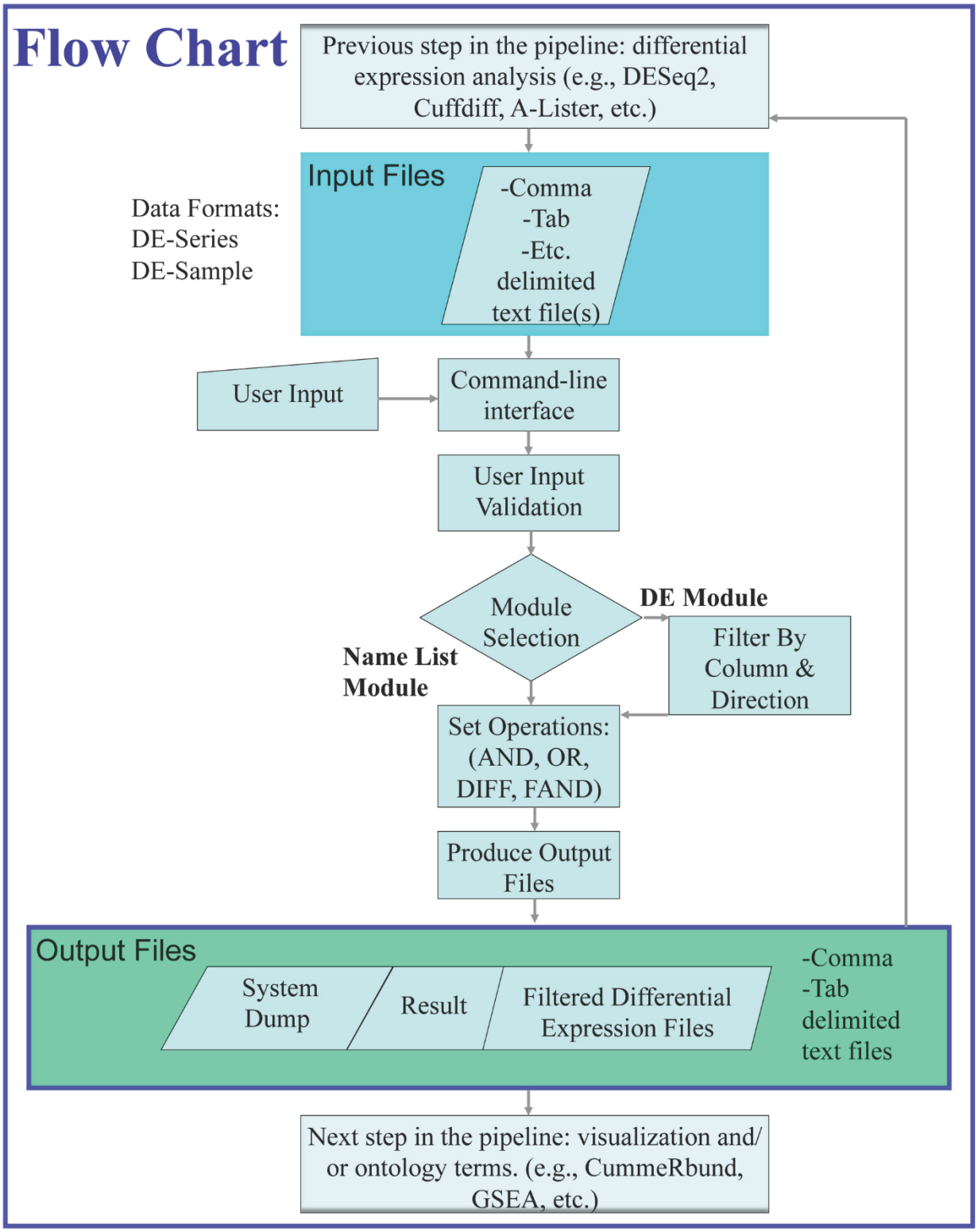

Figure 2.1: Data and control flow diagram of A-Lister.

Input Files

Input files for the name list command

In this mode, A-Lister accepts files containing columns of names delimited by tabs or commas. The header row must contain the group name for each column. An example of a name list file with three groups: control, treated1, and treated2 is shown below (Table 2.1).

Table 2.1: Example of a Name List File

Control	Treated1	Treated2
AADACL2	AADACP1	AADACP1
AADACL4	DUSP5P1	AMICA1

Input files for the differential expression command

In this mode, A-Lister accepts differential expression files containing a primary ID column (e.g. gene name), fold change column(s), and any other columns present. The columns in these files must be delimited by tab, comma, colon, semicolon, or space. A-Lister supports two types of differential expression file formats described below.

Differential Expression Sample Format (DE-Sample) (Row-Format) File: This is a delimited text file containing a primary ID column, single Fold Change column, one Sample1 column, and one Sample2 column. The Sample1 and Sample2 columns identify to which pairwise comparison each row belongs. In this way, multiple pairwise comparisons can be listed within a single DE-Sample file using a single fold change column (Table 2.2). The .diff files that are output from Cuffdiff follow this format (13).

Table 2.2: Example of a DE-Sample File

gene	locus	sample1	sample2	log2(FC)	p-value
FAM3A	chrX:154506158-154,516,242	q1	q2	2.73	0.0023
FAM3A	chrX:154506158-154,516,242	q3	q4	0.0649976	0.81

Differential Expression Series Format (DE-Series) (Column-Format) File: This is a delimited text file containing a single ID column and multiple Fold Change columns. Each Fold Change column contains data pertaining to a single pairwise comparison. In this way, multiple pairwise comparisons can be listed within a single file using multiple fold change columns (Table 2.3). This is the most common format for differential expression files.

Table 2.3: Example of a DE-Series File

gene	locus	log2(FC)	p-value	log2(FC)2	p-value2
FAM3A	chrX:154506158-154,516,242	2.73	0.0023	0.0649976	0.81

A-Lister filtering

A-Lister filtering is performed if the user specifies the optional filter by column (−f) parameter (Table 2.4) for any column (attribute) within a differential expression file. When filtering a DE-Sample file by an attribute the entire file is filtered. When filtering a DE-Series file by an attribute, there are two possible behaviors. First, if the filter attribute belongs to a pairwise comparison, such as p-value2, then only that pairwise comparison is filtered. Second, if the filter attribute belongs to the entire file (e.g. ID column), then the entire file is filtered. Additionally, pairwise comparisons can be filtered by direction (sign of fold change) using the directional query (−dq) argument described below (Table 2.4).

Table 2.4: A-Lister Command Line Interface (CLI). Bolded parameters are mandatory

Command	Argument	Brief Description
name-list	**<input-file>**	Full path to the input file.
	<query>	The query to be performed over input name lists.
	-id	Delimiter used in input file.
	-o	Output directory.
	-od	Delimiter used in output file.
	-v	Verbose flag.
	-e, --examples	Show examples and exit.
	-h, --help	Show help/manual and exit.
diff-expression	**<input-file>**	Full path to the input file.
	-dq, <direct-query>	The directional query to be performed over pairwise comparisons.
	-pc, <pc-mapping>	Specifies the layout of pairwise comparison within the file.
	-n	Specifies the ID column in file.
	-fc	Specifies the fold change column(s) in file.
	-s1	Specifies sample1 column in file.
	-s2	Specifies sample2 column in file.
	-f	Filter parameter used to filter the files by columns/attributes.
	-id	Delimiter used in input file.
	-o	Output directory.
	-od	Delimiter used in output file.
	-v	Verbose flag.
	-e, --examples	Show examples and exit.
	-h, --help	Show help/manual and exit.

A-Lister directional query

A-Lister directional query is composed of pairwise comparisons, set operators, and optional directions. The pairwise comparison names are derived from the pairwise comparison mapping argument (−pc) (Table 2.4). The permitted set operators are: AND, FAND, OR, and

DIFF. Additionally, parenthesis can be used to nest and to set order of operations. A directional query is specified with the (−dq) argument used in the diff-expression command (Table 2.4).

Set operations

Specifying the AND operator on two sets of elements returns a set of all the elements that are present within both sets. The FAND operator applied to two sets returns a set of all the similar elements from within both sets. A customized Jaro-Winkler algorithm is used to calculate similarity. To be considered similar, two strings must have Jaro-Winkler score > 0.84 (14). The OR operator applied to two sets returns all the elements present in either set. The DIFF operator applied to two sets returns all the elements present in the first set, but not in the second. All set operations are implemented using the standard Python library.

Directionality

Specifying the UP keyword in a query selects all entities whose fold change values are positive for a given pairwise comparison. Specifying DOWN in a query selects all entities whose fold change values are negative for a given pairwise comparison. ALL is a special modifier that results in multiple queries. That is, query results are returned as if ALL was specified as all combinations of UP and DOWN. For example, a query containing N ALL directions is transformed into 2^N queries. Each query is then executed and the results for each query are output into the output files in separate directories. NONE is the default direction for all pairwise comparisons. Pairwise comparisons with NONE direction are not filtered by direction.

A-Lister query (non-directional)

A non-directional query is composed of group names and set operators. The set operators are the same as in the directional query (e.g., AND, FAND, OR, DIFF), and can also include

parentheses to nest and order the operators. The group names are derived from the first (header) row of the name list files. The non-directional query argument is used in the name-list command (Table 2.4).

Results

A-Lister can be executed through a command line interface (CLI) or a graphical user interface (GUI). Underlying A-Lister's CLI and GUI is organization into two commands. The two commands are name-list and diff-expression, which represent the generic name list mode and the differential expression mode of execution. Each command has its own set of arguments (Table 2.4). We will first describe the CLI through example use cases to illustrate the parameters and functionality, and then given an overview of the GUI version.

Use case 1: analysis of name list files and fuzzy intersection (FAND) operation

Suppose the user wants to identify all same and similar genes within two sets of genes (Table 2.5). The first set is contained in file A and the second set is contained in file B.

Table 2.5: Example of Name-List Command with Intersection (AND) and Fuzzy Intersection (FAND) Query

Set1	Set2	AND (same)	FAND (similar)
ACTA2	ACTA2	ACTA2	ACTA2
ACTG1	ACTN1	ACTN1	ACTG1
ACTN1	ADAMTS1	ADAMTS1	ACTN1
ACAD9	ERVV-2	FLNB	ADAMTS1
ADAMTS1	FLNB	KLF4	ADAMTS9
ADAMTS9	KDM4A	PIK3R3	FLNA
EXD1	KLF4	SLC20A1	FLNB
FLNA	KLF6	SLC25A25	KLF10
FLNB	MYOM1		KLF2
KLF10	PIK3R3		KLF4
KLF2	SLC19A2		KLF6
KLF4	SLC20A1		PIK3IP1
OR1S1	SLC25A25		PIK3R3
PIK3IP1	TNFRSF12A		SLC12A2
PIK3R3	ZIM2		SLC19A2
SLC12A2			SLC20A1
SLC20A1			SLC25A25
SLC25A25			TNFRSF12A
TNFSF10			TNFSF10
ZFAT-AS1			

The A-Lister command listed below will provide the same genes within the 2 files by using the AND operator:

```
python AListerCLI.py name-list "Set1-AND-Set2" FileA.txt
FileB.txt -o E:/Data/Sample_Output
```

The A-Lister command listed below will provide the similar genes within the 2 files by using the FAND operator:

```
python ALister_CLI.py name-list "Set1-FAND-Set2" FileA.txt
FileB.txt -o E:/Data/Sample_Output
```

The output of these commands is shown in Table 2.5.

Use case 2: analysis of differential expression using a complex query

The data for this use case can be downloaded from NCBI's gene expression omnibus (GEO) database (15). The series number is GSE126785 (16). There are three groups of samples in the study: two types of induced pluripotent stem cells (iPSCs) and embryonic stem cells (ESCs). The gene expression of each group was measured under 5% oxygen and under 20% oxygen. The published files are three DESeq2 files, each containing genes differentially expressed for a single cell line between the 5% oxygen and 20% oxygen conditions. M2 is the ESC line. M4 and M5 are the iPSC lines. Suppose the user wants to know which genes are significantly differentially expressed in the embryonic stem cells (under different oxygen conditions) but are not significantly differentially expressed in either of the induced pluripotent stem cells (Figure 2.2).

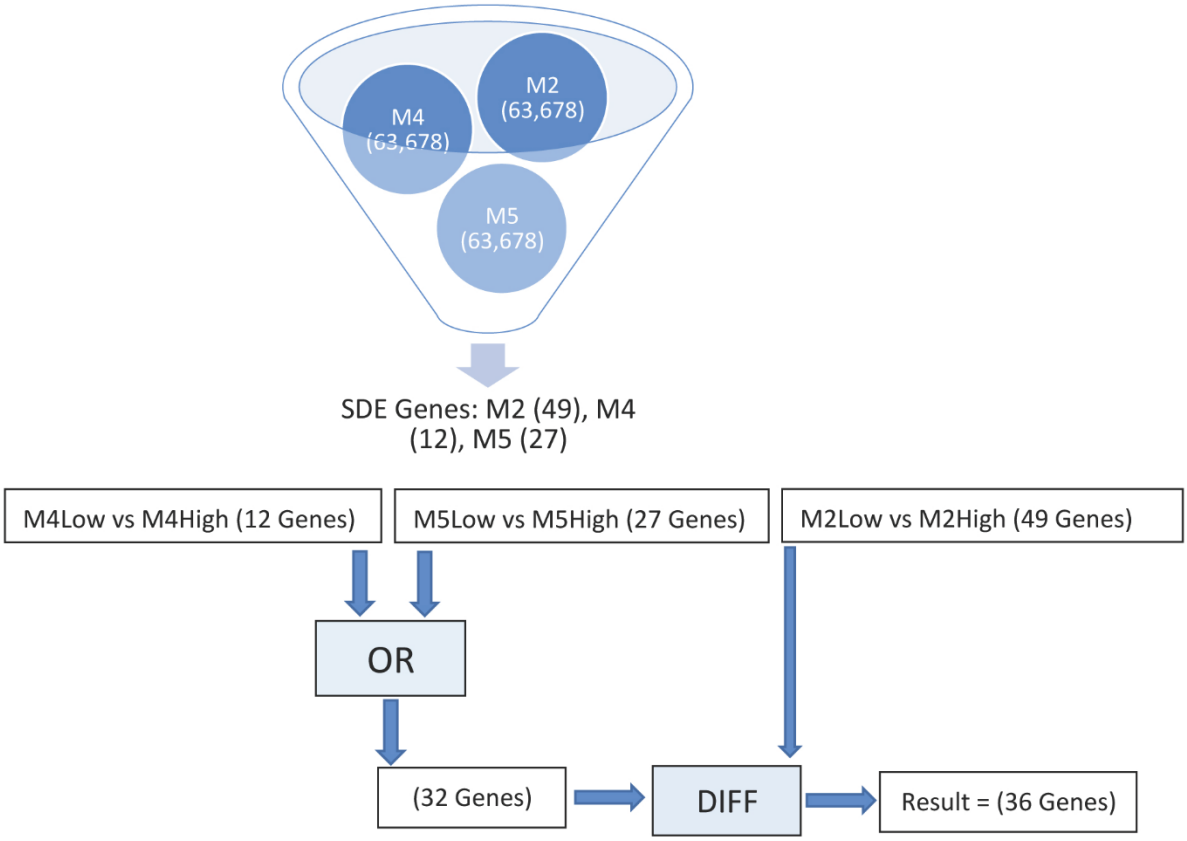

Figure 2.2: Data and process flow chart for use case 2. Input Files M2, M4, and M5 each contain 65,678 rows of differentially expressed genes. A-Lister is used to filter these files by abs(log2(foldchange)) > 1.0. The filtered files are then processed by A-Lister with the OR and DIFF set operators, resulting in 36 genes

The A-Lister command listed below will provide the resulting genes:

```
python ALister_CLI.py diff-expression GSE126785_M2.txt
GSE126785_M4.txt GSE126785_M5.txt -pc "M2Low*M2High->3.log2(FC)"
"M4Low*M4High->3.log2(FC)" "M5Low*M5High->3.log2(FC)" -dq
"M2Low*M2High-DIFF-(M5Low*M5High-OR-M4Low*M4High)" -o
E:/Data/Sample_Output/ -n "1.GeneID" -f "3.log2(FC):agt1.0" -fc
"3.log2(FC)" -v
```

There are three input files. Each file contains a single pairwise comparison that is mapped to its corresponding fold change column within the -pc argument. An output directory is specified using the optional -o argument. The ID and fold change columns are identified for each

file using the -n and -fc flags. Each file is filtered according to the fold change values, which must be greater than 1 or less than − 1. The A-Lister directional query is specified within the -dq argument.

Use case 3: analysis of differential expression using directionality patterns

The data for this use case can be downloaded from National Center for Biotechnology Information's (NCBI's) gene expression omnibus (GEO) database (15). The series number is GSE108643 (17). There are two groups of participants in the study: lean individuals and overweight/obese individuals. Muscle biopsies were collected from both groups before and after exercise. RNA-seq data was generated on the Illumina platform, TopHat was used for sequence alignment, and Cuffdiff was used for differential gene expression analysis.

The Cuffdiff files contain four conditions: LeanPre, LeanPost, OvobPre, OvobPost. Each condition is compared to every other condition resulting in six pairwise comparisons: LeanPre vs. LeanPost, LeanPre vs. OvobPre, LeanPre vs. OvobPost, LeanPost vs. OvobPre, LeanPost vs. OvobPost, and OvobPre vs. OvobPost. Suppose the user wants to examine which genes are significantly upregulated in both lean and overweight/obese individuals post exercise. The A-Lister command listed below will provide the resulting genes:

```
python ALister_CLI.py diff-expression GSE108643_Cuffdiff.txt -pc
"LeanPre->LPE,LeanPost->LPO,OvobPre->OPE,OvobPost->OPO" -dq
"LPE*LPO:UP-AND-OPE*OPO:UP" -f
"log2(fold_change):agt1.0,q_value:lt0.05,value_1:gt1.0,value_2:g
t1.0" -s1 "sample_1" -s2 "sample_2" -n "gene"
```

A diff-expression command will be executed with the *GSE108643_Cuffdiff.txt* input file. Each file specific condition label is mapped to a

globally unique label within the -pc argument. This mapping is important when dealing with multiple files that contain the same condition label names (e.g. q1, q2, q3, etc.) or, as in this example, when the user would like to shorten the name to avoid typing long group names. The -s1, −s2, and -n arguments specify the names of sample1, sample2, and ID columns. In this example, the -f argument will be used to filter the file according to absolute value of log2(fold change) greater than 1.0 (agt1.0), q-value less than 0.05 (lt0.05), values 1 and 2 greater than 1.0 (gt1.0). The A-Lister query is specified within the -dq argument, where LPE*LPO represents lean pre-exercise vs. lean post-exercise, and OPE*OPO represents overweight/obese pre exercise vs. overweight/obese post exercise. Since no output directory was specified, the result is output in the result.txt file within the current working directory.

Now, suppose the user wants to examine all possible directionality patterns for the above-mentioned query. The four possible patterns are up, up; up, down; down, up; and down, down. This could be accomplished by changing the directions within the -dq argument from UP to ALL. This would result in the following A-Lister command:

```
python ALister_CLI.py diff-expression GSE108643_Cuffdiff.txt -pc
"LeanPre->LPE,LeanPost->LPO,OvobPre->OPE,OvobPost->OP2O" -dq
"LPE*LPO:ALL-AND-OPE*OPO:ALL" -f
"log2(fold_change):agt1.0,q_value:lt0.05,value_1:gt1.0,value_2:g
t1.0" -s1 "sample_1" -s2 "sample_2" -n "gene"
```

This is an ALL query (a query containing an ALL directionality) with two ALL directions, so it is effectively translated into four queries: *LPE*LPO:UP-AND-OPE*OPO:UP, LPE*LPO:UP-AND-OPE*OPO:DOWN, LPE*LPO:DOWN-AND-OPE*OPO:UP, LPE*LPO:DOWN-AND-OPE*OPO:DOWN.* Since no output directory was specified, the results for all four queries are output in the result.txt file within the current working directory.

This example found one hundred seven genes are differentially expressed in both LPE*LPO and OPE*OPO pairwise comparisons (Figure 2.3). One hundred genes are upregulated and seven genes are downregulated in both of these pairwise comparisons. Zero genes are upregulated within one of these pairwise comparisons while also being upregulated in another one of these pairwise comparisons.

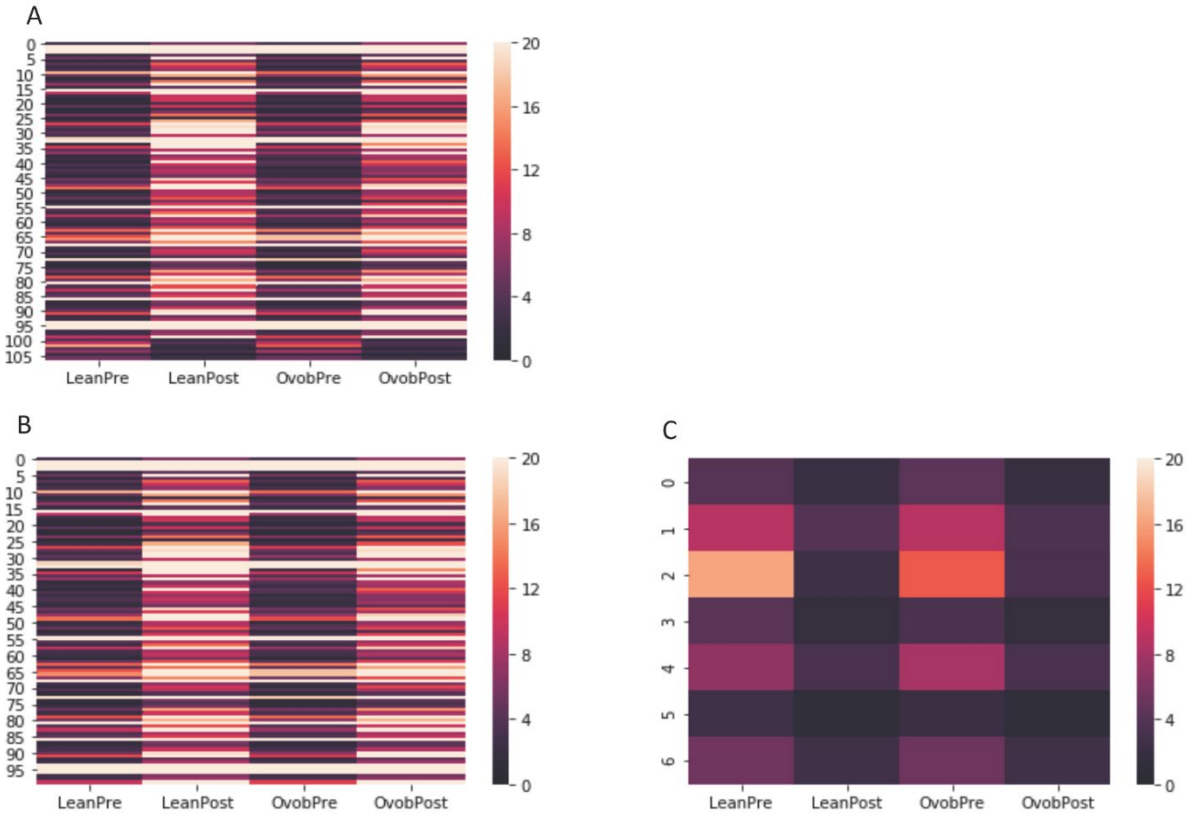

Figure 2.3: Heatmap visualization of significantly differentially expressed genes for use case 3. (**a)** All genes that are significantly differentially expressed for both LeanPre vs. LeanPost and OvobPre vs. OvobPost pairwise comparisons. (**b)** All genes that are significantly upregulated in both pairwise comparisons. (**c)** All genes that are significantly downregulated in both pairwise comparisons

Graphical user Interface (GUI)

The GUI guides the user through creating a command necessary to run A-Lister with desired settings. After selecting the mode (differential expression or name list), the parameters for that mode will be presented. The user will browse for files and preview the column headings and the first few lines of each input file in order to facilitate setting the filtering and mapping parameters. If appropriate for the mode, the GUI will also enable selecting the comparison groups, directionality, and set operators, necessary for creating the query. Once the parameters for all files are set, the user can generate and launch the command. Detailed instructions on the use of the GUI can be found on A-Lister GitHub and an example screenshot of the GUI is shown below (Figure 2.4).

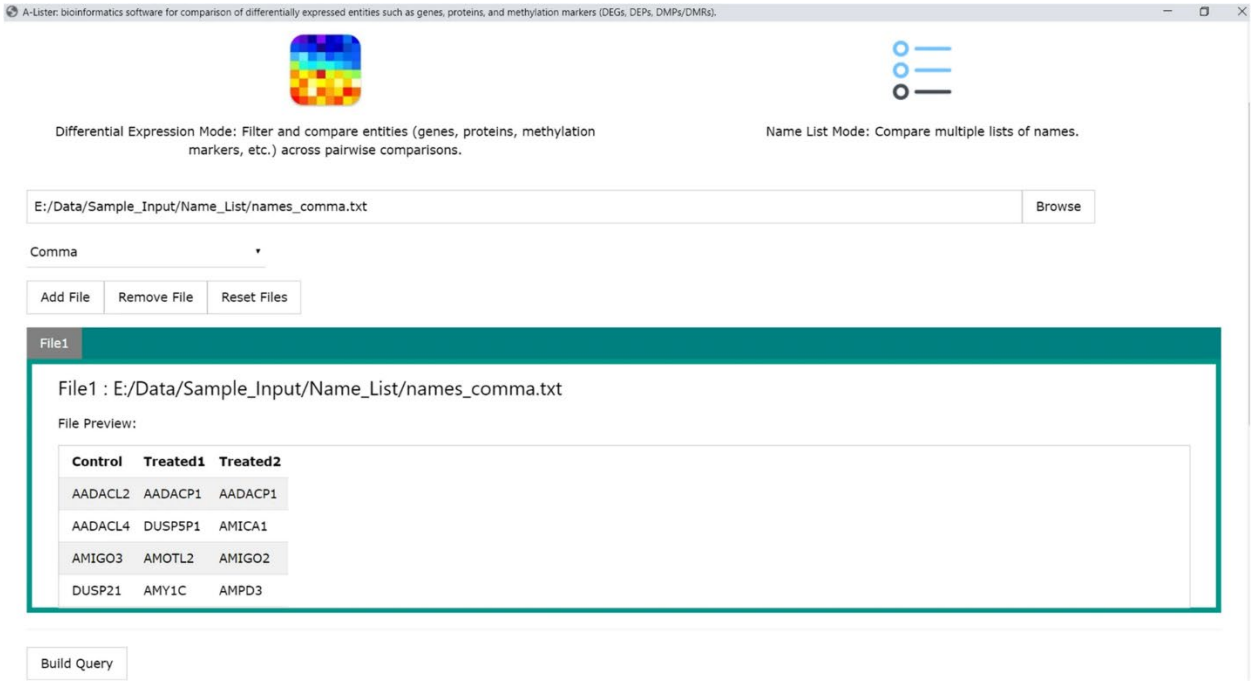

Figure 2.4: Example of screenshot of Graphical User Interface (GUI) version of A-Lister

Discussion

Although several existing bioinformatics tools have some overlapping functionality with A-Lister, none fill the same role as A-Lister. Several such tools are listed in Table 2.6 and are

described below. Intervene is a tool that can compute and visualize intersections of gene sets (or genomic regions) using multiple visualization techniques such as Venn diagrams, UpSet plots, and heatmaps (20). VennPainter and InteractiVenn are similar to Intervene (21, 22). Statistical R packages SuperExactTest and Gene-Overlap package can also be used to compute and visualize intersections of sets (23, 24). Galaxy suite text manipulation tools can be used to filter and compare tab delimited text files (19). The key limitation of these tools is that they are not built to deal with differential expression data specifically. As such the above-mentioned tools lack the means to filter individual pairwise comparisons.

Table 2.6: Comparison to Existing Software

Application	Filters	Query	Set Operators	Input Type	Interface
A-Lister	yes	yes	AND, FAND, OR, DIFF	DE data, lists	Command Line
Functional Heatmap (18)	yes	no	AND (implicit)	DE timeseries data	Web App
Galaxy (Text Manipulation) (19)	yes	no	AND, OR, DIFF	Tabular data	Web App
Intervene (20)	no	no	AND	Genomic regions, binary, counts, lists	Command Line, Web App
VennPainter (21)	no	no	AND	Lists	Graphical User Interface
InteractiVenn (22)	no	no	AND	Lists	Web App
SuperExactTest (23)	no	no	AND, OR	Lists	R
Gene-Overlap (24)	no	no	AND, OR	Lists	R

Functional Heatmap is another novel tool that seeks to make filtering and comparison of lists of differentially expressed entities less cumbersome (18). Similar to A-Lister, Functional Heatmap allows the user to filter differential expression data by columns (e.g. *p*-value) and by direction (sign and magnitude of fold change). However, Functional Heatmap is specialized for analyzing time-series data, specifically analyzing patterns of fold change direction across time. A-Lister, on other hand, can be used to analyze any pairwise comparison differential expression data across conditions, tissues, and timepoints. Moreover, unlike Functional Heatmap, A-Lister supports the notion of queries. The queries allow the user to quickly examine complex relationships between pairwise comparisons (Table 2.6).

In the future we plan to add ID validation and mapping in order to enable integration of different DEE types. Studies containing multiple -omics types are increasingly common, and we would like to be able to seamlessly compare genes, proteins, and methylation markers with each other. However, the names used as IDs for the different data types (genes, proteins, methylation markers) are generally not the same, and rather, are dependent on the data type naming convention and database. The IDs often do not map one to one, but rather one to many or one to none. Even within one data type, such as gene expression data, there are differences in naming due to annotation version, platform used, species, and other characteristics. Currently, UniProt provides a web-based tool (Retrieve/ID mapping tool) to convert IDs between different annotations (25). To maintain A-Lister's lightweight requirements (e.g., the user does not have to download datasets), and offline capabilities for secure human data processing, we did not implement linking with databases through local or outside connections to web services to check or query IDs. We do recommend that users interested in such capability should initially process the data names in their files through a service such as UniProt before executing A-Lister. In the

future, we may implement and host a web server to facilitate this functionality and to make A-Lister more accessible. However, these features will require the addition of complex name mapping functionality and web back-end to A-Lister, which we propose as a future enhancement.

Conclusions

A-Lister allows the user to quickly filter and compare any number of pairwise comparisons across multiple heterogenous differential expression files. Additionally, the A-Lister can be used to examine patterns of fold change direction and to execute complex queries across multiple pairwise comparisons. This tool may be especially useful in the context of data mining applications where dealing with many heterogenous files is common. A-Lister will help researchers to save time spent on writing, maintaining, and adjusting custom differential expression analysis scripts.

CHAPTER 3: Multiclass machine learning diagnostic for liver diseases by transcriptomics of peripheral blood mononuclear cells or liver tissue

Published in JHEP Reports, August 17th 2022:

Listopad S, Magnan C, Asghar A, Stolz A, Tayek JA, Liu Z-X, Morgan TR, and Norden-Krichmar TM. Differentiating between liver diseases by applying multiclass machine learning approaches to transcriptomics of liver tissue or blood based samples. JHEP Reports. 2022;4(10). doi: https://doi.org/10.1016/j.jhepr.2022.100560

Abstract

Background & Aims

Liver disease carries significant healthcare burden and frequently requires a combination of blood tests, imaging, and invasive liver biopsy to diagnose. Distinguishing between inflammatory liver diseases, which may have similar clinical presentations, is particularly challenging. In this study, we implemented a machine learning pipeline for the identification of diagnostic gene expression biomarkers across several alcohol-associated and non-alcohol-associated liver diseases, using either liver tissue or blood-based samples.

Methods

We collected peripheral blood mononuclear cells (PBMCs) and liver tissue samples from participants with alcohol-associated hepatitis (AH), alcohol-associated cirrhosis (AC), non-alcohol-associated fatty liver disease, chronic HCV infection, and healthy controls. We performed RNA sequencing (RNA-seq) on 137 PBMC samples and 67 liver tissue samples. Using gene expression data, we implemented a machine learning feature selection and classification pipeline to identify diagnostic biomarkers which distinguish between the liver disease groups. The liver tissue results were validated using a public independent RNA-seq

dataset. The biomarkers were computationally validated for biological relevance using pathway analysis tools.

Results

Utilizing liver tissue RNA-seq data, we distinguished between AH, AC, and healthy conditions with overall accuracies of 90% in our dataset, and 82% in the independent dataset, with 33 genes. Distinguishing 4 liver conditions and healthy controls yielded 91% overall accuracy in our liver tissue dataset with 39 genes, and 75% overall accuracy in our PBMC dataset with 75 genes.

Conclusions

Our machine learning pipeline was effective at identifying a small set of diagnostic gene biomarkers and classifying several liver diseases using RNA-seq data from liver tissue and PBMCs. The methodologies implemented and genes identified in this study may facilitate future efforts toward a liquid biopsy diagnostic for liver diseases.

Graphical Abstract:

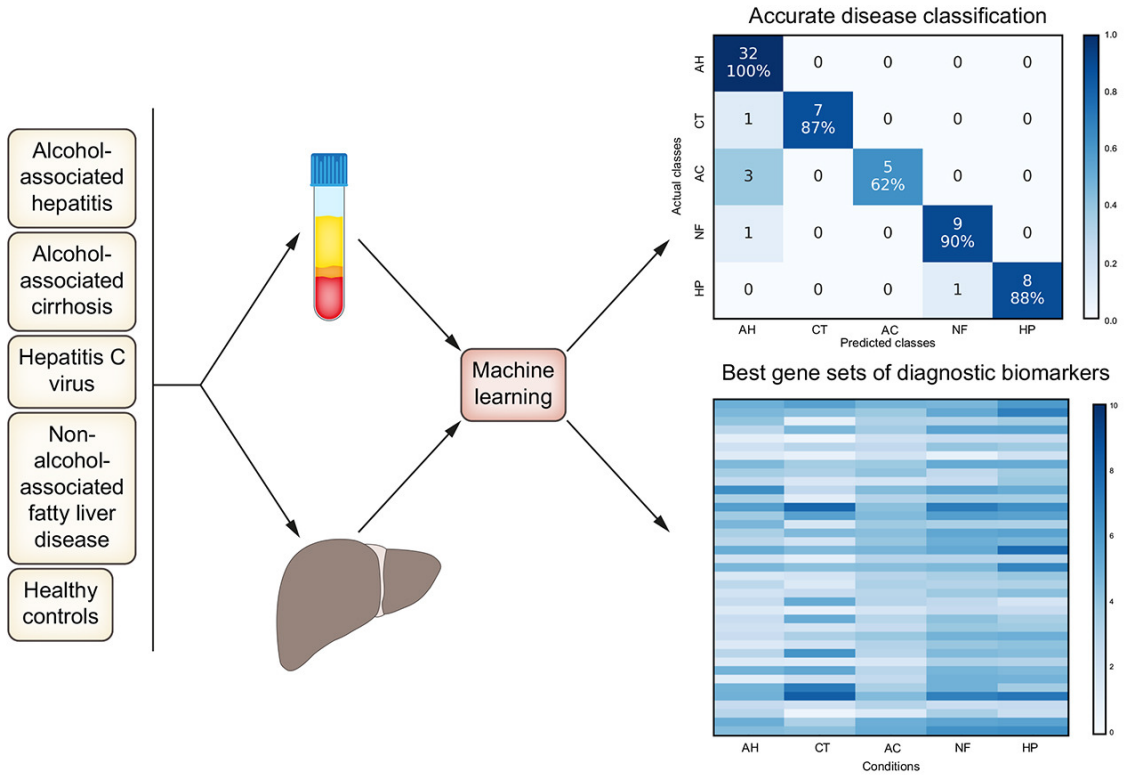

Figure 3.1: Graphical Abstract

Introduction

Liver disease is responsible for 2 million deaths worldwide annually, ranking as one of the leading causes of death in the world (26). Alcohol-associated hepatitis (AH) is one of the deadliest liver diseases (27). Other liver disorders such as alcohol-associated cirrhosis (AC), chronic HCV infection, and non-alcohol-associated fatty liver disease (NAFLD) are less deadly but are more widespread. Distinguishing between various alcohol-associated and non-alcohol-associated liver diseases typically requires multiple lab tests that often culminate in liver biopsy (28). The diagnosis is further complicated because factors that promote liver disease, such as viral hepatitis, obesity, and alcohol misuse, may overlap. Distinguishing AH and AC may be especially difficult and is thus an area of unmet clinical need. Presently, liver biopsy is regarded

as the gold standard for confirming liver disease diagnosis and staging fibrosis severity. This approach has several limitations, such as procedural risk of internal bleeding, high cost, and patient dissatisfaction. While various clinical parameters, blood panels, and imaging tests have been used to supplement liver biopsy, they are not sufficiently effective to fully replace liver biopsy (29). Development of a liquid biopsy that is as accurate as liver biopsy for diagnosis of liver disease would improve quality of patient care and reduce healthcare costs. This process relies on identifying effective blood-based diagnostic biomarkers.

Development of liquid biopsies using blood-based biomarkers holds great promise when used with genomic data. For example, one recent study on epigenetic universal cancer biomarkers utilized DNA methylation markers (30). While the field is expanding, many of the clinically used blood-based biomarkers are cancer-specific (31). There is a shortage of effective diagnostic blood-based biomarkers for liver diseases. Presently many of the established biomarkers for liver disease are proteins found in blood serum such as albumin (32). Circulating microRNAs such as miR-122 and miR-155 have also been identified as diagnostic biomarkers for a range of liver diseases (32). Several previous studies have established that gene expression profiling of peripheral blood mononuclear cells (PBMCs) can be used to characterize HBV, HCV, and primary biliary cholangitis (33,34,35,36). Serum markers have been used to distinguish between alcohol-associated and non-alcohol-associated liver diseases using several machine learning (ML) models (37). Liver tissue gene expression in combination with clinical parameters has been used to establish prognosis in patients with AH and HCV-related early-stage cirrhosis (38, 39).

In this study, we chose to analyze gene expression in PBMCs for a variety of reasons. PBMCs can be extracted from a blood sample, pelleted and flash frozen, and provide ample

material for RNA sequencing (RNA-seq). The differences in gene expression of PBMCs have been shown to reflect disease state. Additionally, we also characterized gene expression of liver tissue. The liver tissue served as a benchmark against which PBMCs could be compared, since pathology of liver tissue is currently the standard for distinguishing between liver diseases.

We were primarily interested in distinguishing between AH and AC, which may have similar clinical presentations. To establish the robustness of our models in discriminating between inflammatory liver diseases, we further sought to distinguish alcohol-associated liver diseases from non-alcohol-associated liver diseases, such as NAFLD and HCV. Therefore, we have trained ML models to differentiate between these liver diseases and healthy controls. As part of the classification process, we have also identified effective diagnostic gene biomarkers.

Like most individual biomedical research studies, ours was limited to a small number of participant samples due to the high costs of recruitment, sequencing, data storage, and data analysis. The gene expression data is also inherently highly dimensional. Datasets that contain more features than samples are difficult to classify. Therefore, it was crucial in our study to use statistical and ML techniques tailored for handling small sample and large feature sizes. In addition to identifying useful PBMC-based diagnostic biomarkers of liver diseases, our secondary goal was to evaluate multiple bioinformatic pipelines in the context of analyzing small sample size RNA-seq data. Special focus was given to feature selection, wherein, we compared several different feature selection approaches. Overall, our ML pipeline demonstrated excellent classification performance across the liver diseases using both liver tissue and PBMCs.

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, NAFLD, HCV, and healthy controls were obtained from the liver tissue cell distribution system (LTCDS) at University of Minnesota. Participant demographics are outlined in Table 3.1, Table 3.2. We summarized the age, MELD (model for end-stage liver disease) score, Maddrey's discriminant function, BMI, sex, and ethnicity of our study population. As expected, the NAFLD group had the highest mean BMI, while the AH group had the highest mean MELD and Maddrey's discriminant function scores.

Table 3.1: Study population demographics (PBMCs)

	PBMC samples				
	AH	CT	AC	NF	HP
	(n = 38)	(n = 20)	(n = 40)	(n = 20)	(n = 19)
Age, mean ± SD	47.3 ± 11.5	35.9 ± 15.6	54.5 ± 9.7	52.2 ± 14.9	58.9 ± 7.4
MELD, mean ± SD	25 ± 3.8	7.3 ± 2.6	13.4 ± 5.8	8.9 ± 4	8.9 ± 2.8
Maddrey's DF, mean ± SD	52.6 ± 20.7	2.4 ± 8.1	21.1 ± 19.1	7.7 ± 14.1	6.7 ± 7.1
BMI, mean ± SD	30 ± 6.2	27 ± 3.5	30.4 ± 5.1	36.5 ± 6	29.6 ± 5.9
Sex, n (%)					
Female	1 (2.6%)	8 (40.0%)	0 (0.0%)	4 (20.0%)	8 (42.1%)
Male	37 (97.4%)	12 (60.0%)	40(100.0%)	16 (80.0%)	11 (57.9%)
Ethnicity, n (%)					
Hispanic	25 (65.8%)	8 (40.0%)	25 (62.5%)	9 (45.0%)	10 (52.6%)
NHW	10 (26.3%)	0 (0.0%)	13 (32.5%)	7 (35.0%)	4 (21.1%)
Black	2 (5.3%)	2 (10.0%)	1 (2.5%)	2 (10.0%)	5 (26.3%)
Other	1 (2.6%)	10 (50.0%)	1 (2.5%)	2 (10.0%)	0 (0.0%)
Source	SCAHC	SCAHC	SCAHC	SCAHC	SCAHC

AC, alcohol-associated cirrhosis; AH, alcohol-associated hepatitis; CT, healthy controls; DF, discriminant function; HP, HCV infection; MELD, model for end-stage liver disease; NF, non-alcoholic fatty liver disease; NHW, non-Hispanic White; SCAHC, Southern California Alcoholic Hepatitis Consortium.

Table 3.2: Study population demographics (Liver)

	Liver tissue samples				
	AH	CT	AC	NF	HP
	(n = 32)	(n = 8)	(n = 8)	(n = 10)	(n = 9)
Age, mean ± SD	43.3 ± 11.3	55.4 ± 4.3*	54.2 ± 6.9*	56.8 ± 11.6	56.8 ± 7.6
MELD, mean ± SD	25.1 ± 5.7	NA	NA	28 ± 5.9*	27.2 ± 7.5*
Maddrey's DF, mean ± SD	52.3 ± 22.1	NA	NA	NA	NA
BMI, mean ± SD	29.4 ± 5.9	NA	NA	NA	NA
Sex, n (%)					
Female	3 (9.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Male	29 (90.6%)	7 (87.5%)	5 (62.5%)	10 (100.0%)	9 (100.0%)
Ethnicity, n (%)					
Hispanic	25 (78.1%)	NA	0 (0.0%)	0 (0.0%)	1 (11.1%)
NHW	5 (15.6%)	NA	4 (50.0%)	7 (70.0%)	5 (55.5%)
Black	1 (3.1%)	NA	0 (0.0%)	1 (10.0%)	2 (22.2%)
Other	1 (3.1%)	NA	0 (0.0%)	0 (0.0%)	0 (0.0%)
Source	SCAHC	LTCDS	LTCDS	LTCDS	LTCDS

The ethnicity and sex percentages may not add up to 100% due to missing data.
AC, alcohol-associated cirrhosis; AH, alcohol-associated hepatitis; CT, healthy controls; DF, discriminant function; HP, HCV infection; LTCDS, liver tissue cell distribution system; MELD, model for end-stage liver disease; NF, non-alcoholic fatty liver disease; NHW, non-Hispanic White; SCAHC, Southern California Alcoholic Hepatitis Consortium.
∗ Missing age for 3 AC participants, MELD for 2 NF participants, and MELD for 4 HP participants.

The biospecimens consisted of 137 PBMC samples and 67 liver tissue (LV) samples. The liver diseases represented were encoded with 2 letter symbols (as presented in the tables and

figures) as follows: alcohol-associated hepatitis (AH), alcohol-associated cirrhosis (AC), NAFLD (NF), chronic HCV (HP), and healthy controls (CT). All PBMC and liver tissue samples were collected from distinct participants except for 19 participants with AH that provided both sample types. Most of the AC participants within the SCAHC study were expected to be in-patients with decompensated cirrhosis. The inclusion and exclusion criteria can be found in the supplementary materials. Best efforts were made during recruitment of the AH and non-AH groups within the SCAHC study to match based on age, sex, and ethnicity. Severity-based matching was not possible due to small sample size.

Sample Collection

The blood samples and liver biopsies from participants with AH were collected before starting treatment. Blood samples from all other groups were collected at entry into the study. PBMCs were freshly isolated from the blood samples by Ficoll-Histopaque (GE Healthcare) gradient centrifugation, flash frozen, and then stored in a liquid nitrogen tank. The AH biopsy sample was placed in a cryovial containing *RNAlater* (Invitrogen) and flash frozen in liquid nitrogen. The liver tissue samples for healthy controls, AC, NALFD, and HCV conditions were obtained from University of Minnesota LTCDS.

Sample Data Processing

RNA sequencing and alignment

Several samples were removed prior to use in our study, due to poor read quality (40). The trimmed, filtered, and decontaminated reads were aligned to the hg38 (GRCh38 assembly) human reference genome using STAR 2.6.0 (41) with default settings (STARCQ), and annotated with Ensembl release 91 (Dec 2017).

Partitioning samples into 4 data sets

We divided our data into 4 datasets, which we refer to as follows: LV 2-Way, LV 3-Way, LV 5-Way, and PBMC 5-Way. LV 2-Way included liver tissue samples from participants with AH (n = 32) and healthy (n = 8) conditions. The LV 3-Way included liver tissue from participants with AH (n = 32), healthy (n = 8), and AC (n = 8) conditions. The LV 5-Way included liver tissue from participants with AH (n = 32), healthy (n = 8), AC (n = 8), NAFLD (n = 10), and HCV (n = 9) conditions. The PBMC 5-Way included PBMC samples from participants with AH (n = 38), healthy (n = 20), AC (n = 40), NAFLD (n = 20), and HCV (n = 19) conditions.

Validation dataset

We validated our liver tissue ML models using the GSE142530 dataset (17). This dataset contained liver tissue RNA-seq data from participants with AH (n = 10), healthy (n = 12), and AC (n = 6) conditions. We utilized the counts data that had been generated with DESeq2 and deposited in GEO (42). Publicly available RNA-seq gene expression data from PBMCs was not available for the conditions in our study, and therefore, only the liver tissue datasets were validated using independent data.

Analysis of gene expression data

For each sample and workflow within our data, standard fragments per kilobase of exon model per million reads mapped (FPKM) values were directly extracted from the corresponding alignment results (BAM files) using the Cuffquant utility of the Cufflinks suite (release 2.2.1) (43). The FPKM counts were then further normalized using Cuffdiff geometric normalization. The RNA-seq counts were transformed using ln(1+count) formula. This transformation greatly reduced count variance and improved classification accuracies (Figure S1 and Figure S2). The validation dataset counts generated by DESeq2 were presumably normalized using DESeq2's

default median of ratios method, which is equivalent to Cuffdiff's geometric normalization. These counts were also transformed using ln(1+count) formula.

Classification and Feature Selection Architecture

Overview of classification and feature selection pipeline

The classification and feature selection pipeline process flow is visualized in Figure 3.2. Feature selection was performed on each training set using differential expression (DE) and information gain (IG) methods. The DE and IG feature selection methods are referred to as filter feature selection methods (44). DE feature selection was performed using Cuffdiff, while the IG feature selection was implemented using scikit-learn (version 0.23.2+) package's implementation of IG algorithm (45).

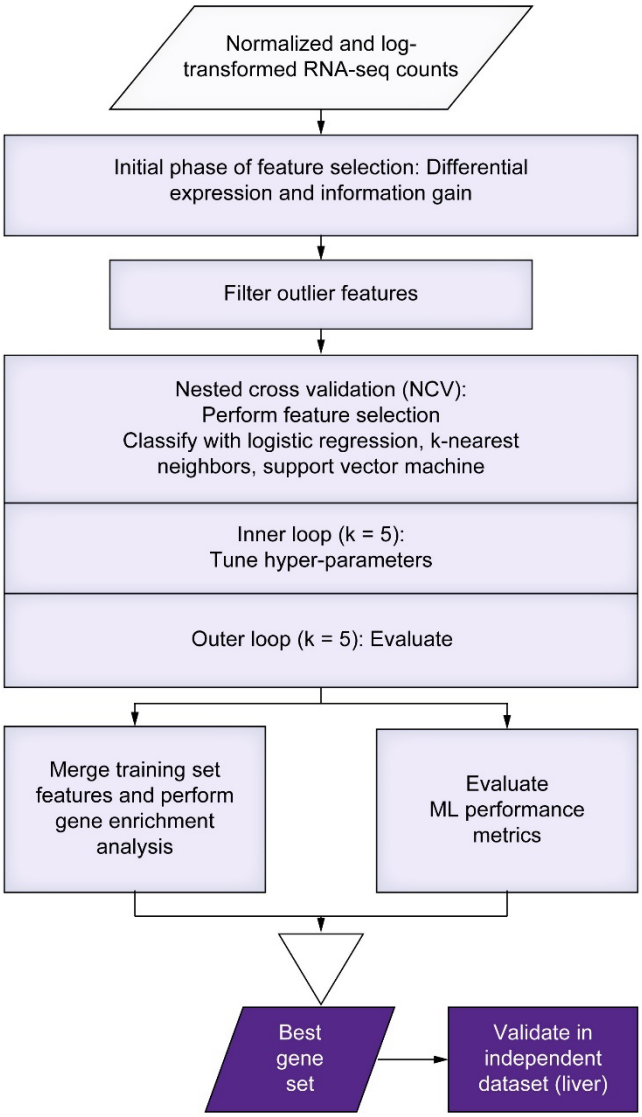

Figure 3.2: Diagram outlining the flow of processes in the machine learning feature selection and classification pipeline

Regardless of the feature selection method used, once the features were selected, the classification process was similar. The classifiers were evaluated using k-fold nested cross-validation (k outer and inner = 5). The feature selection was performed inside of inner and outer loops of nested cross-validation. The classification performance was primarily evaluated using confusion matrices, overall, and per-class accuracies. The features selected in the outer loop of nested cross-validation were merged together to form the candidate gene set, if they appeared in at least 4 out of 5 training sets. The resulting candidate gene sets were then evaluated using gene

enrichment analysis. A combination of feature size, overall accuracy, per-class accuracies, and gene enrichment analysis were then used to pick a best gene set for each dataset. In the case of liver tissue datasets, the best gene sets were then further evaluated in an independent validation dataset. We used Python 3.7+ for all ML analysis, and all of the classifiers were implemented in scikit-learn package. The power size calculation was performed in R.

ML Classifiers

The ML analysis for all 4 of our datasets was performed and was reported in this study using logistic regression (LR), k-nearest neighbors (kNN), and support vector machine (SVM) classifiers. The corresponding hyper-parameters used during grid search can be found in the codebase.

In Silico Biological Validation and Best Gene Selection

The genes selected during feature selection were computationally evaluated for biological relevance using gene enrichment analysis via Enrichr with pathway, tissue, and disease Enrichr libraries (46). The resulting hits were filtered using an adjusted *p* value cut-off of 0.05 and regular expression matching. The terms used for pathways regular expression matching included names of various immune system pathways. The terms used for tissue regular expression matching included names of various cell types that comprise blood and liver tissues. The terms used for disease regular expression matching included the conditions within this study (AH, AC, NAFLD, HCV) along with a few other liver and blood disorders.

To compare the *in silico* biological relevance of many different gene sets, we devised a simple tallying system to count the number of hits within pathway, tissue, and disease libraries that passed the adjusted *p* value cut-off and regular expression matching. For each of the 4 datasets, we identified a gene set (Figure 3.3) that exhibited both high classification accuracy and

highly relevant *in silico* biological validation results using Enrichr. We have also provided the fold changes of the best genes for Liver 5-Way and PBMC 5-Way datasets (Tables S6 and S7).

Liver 2-Way	*AKR1B10, FITM1, KRT23, MMP7, MT1M, PLA2G2A, PPP1R1A, SCTR, TREM2.*
Liver 3-Way	*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.*
Liver 5-Way	*AC025259.3, AKR1B10, ATF3, CYP2A7, CYP2B6, DOCK7, DUSP1, EPS8L1, GADD45B, GADD45G, GSTA2, HBA2, IFI6, IFI27, IFI44L, IFITM1, IGFBP1, IGHV3-23, ISG15, KRT7, KRT23, LINC01554, MMP7, MT1G, MT1M, MUC1, MUC6, NR4A1, OASL, PLA2G2A, PPP1R1A, RGS1, S100A8, SAA2-SAA4, SCTR, SERHL2, SLC2A3, SPINK1, SYT8.*
PBMC 5-Way	*AHSP, ALAS2, ALPL, ANXA3, AQP9, ATF7IP2, AZU1, BCAT1, C1QA, C1QB, CAMP, CCR2, CD180, CEACAM3, CEACAM8, CHI3L1, CRISP3, CTSG, CXCL5, CXCR1, DEFA3, DEFA4, DSC2, DYSF, ELANE, FCGR3B, FFAR2, FLVCR2, FPR2, GTF2IRD2B, HBD, HBM, HBQ1, HP, IFITM3, IGHG3, IGHG4, IGKV1-12, IGKV1-39, IGKV1D-13, IGLC3, IGLV3-10, KCNJ15, LCN2, LTF, MME, MMP8, MPO, MPZL2, NLRC4, NRP1, ORM1, OSBPL10, PGLYRP1, PLA2G4C, PRRG4, PTK7, RAB10, RETN, RNASE2, RNASE3, S100B, S100P, SC5D, SIGLEC6, SLC25A37, SLPI, TCF7L2, TLR8, TMEM144, TMEM150B, TMEM170B, TNFSF10, VSIG4, ZNF683.*

Figure 3.3: Best gene sets for Liver 2-Way, Liver 3-Way, Liver 5-Way, and PBMC 5-Way datasets

Additionally, we evaluated the best gene sets for Liver 5-Way and PBMC 5-Way datasets using Ingenuity Pathway Analysis (IPA), gene set-enrichment analysis (GSEAPreranked), and blood transcription module (BTM) analysis (BloodGen3Module) tools (47, 48, 49). Blood transcription module analysis was performed with the PBMC 5-Way dataset only, since this method is specific to blood-based samples. Notably, this technique was recently utilized to analyze RNA-seq data from PBMCs to predict response to corticosteroid therapy in patients with AH (50). Since these tools utilize different knowledgebases and statistical methods, they provided complementary pathway annotations. The methods and results for these tools are provided in the supplementary information.

Independent Validation Dataset

After the best gene set was selected for each of our 3 liver tissue datasets, the independent validation dataset was utilized as follows. The ML classifier that performed best with the selected gene set was trained on the entirety of the corresponding liver dataset (*i.e.*, LV 2-Way, LV 3-Way, or LV 5-Way), using only the best genes selected for that dataset. The hyper-parameters for this classifier were selected by performing a regular cross-validation over the

entirety of the corresponding liver dataset. The trained model was then tested in the independent dataset. While the PBMC 5-Way model could not be tested in an independent dataset set due to lack of appropriate public data, the methods prior to the independent dataset evaluation were the same for both liver and PBMC tissues. Therefore, we are confident that the PBMC genes identified in this study will have reasonable generalization. Additionally, the PBMC dataset had twice as many samples available for training and testing as the liver dataset, thereby also strengthening confidence in the best PBMC gene set. For additional details regarding methods, please refer to the supplementary methods.

Results

Classification of LV 2-Way (AH vs. Healthy)

We developed many of our approaches described in the Methods section while first analyzing the binary dataset of AH *vs.* healthy samples. The task of distinguishing between AH and healthy samples proved simple, with accuracy as high as 100% depending on feature size, classifier, and feature selection methods. Based on their classification performance and runtime in the LV 2-Way dataset we chose to use LR, kNN, and SVM classifiers for the remaining datasets. The gene sets produced via various feature selection and outlier filtering strategies were also computationally evaluated for biological relevancy using Enrichr (Table S18). We selected the best gene set for our LV 2-Way dataset and then validated it in the independent test dataset. Using the best gene set of only 9 genes, we attained 97% classification accuracy within the LV 2-Way dataset, and 95% accuracy in the validation dataset, as visualized using confusion matrices (Figure 3.4). Heatmaps of the RNA-seq counts per condition as an average and for each replicate show that the 2 conditions are very distinct from each other in both our LV 2-Way

dataset and the independent dataset (Figure 3.3). The best gene set for each of the 4 datasets is shown in Figure 3.3.

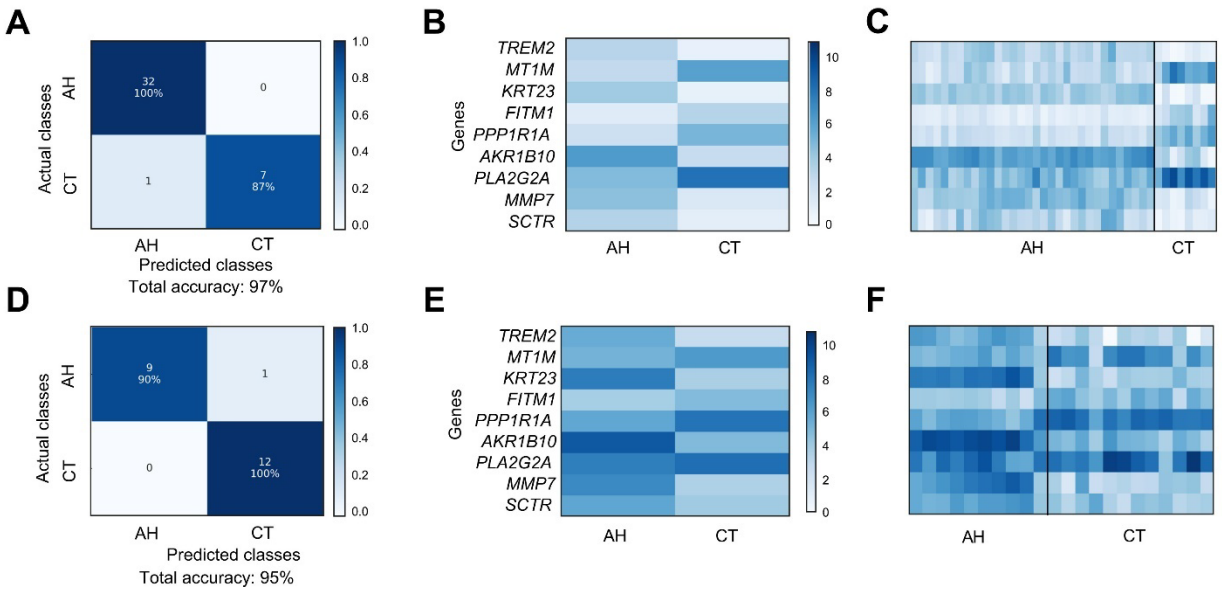

Figure 3.4: Confusion matrices and RNA-seq count heatmaps corresponding to the best gene set of LV 2-Way dataset. (**A**) Confusion matrix for classification of LV 2-Way dataset using best gene set. The diagonal contains the number and percentage of the correctly predicted samples. (**B**) Heatmap of best LV 2-Way gene set averaged per condition. (**C**) Per replicate heatmap of best LV 2-Way gene set. (**D**) Confusion matrix for classification of AH and CT samples within validation dataset. (**E**) Heatmap of best gene set within validation dataset averaged per condition. (**F**) Per replicate heatmap of best gene set within validation dataset. AH, alcohol-associated hepatitis; CT, healthy controls; LV, liver tissue; RNA-seq, RNA sequencing

Classification of LV 3-Way (AH vs. Healthy vs. AC)

Having successfully distinguished between AH and healthy samples with high accuracy, we proceeded to the more difficult multiclass classification task of discriminating between multiple liver diseases at once. Our classifiers peaked around 90% overall accuracy within our LV 3-Way dataset (Table S19). We identified the best gene set by examining the accuracies and *in silico* biological validation scores of each gene set produced by various feature selection configurations (Table S19 and S20). The top Enrichr hits for the LV 3-Way dataset are shown in Table S21. Using the best gene set comprised of 33 genes, we attained 90% overall accuracy

in the LV 3-Way dataset (via nested cross-validation) and 82% overall accuracy in the independent validation dataset. The confusion matrices and the heatmaps of RNA-seq counts corresponding to the best gene set within LV 3-Way and the independent validation datasets are displayed in Figure 3.5.

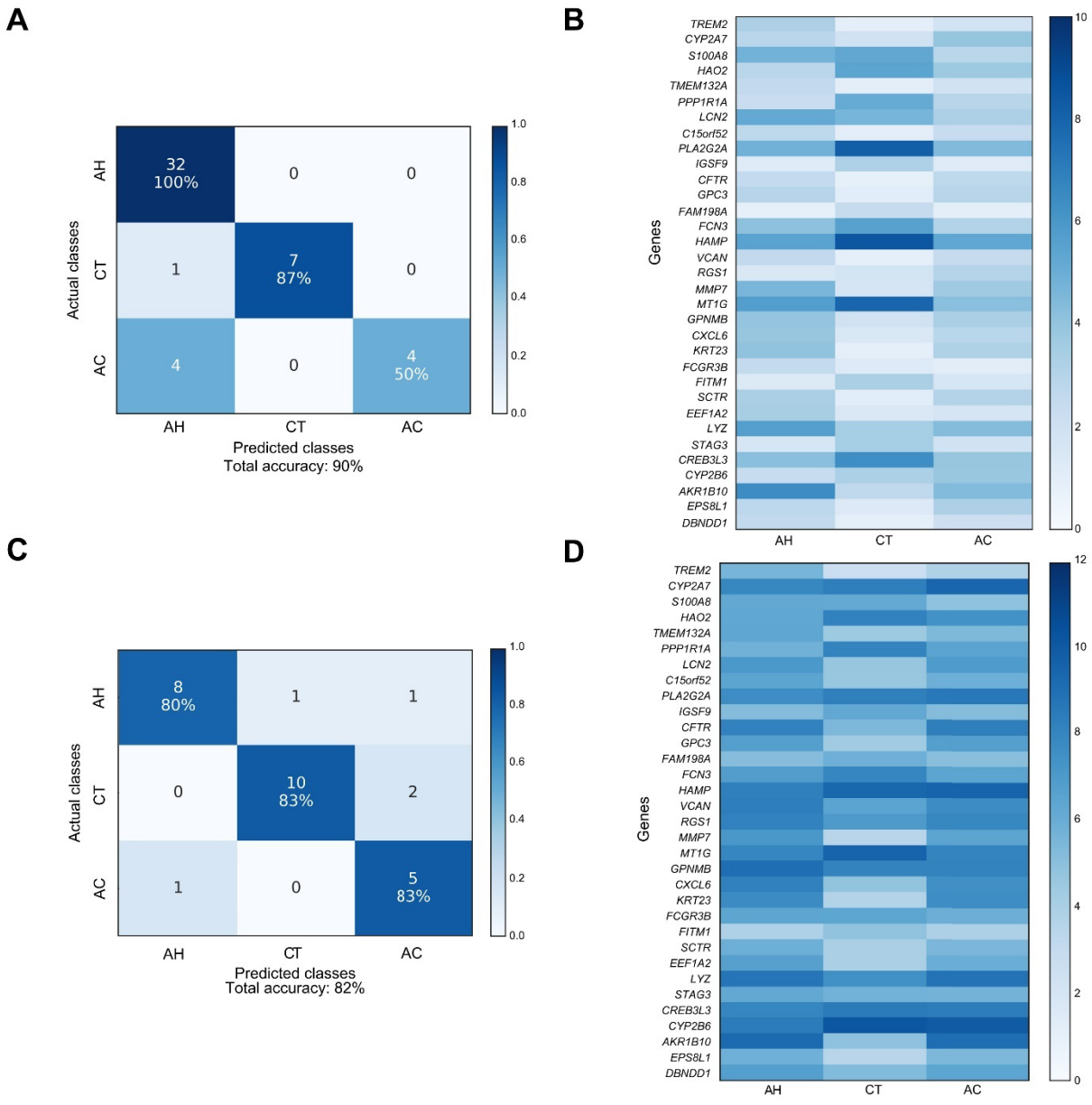

Figure 3.5: Confusion matrices and RNA-seq count heatmap corresponding to the best gene set of LV 3-Way dataset. (**A**) Confusion matrix for classification of LV 3-Way dataset using best gene set identified by filter feature selection. (**B**) RNA-seq count heatmap of best LV 3-Way gene set averaged per condition. (**C**) Confusion matrix for classification of AH, AC, and CT samples within independent validation dataset. (**D**) RNA-seq count heatmap of best gene set within independent validation dataset (AH, AC, and CT) averaged per condition. AC, alcohol-associated cirrhosis; AH, alcohol-associated hepatitis; CT, healthy controls; LV, liver tissue; RNA-seq, RNA sequencing

Classification of LV 5-Way (AH vs. Healthy vs. AC vs. NAFLD vs. HCV)

The LV 5-Way dataset was the most complex liver tissue dataset in the study. While AH and healthy groups were generally classified with high accuracy, the remaining conditions proved to be more challenging to appropriately classify (Figure 3.6). The classifiers peaked at around 90% overall accuracy within the LV 5-Way dataset (Table S22). We identified the best gene set using a combination of classification performance and *in silico* biological validation metrics (Tables S22 and S23). For the annotations of the best gene set for LV 5-Way, the top hits using Enrichr are shown in Table S24, IPA in Table S28, and GSEA in Table S30. Using the best gene set comprised of 39 genes, we attained 91% overall accuracy within the LV 5-Way dataset (via nested cross-validation) and 64% overall accuracy in the validation dataset. While the overall classification accuracy in the independent dataset was lower than in the LV 3-Way testing, this was expected since the LV 5-Way gene set was based on 2 additional liver diseases (NAFLD and HCV), which were not present in the independent dataset. Notably, there were no samples from the independent dataset that were misclassified as NAFLD or HCV. The confusion matrix and the heatmap of RNA-seq counts corresponding to the best gene set within LV 5-Way and the independent validation datasets are shown in Figure 3.6.

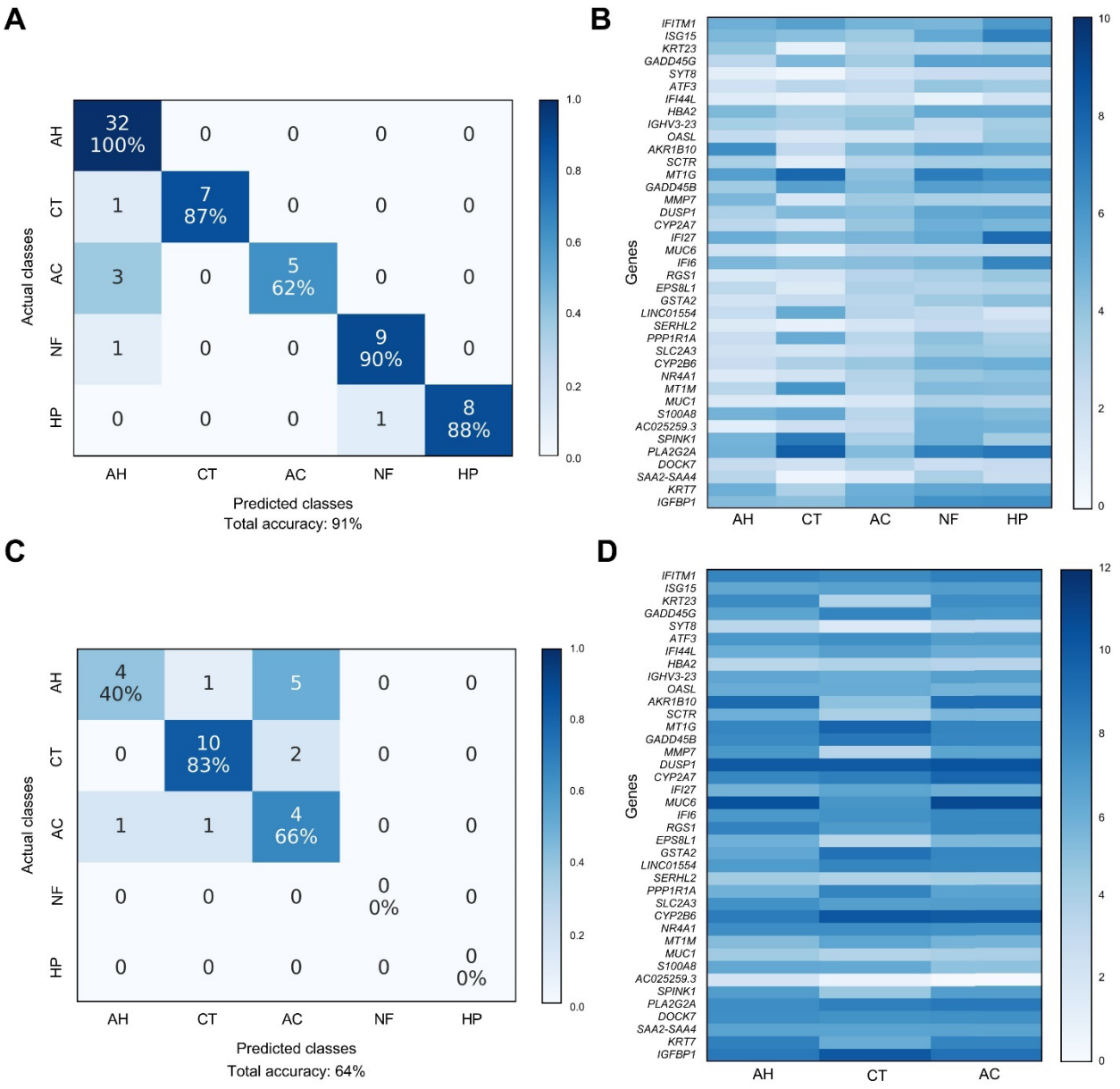

Figure 3.6: Confusion matrices and RNA-seq count heatmaps corresponding to the best gene set of LV 5-Way dataset. (**A**) Confusion matrix for classification of LV 5-Way dataset using best gene set identified by filter feature selection. (**B**) RNA-seq count heatmap of best LV 5-Way gene set averaged per condition. (**C**) Confusion matrix for classification of AH, AC, and CT samples within independent validation dataset. (**D**) RNA-seq count heatmap of best gene set within independent validation dataset (AH, AC, and CT) averaged per condition. AC, alcohol-associated cirrhosis; AH, alcohol-associated hepatitis; CT, healthy controls; HP, chronic HCV infection; LV, liver tissue; NF, non-alcohol-associated fatty liver disease; RNA-seq, RNA sequencing

Classification of PBMC 5-Way (AH vs. Healthy vs. AC vs. NAFLD vs. HCV)

Having achieved high classification accuracies in liver datasets, we broadened the scope of our study by applying these same ML models and strategies to our PBMC dataset. The classifiers tested peaked at 75% overall accuracy (Table S25). We identified the best gene set using a combination of classification performance and *in silico* biological validation metrics (Tables S25 and S26). For the annotations of the best gene set for PBMC 5-Way, the top hits using Enrichr are shown in Table S27, IPA in Table S29, GSEA in Table S31, and BloodGen3Module in Table S32. Using the best gene set comprised of 75 genes, we attained 75% overall accuracy in PBMC 5-Way dataset (via nested cross-validation). Because we could not obtain public RNA-seq data from PBMCs for several of our liver diseases, we could not validate the PBMC genes and classification performance in an independent data set. However, since the methods used to identify the best gene set were identical for both liver and PBMC datasets, we are confident of our results. The confusion matrix and the heatmap of RNA-seq counts corresponding to this gene set are shown in Figure 3.7.

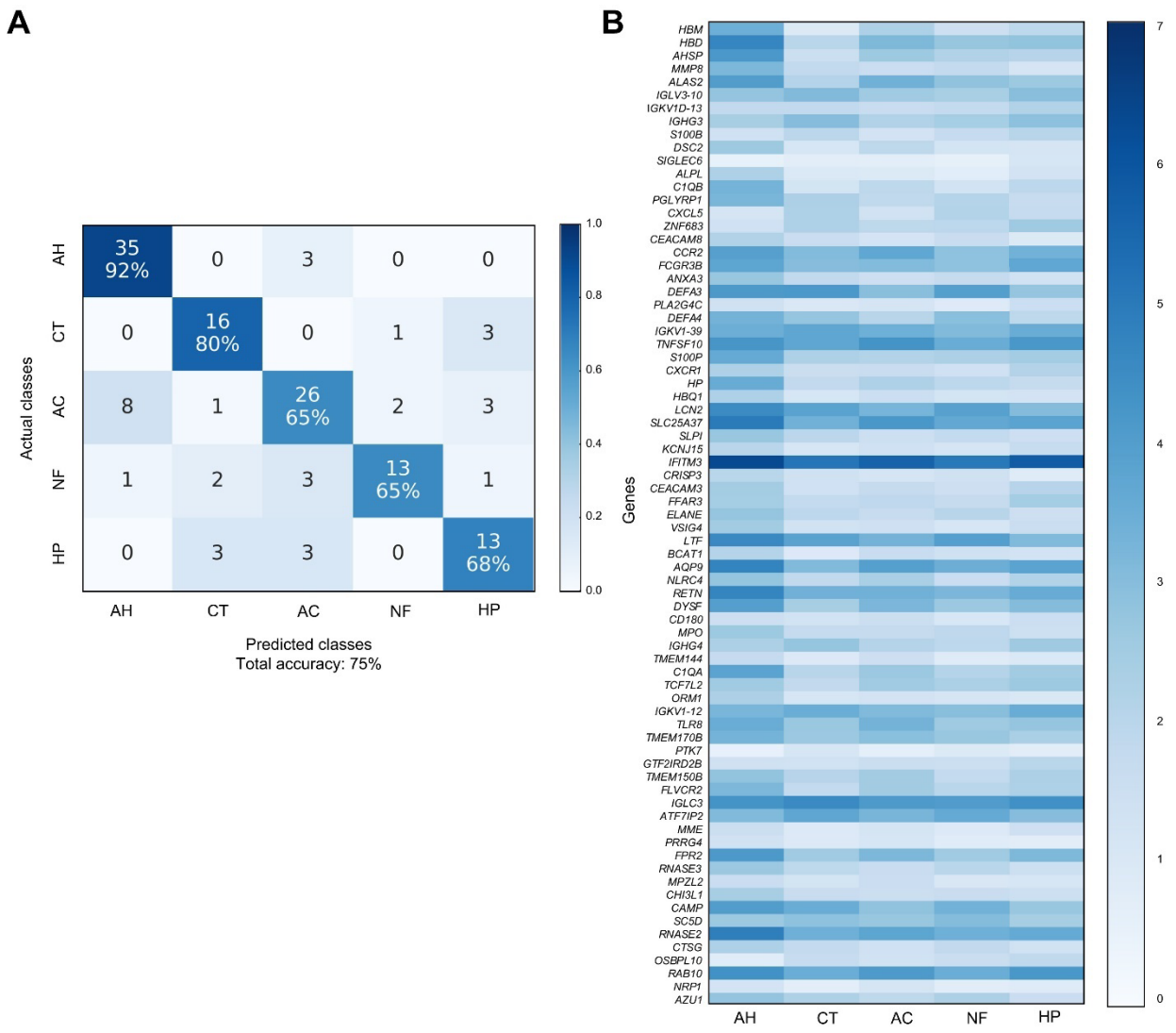

Figure 3.7: Confusion matrices and RNA-seq count heatmaps corresponding to the best gene set of PBMC 5-Way dataset. (A) Confusion matrix for classification of PBMC 5-Way dataset using best gene set identified by filter feature selection. (B) RNA-seq count heatmap of best PBMC 5-Way gene set averaged per condition. AC, alcohol-associated cirrhosis; AH, alcohol-associated hepatitis; CT, healthy controls; HP, chronic HCV infection; NF, non-alcohol-associated fatty liver disease; PBMC, peripheral blood mononuclear cells; RNA-seq, RNA sequencing

Discussion

To the best of our knowledge, this is the first study to utilize ML approaches with liver tissue and PBMC gene expression data to distinguish among several alcohol-associated and non-alcohol-associated liver diseases simultaneously with overall classification accuracies above

75%. Optimization of gene feature selection played a key role in attaining high accuracies. We have also identified gene signatures that were enriched for various inflammation and metabolism pathways, which thus show promise as diagnostic biomarkers for the liver diseases included in the study.

We found that the use of feature selection was one of the most crucial components of successful classification. The feature space of a typical RNA-seq experiment consists of thousands of genes. While exploring every possible subset of genes is computationally infeasible, we found that it was crucial to experiment with at least a small number of intelligently selected gene subsets. The filter feature selection proved to be the most effective and runtime efficient approach. While DE and IG filters attained similar classification accuracies, the DE filter resulted in more biologically relevant gene sets. The choice of ML classifier had minor impact on classification accuracy with LR, kNN, and SVM classifiers proving to be most effective for our datasets.

The outlier feature removal proved useful toward establishing adequate *in silico* biological relevance. Small sample size RNA-seq datasets are typically noisy and highly impacted by batch effects. RNA-seq data also often contains many aberrantly expressed non-coding genes. The removal of these genes resulted in gene signatures with more biologically relevant terms. In addition to using Enrichr for *in silico* biological validation, we also performed pathway analysis of best gene signatures for the 5-Way datasets using IPA, GSEA, and BTM analysis software, which highlighted relevant pathways in these gene sets on pairwise comparison basis (Tables S28–S32).

Using the best gene signature identified in the PBMC 5-Way dataset (AH, Healthy, AC, NAFLD, HCV), we examined significantly enriched pathways with IPA for each pairwise

comparison. The significantly enriched pathways mainly fell into 2 categories: iron homeostasis and immune system processes. Iron homeostasis pathways included heme biosynthesis, tetrapyrrole biosynthesis, and erythropoietin signaling. Iron homeostasis is one of the principal liver functions, while most of the functional iron in the body is stored in hemoglobin within red blood cells. Large amounts of iron are recycled from senescent erythrocytes by macrophages (51). Chronic liver disease has been extensively linked to iron deficiency anemia (52). Therefore, it would be expected that PBMCs demonstrate altered expression of genes that play crucial roles in iron homeostasis in patients with chronic liver diseases. Erythropoietin plays a crucial role in regulation of erythropoiesis and has been shown to ameliorate fatty liver disease in animal models (53). Immune system processes included signaling pathways (*e.g.*, TREM1, IL-8, IL-17A, B cell receptor, and acute phase), complement system, and agranulocyte adhesion and diapedesis. TREM1 expression in resident and infiltrating immune system cells promotes inflammation during the course of liver disease (54). The IL-8 signaling pathway is enriched by differential expression of the *CXCR1* gene within the PBMC 5-Way dataset. Altered expression of CXCR1 in circulating monocytes of patients with cirrhosis has previously been established (55). Increased expression of IL-17A within a range of immune cells has previously been shown to be an indicator of chronic liver disease (56). In addition to pathway analysis with IPA, we also performed GSEA and BTM analyses of the PBMC 5-Way best gene signature. The most enriched GSEA pathways per pairwise comparison reflected immune response and homeostatic processes (Table S31). Differentially enriched BTMs primarily involved immune response, inflammatory response, oxygen transport, and hemopoiesis (Table S32). Thus, the results of the GSEA and BTM analyses provided additional confirmation of the IPA analysis, and insights into the directionality of the enriched pathways. While alterations in the expression of immune and

inflammatory genes in PBMCs due to liver diseases were expected, it was intriguing that the expression levels of these genes could be used to differentiate between these diverse liver diseases.

Pathway analysis of the Liver 5-Way dataset identified many pathways related to metabolism, biosynthesis, and degradation. For example, when comparing disease groups in the liver dataset (AH, AC, NAFLD, HCV) to healthy controls, some commonly and significantly enriched pathways involved degradation of bupropion, methylglyoxal, tryptophan, acetone, nicotine, and melatonin. Retinoate, retinol, and estrogen biosynthesis pathways were also highly enriched. Abnormal estrogen metabolism due to liver disease has been established previously (57). Abnormal vitamin A metabolism has been heavily implicated in liver disease, especially NAFLD (58, 59). The retinoate and retinol pathways were enriched by differential expression of aldo-keto reductase family 1 member B10 (AKR1B10). AKR1B10 has been reported as an effective biomarker of advanced liver fibrosis and liver cancer (60, 61). The pregnane X receptor activation pathway was also highly enriched across many pairwise comparisons and has been implicated in chronic liver disease (62). The pairwise comparisons involving AH and AC conditions were enriched for ethanol degradation pathways (63) by differential expression of CYP2A7 in our gene signature. Changes in expression of CYP2A genes in liver tissue have been linked with NAFLD and alcohol-associated liver disease (64). These enriched pathways and genes suggest that alterations in the liver's ability to degrade and synthesize these compounds may be related to the liver diseases in the study.

Both PBMC 5-Way and LV 5-Way datasets were enriched for several common immune system pathways, such as: inhibition of matrix metalloproteases (MMPs), macrophage migration inhibitory factor regulation of innate immunity, and interferon signaling pathways. As reported

by IPA, these pathways were enriched by *MMP8, PLA2G4C,* and *IFITM3* genes, respectively, in the PBMC 5-Way dataset. In the LV 5-Way dataset, these pathways were enriched by *MMP7, PLAG2GA*, and a combination of *IFITM1, IFI6,* and *ISG15* genes, respectively. Genes in the MMP family have been established as key actors in liver regeneration and fibrosis (65). PLA2G4C has been reported to play a role in HCV replication (66). Interferon genes have long been implicated in both HCV and viral infections broadly (67). As expected, the interferon signaling pathway had higher enrichment in pairwise comparisons involving HCV in both the PBMC and liver tissue datasets.

We further analyzed the gene expression data from the 19 participants with AH who donated both liver tissue and PBMCs. We identified several genes and gene families that were similarly up- or downregulated within both AH sample types, when compared with healthy controls (Table S34, Figures S12-S14). The genes fell into 4 groups: interferon (*IFITM1, IFI44L*), MMP (*MMP7, MMP8, MMP14*), iron homeostasis (*SLC25A37, SLC11A1*), and tumor necrosis factor (*TNFS10, TNFRSF21, TNFSF13B*) genes. Notably, these findings are similar to our results when comparing the best gene sets across 5-Way PBMCs and 5-Way LV datasets. The similarities in directionality of gene expression between liver and PBMC samples lend credence to using blood-based biomarkers for AH.

While we achieved excellent classification performance and the identification of biologically relevant gene signatures, there were several limitations to our study. Use of independent datasets is crucial in ML and biomarker discovery, however, we could not find any publicly available data on gene expression in PBMCs attained from individuals with AH or AC. Therefore, only our liver tissue dataset results could be independently validated at this time. A larger study with more samples is necessary to validate the biomarkers identified. Our

classification performance could also be improved with the use of more advanced feature selection methodologies such as multi-objective genetic algorithms (68).

In conclusion, our machine learning approach using gene expression data from PBMCs and liver tissue was effective at distinguishing among multiple liver diseases and healthy controls. Additionally, our models were able to distinguish between clinically similar alcohol-associated liver conditions, such as AH and AC. Notably, the AC group for our PBMC samples included both recently drinking and abstinent individuals with AC. AC in patients reporting recent drinking is especially difficult to distinguish from AH clinically, which further demonstrates the utility of this study. While the gene expression data from liver tissue had better classification performance than that of PBMCs, the attainment of liver biopsy is difficult and not standard of care at many healthcare facilities. PBMCs from blood samples, on the other hand, can be easily attained and stored. Based on the outcome of this study, we have demonstrated that blood-based biomarkers from gene expression can be utilized with machine learning methods for the diagnosis of liver disease, paving the way toward the clinical application of liquid biopsy.

CHAPTER 4: Identification of integrated proteomics and transcriptomics signature of alcohol-associated liver disease using machine learning approaches

In preparation for publication.

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

Abstract

Background & Aims

Distinguishing between alcohol-associated hepatitis (AH) and alcohol-associated cirrhosis (AC) remains a challenge in the clinical setting. In this study, we used transcriptomics and proteomics data from liver tissue and peripheral mononuclear blood cells (PBMCs) to classify patients with alcohol-associated liver disease, and identify effective gene and protein expression biomarkers.

Methods

The conditions involved in the study were AH, AC, and healthy controls. We processed 98 PBMC RNAseq samples, 56 PBMC proteomic samples, 67 liver tissue RNAseq samples, and 54 liver tissue proteomic samples. We have built classification and feature selection pipelines for transcriptomics and proteomics data. We have used each type of genomic data separately to classify samples. The liver tissue models were validated in independent liver tissue datasets. Next, we built an integrated gene and protein expression model that allowed us to identify a combined gene-protein biomarker panel for alcohol-associated liver disease. We have also identified several matching gene-protein biomarkers for liver tissue and PBMC integrated models.

Results

For liver tissue, we attained 90% accuracy in test dataset and 82% accuracy in the independent validation dataset using transcriptomic data alone. Similarly, we attained 100% accuracy in the test dataset and 61% accuracy in the independent validation dataset using proteomic data alone. For PBMCs, we attained 83% and 85% accuracy with transcriptomic and proteomic data, respectively. The integration of the two data types did not increase the classification performance in the liver tissue data, but resulted in improvement for PBMCs. We also identified the following gene-protein matches within the gene-protein biomarker panel for liver tissue (presented as genes): CLEC4M, GSTA1, and GSTA2. The matches for PBMC were: SELENBP1, HBD, HBM, and HBZ.

Conclusions

In this study, machine learning models had high classification accuracy for both transcriptomics and proteomics data, across liver tissue and PBMC sample types. The integration of transcriptomics and proteomics into a multi-omics model yielded improvement in classification accuracy for the PBMC data. The set of integrated gene-protein biomarkers for PBMCs showed promise toward developing a liquid biopsy for alcohol-associated liver disease.

Introduction

We have previously 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 (6). However, in addition to transcriptomic data, we also obtained proteomic data for patients from the same cohort. In this study, unlike the previous one, we focused our analysis on three conditions instead of five:

alcohol-associated hepatitis (AH), alcohol-associated cirrhosis (AC), and healthy controls. 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.

Integrating two -omics datatypes further amplified the challenges we encountered in our earlier work (6). 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 as given complex enough model any given set of samples can be perfectly separated. Some of the other challenges were: ensuring that the integrated model does 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 same individuals, but some were not).

We chose to focus on AH and AC because these are deadly conditions with similar clinical presentation. In 2019 there were 23,780 deaths from alcohol-associated cirrhosis (AC) in United States (69). 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 (70). However, attaining a liver transplant as an ALD patient is difficult, since donor livers are scarce and there are concerns about allocating them to individuals with alcohol addiction (70). Typically, a 6 month abstinence from alcohol is required to be a candidate for liver transplant (70). Many of ALD patients have alcohol-associated hepatitis (AH) a condition

which carries mortality of as high as 50% at 3 months (71). For the severe AH patients, the 6 month abstinence requirement can be tantamount to death sentence (70). When carefully selected, AH patients can benefit from liver transplantation (72,73,74,75). Thus, distinguishing between AH and AC patients is an important issue, since AH patients can benefit from urgent liver transplantation. Currently establishing AH diagnosis can require liver biopsy, typically done using transjugular route (71). 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 is a large number of imaging and blood tests for diagnosis of liver cirrhosis (76). However, liver biopsy remains the gold standard for diagnosis (77). Furter improvement in accuracy of non-invasive tests is necessary to reduce the need for liver biopsy (78). The gene and protein biomarkers identified in this study, with further validation, could be used to develop new highly accurate blood tests for 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 demographic tables of the patients that donated samples for RNAseq analysis can be found in Chapter 3 (6) methods. The demographics of patients that donated liver tissue and PBMC samples for proteomic analyses can be found in Tables 4.1 and 4.2.

Table 4.1: Demographics of patients that donated liver tissue for proteomic analysis.

	Liver tissue samples (proteomics)		
	AH	CT	AC
	n=34	n=10	n=10
Age: mean ± std (range)	42.3 ± 11.5	56 ± 8.6	51.9 ± 13.1
MELD: mean ± std (range)	25.2 ± 5.6	NA	32 ± 6.1*
Maddrey's DF: mean (range)	53.3 ± 21.8	NA	NA
BMI: mean ± std (range)	28.8 ± 5.3	NA	25.6 ± 8.4*
Gender: N (percent)			
Female	3(8.8%)	0(0.0%)	0(0.0%)
Male	31(91.2%)	10(100%)	9(90%)
Ethnicity: N (percent)			
Hispanic	26(76.5%)	NA	0(0.0%)
NHW	5(14.7%)	NA	5(50%)
Black	2(5.9%)	NA	0(0.0%)
Other	1(2.9%)	NA	0(0.0%)
Source	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.

Table 4.2: Demographics of patients that donated PBMCs for proteomic analysis.

	PBMC samples (proteomics)		
	AH	CT	AC
	(n=21)	(n=22)	(n=13)
Age: mean ± std (range)	47.7 ± 12.2	34.8 ± 15.1	54.2 ± 11.2
MELD: mean ± std (range)	24.4 ± 3.5	7.5 ± 2.5	13.6 ± 6.7
Maddrey's DF: mean (range)	49.5 ± 16.9	2.5 ± 7.8	22.1 ± 23.3
BMI: mean ± std (range)	29.4 ± 5.5	27.1 ± 4	30 ± 4.8
Gender: N (percent)			
Female	1(4.8%)	10(45.4%)	0(0.0%)
Male	20(95.2%)	12(54.6%)	13(100%)
Ethnicity: N (percent)			
Hispanic	13(61.9%)	12(54.5%)	10(76.9%)
NHW	5(23.8%)	0(0.0%)	2(15.4%)
Black	2(9.5%)	1(4.5%)	0(0.0%)
Other	1(4.8%)	12(54.5%)	1(7.7%)
Source	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.

The biospecimens consisted of 98 PBMC RNAseq samples, 56 PBMC proteomic samples, 67 liver tissue RNAseq samples, and 54 liver tissue proteomic samples. The liver diseases represented were encoded with two letter symbols as follows: alcohol-associated hepatitis (AH) and alcohol-associated cirrhosis (AC). Most of the AC participants within SCAHC study were expected to be in-patients with decompensated cirrhosis. The inclusion and exclusion criteria can be found in the Supplementary Materials. 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.

Sample collection

The blood samples and liver biopsies from participants with AH were collected before starting treatment. Blood samples from all other groups were collected at entry into the study. PBMCs were freshly isolated from the blood samples by Ficoll-Histopaque (GE Healthcare) gradient centrifugation, flash frozen, and then stored in a liquid nitrogen tank. The AH biopsy sample were flash frozen in liquid nitrogen. The liver tissue samples for healthy controls and AC patients were obtained from University of Minnesota LTCDS.

Sample data preprocessing

Sample processing is described in full in Supplemental Methods.

Partitioning samples into datasets:

Some proteomic and transcriptomic samples came from the same patients, while others did not.

Tables 4.3,4.4,4.5, and 4.6 summarize the degree of matching between proteomic and transcriptomic samples in liver tissue and PBMC. For the purposes of analyses some of the unmatched subsets were too small. Therefore, we moved some matched samples into unmatched

sample categories, and called these new categories balanced matched and balanced unmatched subsets. We divided our data into the following dataset categories. We refer to the datasets below as our test data.

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.

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.

Table 4.3: Number of PBMC RNAseq samples that match PBMC proteomic samples

PBMC RNAseq	AH	CT	AC
Total	38	20	40
Matched	18	19	13
Unmatched	20	1	27
Matched Balanced	9(-9)	12(-7)	6(-7)
Unmatched Balanced	29(+9)	8(+7)	34(+7)

Table 4.4: Number of PBMC proteomic samples that match PBMC RNAseq samples

PBMC Proteomics	AH	CT	AC
Total	21	22	13
Matched	18	19	13
Unmatched	3	3	0
Matched Balanced	9(-9)	12(-7)	6(-7)
Unmatched Balanced	12(+9)	10(+7)	7(+7)

Table 4.5: Number of liver RNAseq samples that match liver proteomic samples

Liver RNAseq	AH	CT	AC
Total	32	8	8
Matched	29	3	5
Unmatched	3	5	3
Matched Balanced	24(-5)	3	3(-2)
Unmatched Balanced	8(+5)	5	5(+2)

Table 4.6: Number of liver proteomic samples that match liver RNAseq samples

Liver Proteomics	AH	CT	AC
Total	34	10	10
Matched	29	3	5
Unmatched	5	7	5
Matched Balanced	24(-5)	3	3(-2)
Unmatched Balanced	10(+5)	7	7(+2)

Validation dataset

We validated our proteomic liver tissue machine learning (ML) models using data obtained from MassIVE repository (accession number MSV000089168) (80). This dataset contained liver tissue proteomic data from participants with AH (n=6) and healthy controls (n=12). We utilized proteomic counts generated by collaborator. 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 RNAseq liver tissue validation dataset can be found in Chapter 3 (6).

RNAseq Classification and Feature Selection Pipeline

The methods used to classify RNAseq counts and identify best genes are covered in (1).

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 (6).

Imputation

We used median and replacement with zero imputation strategies. Median: replace missing values using the median along each column (feature, in this case protein). Zero: replace all missing values with zeros.

We only imputed values for proteins that were missing data for 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 < 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

We used INFERNORDN to perform differential expression analysis with proteomic counts (81). Proteins were filtered by q-value < 0.05 and based on imputation threshold.

In silico Biological Validation and Best Gene Selection

The Enrichr (46) was replaced with AGOTOOL (82) for enrichment analysis of proteins. When selecting best gene set identical algorithm was used for both transcriptomic and proteomic data with one exception. For proteomic data, gene sets produced by configurations with least imputation were preferred.

Analysis Outline

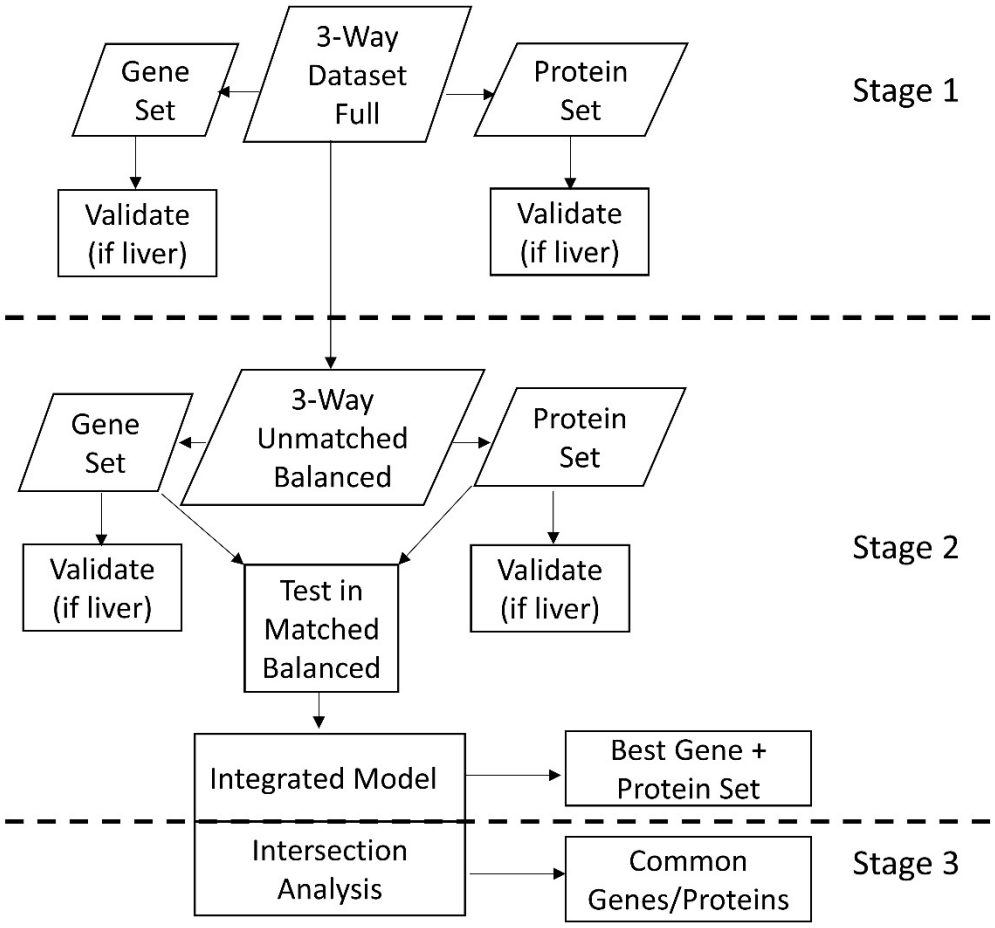

Figure 4.1: Flowchart demonstrating 3 stages of the analysis. Stage 1: Analysis of all proteomic and RNAseq samples separately. Stage 2: Training ML models in unmatched balanced data with subsequent testing and integration in matched balanced data. Stage 3: Intersection analysis of the combined best gene-protein sets for liver and PBMC tissues

Stage 1:

We began our analysis by classifying the Liver 3-Way Full and PBMC 3-Way Full datasets and identifying best genes and proteins for both tissue types using our RNAseq and proteomic pipelines. The entirety of analysis is summarized in the flowchart (Figure 4.1).

Stage 2:

We then performed the same type of analysis on the Liver 3-Way Unmatched Balanced and PBMC 3-Way Unmatched Balanced datasets. Afterward we trained RNAseq and proteomic ML models on entirety of unmatched balanced data. These models were then tested in matched balanced data. Finally, the RNAseq and proteomic models were integrated and tested in matched balanced data using cross-validation. The integration of RNAseq and proteomic models was done by supplying the prediction probabilities they output as input into a third model (Figure 4.2). This integrated model was then evaluated with combinations of gene and protein sets obtained during analysis of Unmatched Balanced datasets. The pair of gene and protein sets that attained the best performance in the integrated model was reported as the best combined gene and protein panel.

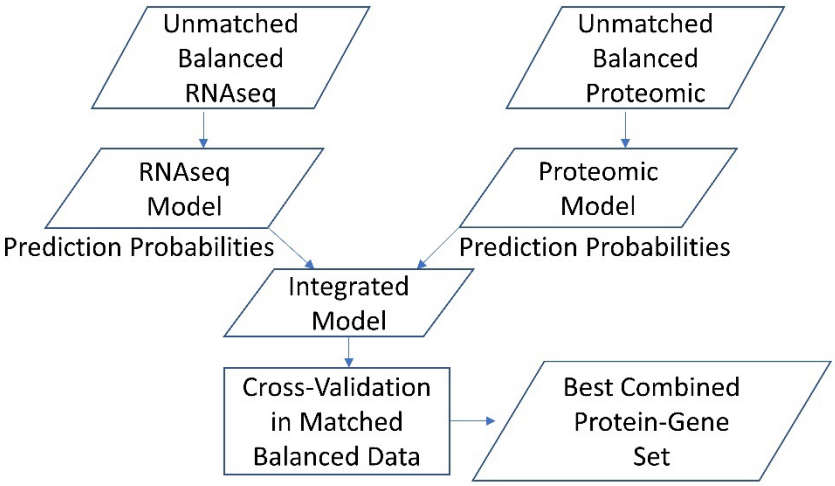

Figure 4.2: Flowchart representation of integrated RNAseq – Proteomic ML model and its application in Matched Balanced data

Stage 3:

As a last step 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.

Validation in independent liver tissue data

All liver tissue ML models (RNAseq and proteomic) were validated in independent liver tissue validation data. The methods for independent validation were identical for both RNAseq and proteomic datatypes. The description of these methods can be found in Chapter 3 (6) methods.

Machine Learning Classifier

The classifiers used in analysis of RNAseq data are k nearest neighbors (kNN), logistic regression (LR), and support vector machine (SVM). For the purposes of analyzing proteomic data we solely used logistic regression (LR) classifier. The LR model has been shown to be well

suited for small sample size proteomic data previously (83). Both LR and SVM classifiers were regularized.

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 best gene and protein sets. The best gene set contained 33 genes, attained 90% accuracy in main data and 82% accuracy in validation data (Figure 4.3A, 4.3C). The best protein set contained 27 proteins, attained 100% accuracy in main data and 61% accuracy in validation data (Figure 4.3B, 4.3D). 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 Supplemental (Fig. S1-S8). The enriched pathways, tissues, and diseases for best gene and protein sets can be found in the supplemental (Tables S3 and S4). The best gene and protein sets for each dataset are shown in Table 4.7.

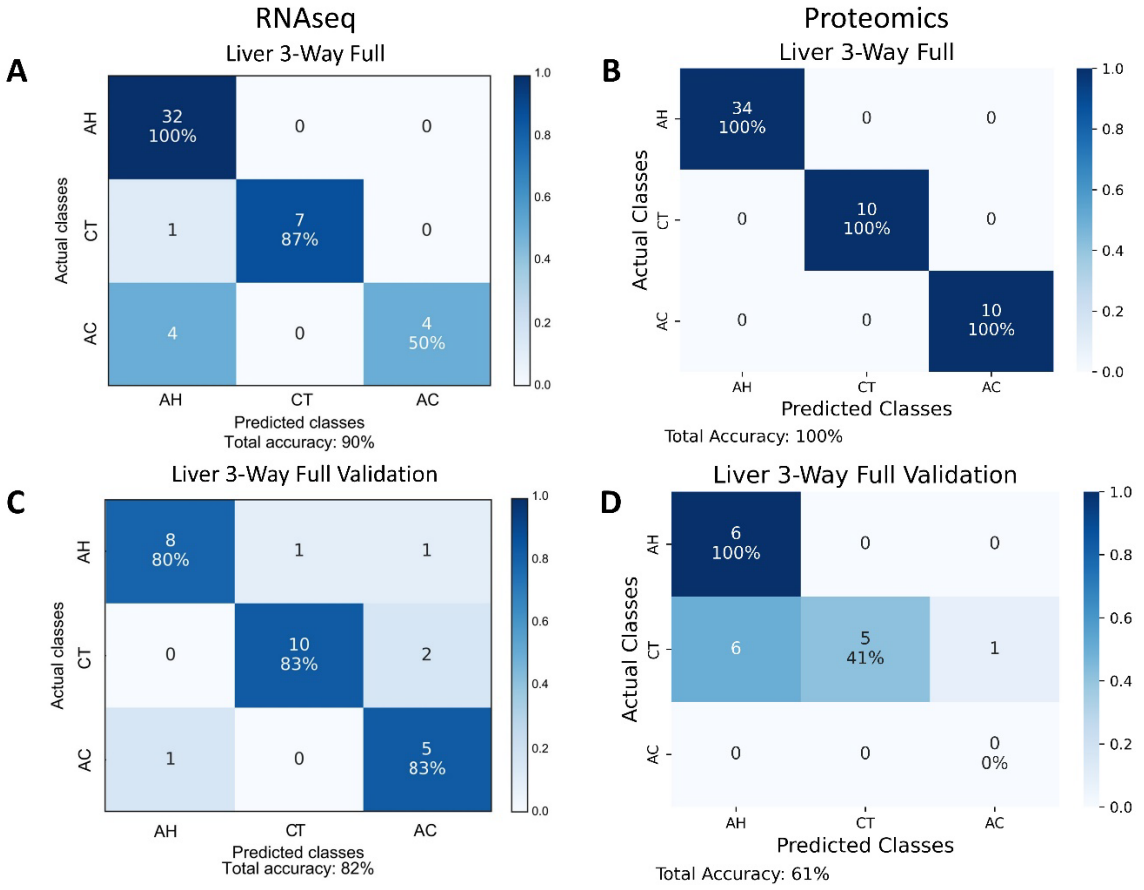

Figure 4.3: Confusion matrices corresponding to the best gene and protein sets of Liver 3-Way Full datasets. (**A**) Confusion matrix for classification of Liver 3-Way Full RNAseq dataset using best gene set identified by filter feature selection. The diagonal contains the number and percentage of the correctly predicted samples. (**B**) Confusion matrix for classification of Liver 3-Way Full proteomic dataset using best protein set identified by filter feature selection. (**C**) Confusion matrix for classification of AH, AC, and healthy control (CT) samples within independent validation RNAseq dataset. (**D**) Confusion matrix for classification of AH, AC, and CT samples within independent validation proteomic dataset

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 (Figure 4.4A). The best protein set contained 46 proteins and attained 85% accuracy in main data (Fig. 4.4B). RNAseq and proteomic data proved equally effective at classifying our PBMC 3-Way samples. The heatmaps of RNAseq and proteomic counts can be found in Supplemental (Fig. S9-

S12). The enriched pathways, tissues, and diseases for best gene and protein sets can be found in the supplemental (Tables S5 and S6). The best gene and protein sets for each dataset are shown in Table 4.7.

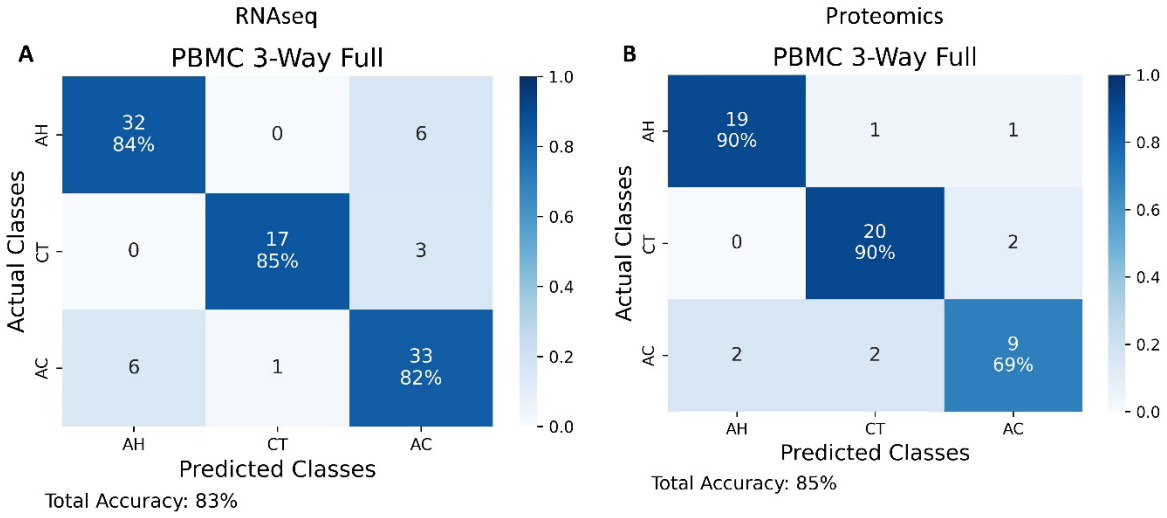

Figure 4.4: Confusion matrices corresponding to the best gene and protein sets of PBMC 3-Way Full datasets. (**A**) Confusion matrix for classification of PBMC 3-Way Full RNAseq dataset using best gene set identified by filter feature selection. (**B**) Confusion matrix for classification of PBMC 3-Way Full proteomic dataset using best protein set identified by filter feature selection

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

Integration

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 (Figure 4.5A). Using the best protein set of 27 proteins we attained 100% classification accuracy within matched balanced proteomic data (Figure 4.5B). Using a combination of best gene and protein sets, we attained 96% accuracy in matched balanced integrated data (Figure 4.5C).

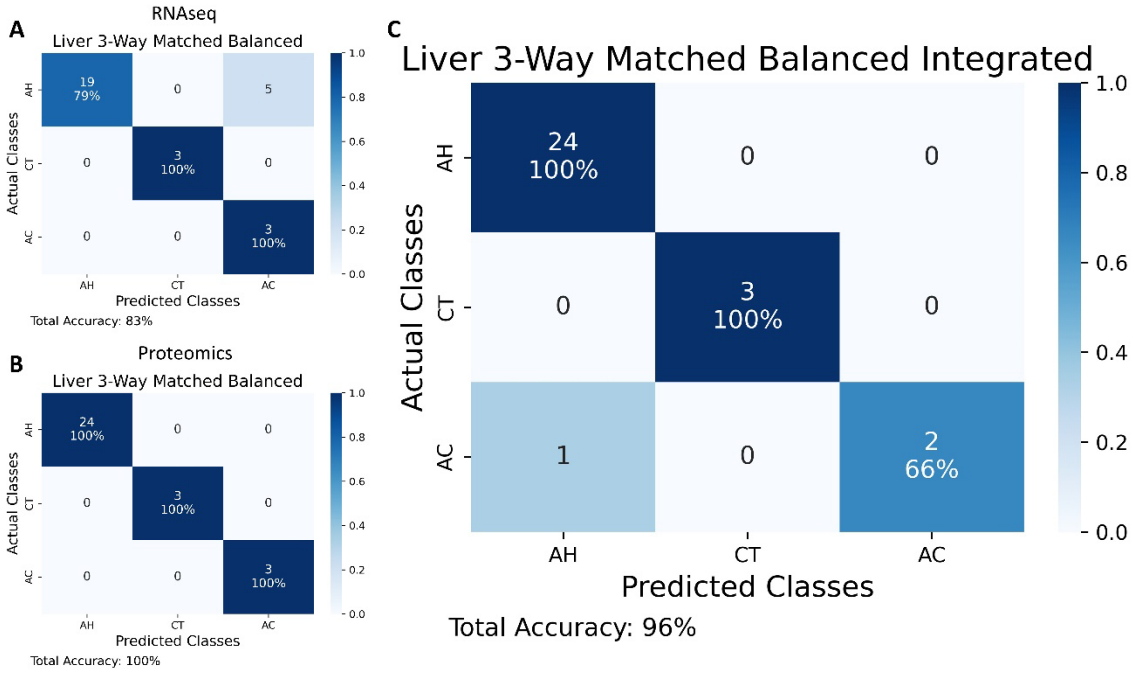

Figure 4.5: Confusion matrices corresponding to the best gene and protein sets evaluated within Liver 3-Way Matched Balanced data. (**A**) Confusion matrix for classification of Liver 3-Way Matched Balanced RNAseq dataset using best gene set identified by filter feature selection. (**B**) Confusion matrix 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

Intersection

Additionally, we examined which biomarkers were shared between the best gene and protein sets (Figure 4.6). The CLEC4M, GSTA1, and GSTA2 were found in common. The CLEC4M 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 0.18 would be shared. Calculation of expected value can be found in Supplemental. 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.

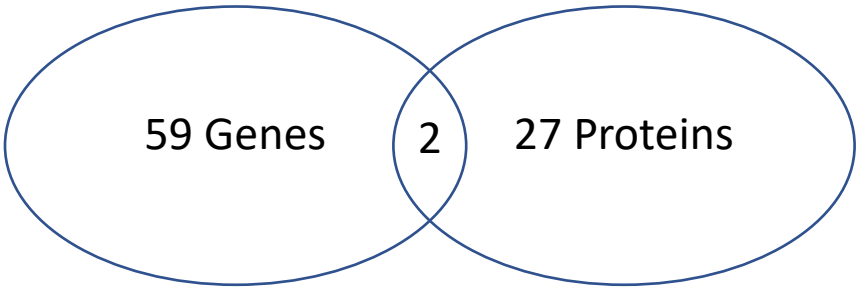

Figure 4.6: Venn diagram of overlap between best genes and proteins of Liver 3-Way

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

Integration

The best gene set and protein set 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 (Figure 4.7A). Using the best protein set of 24 proteins we attained 59% classification accuracy within matched balanced proteomic data (Figure 4.7B). Using a combination of best gene and protein sets, we attained 77% accuracy in matched balanced integrated data (Figure 4.7C).

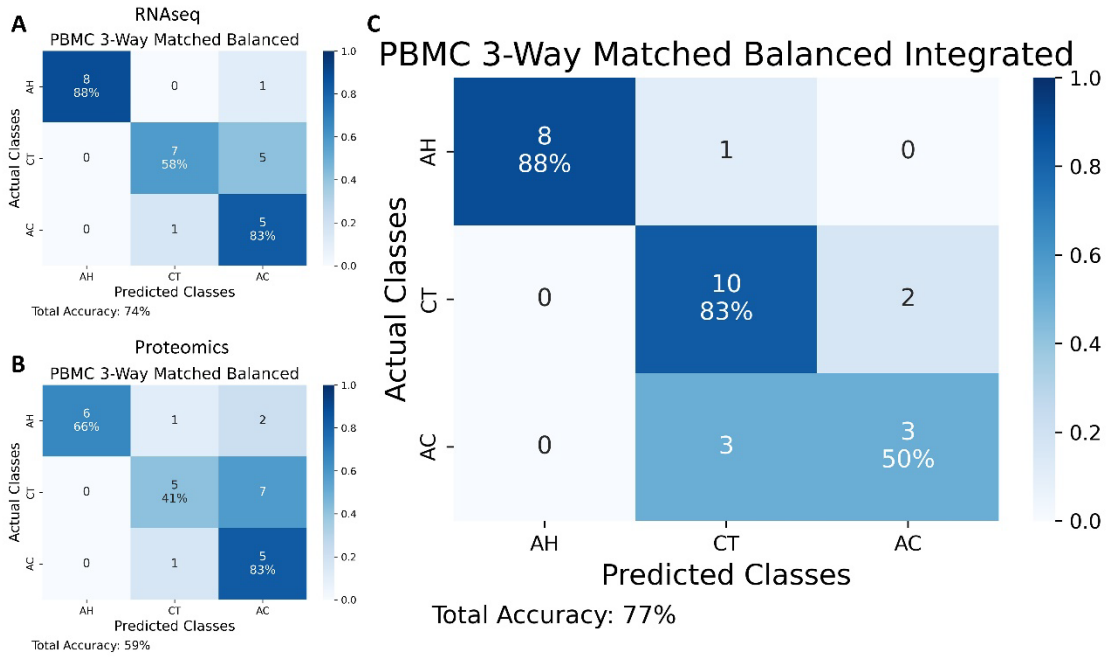

Figure 4.7: Confusion matrices corresponding to the best gene and protein sets evaluated within PBMC 3-Way Matched Balanced data. (**A**) Confusion matrix for classification of PBMC 3-Way Matched Balanced RNAseq dataset using best gene set identified by filter feature selection. (**B**) Confusion matrix for classification of PBMC 3-Way Matched Balanced proteomic dataset using best protein set identified by filter feature selection. (**C**) Confusion matrix for classification of PBMC 3-Way Matched Balanced dataset using a combination of best gene and protein sets

Intersection

Additionally, we examined which biomarkers were shared between the best gene and protein sets (Figure 4.8). The SELENBP1, HBZ, HBM, and HBD were found in common. The SELENBP1 was a direct match, while the HBZ (HBAZ protein) was a familial match with HBM and HBD (genes). If the genes and proteins had been selected randomly from among significantly differentially expressed genes and proteins, an expected 0.04 would be shared. Calculation of expected value can be found in Supplemental. Therefore, we have identified more biomarkers in common than expected. Best gene and protein sets were commonly enriched for several different inflammation and cancer related pathways.

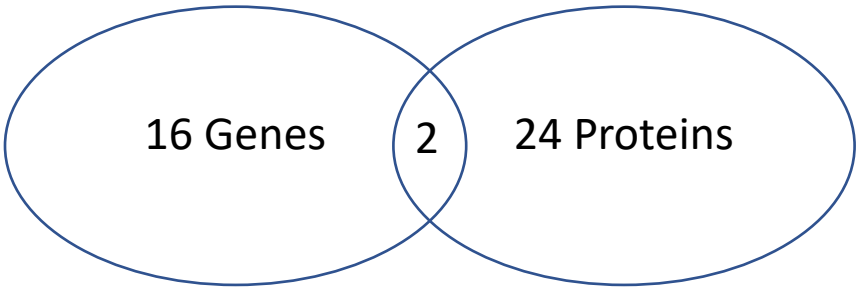

Figure 4.8: Venn diagram of overlap between best genes and proteins of PBMC 3-Way

Table 4.7: 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.	ACTN1, APOA1, APOA2, BLVRB, CATS, CATZ, CCL5, CSCL1, CSRP1, EST1, FHL1, FIBA, FIBB, FIBG, GELS, GP1BB, GPIX, HBD, ILK, ITA2B, ITA6, ITB1, ITB3, LIMS1, LRC25, LTBP1, MYL6, MYL9, MYLK, NACA2, PMGE, RAP1A, RAP1B, RGS18, RS4Y1, SDPR, SRC, TAGL2, TBA4A, TOR4A, TSN9, TSP1, VINC, VTDB, VWF, ZYX
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, ADH1G, ADH4, ADH6, AL1L1, ALBU, CD34, CES1P, **CLC4M**, CO1A2, CP1A2, CYB5, DHCR7, GBA3, **GSTA1**, GSTM1, H2B1C, HBAZ, HBA, HS71L, LDH6A, MGST1, PEBP1, RDH16, SAA1.
PBMC 3-Way Matched Balanced Integrated	AHSP, ALAS2, CA1, CD177, CDK10, EHMT1, **HBD**, **HBM**, IFI27, IL1R2, MECP2, MMP8, MMP9, **SELENBP1**, SLC4A1, TANGO2.	ALBU, CXCL7, FHL1, FIBA, FIBB, FIBG, FSTL1, GP1BB, **HBAZ**, ILK, ITA2B, ITB3, LIMS1, LYSC, MYL9, RAP1A, **SBP1**, SDPR, TBA4A, TSN15, TSP1, URP2, VINC, VTDB.

Discussion

Healthy controls, AHs, and ACs were effectively classified using either transcriptomics or proteomics data. Liver tissue models outperformed PBMC models by a small margin in test data. However, PBMC models still performed well enough to warrant future examination. Both

types of liver tissue ML models generalized relatively well in the independent validation data. Overall, the transcriptomic and proteomic models performed similarly well in each tissue.

The integration of proteomic and transcriptomic data did not increase classification accuracy with liver tissue. For PBMCs, on the other hand, the integration improved classification accuracy. While the performance of PBMC biomarkers still lags behind 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 time 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 (84). GSTA1 and GSTA2 have been previously identified as biomarkers of liver injury (including ethanol injury) and hepatocellular carcinoma respectively (85, 86). Less is known about the role of matched PBMC genes in liver disease. Differential expression of SELENBP1 in PBMCs of hepatocellular carcinoma patients has been established previously (87). Meanwhile, HBD, HBM, and HBZ are all hemoglobin genes connected to anemias, blood disorders, and blood cancer (88, 89). Chronic liver disease has been connected to anemia (22). Therefore, altered expression of these hemoglobin genes in PBMCs due to liver disease seems plausible.

We have discussed the importance of using appropriate ML methods for analysis of small sample size RNAseq data (6) 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 (90).

The liver tissue proteomics model's performance in independent validation data was lower than expected. The healthy control samples in independent validation proteomic dataset came from two different 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. A larger sample size and an independent integrated validation cohort are needed to further investigate these biomarkers.

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

CHAPTER 5: CONCLUSION

Throughout my doctoral studies, I have developed a bioinformatics software tool, built a classification and feature selection pipeline for analysis of small sample size genomic data, and applied it to both transcriptomics and proteomics human liver disease data to identify actionable diagnostic biomarkers. Several of the key computational findings of this research included the suitability of filter feature selection for analysis of small sample size genomic data, and the importance of incorporation of functional relevance scoring toward biomarker discovery. Filter feature selection methods, such as differential expression, minimize risk of overfitting compared to more complex feature selection methods. Functional relevance scoring further reduces risk of overfitting by incorporating genomic knowledge into feature selection.

As future directions, I plan to expand on my doctoral work. In particular, to enhance my first aim, I have observed over the last several years that bioinformatics software has been increasingly shifting online. Web applications are much more convenient to use than desktop applications. For use with data that does not have HIPAA restrictions, I plan to re-implement A-Lister as a web application. A-Lister could also be packaged as a Docker or Singularity container, which would allow its installation on a user's secure server. I also have interest in further improving feature selection in context of genomic data analysis. This is important for the identification of actionable biomarkers. Use of multi-omics data is also becoming more feasible as the amount of genomic data increases, and novel methods for analysis of such data are required.

Throughout my dissertation research, I have shown 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 methods to a liver disease genomics data set, I 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.

REFERENCES

1. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai CX, Efron MJ, et al. Big Data: Astronomical or Genomical? Plos Biol. 2015;13(7).
2. McDermott JE, Wang J, Mitchell H, Webb-Robertson B, Hafen R, Ramey J, et al. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. Expert Opinion on Medical Diagnostics. 2013;7(1):37-51.
3. Listopad S and Norden-Krichmar TM. A-Lister: a tool for analysis of differentially expressed omics entities across multiple pairwise comparisons. BMC Bioinformatics. 2019;20.
4. Gao Y, Liu S, Baldwin VI RL, Conno EE, Cole JB, Ma L, et al. Functional annotation of regulatory elements in cattle genome reveals the roles of extracellular interaction and dynamic change of chromatin states in rumen development during weaning. Genomics. 2022;114(2).
5. Cheng X, Yan J, Liu Y, Wang J, and Taubert S. eVITTA: a web-based visualization and inference toolbox for transcriptome analysis. Nucleic Acids Research. 2021;49(W1):W207-W215.
6. Listopad S, Magnan C, Asghar A, Stolz A, Tayek JA, Liu Z, Morgan, T.R., and Norden-Krichmar, T.M. Differentiating between liver diseases by applying multiclass machine learning approaches to transcriptomics of liver tissue or blood based samples. JHEP Reports. 2022;4(10).
7. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;17(13).
8. Efstathiou G, Antonakis AN, Pavlopoulos GA, Theodosiou T, Divanach P, Trudgian DC, et al. ProteoSign: an end-user online differential proteomics statistical analysis platform. Nucleic Acids Research. 2017;45(W1):W300-W306.
9. Love MI, Huber W, and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12).
10. Robinson MD, McCarthy DJ, and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139-140.
11. Yassi M, Davodly ES, Shariatpanahi AM, Heidari M, Dayyani M, Heravi-Moussavi A, et al. DMRFusion: A differentially methylated region detection tool based on the ranked fusion method. Genomics 2018;110(6):366-374.
12. Listopad S: A-Lister. https://github.com/staslist/A-Lister; 2019.
13. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, and Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 2013;31(1):46-+.

14. Christen P. A comparison of personal name matching: Techniques and practical issues. Icdm 2006: Sixth Ieee International Conference on Data Mining, Workshops. 2006;290-294.

15. Edgar R, Domrachev M, and Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 2002; 30(1):207-210.

16. Spyrou J, Gardner DK, and Harvey AJ: Metabolism Is a Key Regulator of Induced Pluripotent Stem Cell Reprogramming. Stem Cells International 2019.

17. Deyarshi PM, Jones AD, Campbell WW, Taylor EM, and Henagan TM. Effects of Acute Aerobic Exercise on Whole Genome Nucleosome Maps and Gene Expression in Skeletal Muscle of Lean Vs Overweight/Obese Men. Faseb J 2017;31.

18. Williams JR, Yang RT, Clifford JL, Watson D, Campbell R, Getnet D, et al. Functional Heatmap: an automated and interactive pattern recognition tool to integrate time with multi-omics assays. Bmc Bioinformatics 2019;20.

19. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 2018;46(W1):W537-W544.

20. Khan A and Mathelier A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. Bmc Bioinformatics 2017;18.

21. Lin GL, Chai J, Yuan S, Mai C, Cai L, Murphy RW, et al. VennPainter: A Tool for the Comparison and Identification of Candidate Genes Based on Venn Diagrams. Plos One 2016;11(4).

22. Heberle H, Meirelles GV, da Silva FR, Telles GP, and Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. Bmc Bioinformatics 2015;16.

23. Wang MH, Zhao YZ, and Zhang B. Efficient Test and Visualization of Multi-Set Intersections. Sci Rep-Uk 2015;5.

24. Shen L. GeneOverlap: An R package to test and visualize gene overlaps. https://bioconductor.org/packages/release/bioc/html/GeneOverlap.html. 2014.

25. Bateman A, Martin MJ, Orchard S, Magrane M, Alpi E, Bely B, et al. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research. 2019;47(D1):D506-15.

26. Asrani SK, Devarbhavi H, Eaton J, and Kamath PS. Burden of liver diseases in the world. Journal of Hepatology 2019;70:151-171.

27. Shasthry SM and Sarin SK. New treatment options for alcoholic hepatitis. World Journal of Gastroenterology 2016;22:3892-3906.

28. Torruellas C, French SW, and Medici V. Diagnosis of alcoholic liver disease. World Journal of Gastroenterology 2014;20:11684-11699.

29. Singh T, Allende DS, and McCullough AJ. Assessing liver fibrosis without biopsy in patients with HCV or NAFLD. Cleveland Clinic Journal of Medicine 2019;86:179-186.

30. Ibn Sina A, Carrascosa LG, Liang ZY, Grewal YS, Wardiana A, Shiddiky MJA, et al. Epigenetically reprogrammed methylation landscape drives the DNA self-assembly and serves as a universal cancer biomarker. Nature Communications 2018;9.

31. Berdasco M and Esteller M. Clinical epigenetics: seizing opportunities for translation. Nature Reviews Genetics 2019;20:109-127.

32. Nallagangula KS, Nagaraj SK, Venkataswamy L, and Chandrappa M. Liver fibrosis: a compilation on the biomarkers status and their significance during disease progression. Future Science Oa 2018;4.

33. Ding WC, Xin JJ, Jiang LY, Zhou Q, Wu TZ, Shi DY, et al. Characterisation of peripheral blood mononuclear cell microRNA in hepatitis B-related acute-on-chronic liver failure. Scientific Reports 2015;5.

34. Waldron PR and Holodniy M. Peripheral Blood Mononuclear Cell Gene Expression Remains Broadly Altered Years after Successful Interferon-Based Hepatitis C Virus Treatment. Journal of Immunology Research 2015;2015.

35. Zhang L, Ma DX, Li X, Deng CW, Shi Q, You X, et al. Gene expression profiles of peripheral blood mononuclear cells in primary biliary cirrhosis. Clinical and Experimental Medicine 2014;14:409-416.

36. Zhou Q, Ding WC, Jiang LY, Xin JJ, Wu TZ, Shi DY, et al. Comparative transcriptome analysis of peripheral blood mononuclear cells in hepatitis B-related acute-on-chronic liver failure. Scientific Reports 2016;6.

37. Sowa JP, Atmaca O, Kahraman A, Schlattjan M, Lindner M, Sydor S, et al. Non-Invasive Separation of Alcoholic and Non-Alcoholic Liver Disease with Predictive Modeling. Plos One 2014;9.

38. Trepo E, Goossens N, Fujiwara N, Song WM, Colaprico A, Marot A, et al. Combination of Gene Expression Signature and Model for End-Stage Liver Disease Score Predicts Survival of Patients With Severe Alcoholic Hepatitis. Gastroenterology 2018;154:965-975.

39. Hoshida Y, Villanueva A, Sangiovanni A, Sole M, Hur C, Andersson KL, et al. Prognostic Gene Expression Signature for Patients With Hepatitis C-Related Early-Stage Cirrhosis. Gastroenterology 2013;144:1024-1030.

40. Sheng QH, Vickers K, Zhao SL, Wang J, Samuels DC, Koues O, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. Briefings in Functional Genomics 2017;16:194-204.

41. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.

42. Massey V, Parrish A, Argemi J, Moreno M, Mello A, García-Rocha M, et al. Integrated Multiomics Reveals Glucose Use Reprogramming and Identifies a Novel Hexokinase in Alcoholic Hepatitis. Gastroenterology. 2021;160(5):1725-1740.

43. 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;7:562-578.

44. Saeys Y, Inza I, and Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507-2517.

45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825-2830.

46. 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;14.

47. Krämer A, Green J, Pollard J, and Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30(4):523–30.

48. Mootha V, Lindgren C, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 2003;34:267–273.

49. Rinchai D, Roelands J, Toufiq M, Hendrickx W, Altman MC, and Bedognetti D. BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R. *Bioinformatics.* 2021;37(16):2382–2389.

50. Sharma S, Baweja S, Maras JS, Shasthry SM, Moreau R, and Sarin SK. Differential blood transcriptome modules predict response to corticosteroid therapy in alcoholic hepatitis. JHEP Reports. 2021;3(3).

51. Anderson ER and Shah YM. Iron Homeostasis in the Liver. Comprehensive Physiology 2013;3:315-330.

52. Gkamprela E, Deutsch M, and Pectasides D. Iron deficiency anemia in chronic liver disease: etiopathogenesis, diagnosis and treatment. Annals of Gastroenterology 2017;30:405-413.

53. Hong T, Ge ZJ, Zhang BJ, Meng R, Zhu DL, and Bi Y. Erythropoietin suppresses hepatic steatosis and obesity by inhibiting endoplasmic reticulum stress and upregulating fibroblast growth factor 21. International Journal of Molecular Medicine 2019;44:469-478.

54. Sun H, Feng J, and Tang L. Function of TREM1 and TREM2 in Liver-Related Diseases. *Cells*. 2020; 9(12):2626.

55. Zimmermann HW, Seidler S, Gassler N, Nattermann J, Luedde T, et al. Interleukin-8 Is Activated in Patients with Chronic Liver Diseases and Associated with Hepatic Macrophage Accumulation in Human Liver Fibrosis. PLOS ONE 2011;6(6):e21381.

56. Tan Z, Qian X, Jiang R, Liu Q, Wang Y, Chen C, et al. IL-17A Plays a Critical Role in the Pathogenesis of Liver Fibrosis through Hepatic Stellate Cell Activation. The Journal of Immunology 2013;191(4):1835-1844.

57. Chai X, Zeng S, and Xie W. Estrogen-Metabolizing Enzymes in Systemic and Local Liver Injuries: A Case Study of Disease-Drug Interaction. Drug Metabolism in Diseases 2017:241-255.

58. Pettinelli P, Arendt BM, Teterina A, McGilvray I, Comelli EM, Fung SK, et al. Altered hepatic genes related to retinol metabolism and plasma retinol in patients with non-alcoholic fatty liver disease. Plos One 2018;13.

59. Saeed A, Dullaart RPF, Schreuder TCMA, Blokzijl H, and Faber KN. Disturbed Vitamin A Metabolism in Non-Alcoholic Fatty Liver Disease (NAFLD). Nutrients 2018;10.

60. Kanno M, Kawaguchi K, Honda M, Horii R, Takatori H, Shimakami T, et al. Serum aldo-keto reductase family 1 member B10 predicts advanced liver fibrosis and fatal complications of nonalcoholic steatohepatitis. Journal of Gastroenterology 2019;54:549-557.

61. Ye X, Li CY, Zu XY, Lin ML, Liu Q, Liu JH, et al. A Large-Scale Multicenter Study Validates Aldo-Keto Reductase Family 1 Member B10 as a Prevalent Serum Marker for Detection of Hepatocellular Carcinoma. Hepatology 2019;69:2489-2501.

62. Sayaf K, Zanotto I, Russo FP, Gabbia D, and De Martin S. The Nuclear Receptor PXR in Chronic Liver Disease. *Cells*. 2022;11(1):61.

63. Teschke R. Alcoholic Liver Disease: Alcohol Metabolism, Cascade of Molecular Mechanisms, Cellular Targets, and Clinical Aspects. Biomedicines 2018;6.

64. Wang KS, Chen X, Ward SC, Liu Y, Ouedraogo Y, Xu C, et al. CYP2A6 is associated with obesity: studies in human samples and a high fat diet mouse model. International Journal of Obesity 2019;43:475-486.

65. Naim A, Pan QW, and Baig MS. Matrix Metalloproteinases (MMPs) in Liver Diseases. Journal of Clinical and Experimental Hepatology 2017;7:367-372.

66. Su X, Liu S, Zhang X, Lam SM, Hu X, Zhou Y, et al. Requirement of cytosolic phospholipase A2 gamma in lipid droplet formation. Biochim Biophys Acta Mol Cell Biol Lipids. 2017;1862(7):692-705.

67. Narayana S, Helbig K, McCartney E, Eyre N, Bull R, Eltahla A, et al. The Interferon-induced Transmembrane Proteins, IFITM1, IFITM2, and IFITM3 Inhibit Hepatitis C Virus Entry. Journal of Biological Chemistry. 2015;190(43):25946-25959.

68. Vafaee F, Diakos C, Kirschner MB, Reid G, Michael MZ, Horvath LG, et al. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. Npj Systems Biology and Applications 2018;4.

69. 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;135(10):1263-1266.

70. Mellinger JL and Volk ML. Transplantation for alcohol-related liver disease: is it fair? Alcohol and Alcoholism. 2017;53(2):173-177.

71. Thursz M and Morgan TR. Treatment of severe alcoholic hepatitis. Gastroenterology. 2016;150(8):1823-1834.

72. 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.

73. 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;16(3):841-849.

74. 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;265(1):20-29.

75. Singal AK, Bashar H, Anand BS, Jampana SC, Singal V, and Kuo Y. Outcomes after liver transplantation for alcoholic hepatitis are similar to alcoholic cirrhosis: exploratory analysis from the UNOS database. Hepatology. 2011;55(5):1398-1405.

76. Soresi M, Giannitrapani L, Cervello M, Licata A, and Montalto G. Non invasive tools for the diagnosis of liver cirrhosis. World Journal of Gastroenterology. 2014;20(48):18131-18150.

77. Berger D, Desai V, and Janardhan S. Con: liver biopsy remains the gold standard to evaluate fibrosis in patients with nonalcoholic fatty liver disease. Clinical Liver Disease. 2019;13(4):114-116.

78. Lambrecht J, Verhulst S, Mannaerts I, Reynaert H, and Grunsven LA. Prospects in non-invasive assessment of liver fibrosis: liquid biopsy as the future gold standard? Molecular Basis of Disease. 2018:1864(4):1024-1036.

79. Hardesty J, Day L, Warner J, Warner D, Gritsenko M, Asghar A, Stolz A, et al. Hepatic protein and phosphoprotein signatures of alcohol-associated cirrhosis and hepatitis. The American Journal of Pathology. 2022;192(7):1066-1082.

80. Argemi J, Kedia K, Gritsenko M, Clemente-Sanchez A, Asghar A, Herranz J, et al. Integrated transcriptomic and proteomic analysis identifies plasma biomarkers of hepatocellular failure in alcohol-associated hepatitis. The American Journal of Pathology. 2022.

81. 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; 24(13):1556-8.

82. 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;12(11):1003-4.

83. 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;28:1277-1287.

84. 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;19(3):1711-1720.

85. 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;14(4):3798-3804.

86. 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;188.

87. 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;17797.

88. Harteveld CL, Wijermans PW, Arkesteijn SGJ, Delft PV, Kerkhoffs JL, and Giordano PC. Hb Lepore-Leiden: A New δ/β Rearrangement Associated with a β-Thalassemia Minor Phenotype. Hemoglobin. 2008;32(5):446-453.

89. Zhao T and Matsuoka M. HBZ and its roles in HTLV-1 oncogenesis. Frontiers in Microbiology. 2012;3.

90. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights into Imaging. 2021;172.

APPENDIX A: CHAPTER 3 SUPPLEMENTAL

1. SUPPLEMENTARY METHODS

Sections **a-j** below describe the collection and processing of the samples that were RNA sequenced in the current study.

The study was approved by the Department of Veterans Affairs VA Long Beach Healthcare Systems Institutional Review Board (IRB# 1254), by the Human Subjects Committee, Los Angeles Biomedical Research Institute (Project No. 20607-0), University of Southern California Health Sciences Campus Institutional Review Board (Project # HS-13-00815), and by the University of California, Irvine Institutional Review Board, HS #2016-3064. All participants signed written consents prior to providing biospecimens.

For information about the independent RNA-seq liver tissue dataset used for external validation, please refer to GSE142530 (1).

a. Inclusion and Exclusion Criteria:

Alcohol-associated Liver Disease (AH, AC) Donors:

Common Inclusion Criteria: History of chronic alcohol consumption sufficient to cause liver damage. Generally, this is considered to be >40 g/day for women and >60 g/day for men, for many years.

Common Exclusion Criteria: Liver disease significantly caused by hemochromatosis, autoimmune liver disease, Wilson disease, NAFLD, hepatitis C, or hepatitis B.

Specific to Alcohol-Associated Hepatitis Donors (AH):

Inclusion Criteria: A clinical diagnosis of possible alcoholic hepatitis. Serum total bilirubin >3 mg/dL.

Specific to Alcohol-Associated Liver Cirrhosis Donor (AC):

Inclusion Criteria: This group contained both abstinent and recently drinking alcohol associated cirrhosis. Inclusion Criteria for Abstinent donors: Abstinent (consumption of less than one standard drink*/week) during the 6 months prior to enrollment. Inclusion Criteria for Recently drinking donors: Heavy alcohol use until recently (stopped/reduced alcohol use within past 60 days). For the current study, both groups were combined into a single group for analysis.

Non-Alcohol-Associated Fatty Liver Disease Donors:

Inclusion Criteria: A clinical diagnosis of non-alcoholic fatty liver disease (NAFLD) with at least two of the following criteria: a) A history of diabetes mellitus or use of medicines to treat diabetes (e.g., metformin, insulin, etc.) b) Liver biopsy consistent with NAFLD or NASH c) BMI>30 d) Fasting triglycerides >250 mg/dL or receiving treatment for high triglycerides e) CT or MRI imaging consistent with NAFLD ALT >50 IU/ml at baseline. Abstinent (consumption of less than one standard drink*/week) during the 6 months prior to enrollment.

Exclusion Criteria: Liver disease caused by hemochromatosis, autoimmune liver disease, Wilson disease, hepatitis C, or hepatitis B. Participants currently receiving treatment for NAFLD.

Chronic Hepatitis C Donors:

Inclusion Criteria: Chronic hepatitis C diagnosis. Evidence of cirrhosis based on at least one of the following criteria: a) Fibroscan stiffness >12.5 kPa b) Liver biopsy showing Metavir F3 or F4 or Ishak fibrosis stage 4, 5, or 6 c) Nodular liver on ultrasound, CT or MRI d) FIB-4 score >3.25 e) Platelet count <150,000 /mm^3. Abstinent (consumption of less than one standard drink*/week) during the 6 months prior to enrollment.

Exclusion Criteria: Clinical evidence for NAFLD or laboratory evidence of hemochromatosis, autoimmune liver disease, Wilson disease, or hepatitis B. Has received or currently receiving treatment for HCV infection.

Healthy Donors:

Inclusion Criteria: AUDIT-C scores of <4 for men and <3 for women (signifying no alcohol misuse). Abstinent (consumption of less than one standard drink*/week) during the 6 months prior to enrollment.

Exclusion Criteria: Clinical history or laboratory evidence of liver disease including alcoholic liver disease, NAFLD, hemochromatosis, alcoholic hepatitis, autoimmune liver disease, Wilson disease, hepatitis C, or hepatitis B. BMI>32. Any of the following laboratory abnormalities within 90 days prior to signing the consent. - Creatinine: >1.5 mg/dL; - Hemoglobin: <12 g/dL; Total bilirubin: >1.5 mg/dL; - AST: >40 IU/mL; - ALT: >40 IU/mL.

b. Reference Genome:

To determine if the reference genome influenced our results, gene expression analyses were performed using both the hg19 (GRCh37 assembly) and hg38 (GRCh38 assembly) human reference genomes, downloaded from the UCSC Genome Browser. In both cases, chrM was not included in the assembly.

c. Gene Annotations:

For each reference genome, we performed the gene expression analyses using four distinct sets of gene annotations for comparison purposes. In particular, we used the following four versions of the gene annotations: 1) RefSeq from the UCSC Genome Browser (Dec 2017); 2) GENCODE release 28 (Apr 2018); 3) Ensembl release 91 (Dec 2017); and 4) a merged set of gene annotations curated from these versions of RefSeq, GENCODE, and Ensembl annotations.

d. Short-read alignment to reference genome and transcriptome:

The filtered and decontaminated reads were aligned to the reference genome and transcriptome for each of the 8 combinations of reference genome and gene annotations described in the previous sections. Three short-read aligners were used during this step for comparison purposes: 1) TopHat release 2.1.1. (2) in combination with Bowtie2 2.3.4.1 (3) with default settings (TUXEDO); 2) HiSat2 2.1.0 (4) with default settings (HISAT2); and 3) STAR 2.6.0 (5) with default settings (STARCQ).

e. Sample Sequencing:

RNA was isolated from the cell pellets and liver tissue according to total RNA extraction kit instructions (Qiagen RNAeasy kit). Total RNA was monitored for quality control using the Agilent Bioanalyzer Nano RNA chip and Nanodrop absorbance ratios for 260/280nm and 260/230nm. Library construction was performed according to the Illumina TruSeq mRNA stranded protocol.

All samples included in this study were RNA sequenced on an Illumina platform by the Genomics High-Throughput Facility (GHTF) at the University of California, Irvine (UCI), except for one healthy liver sample for which the sequencing data was directly downloaded from the European Bioinformatics Institute (EBI) ArrayExpress database (accession number E-MTAB-1733) (6). The number of paired or single reads per sample was approximately 140M before filtering and decontamination.

f. Read Trimming & Quality Filters:

The sequencing reads in each dataset were first filtered to remove low quality reads and trim all 3' regions matching with the Illumina sequencing primers or 5' regions with skewed base distributions. The following is each step of the protocol:

1) Sequencing primers attached to short inserts were removed using Trimmomatic release 0.38 (7).

2) Reads not passing the standard Illumina quality tests as reported in the header line of each entry in a FastQ file were removed.

3) Reads with any number of uncalled bases (N) were discarded with a few exceptions for some positions observed with more than 3% uncalled bases in the corresponding dataset. In these cases, 1 uncalled base max was allowed in the reads.

4) Reads were trimmed on the 5' end to remove the positions observed with highly variable base distribution following this protocol. First, the standard deviation of the base distribution was calculated for each position. Second, the mean standard deviation was calculated for every contiguous set of 5 positions in the reads. Finally, positions on the 5' end were trimmed as long as the mean standard deviation of the first 5 bases in the reads was greater than twice the lowest mean standard deviation observed in the reads during the previous step. In most cases, 5 to 15 positions were trimmed on the 5' end of the reads in each dataset following this protocol.

5) Reads were trimmed on the 3' end using a fixed number of positions = 1 except for three datasets for which between 25 and 30 positions were trimmed on the 3' end to account for sequencing issues specific to these samples.

6) Reads shorter than 60 bases after trimming were discarded from the datasets.

7) A min PHRED quality score per position of 20 was used to further filter the reads with several positions allowed below this threshold ranging from 2 to 10 such that the lowest number of exceptions not discarding more than 20% of the reads was selected. No more than 10 exceptions were allowed during this step.

8) A min average PHRED quality score per read was used as an additional filter, with a value ranging from 24 to 36 such that the highest mean quality score not discarding more than 20% of the reads was selected.

On average, 9.62% of the original reads were discarded during this step and 15.43% of the paired reads were orphaned. The mean PHRED quality score of the remaining reads was approximately 40.

g. Sample Decontamination:

The remaining quality-filtered and trimmed reads for each dataset were then further filtered to remove possible contaminants in each sample such as PhiX control reads or bacterial contamination. In addition, both the human mitochondrial genome and ribosomal DNA/RNA sequences were treated as contaminants during this step due to highly variable quantities of these reads in the various datasets generated during the experiment, ranging from a few percent of the reads in most cases to about 80% of the reads for some highly contaminated samples. Such differences significantly impact gene expression results, notably the FPKM values calculated during the next step of our analysis, so this bias was removed prior to the gene expression analysis by simply removing the corresponding reads from all the datasets. This step was performed using the following protocol. The reads were first aligned to all the contaminant sequences using Bowtie2 release 2.3.4.1. Any read successfully located on any contaminant sequence was then aligned against the human transcriptome using the same short-read aligner. Reads not matched with any known human transcript (i.e. only matched to a contaminant sequence) and reads with a better alignment score to a contaminant sequence than the best alignment score with the human transcriptome were discarded, the remaining reads were kept for

the rest of the analysis. On average, approximately 115M paired and single reads were left per sample and used for the gene expression analysis described in the next sections.

h. Normalized RNA-seq counts before and after application of log transformation:

Fig. S1 shows the relationship between variance and mean of the RNA-seq counts for the PBMC Alcoholic Hepatitis (AH) samples. It can be readily observed that there is a linear relationship between the two. This is usually an undesirable property for machine learning (ML) algorithms. After transforming the RNA-seq counts using the ln(1+count) formula we can observe that there is no longer a linear relationship between mean and variance of the RNA-seq counts (Fig. S2). Moreover, the variance and mean values are much smaller and more consistent. The log transformation improved the classification accuracy by approximately 5% for logistic regression classifier when tested with LV 2-Way dataset. Therefore, we used log transformed counts with all four of our datasets.

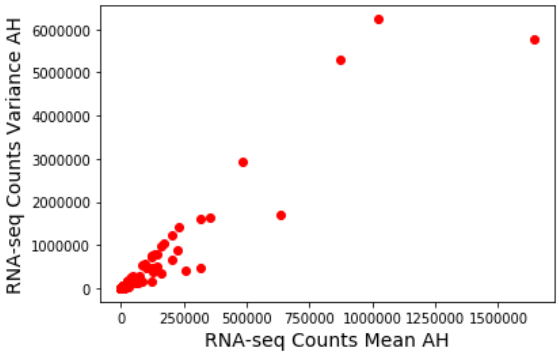

Figure S1: Geometrically Normalized RNA-seq counts

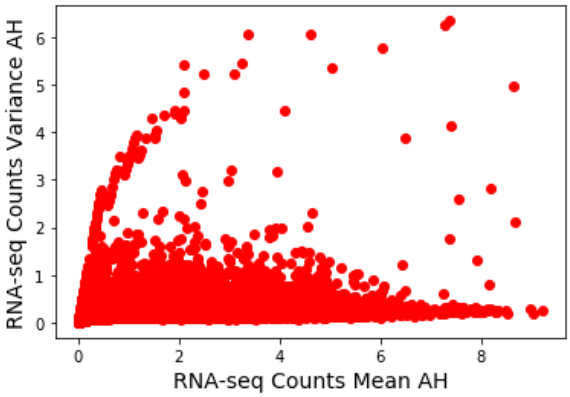

Figure S2: RNA-seq counts above after being transformed using ln(1+count) formula

i. Alignment Pipeline Selection:

We compared the results from 24 different alignment pipelines using the PBMC AH and CT conditions. These 24 pipelines were formed using two human reference genomes (hg19, hg38), four different genome annotations (Curated, Ensembl, Gencode, and Refflat), and three different genome aligners (Tuxedo, Hisat2, and Starcq). The PBMC AH and CT counts from each of the genome pipelines were then utilized in our classification and feature selection pipeline using differential expression feature selection only. No alignment pipeline proved to be advantageous over others according to classification performance. We then compared the alignment pipelines according to the in silico biological validation of the selected genes utilizing

Ingenuity Pathway Analysis (IPA) software (Fig. S3). Ensembl annotation resulted in the most biologically relevant genes according to IPA. The choice of human reference genome and aligner did not seem significant and therefore we decided to utilize the more recent hg38 reference genome and Starcq aligner along with Ensembl annotation for our four datasets.

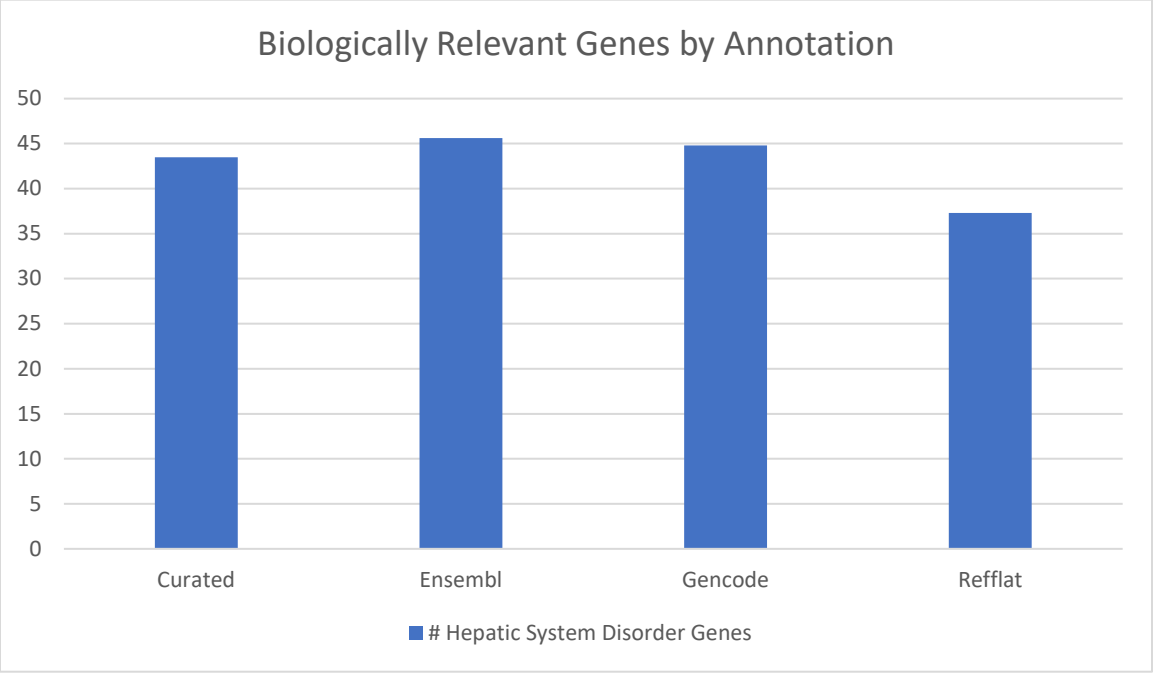

Figure S3: Comparison of annotations by number of hepatic system disorder related genes using the PBMC 2-Way dataset

j. Nested Cross-Validation Setup:

We utilized nested cross-validation to attain the estimates of classification performance for various feature selection (FS) strategies, classifiers, and feature sizes within our data. The best feature (gene) sets selected for each of the four datasets were then validated in the independent test set. The nested cross-validation was implemented in the standard configuration with k = 5 in both the inner and outer loops. The outer loop was used for model evaluation (i.e., classification performance), while the inner loop was used for model selection (i.e., hyper-parameter tuning). The feature selection was done within both inner and outer loops. That is FS was done for each training set in inner and outer loops. This means that effectively there were 30

training sets (25 in inner loop, 5 in outer loop) as part of a single nested cross-validation execution. Feature selection occurred for each of these training sets.

Since one of our classification strategies relied on differential expression as computed by Cuffdiff (8), the feature selection process within nested cross validation was time consuming. A single Cuffdiff analysis could require anywhere from 30 minutes to 5 hours depending on the number of samples. In order to keep runtime reasonable, all folds were pre-defined, and only a single splitting of samples into folds (for both inner and outer loops) was used within each dataset. Typically, multiple repeated data splits of samples to folds are desired to obtain best estimate of classifier's performance. However, due to Cuffdiff's large runtime performing multiple data splits proved to be prohibitive.

Cuffdiff produced three key files: genes.read_group_tracking containing the normalized RNA-seq counts, gene_exp.diff containing the differential expression analysis data over all input samples, and the read_groups.info containing the names of input CXB files (samples). CXB files (samples).

k. Feature Selection Strategies:

Feature selection for gene expression data was essential, since our datasets contained tens of thousands of genes, far more than the number of samples. ML algorithms typically perform very poorly when given significantly more features than samples. Initially, we briefly compared three types of feature selection strategies within our study: filter feature selection via differential expression (DE) and information gain (IG) algorithms, hybrid feature selection (filter + wrapper), and embedded feature selection via random forest (RF) algorithm. All three strategies resulted in similar classification performance (Table S1). However, the filter feature selection had much lower runtime then the hybrid and embedded FS strategies. Additionally, we had

concerns that both hybrid and embedded feature selection strategies were prone to overfitting based on past analyses. Therefore, we decided to use only the filter feature selection strategies within the remainder of the study.

Table S1: Comparison of three feature selection architectures: filter, hybrid, and embedded using LV 2-Way dataset

Feature Size	Filter: DE	Hybrid: DE + SFS	Embedded: RF
2	0.91	0.88	0.95
3	0.95	0.97	0.91
4	0.97	0.88	0.95
5	0.97	0.95	0.95
10	0.97	0.97	0.95
15	0.97	1.0	0.97
20	1.0	0.97	1.0

*Used LR classifier with Filter and Hybrid architectures. Used Union filter with threshold of 3.0 with all architectures.

The hybrid feature selection was done by pairing the filter feature selection strategies (DE, IG) with forward sequential feature selection algorithm (forward-SFS) as described in scikit-learn documentation. The features were first selected by filter feature selection and then halved using forward-SFS. The forward-SFS was performed using logistic regression classifier.

The embedded feature selection was performed using random forest. Specifically, the RF classifier was simply given data with all features included. We then extracted the feature rankings from the RF models to determine which features it valued the most.

l. Differential Expression (DE) Feature Selection:

For every training set all pairwise comparisons (within gene_exp.diff files) were filtered by normalized FPKM (> 1.0) and q-values (< 0.05). All of the genes belonging to each pairwise comparison were then sorted by absolute log2(fold change) value, and the top gene for each pairwise comparison was taken. If that gene was not already in the top genes list, the gene was added to the list. The algorithm continued to cycle through the pairwise comparisons until the desired number of genes was reached. This procedure was used for all the datasets. The best features for each training set were then stored in text files.

Other DE feature selection approaches were implemented and tested by us as well. However, we found that pairwise DE selection was best performer since other DE feature selection approaches, we tested were too easily biased by the most strongly differentially expressed pairwise comparisons.

m. Information Gain (IG) Feature selection:

For every training set, the genes within normalized RNA-seq counts were ranked using the scikit-learn's mutual_info_classif function.

n. Feature Sizes:

We refer to the number of features selected during filter feature selection as "feature size". The feature sizes used with DE & IG feature selection were: 2, 3, 4, 5, 10, 15, 20, 25, and, 50 for LV 2-Way dataset and 10, 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 for the other three datasets. The feature sizes denote the number of features selected within each training set. We found during preliminary testing that we required at least 5-10 features per training set to attain reasonable classification performance and that we generally did not see benefit in using more than 500 features per training set. The maximum feature size was also influenced by our

power size calculation (that is number of significantly differentially expressed genes within our datasets).

o. Performance Metrics:

Several different ML performance metrics were evaluated for use in this project including overall accuracy, per-class accuracy, balanced accuracy, confusion matrices, Matthews Correlation Coefficient (MCC), and F1-score. Balanced accuracy, MCC, and F1-score attempt to account for class sizes when evaluating performance, while the confusion matrices provide information about both class sizes and also per-class accuracies. Therefore, we chiefly reported our classification performance in the form of confusion matrices.

p. Machine Learning Classifiers.

We initially tested 7 classifiers: Adaptive Boosting Algorithm (ADA) using decision tree, decision tree (DT), gaussian naïve bayes (GNB), logistic regression (LR), k nearest neighbors (kNN), support vector machine (SVM), and random forest (RF). Based upon comparison of their performance and run time, we narrowed down our selection to LR, kNN, and SVM only. Table S2 demonstrates the performance of all classifiers using the LV 2-Way dataset with filter feature selection, with the exception of RF, which belongs to embedded feature selection architecture.

Table S2: Comparison of six ML classifiers in LV 2-Way dataset

	ADA	DT	GNB	kNN	LR	SVM
2	0.77	0.79	0.84	0.82	0.82	0.82
3	0.92	0.79	0.92	0.86	0.9	0.86
4	0.9	0.93	0.94	0.97	0.88	0.95
5	0.97	0.85	0.95	0.95	0.97	0.92
10	0.95	0.95	0.97	0.97	0.97	0.95
15	0.97	0.93	0.97	0.97	0.97	0.95
20	0.93	0.93	0.97	0.97	0.97	0.95

*Classifiers in Table S2 were used in conjunction with DE feature selection and Intersection filter with threshold of 3.0.

q. Sample Size Calculation:

There are few established guidelines for calculating sample size for RNA-seq experiments. Recommendations vary from having at least 3, 6, or 12 biological replicates per condition depending on sequencing depth and fold change cutoff. All selected conditions within our PBMC dataset contain more than 12 biological replicates. All selected conditions within the liver dataset contain more than 6 biological replicates. The average number of reads per sample is approximately 115 million after filtering and decontamination. We utilized the RNASeqPower R package (9) to establish the best fold change cutoff and the expected number of significantly differentially expressed genes (SDEGs) for our LV 2-Way dataset. According to the output of the package for our dataset, there are approximately 450 SDEGs in the LV 2-Way dataset. We assumed that the number of SDEGs is approximately similar across all datasets. This helped us to determine the upper bound on useful feature sizes.

r. Enrichr Libraries:

The genes selected during feature selection were computationally evaluated using gene enrichment analysis via Enrichr (10) with pathway, tissue, and disease Enrichr libraries listed below. Custom code was written using regular expressions to match: a) immune system pathways; b) cell types that comprise blood and liver tissues; c) diseases included the conditions within this study (AH, AC, NAFLD, HCV) along with several other liver and blood disorders.

In order to attain the top three Enrichr hit tables (Tables S18, S21, S24, and S27) we performed the following steps. Enrichr hits for the best gene sets, after matching using the regular expressions, were sorted by adjusted p-value with a cutoff of 0.05. We removed entries with redundant term names or genes. We then displayed up to three top entries for each category: pathway, tissue, disease.

Enrichr Libraries used:

Pathways: 'BioPlanet_2019', 'WikiPathways_2019_Human', 'KEGG_2019_Human', 'GO_Biological_Process_2018'.

Tissues: 'ARCHS4_Tissues', 'Human_Gene_Atlas'.

Diseases: 'Disease_Perturbations_from_GEO_up', 'Disease_Perturbations_from_GEO_down'.

s. Regular Expression (Regex) Patterns for Enrichr Libraries:

The regular expression (regex) patterns used for filtering the results returned by Enrichr are listed below.

Disease Regex:

'hepa|liver|cirrhosis|NAFLD|liver fibrosis|NASH|steatohepatitis|HCV|alcohol|sepsis|septic shock|hypercholesterolemia|hyperlipidemia|obesity'

Tissue Regex:

'Blood|Macrophage|Erythro|Platelet|Basophil|Neutrophil|Eosinophil|Cytokine|Tumor Necrosis Factor|Monocyte|Lymphocyte|Granulocyte|Dendritic|Megakaryocyte|T Cell|B Cell|NK Cell|Toll-like receptor|Fc receptor|Liver|Hepatocyte|Stellate|Kupffer|Sinusoidal Endothelial Cells|CD34+|Natural Killer Cell|PBMC|Tcell|Bcell|lymphoblast|CD8+|CD19+|CD4+|CD71+|Omentum'

Pathway Regex:

'Interferon|Immun|Interleukin|Prolactin|Complement|Chemokine|Oncostatin M|Rejection|Inflamma|IL1|IL-

|selenium|osteopontin|circulation|coagulation|clotting|biosynthesis|degradation|cholesterol|lipid|TNF|steroid|metal ion|heme|metallo|CXCR|LDL|Phagocytosis|metabolism|TYROBP|AP-1|'

Additionally, the pathway regex included all of the disease and tissue terms.

t. Impact of Outlier Gene (Feature) Removal – Variance, Intersection, and Union Filtering:

RNA-seq serves as a proxy for the level of gene expression in a biological sample. One challenge with interpretation of RNA-seq output, however, involves expression of non-coding genes that were presumed to be removed via poly(A)-selection. It is also common to observe genes with aberrant expression that poorly distinguish between the study conditions, thereby hindering classification performance. As an example, in Fig. S4 the RNA-seq counts of the LV 2-Way dataset are visualized as a heatmap. The genes selected were chosen by differential expression analysis. We observed that genes such as SNHG25, RNY1, RNU6ATAC, and UBA3 are all highly variant. Moreover, three of these are non-coding. The Fig. S5 shows the same dataset after genes were filtered using the Union filter with threshold of 3.0. In this example, the genes removed were replaced with other top DE genes such that the total number of genes remained the same. The latter heatmap is much more visually distinct between the AH and CT conditions.

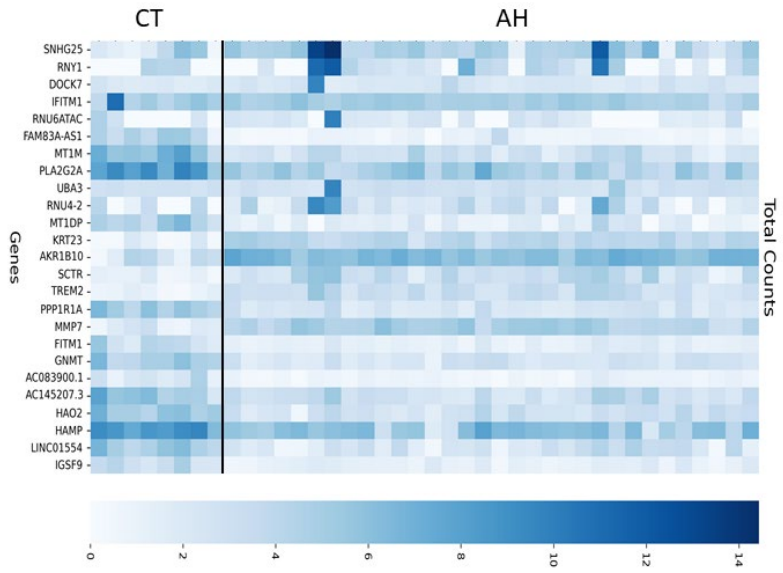

Figure S4: LV 2-Way RNA-seq counts – no filter, 25 genes total

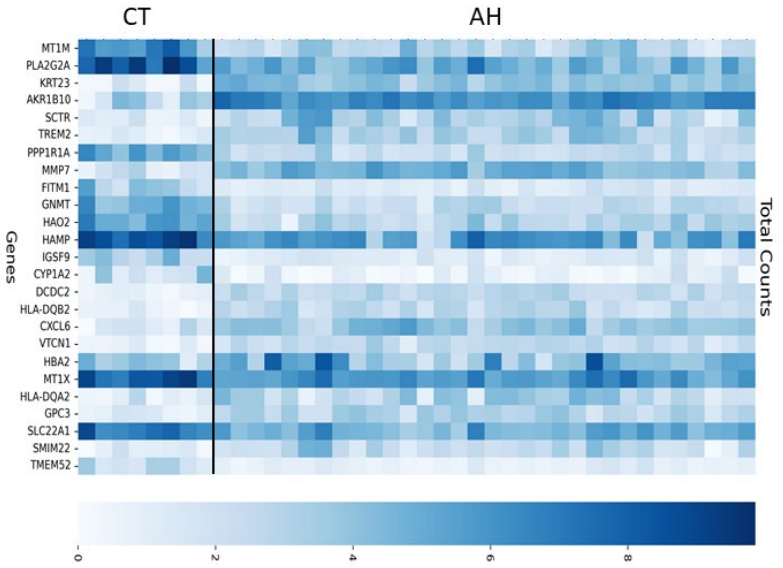

Figure S5: LV 2-Way RNA-seq counts – Union filter with threshold of 3.0, 25 genes total

Based on our observations and explanation above, we developed three strategies for removing undesirable genes: Variance, Intersection, and Union filtering. Variance filtering was implemented by removing genes in which the RNA-seq counts for at least one sample were

further than a standard deviation multiplied by the threshold from the mean in any of the conditions (AH, CT, etc.). Throughout the study, we used three threshold values: 2.5, 3.0, and 3.5. Lower thresholds resulted in more genes being eliminated, while higher thresholds resulted in less genes being eliminated. The filtered-out genes were not used in the subsequent feature selection process. The Union filter built upon the Variance filter by removing all genes that were either highly variant (as defined above) *or* non-coding as determined by ENSEMBL database's gene "biotype" column. The Intersection filter was similar to the Union filter, except that only the genes that were both highly variant *and* non-coding were removed. In addition to improving the odds of successful classification, the outlier feature filtering was also found to improve in silico biological validation of identified gene signatures, since protein coding genes are more extensively annotated than non-coding ones. These three filters also removed all genes whose counts were mostly zeroes across all samples.

We applied the three filter procedures (each paired with three possible threshold values of 2.5, 3.0, and 3.5, for a total of 9 filter configurations) to the LV 2-Way dataset. Tables S3 and S4 show the impact of each filtering strategy (with threshold of 3.0) on the overall classification accuracy and biological relevance of LV 2-Way dataset. The biological relevance was determined by performing gene enrichment analysis using Enrichr. The in silico biological validation results are reported as follows: pathway hits / tissue hits / disease hits. In the example below, the genes removed by the outlier filtering strategy were replaced with the next highly-ranked DE genes. The classification accuracies were attained using nested cross-validation. The feature size in the Table S3 is the feature size within each individual training set of the nested cross-validation. Before commencing with in silico biological validation, we merged the gene sets produced by training sets in the outer loop of nested cross-validation. The features sizes in

Table S4 are listed with the following notation: feature size of training set in nested cross validation – feature size of merged gene set using first filter procedure / feature size of merged gene set using second filter procedure / and so on. For example, in Table S4, the numbers in the first column are the feature size of the training set in nested cross validation, followed by the number of genes in the four filter procedures: No Filter/Variance 3.0/Union 3.0/Intersection3.0. Based on further analysis in the LV 3-Way dataset, we decided to use Intersection and Union filters with thresholds of 2.5, 3.0, and 3.5 for all four of our datasets. The gene sets of size 0 could not be analyzed using Enrichr. Therefore, cells corresponding to empty gene sets in Table S4 are labeled as NA.

Table S3: The impact of outlier feature removal strategies on classification accuracies of k nearest neighbors (kNN) classifier and DE feature selection within LV 2-Way dataset

Feature Size	No Filter	Variance 3.0	Union 3.0	Intersection 3.0
2	0.8	0.95	0.93	0.82
3	0.76	0.97	0.95	0.86
4	0.86	0.95	0.97	0.97
5	0.89	0.95	0.97	0.95
10	0.94	0.97	0.95	0.97
15	0.97	0.97	0.95	0.97
20	0.97	0.97	0.97	0.97
25	0.97	0.97	1	0.97
50	0.97	0.97	0.95	0.97

Table S4: The impact of outlier feature removal strategies on in silico biological relevancy of LV 2-Way gene sets with DE feature selection

Feature Sizes	No Filter	Variance 3.0	Union 3.0	Intersection 3.0
2 – 0/0/0/1	NA	NA	NA	1/0/0
3 – 1/0/0/2	1/0/0	NA	NA	16/0/12
4 – 1/0/1/2	1/0/0	NA	23/0/8	16/0/12
5 – 2/0/3/3	16/0/12	NA	30/2/4	18/0/0
10 – 5/7/9/9	34/0/3	31/2/5	28/3/6	22/1/3
15 – 11/9/9/12	13/1/4	28/3/6	28/3/6	17/2/5
20 – 13/11/13/13	1/1/5	19/3/5	1/4/5	18/3/7
25 – 16/14/14/18	1/3/7	1/4/5	1/4/6	1/3/7

50 – 31/30/34/33	1/4/7	0/4/5	1/5/12	1/3/7

*Numbers in Filter columns denote number of Enrichr hits in: pathway/tissue/disease. Higher scores for the number of pathway/tissue/disease hits suggest that the genes were more biologically relevant. NA represents the case where there were no genes in the set for input to Enrichr.

u. Summary of Methods:

While we experimented with a large number of methods with the LV 2-Way (and sometimes LV 3-Way) datasets, we were able to reduce the set of methods applied to our four datasets as shown in Table S5.

Table S5: The methods used within the study initially and as part of final configuration

Methods	Feature Selection	Outlier Feature Removal	ML Classifiers
Briefly Examined	Filter (DE, IG), Hybrid (Filter + Wrapper), Embedded (RF).	Variance, Intersection, and Union filtering. (Standard deviation thresholds: 2.5, 3.0, 3.5)	LR, kNN, SVM, GNB, DT, ADA, RF.
Final Configuration	Filter (DE, IG).	Intersection and Union filtering. (Standard deviation thresholds: 2.5, 3.0, 3.5)	LR, kNN, SVM.

Therefore, the final analysis included the following method configurations for each of the four datasets: 2 feature selection strategies (DE, IG), 2 outlier feature removal strategies (Intersection, Union) each paired with three different thresholds (2.5, 3.0, 3.5), and 3 ML classifiers (LR, kNN, SVM). This resulted in a total of 36 configurations. For each configuration there was also a range of possible feature sizes as described in feature size section above. The

nested cross-validation ML metrics were recorded for each of these configurations, for each feature size.

v. Candidate Gene Sets:

Since one of the overarching goals of this study was to identify characteristic gene expression signatures to diagnose liver disease using liver tissue and PBMC RNA-seq data, the next step of our pipeline involved selecting the best gene sets for our datasets. Within nested cross-validation, feature selection was performed for every training set in both inner and outer loops, resulting in 30 total gene sets (5 in outer, 25 in inner) for each feature size. The gene sets selected in the inner loops are not relevant, since the inner loop was only used for hyper-parameter tuning. Therefore, we developed a method of merging the gene sets produced for each of the outer loop training sets. The strategy used was as follows: if a given gene appeared in N out of the 5 (k = 5 in outer loop) gene sets it was added to the merged gene set. After examining the results, we determined that N = 4 and N = 5 yielded our best results. The candidate gene sets were analyzed using Enrichr to establish their biological relevancies. The classification accuracy attained from the associated instance of the nested cross-validation of each candidate gene set was also examined.

w. Best Gene Set Selection:

From the large collection of candidate gene sets attained by running the 36 different method configurations for each dataset across multiple feature sizes, we used the following strategy to select a single best candidate gene set for each of the four datasets. This process involved the evaluation of a combination of candidate gene set's size, classification performance, and biological relevancy metrics. The algorithm for picking best gene sets is described below.

1) The candidate gene set size was restricted between 5 (genes per pairwise comparison) to 100 total genes, if possible. Gene set sizes of between 100 and 200 were also considered, if suitable performance was not observed in candidate gene sets below 100 genes. The LV 2-Way dataset contains 1 pairwise comparison, LV 3-Way dataset contains 3 pairwise comparisons, and 5-Way datasets contain 10 pairwise comparisons. Therefore, the candidate gene set sizes, using the range guidelines above for each dataset, are as follows: 5-100 genes for LV 2-Way, 15-100 genes for LV 3-Way, and 50-100 genes for LV 5-Way and PBMC 5-Way. The gene set size guidelines were developed to minimize the chance of either under- or overfitting.
2) Biological relevancy as indicated by Enrichr was prioritized slightly higher than the classification accuracy. That is, gene sets with highest number of pathway, tissue, and disease hits were examined in detail first. Gene sets were only considered if they included at least 10 pathway hits, 1 tissue hit, and 3 disease hits. The tissue, pathway, and disease hits were examined to verify that they were appropriate and relevant to the disease groups.
3) Total and per-class classification accuracies were considered after the in silico biological relevancy. In general, only gene sets within 10% of the best recorded performance (for a given dataset) were considered.

Once a single gene set that best satisfied all 3 criteria was selected, it was used to generate the heatmaps, confusion matrices, and pathway analysis. The liver tissue gene sets selected from our data set were evaluated with the independent validation dataset.

x. Additional in silico biological validation methods.

In order to further analyze and validate our gene sets we performed additional annotation enrichment analysis using Ingenuity Pathway Analysis (IPA), Gene Set Enrichment Analysis (GSEAPreranked), and blood transcription module analysis (BloodGen3Module) tools (11, 12, 13). Since these tools use different knowledgebases and statistical methods, a more complete view of biological annotation is produced. There was generally a large overlap between the results of the different annotation enrichment tools. Additionally, the different visualizations offered by each tool proved to be complementary.

Ingenuity Pathway Analysis (IPA):

The best gene sets for LV 5-Way and PBMC 5-Way datasets (as shown in Table 3 of main text) were analyzed using IPA. The analysis was performed using only the best gene sets, i.e., 75 genes for PBMC 5-Way, and 39 genes for LV 5-Way. A fold change cutoff of 1.0 was used for PBMC 5-Way, and cutoff of 1.5 was used for LV 5-Way during the analysis. The top enriched pathways were identified in each dataset on per pairwise comparison group basis. The top pathways for each pairwise comparison were sorted using p-value and organized into Tables S29 and S30. The dot plots (Figs. S10 and S11) were generated using the pathways and p-values from the tables, with pathways on the y-axis and pairwise comparisons on x-axis. The dots are color-coded by p-value significance, with blue dots representing lower significance and red representing higher significance.

Gene Set Enrichment Analysis (GSEAPreranked):

The GSEAPreranked analysis was performed using GSEA software version 4.2.3 with only the best gene sets identified during PBMC 5-Way (75 genes) and LV 5-Way (39 genes) analysis. The analysis struggled to attain significant p-values with such a small number of genes. The following gene set libraries were used: c2.reactome, c2.wikipathways, c2.kegg, c5.GO:

biological processes, c8.all (cell type signatures) v 7.5.1. The required parameters were set as follows: number of permutations: 1000; minimum set size: 10; and maximum set size: 1000. The ranking metric used was log2(FC). Similar to IPA and Enrichr, the most significantly enriched pathways involved immune system and inflammation processes.

Blood Transcription Module (BTM) Analysis (BloodGen3Module):

In order to obtain a more complete annotation of the blood-based 5-Way PBMC best gene set, blood transcription module (BTM) analysis was performed using R BloodGen3Module version 1.4.0 package. Only the best gene set comprised of 75 genes from the PBMC 5-way analysis were input into the BloodGen3Module software. This analysis resulted in the differential module response status of 39 different BTM modules for each of the pairwise comparisons (Table S32). Cells in shades of red are upregulated for the condition listed first, and shades of green if downregulated for the condition listed first.

y. Codebase:

Github: https://github.com/staslist/Liver-Disease-Diagnostic The repository contains the code used to perform the analysis. Directories and sample names have been removed from the codebase.

2. SUPPLEMENTARY RESULTS

a. Best Gene Sets Fold Changes.

Listed below are fold changes corresponding to the best gene sets for LV 5-Way and PBMC 5-Way datasets as provided in the main text. The fold change (FC) is computed by taking the log2(q1/q2) wherein q1 is the first condition listed in the q1 v q2 format and q2 is the second listed condition. The bolded entries are significant according to false discovery rate (q-value) metric. For brevity only the pairwise comparisons involving controls (CT) are shown.

LV 5-Way.

Table S6: LV 5-Way best gene set directionality table

	AH v CT		AC v CT		NF v CT		HP v CT	
	FC	Q-Value	FC	Q-Value	FC	Q-Value	FC	Q-Value
AKR1B10	**4.635**	9.89E-04	**2.15**	7.21E-03	**3.742**	9.89E-04	**3.008**	9.89E-04
ATF3	**-2.278**	9.89E-04	-1.147	1.74E-01	0.715	3.62E-01	0.297	7.30E-01
CYP2A6 (includes others)	**2.708**	7.91E-03	**3.206**	9.89E-04	**4.847**	9.89E-04	**4.463**	9.89E-04
CYP2B6	**-2.177**	9.89E-04	0.672	5.37E-01	1.343	8.08E-02	1.277	1.02E-01
DOCK7	**6.367**	9.89E-04	**4.327**	5.06E-03	0.444	7.21E-01	0.931	3.95E-01
DUSP1	**-1.86**	9.89E-04	0.455	5.95E-01	0.95	8.27E-02	**1.165**	3.41E-02
EPS8L1	**2.704**	9.89E-04	**3.202**	9.89E-04	**3.056**	9.89E-04	**3.355**	9.89E-04
GADD45B	**-2.835**	9.89E-04	-0.819	2.18E-01	0.292	6.80E-01	-0.052	9.56E-01
GADD45G	**-2.428**	9.89E-04	-0.559	4.28E-01	**1.182**	1.89E-03	**1.406**	2.73E-03
GSTA2	-2.796	2.92E-01	-0.485	7.60E-01	0.617	7.36E-01	0.983	5.86E-01
HBA1/HBA2	**3.437**	1.12E-02	0.104	9.20E-01	**2.493**	5.80E-03	**2.885**	9.89E-04
IFI27	1.09	5.80E-02	0.261	7.74E-01	0.992	1.12E-01	**4.941**	9.89E-04
IFI44L	1.097	6.34E-02	**2.944**	9.89E-04	-0.066	9.64E-01	**2.499**	9.89E-04
IFI6	0.628	2.45E-01	0.147	8.76E-01	0.534	4.65E-01	**4.075**	9.89E-04
IFITM1	**-5.535**	9.89E-04	**-5.466**	9.89E-04	**-6.084**	9.89E-04	-0.262	9.15E-01
IGFBP1	-0.793	2.02E-01	1.37	1.13E-01	**1.948**	2.73E-03	**2.047**	1.89E-03
IGHV3-23	0.492	6.21E-01	1.72	9.36E-02	-1.422	1.61E-01	0.666	4.98E-01
ISG15	0.695	1.55E-01	-0.398	5.87E-01	**1.365**	4.31E-03	**4.392**	9.89E-04
KRT23	**4.651**	9.89E-04	**3.702**	9.90E-03	**3.253**	8.59E-03	**4.026**	5.80E-03
KRT7	**2.401**	9.89E-04	**2.376**	9.89E-04	**3.173**	9.89E-04	**3.1**	9.89E-04
LINC01554	**-3.797**	9.89E-04	**-2.641**	9.89E-04	**-3.532**	9.89E-04	**-4.426**	9.89E-04
MMP7	**4.246**	9.89E-04	**3.479**	9.89E-04	**2.503**	9.89E-04	**2.085**	6.51E-03
MT1G	**-2.648**	9.89E-04	**-3.919**	2.73E-03	0.038	9.78E-01	-1.088	4.20E-01
MT1M	**-5.155**	9.89E-04	**-5.243**	9.89E-04	**-2.021**	1.89E-03	**-3.182**	9.89E-04
MUC1	0.214	9.05E-01	0.877	6.96E-01	**3.22**	2.44E-02	**3.126**	3.74E-02
MUC6	**2.099**	9.89E-04	**4.06**	9.89E-04	**3.782**	9.89E-04	**3.62**	9.89E-04
NR4A1	**-1.789**	4.96E-02	1.587	1.61E-01	1.882	7.04E-02	**2.16**	3.97E-02
OASL	**1.637**	2.00E-02	0.447	6.96E-01	1.235	1.52E-01	**3.753**	9.89E-04
PLA2G2A	**-5.022**	9.89E-04	**-4.848**	9.89E-04	**-2.545**	2.58E-02	-1.763	1.83E-01
PPP1R1A	**-4.325**	9.89E-04	**-3.138**	9.89E-04	**-1.577**	7.91E-03	**-2.89**	9.89E-04
RGS1	-0.885	2.53E-01	**2.52**	3.53E-03	**2.093**	3.53E-03	**3.005**	9.89E-04
S100A8	0.228	7.57E-01	**-3.269**	9.89E-04	-0.617	4.07E-01	-0.997	8.69E-02
SAA2-SAA4	1.961	5.58E-01	-1.113	7.18E-01	2.711	5.42E-01	0.727	8.02E-01
SCTR	**4.567**	9.89E-04	**3.488**	2.54E-02	**4.217**	9.89E-04	**4.326**	9.89E-04
SERHL2	1.435	2.42E-01	1.65	3.31E-01	**3.623**	4.31E-03	**2.8**	3.86E-02
SLC2A3	0.246	8.18E-01	1.056	2.71E-01	**2.805**	9.89E-04	**2.905**	9.89E-04
SPINK1	-2.443	2.28E-01	**-4.313**	9.89E-04	**-2.898**	9.89E-04	**-4.821**	9.89E-04
SYT8	1.626	1.51E-01	**4.065**	9.89E-04	**4.128**	9.89E-04	**4.443**	9.89E-04

*Green shading highlights positive fold change (up-regulation), and red shading highlights negative fold change (down-regulation). Bolded entries are significant according to q-value.

PBMC 5-Way.

Table S7: PBMC 5-Way best gene set directionality table

	AH v CT		DAAA v CT		NF v CT		HP v CT	
	FC	Q-Value	FC	Q-Value	FC	Q-Value	FC	Q-Value
AHSP	**6.398**	1.48E-03	**3.446**	1.48E-03	**1.557**	1.48E-03	**2.137**	1.48E-03
ALAS2	**4.577**	1.48E-03	**3.697**	1.48E-03	**1.394**	1.58E-02	**2.351**	1.48E-03
ALPL	**2.742**	1.48E-03	-0.042	9.85E-01	-0.889	1.37E-01	0.759	2.64E-01
ANXA3	**2.434**	1.48E-03	-0.325	6.26E-01	-0.019	9.92E-01	-0.773	6.73E-02
AQP9	**2.335**	1.48E-03	**1.2**	1.48E-03	0.487	7.49E-02	**1.211**	1.48E-03
ATF7IP2	**-0.874**	2.14E-02	-0.612	1.78E-01	-0.086	9.58E-01	**-1.093**	1.20E-02
AZU1	**1.49**	1.48E-03	-0.047	9.79E-01	0.368	6.16E-01	-0.658	3.03E-01
BCAT1	**2.263**	1.48E-03	**1.378**	1.48E-03	**1.017**	2.72E-03	**0.881**	3.19E-02
C1QA	**2.994**	1.48E-03	**1.403**	1.48E-03	0.214	8.04E-01	**1.094**	1.48E-03
C1QB	**3.697**	1.48E-03	**1.786**	1.48E-03	0.17	8.95E-01	**1.633**	1.48E-03
CAMP	**1.389**	1.48E-03	-0.344	4.89E-01	-0.006	9.97E-01	**-0.874**	2.53E-02
CCR2	**1.36**	1.48E-03	**1.064**	1.48E-03	-0.202	7.65E-01	0.557	8.00E-02
CD180	0.521	7.33E-02	**0.668**	8.64E-03	-0.428	2.83E-01	0.551	1.40E-01
CEACAM3	**1.964**	1.48E-03	0.725	6.81E-02	0.259	8.23E-01	**1.222**	1.48E-03
CEACAM8	**2.11**	1.12E-02	-0.117	9.64E-01	0.293	8.66E-01	-1.189	2.85E-01
CHI3L1	**2.149**	1.48E-03	0.217	8.03E-01	0.046	9.77E-01	-0.065	9.66E-01
CRISP3	**1.838**	1.48E-03	0.051	9.70E-01	0.478	3.48E-01	-1.051	5.35E-02
CTSG	**1.616**	1.48E-03	-0.266	8.24E-01	0.345	6.78E-01	-0.433	6.66E-01
CXCL5	**-1.636**	1.48E-03	**-1.559**	1.48E-03	-0.051	9.69E-01	**-0.818**	1.79E-02
CXCR1	**1.249**	1.48E-03	0.163	8.34E-01	-0.268	6.91E-01	**1.079**	1.48E-03
DEFA1 (includes others)	0.615	9.67E-02	**-1.274**	1.48E-03	**-0.72**	2.47E-02	**-1.728**	1.48E-03
DEFA4	**2.227**	1.48E-03	-0.469	4.98E-01	0.546	1.77E-01	-0.873	2.12E-01
DSC2	**2.649**	1.48E-03	**1.505**	1.48E-03	0.246	7.98E-01	0.093	9.54E-01
DYSF	**2.356**	1.48E-03	**1.278**	1.48E-03	0.443	2.18E-01	**1.04**	1.48E-03
ELANE	**2.347**	1.48E-03	-0.002	9.99E-01	0.457	4.76E-01	-0.51	5.09E-01
FCGR3A/FCGR3B	**0.804**	2.72E-03	-0.125	8.93E-01	-0.546	1.66E-01	**0.894**	5.87E-03
FFAR2	**1.161**	1.48E-03	0.322	5.51E-01	-0.25	7.68E-01	**1.199**	1.48E-03
FLVCR2	**2.292**	1.48E-03	**1.308**	1.48E-03	0.399	6.08E-01	0.871	5.30E-02
FPR2	**2.303**	1.48E-03	1.058	1.10E-01	0.018	9.95E-01	1.022	1.88E-01
GTF2IRD2/GTF2IRD2B	0.24	7.50E-01	0.547	1.75E-01	0.474	3.27E-01	**1.185**	6.82E-03
HBD	**6.483**	1.48E-03	**3.388**	1.48E-03	**1.76**	1.48E-03	**2.69**	1.48E-03
HBM	6.907	2.33E-01	**3.941**	1.48E-03	**1.618**	1.48E-03	**3.352**	1.48E-03
HBQ1	**3.467**	1.48E-03	**1.568**	1.48E-03	0.298	8.62E-01	0.834	2.18E-01
HP	**3.353**	1.48E-03	1.253	1.43E-01	0.642	5.93E-01	0.285	8.97E-01
IFITM3	**1.709**	1.48E-03	**0.679**	1.48E-03	-0.205	7.20E-01	**1.175**	1.48E-03
IGHG3	-0.749	1.19E-01	**-1.688**	1.48E-03	**-1.366**	1.48E-03	-0.813	5.97E-02
IGHG4	0.023	9.88E-01	**-0.857**	1.20E-02	**-1.307**	1.48E-03	-0.21	8.40E-01
IGKV1-12	0.587	2.73E-01	-0.389	4.86E-01	**-1.004**	2.27E-02	-0.032	9.85E-01
IGKV1-39	-0.292	7.39E-01	-0.609	1.84E-01	**-1.337**	1.48E-03	-0.615	2.33E-01

IGKV1D-13	-0.704	1.76E-01	**-1.402**	1.48E-03	-1.073	5.21E-02	-0.027	9.91E-01
IGLC3	-0.164	7.97E-01	**-0.793**	4.89E-03	**-1.085**	1.48E-03	-0.555	1.04E-01
IGLV3-10	-0.42	7.32E-01	-1.089	7.37E-02	**-1.624**	4.89E-03	-0.103	9.71E-01
KCNJ15	**1.617**	1.48E-03	0.585	1.44E-01	-0.036	9.85E-01	**1.331**	1.48E-03
LCN2	**2.224**	1.48E-03	-0.276	6.28E-01	0.235	7.44E-01	**-0.905**	2.53E-02
LTF	**2.084**	1.48E-03	0.06	9.61E-01	0.386	3.91E-01	-0.708	5.55E-02
MME	**1.325**	1.48E-03	**0.677**	3.31E-02	-0.168	8.83E-01	**1.107**	1.48E-03
MMP8	**3.867**	1.48E-03	0.199	7.34E-01	0.205	7.77E-01	**-1.043**	1.48E-03
MPO	**2.354**	1.48E-03	0.162	8.29E-01	0.526	1.65E-01	-0.43	3.09E-01
MPZL2	**0.754**	1.12E-02	**0.737**	1.94E-02	-0.337	6.70E-01	-0.199	8.61E-01
NLRC4	**1.635**	1.48E-03	**1.004**	1.48E-03	-0.195	8.44E-01	**0.648**	4.81E-02
NRP1	**1.329**	1.48E-03	**1.25**	1.48E-03	0.151	9.25E-01	0.403	6.59E-01
ORM1	**2.565**	1.48E-03	0.662	3.56E-01	0.547	5.99E-01	0.025	9.92E-01
OSBPL10	**-1.804**	1.48E-03	-0.495	2.20E-01	-0.132	9.09E-01	0.318	6.22E-01
PGLYRP1	**2.619**	1.48E-03	-0.234	9.18E-01	-0.013	9.97E-01	-0.64	7.21E-01
PLA2G4C	**0.916**	2.72E-03	0.553	1.32E-01	-0.265	7.49E-01	**1.12**	2.72E-03
PRRG4	**0.766**	1.03E-02	0.364	3.74E-01	-0.376	5.22E-01	-0.6	1.59E-01
PTK7	**-1.44**	1.48E-03	**-1.174**	1.48E-03	-0.372	6.15E-01	-0.971	1.20E-02
RAB10	**1.292**	1.48E-03	**0.919**	1.48E-03	0.029	9.83E-01	**0.997**	1.48E-03
RETN	**2.55**	1.48E-03	0.223	6.90E-01	-0.123	9.11E-01	0.645	6.37E-02
RNASE2	**2.366**	1.48E-03	**0.594**	2.59E-02	0.133	8.98E-01	0.458	2.57E-01
RNASE3	**1.712**	1.48E-03	-0.305	7.78E-01	0.447	4.88E-01	-0.669	4.08E-01
S100B	**-0.998**	3.65E-02	**-1.145**	2.07E-02	0.107	9.58E-01	0.409	5.49E-01
S100P	**2.345**	1.48E-03	0.138	9.16E-01	0.11	9.48E-01	0.864	1.15E-01
SC5D	-0.349	3.55E-01	-0.179	8.04E-01	0.324	5.39E-01	**-0.809**	8.64E-03
SIGLEC6	-0.576	2.70E-01	-0.094	9.52E-01	-0.446	5.19E-01	**1.223**	1.48E-03
SLC25A37	**3.529**	1.48E-03	**1.824**	1.48E-03	0.464	1.26E-01	**1.226**	8.64E-03
SLPI	**2.081**	1.48E-03	-0.535	3.89E-01	-0.027	9.91E-01	-0.384	7.13E-01
TCF7L2	**1.286**	1.48E-03	**1.274**	1.48E-03	**0.873**	1.48E-03	**1.365**	1.48E-03
TLR8	**1.305**	1.48E-03	**1.011**	1.48E-03	-0.109	8.94E-01	0.13	8.58E-01
TMEM144	**1.579**	1.48E-03	**1.126**	1.48E-03	-0.23	9.10E-01	0.081	9.69E-01
TMEM150B	**1.227**	1.48E-03	**0.88**	2.21E-02	-0.274	8.36E-01	0.442	5.57E-01
TMEM170B	**1.075**	1.48E-03	**0.508**	9.49E-03	0.206	6.10E-01	**-0.474**	4.08E-02
TNFSF10	**0.687**	1.48E-03	**0.785**	1.48E-03	-0.396	3.27E-01	0.588	6.69E-02
VSIG4	**2.567**	1.48E-03	0.27	6.97E-01	-0.321	7.05E-01	0.553	2.93E-01
ZNF683	**-1.473**	1.48E-03	-0.517	2.47E-01	-0.356	6.77E-01	0.76	7.45E-02

*Green shading highlights positive fold change (up-regulation), and red shading highlights negative fold change (down-regulation). Bolded entries are significant according to q-value.

b. Classification Performance, In Silico Biological Validation, and Top Enrichr Hits Tables.

Classification Performance Tables Description:

Listed below are the classification accuracies using nested cross-validation for our four datasets. For the LV 2-Way dataset, 36 configurations are given. For the 3 other datasets (LV 3-Way, LV 5-Way, and PBMC 5-Way), only the configuration that resulted in the best gene set are given. In the tables below, each dataset has a single entry highlighted in green and bolded. This denotes the configuration and feature size that produced the best gene set.

The classification performance tables are formatted as follows. The headings in the table indicate FS method (DE, IG) and Outlier Filter Threshold (2.5, 3.0, 3.5). Configurations are represented as: ML Classifier / Outlier Feature Filter Method.

Enrichr In Silico Biological Validation Tables Description:

The *in silico* biological validation tables contain the tallies that were attained via Enrichr for each dataset. Each of the 36 method configurations produced two gene sets, one attained via (4 out of 5) and (5 out of 5) gene set intersection, as described within Candidate Gene Set section of Supplemental Methods. The choice of the classifier did not impact the gene set generated. Therefore, there were only a total of 24 distinct gene set configurations (i.e., resulting from multiplying 2 FS methods (DE, IG) by 2 outlier filtering strategies (Intersection, Union) by 3 filter thresholds (2.5, 3.0, 3.5) by 2 merge strategies (4 out of 5 merge, and 5 out of 5 merge).

The configurations are listed above the tables: Outlier Filtering Strategy / Merge Strategy. The entries in the 1st column are: training set feature size (in outer loop of nested cross validation) – gene set size after merge for configuration #1/ #2/ … / #6. Cells containing "NA" indicate that the gene set size was zero after merge, and therefore, enrichment analysis could not be performed. The headings in the table indicate FS method (DE, IG) and Outlier Filter Standard

Deviation Threshold (2.5, 3.0, 3.5). The values that are bolded and highlighted in green correspond to the best gene set for a given dataset. For all datasets, other than LV 2-Way, we only provided the configuration that resulted in the best gene set.

Top Enrichr Hits Tables Description:

The respective hits were sorted by adjusted p-value and filtered using the regular expression described in the methods, then the top 3 were selected for each category, for each one of the four datasets. Highly redundant entries (in either gene list or function) were removed.

LV 2-Way.

Classification Performance:

kNN / Intersection:

Table S8: The classification accuracies for kNN / Intersection configuration in LV 2-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2	0.79	0.82	0.82	0.94	0.92	0.92
3	0.92	0.86	0.81	0.97	0.95	0.86
4	0.97	0.97	0.93	0.97	1	0.97
5	0.97	0.95	0.97	0.97	0.97	0.97
10	0.97	0.97	0.97	0.97	0.97	0.95
15	0.97	0.97	0.97	0.97	0.97	0.97
20	0.97	0.97	0.97	0.97	0.97	0.97
25	0.97	0.97	0.97	0.97	0.97	0.97
50	0.97	0.97	0.97	0.97	0.97	0.97

kNN / Union:

Table S9: The classification accuracies for kNN / Union configuration in LV 2-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2	0.93	0.93	0.9	0.92	0.95	0.97

3	0.95	0.95	0.95	0.93	0.95	0.95
4	0.97	0.97	0.97	0.93	0.95	0.95
5	0.95	0.97	0.97	0.95	0.97	0.95
10	0.92	0.95	0.95	0.97	0.97	0.95
15	0.97	0.95	0.95	0.97	0.97	0.95
20	0.97	0.97	0.97	0.97	1	1
25	0.97	1	1	1	1	1
50	0.95	0.95	0.95	1	1	1

LR / Intersection:

Table S10: The classification accuracies for LR / Intersection configuration in LV 2-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2	0.82	0.82	0.82	0.87	0.86	0.92
3	0.95	0.9	0.77	0.93	0.88	0.92
4	0.92	0.88	0.88	0.95	0.95	0.92
5	0.92	0.97	0.97	0.95	0.95	0.97
10	0.97	0.97	0.97	0.97	0.97	0.97
15	0.97	0.97	0.97	0.97	0.97	0.97
20	0.97	0.97	0.97	0.97	0.97	0.97
25	0.97	0.97	0.97	0.97	0.97	0.97
50	1	1	1	0.97	0.97	0.97

LR / Union:

Table S11: The classification accuracies for LR / Union configuration in LV 2-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2	0.83	0.91	0.88	0.92	0.93	0.92
3	0.95	0.95	0.92	0.95	0.97	0.95
4	0.95	0.97	0.97	0.95	0.95	0.97
5	0.92	0.97	0.97	0.95	0.95	0.95
10	0.97	**0.97**	0.97	0.95	0.97	0.95
15	1	0.97	0.97	0.97	1	0.95
20	1	1	1	0.97	1	1
25	1	1	1	0.97	1	1
50	1	1	1	0.97	0.97	1

SVM / Intersection:

Table S12: The classification accuracies for SVM / Intersection configuration in LV 2-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2	0.82	0.82	0.82	0.93	0.92	0.97
3	0.87	0.86	0.74	0.97	0.9	0.97
4	0.95	0.95	0.95	0.97	0.95	0.95
5	0.93	0.92	0.9	0.97	0.95	0.95
10	0.95	0.95	0.95	1	0.97	0.97
15	0.95	0.95	0.95	0.97	0.97	0.97
20	0.95	0.95	0.95	0.97	0.97	0.97
25	0.97	0.95	0.95	0.97	0.97	0.97
50	1	0.97	0.97	0.97	0.97	0.97

SVM / Union:

Table S13: The classification accuracies for SVM / Union configuration in LV 2-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2	0.93	0.93	0.87	0.93	0.9	0.91
3	0.93	0.93	0.93	0.95	0.97	0.95
4	0.97	0.95	0.92	0.97	1	0.93
5	0.97	0.97	0.95	0.93	1	1
10	0.95	0.97	0.95	0.95	1	0.97
15	0.97	0.97	0.97	0.97	1	0.97
20	1	0.97	0.97	1	1	0.97
25	1	0.97	0.97	1	1	1
50	0.97	1	1	1	1	1

In Silico Biological Validation:

Intersection / 4 out of 5 Merge:

Table S14: The Enrichr hits Intersection / 4 out of 5 Merge configuration in LV 2-Way dataset

	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2 - 1/1/1/0/0/0	1/0/0	1/0/0	1/0/0	NA	NA	NA
3 - 2/2/1/0/0/0	16/0/12	16/0/12	1/0/0	NA	NA	NA
4 - 3/2/2/0/0/1	18/0/0	16/0/12	16/0/12	NA	NA	0/0/0
5 - 3/3/3/1/1/1	18/0/0	18/0/0	18/0/0	0/0/0	0/0/0	0/0/0
10 - 9/9/7/1/2/2	22/1/3	22/1/3	33/1/6	0/0/0	0/0/0	0/0/0
15 - 12/12/12/3/2/2	17/2/5	17/2/5	17/2/5	0/0/0	0/0/0	0/0/0
20 - 13/13/13/4/3/4	18/3/7	18/3/7	18/3/7	7/0/0	0/0/0	1/0/0
25 - 18/18/18/5/6/5	1/3/7	1/3/7	1/3/7	5/0/0	3/0/0	2/0/0
50 - 34/33/32/16/13/18	1/3/7	1/3/7	1/4/7	7/0/0	5/0/0	0/0/0

Intersection / 5 out of 5 Merge:

Table S15: The Enrichr hits Intersection / 5 out of 5 Merge configuration in LV 2-Way dataset

	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
10 - 2/1/1/1/0/1	22/0/10	23/0/8	23/0/8	0/0/0	NA	0/0/0
15 - 6/6/5/2/0/1	32/2/6	32/2/6	30/1/4	0/0/0	NA	0/0/0
20 - 7/7/6/2/0/3	31/1/5	31/1/5	32/2/6	0/0/0	NA	1/0/0
25 - 8/8/7/2/0/3	32/1/4	32/1/4	31/1/5	0/0/0	NA	1/0/0
50 - 16/17/16/4/4/6	3/3/4	1/3/3	1/3/4	1/0/0	1/0/0	6/0/0

*Feature Sizes 2-5 resulted in gene sets of size 0 and were therefore excluded.

Union / 4 out of 5 Merge:

Table S16: The Enrichr hits Union / 4 out of 5 Merge configuration in LV 2-Way dataset

	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
2 - 0/0/1/0/0/0	NA	NA	19/0/12	NA	NA	NA
3 - 0/0/1/0/0/0	NA	NA	19/0/12	NA	NA	NA
4 - 3/1/2/0/0/0	10/2/1	23/0/8	37/0/18	NA	NA	NA
5 - 3/3/3/0/0/0	10/2/1	30/2/4	44/0/5	NA	NA	NA

10 - 6/9/9/1/1/1	19/2/6	**28/3/6**	32/2/10	1/2/1	0/0/0	0/0/0
15 - 8/9/11/3/1/1/2	11/3/4	28/3/6	20/3/7	8/0/0	0/0/0	0/0/0
20 - 9/13/16/4/4/3	5/4/2	1/4/5	1/4/7	10/0/0	1/0/0	0/0/0
25 - 10/14/16/6/7/6	12/4/4	1/4/6	1/4/7	7/0/0	4/0/0	3/0/0
50 - 26/34/34/16/18/17	2/5/8	1/5/12	3/5/11	0/0/0	0/0/0	0/1/0

Union / 5 out of 5 Merge:

Table S17: The Enrichr hits Union / 5 out of 5 Merge configuration in LV 2-Way dataset

	DE 2.5	DE 3	DE 3.5	IG 2.5	IG 3	IG 3.5
5 - 0/1/0/0/0/0	NA	23/0/8	NA	NA	NA	NA
10 – 1/4/4/0/0/0	4/0/5	32/1/4	32/1/4	NA	NA	NA
15 - 3/6/7/0/0/0	9/0/0	32/1/5	31/1/5	NA	NA	NA
20 - 5/7/7/2/0/0	15/2/1	34/1/4	31/1/5	2/1/1	NA	
25 - 5/8/10/3/0/1	15/2/1	32/2/4	13/3/3	2/0/1	NA	0/0/0
50 - 12/17/20/7/3/3	6/3/1	1/3/4	1/3/4	4/0/0	1/0/0	1/0/0

*Feature Sizes 2-4 resulted in gene sets of size 0 and were therefore excluded.

Top Enrichr Hits:

Table S18: The Enrichr top hits for LV 2-Way best gene set

Pathway		
Term	Adjusted P-Value	Genes
Linoleic acid metabolism	0.00542024	AKR1B10;PLA2G2A
phospholipid metabolic process (GO:0006644)	0.0309956	PLA2G2A;FITM1
primary alcohol catabolic process (GO:0034310)	0.0309956	AKR1B10
Tissue		
HEPATOCYTE	1.88588e-05	AKR1B10;PPP1R1A;MT1M;PLA2G2A;SCTR;FITM1;KRT23;TREM2
LIVER (BULK TISSUE)	0.0212485	AKR1B10;MT1M;PLA2G2A;SCTR;FITM1
OMENTUM	0.0212485	MMP7;PPP1R1A;MT1M;PLA2G2A;TREM2
Disease		

Alcoholic Hepatitis DOID-12351 human GSE28619 sample 477	1.27297e-05	MMP7;AKR1B10;PLA2G2A;KRT23;TREM2
hepatocellular carcinoma DOID-684 human GSE57957 sample 660	0.00078694	MMP7;AKR1B10;MT1M;PLA2G2A
Carcinoma, Hepatocellular C0019204 human GSE6764 sample 407	0.00966049	AKR1B10;PLA2G2A;KRT23

LV 3-Way.

Classification Performance:

kNN / Union:

Table S19: The classification accuracies for kNN / Union configuration in LV 3-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5
10	0.82	0.86	0.83
25	0.85	0.86	0.83
50	0.86	0.85	0.9
100	0.88	0.9	0.85
150	0.88	0.9	**0.9**
200	0.9	0.9	0.9
250	0.88	0.9	0.91
300	0.88	0.88	0.9
350	0.88	0.88	0.9
400	0.9	0.88	0.88
450	0.9	0.9	0.88
500	0.9	0.9	0.9

Enrichr In Silico Biological Validation:

Union / 5 out of 5 Merge:

Table S20: The Enrichr hits Union / 5 out of 5 Merge configuration in LV 3-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5
10 - 0/1/0	NA	23/0/8	NA
25 - 2/4/4	4/0/6	34/1/5	30/1/4
50 - 5/8/10	16/0/0	34/4/10	13/4/11

100 - 12/19/21	2/4/0	11/4/13	10/4/14
150 - 19/32/33	0/4/7	9/5/15	**13/5/17**
200 - 27/44/46	1/4/6	6/5/19	2/5/21
250 - 37/55/58	1/4/9	4/5/28	2/5/19
300 - 43/65/71	0/5/9	4/6/30	5/6/26
350 - 54/74/84	1/5/11	4/6/28	5/6/30
400 - 76/89/94	4/5/17	2/6/31	7/6/29
450 - 93/95/109	5/6/24	2/6/31	5/6/32
500 - 111/115/117	12/6/22	11/7/35	11/6/33

Top Enrichr Hits:

Table S21: The Enrichr top hits for LV 3-Way best gene set

Pathway		
Term	Adjusted P-Value	Genes
Oncostatin M	0.0181233	CXCL6;AKR1B10;LCN2;HAMP;S100A8
IL-17 signaling pathway	0.022097	CXCL6;LCN2;S100A8
Endogenous Toll-like receptor signaling	0.0259474	VCAN;S100A8
Tissue		
HEPATOCYTE	2.53723e-07	FCN3;PLA2G2A;SCTR;FITM1;KRT23;TREM2;IGSF9;FAM198A;DBNDD1;CYP2A7;AKR1B10;CYP2B6;PPP1R1A;CREB3L3;LCN2;GPC3;MT1G;HAO2
LIVER (BULK TISSUE)	4.02437e-05	FCN3;PLA2G2A;SCTR;FITM1;IGSF9;FAM198A;CYP2A7;AKR1B10;CYP2B6;CREB3L3;GPC3;MT1G;HAMP;HAO2;CFTR
OMENTUM	0.00010671	CXCL6;FCN3;MMP7;PLA2G2A;TREM2;IGSF9;FAM198A;PPP1R1A;GPNMB;RGS1;GPC3;MT1G;EPS8L1;S100A8
Disease		
Alcoholic Hepatitis DOID-12351 human GSE28619 sample 477	8.50532e-09	CXCL6;VCAN;MMP7;AKR1B10;GPNMB;PLA2G2A;EEF1A2;LCN2;KRT23;TREM2

hepatocellular carcinoma DOID-684 human GSE39791 sample 663	4.53136e-05	CYP2A7;CXCL6;FCN3;MMP7;PPP1R1A;MT1G;HAMP;S100A8
Carcinoma, Hepatocellular C0019204 human GSE6764 sample 407	9.65272e-05	FCN3;CYP2B6;PPP1R1A;MT1G;HAMP;HAO2;S100A8

LV 5-Way.

Classification Performance:

SVM / Intersection:

Table S22: The classification accuracies for SVM / Intersection configuration in LV 5-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5
10	0.84	0.75	0.75
25	0.84	0.81	0.83
50	0.89	0.86	0.88
100	0.86	0.85	0.84
150	**0.91**	0.87	0.89
200	0.86	0.86	0.85
250	0.86	0.85	0.83
300	0.85	0.86	0.83
350	0.85	0.83	0.83
400	0.85	0.85	0.83
450	0.85	0.85	0.83
500	0.85	0.83	0.86

In Silico Biological Validation:

Intersection / 5 out of 5 Merge:

Table S23: The Enrichr hits Intersection / 5 out of 5 Merge configuration in LV 5-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5
10 -3/1/1	17/0/4	19/0/12	19/0/12
25 -6/6/6	20/0/6	20/0/6	20/0/6
50 - 16/14/14	10/2/5	11/2/6	11/2/6

100 - 29/27/26	21/7/19	14/5/18	13/5/18
150 - 39/38/38	**25/7/27**	26/7/26	26/7/26
200 - 52/51/49	22/7/22	19/7/21	25/5/21
250 - 70/70/66	26/9/30	26/9/29	26/9/28
300 - 85/86/85	38/9/35	37/9/35	36/9/35
350 - 100/100/97	72/9/41	72/8/41	45/8/40
400 - 121/117/114	84/9/39	71/9/39	75/9/38
450 - 140/138/138	83/10/43	81/9/40	81/9/40
500 - 160/161/159	98/9/48	94/9/47	94/9/47

Top Enrichr Hits:

Table S24: The Enrichr top hits for LV 5-Way best gene set

Pathway		
Term	Adjusted P-Value	Genes
Interferon alpha/beta signaling	2.59005e-05	IFITM1;IFI27;IFI6;ISG15;OASL
cytokine-mediated signaling pathway (GO:0019221)	0.00940058	IFITM1;MUC1;IFI27;GSTA2;IFI6;ISG15;OASL
Drug metabolism: cytochrome P450	0.0193383	CYP2A7;CYP2B6;GSTA2
Tissue		
LIVER (BULK TISSUE)	6.9614e-08	IGFBP1;IFITM1;SPINK1;GADD45B;PLA2G2A;MT1M;SCTR;KRT7;ISG15;SAA2-SAA4;IFI44L;OASL;CYP2A7;AKR1B10;CYP2B6;IFI27;GSTA2;MT1G;ATF3;MUC6
OMENTUM	6.9614e-08	IGFBP1;MMP7;GADD45B;DUSP1;PLA2G2A;MT1M;IGHV3-23;KRT7;HBA2;SAA2-SAA4;GADD45G;NR4A1;MUC1;IFI27;PPP1R1A;RGS1;MT1G;EPS8L1;S100A8;ATF3
HEPATOCYTE	3.61882e-07	IGFBP1;SPINK1;GADD45B;PLA2G2A;MT1M;SCTR;KRT7;KRT23;SAA2-SAA4;SYT8;GADD45G;CYP2A7;AKR1B10;CYP2B6;IFI27;PPP1R1A;GSTA2;MT1G;MUC6
Disease		

Alcoholic Hepatitis DOID-12351 human GSE28619 sample 477	1.07891e-08	IGFBP1;NR4A1;GADD45B;PPP1R1A;DUSP1;RGS1;MT1M;MT1G;IFI44L;ATF3;GADD45G
hepatocellular carcinoma DOID-684 human GSE39791 sample 663	3.37686e-07	IGFBP1;CYP2A7;IFITM1;MMP7;GADD45B;PPP1R1A;MT1M;MT1G;HBA2;S100A8
Carcinoma, Hepatocellular C0019204 human GSE6764 sample 407	4.39099e-07	IFITM1;SPINK1;AKR1B10;IFI27;PLA2G2A;IFI6;KRT23;ISG15;IFI44L

PBMC 5-Way.

Classification Performance:

LR / Union:

Table S25: The classification accuracies for LR / Union configuration in PBMC 5-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5
10	0.5	0.44	0.56
25	0.59	0.62	0.54
50	0.64	0.67	0.63
100	0.66	0.64	0.66
150	0.66	0.66	0.67
200	0.67	**0.75**	0.69
250	0.66	0.72	0.68
300	0.67	0.72	0.71
350	0.68	0.72	0.74
400	0.65	0.72	0.71
450	0.65	0.72	0.72
500	0.64	0.72	0.72

In Silico Biological Validation:

Union / 5 out of 5 Merge:

Table S26: The Enrichr hits Union / 5 out of 5 Merge configuration in PBMC 5-Way dataset

Feature Size	DE 2.5	DE 3	DE 3.5
10 -1/1/3	0/1/0	0/0/2	10/0/1
25 -4/6/9	19/0/1	14/4/4	27/3/6
50 -6/13/18	25/0/3	27/6/3	24/4/7

100 - 13/36/39	28/0/4	35/7/11	26/7/7
150 - 16/51/67	3/0/9	33/7/15	35/9/18
200 - 22/75/89	6/2/14	**41/10/17**	49/11/21
250 - 25/91/122	28/4/15	48/10/22	44/12/27
300 - 31/109/148	18/3/9	47/10/21	60/12/30
350 - 35/131/168	15/3/10	40/10/17	69/12/26
400 - 39/153/191	12/4/10	37/8/18	78/12/32
450 - 45/170/213	19/1/16	33/9/18	72/13/31
500 - 52/192/240	13/0/13	56/10/21	73/14/29

Top Enrichr Hits:

Table S27: The Enrichr top hits for PBMC 5-Way best gene set

Pathway		
Term	Adjusted P-Value	Genes
neutrophil mediated immunity (GO:0002446)	5.36008e-22	ORM1;CRISP3;FPR2;RETN;MPO;CXCL5;FCGR3B;CXCR1;CTSG;PGLYRP1;CAMP;ELANE;MME;ANXA3;DEFA4;AZU1;RNASE3;MMP8;RNASE2;CEACAM3;RAB10;SLPI;LCN2;CHI3L1;S100P;CEACAM8;LTF
Innate immune system	3.02077e-08	C1QB;C1QA;DEFA4;CD180;DEFA3;NLRC4;S100B;IGHG3;IGHG4;IGKV1-39;IGLC3;TLR8;CCR2
mucosal immune response (GO:0002385)	3.11393e-08	DEFA4;DEFA3;FFAR2;RNASE3;CAMP;LTF
Tissue		
PERIPHERAL BLOOD	7.95747e-24	ALAS2;ORM1;CRISP3;DYSF;HBD;FPR2;NLRC4;RETN;CXCL5;IGHG3;IGHG4;HBM;FCGR3B;CXCR1;IGKV1-12;TNFSF10;IGLC3;FLVCR2;FFAR2;AHSP;HBQ1;PGLYRP1;CAMP;CCR2;MPZL2;ZNF683;TMEM150B;MME;DEFA4;CD180;DEFA3;RNASE3;MMP8;TMEM170B;RNASE2;IGKV1D-13;CEACAM3;SLC25A37;SLPI;IGLV3-10;LCN2;ALPL;TLR8;CHI3L1;S100P;SIGLEC6;LTF
GRANULOCYTE	8.46854e-12	ORM1;CRISP3;DYSF;FPR2;NLRC4;PRRG4;FCGR3B;CXCR1;TNFSF10;FFAR2;VSIG4;PGLYRP1;CAMP;ELANE;MPZL2;MME;ANXA3;DEFA4;KCNJ15;DEFA3;RNAS

		E3;MMP8;TMEM170B;RNASE2;CEACAM3;SLC25A37;SLPI;LCN2;ALPL;TLR8;CHI3L1;S100P;CEACAM8;LTF
WholeBlood	8.32759e-06	CEACAM3;FCGR3B;CXCR1;AQP9;KCNJ15;DYSF;TNFSF10;ALPL;TLR8;CHI3L1;FFAR2;FPR2
Disease		
Septic Shock C0036983 human GSE9692 sample 307	1.67821e-26	C1QB;C1QA;ORM1;CRISP3;AQP9;HP;DYSF;FPR2;NLRC4;RETN;MPO;CXCR1;FFAR2;VSIG4;PGLYRP1;CCR2;ANXA3;DEFA4;KCNJ15;RNASE3;MMP8;RNASE2;SLPI;LCN2;ALPL;TLR8;S100P;CEACAM8;BCAT1;LTF
familial combined hyperlipidemia DOID-13809 human GSE11393 sample 773	2.72273e-21	IFITM3;MME;ANXA3;DEFA4;AQP9;DYSF;DEFA3;FPR2;RNASE3;MMP8;RNASE2;CEACAM3;SLC25A37;FCGR3B;CXCR1;LCN2;ALPL;CHI3L1;S100P;FFAR2;CEACAM8;CAMP;LTF
hepatitis C virus related hepatocellular carcinoma UMLS CUI-C1333978 human GSE58208 sample 736	3.48243e-08	DEFA4;DYSF;DEFA3;HBD;MPO;HBM;SLPI;CXCR1;LCN2;S100P;FFAR2;AHSP;CEACAM8;LTF

c. Per Replicate RNA-seq Count Heatmaps:

Per replicate RNA-seq count heatmaps are provided to visualize the gene expression counts for each individual sample. The heatmaps are displayed using the best gene sets shown in Table 3 of the main text. In the figures below, the first heatmaps display the gene expression in our data, and the second heatmaps display the gene expression in the independent test dataset. For PBMC 5-Way dataset, only the heatmap of our data is provided.

LV 2-Way.

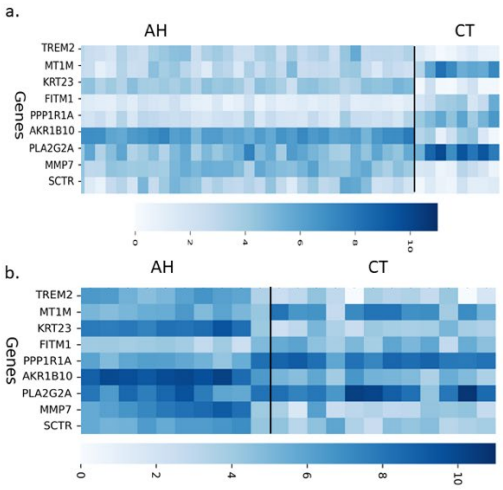

Figure S6: LV 2-Way; Per Replicate Heatmap of Counts for Best Gene Set. a. Per replicate heatmap of best LV 2-Way gene set. b. Per replicate heatmap of best gene set within validation dataset.

LV 3-Way.

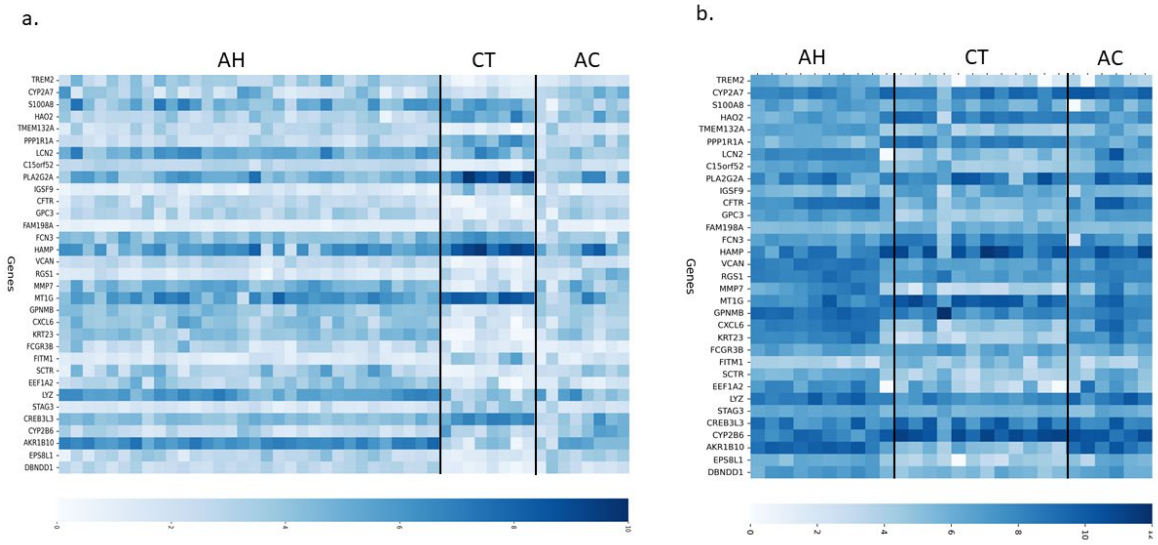

Figure S7: LV 3-Way; Per Replicate Heatmap of Counts for Best Gene Set. a. Per replicate heatmap of best LV 3-Way gene set. b. Per replicate heatmap of best gene set within validation dataset.

LV 5-Way.

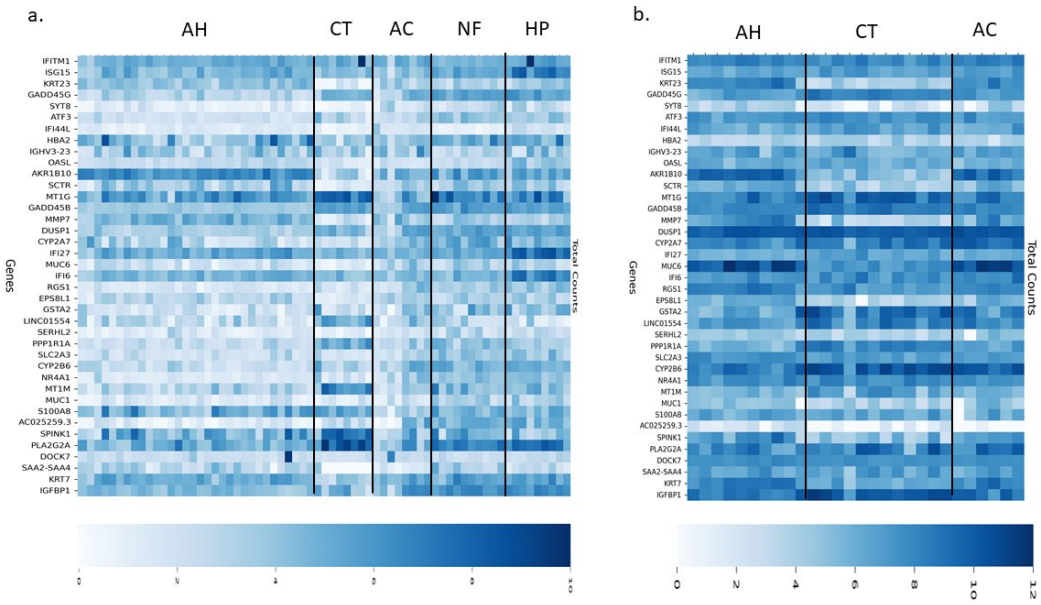

Figure S8: LV 5-Way; Per Replicate Heatmap of Counts for Best Gene Set. a. Per replicate heatmap of best LV 5-Way gene set. b. Per replicate heatmap of best gene set within validation dataset.

PBMC 5-Way.

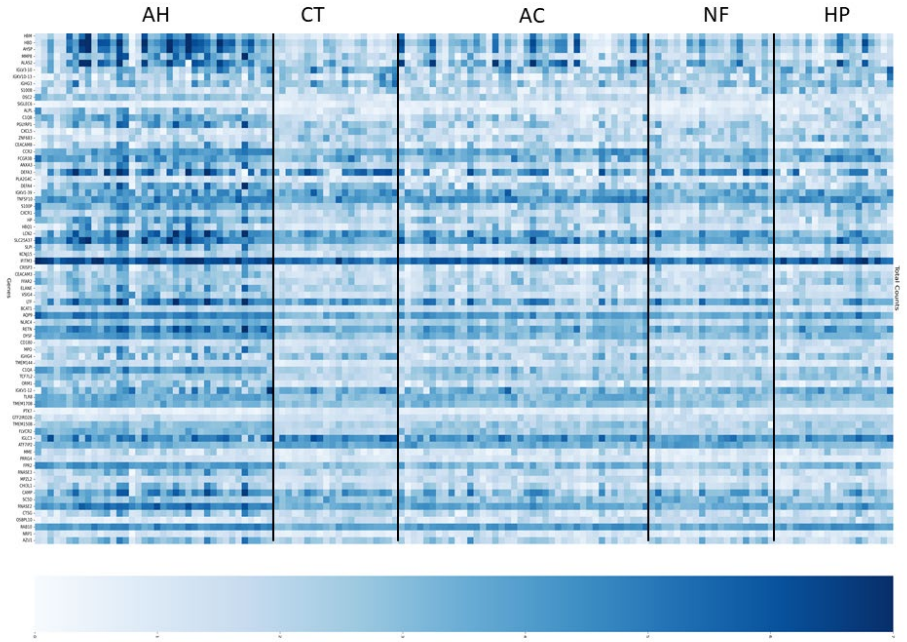

Figure S9: PBMC 5-Way; Per replicate heatmap of best PBMC 5-Way gene set.

d. Comparison of additional in silico biological validation approaches:

IPA.

1. *LV 5-Way:*

Table S28: Top enriched IPA pathways per pairwise comparison in LV 5-Way best gene set

Ingenuity Canonical Pathways	Pairwise Comparison	p-value	Molecules
PXR/RXR Activation	AH vs AC	1.10E-05	CYP2B6,GSTA2,IGFBP1
Acetone Degradation I (to Methylglyoxal)	AH vs AC	2.29E-04	AKR1B10,CYP2B6
Estrogen Biosynthesis	AH vs AC	3.89E-04	AKR1B10,CYP2B6
GADD45 Signaling	AH vs AC	7.76E-04	GADD45B,GADD45G
Senescence Pathway	AH vs AC	1.05E-03	GADD45B,GADD45G,SAA2-SAA4
SPINK1 General Cancer Pathway	CT vs AC	1.78E-05	MT1G,MT1M,SPINK1
Acetone Degradation I (to Methylglyoxal)	CT vs AC	3.31E-04	AKR1B10,CYP2A6 (includes others)
Estrogen Biosynthesis	CT vs AC	5.75E-04	AKR1B10,CYP2A6 (includes others)
Atherosclerosis Signaling	CT vs AC	5.01E-03	PLA2G2A,S100A8
Glucocorticoid Receptor Signaling	CT vs AC	1.00E-02	KRT23,KRT7,PLA2G2A
Acetone Degradation I (to Methylglyoxal)	CT vs AH	4.68E-06	AKR1B10,CYP2A6 (includes others),CYP2B6
Estrogen Biosynthesis	CT vs AH	1.07E-05	AKR1B10,CYP2A6 (includes others),CYP2B6
Bupropion Degradation	CT vs AH	2.75E-04	CYP2A6 (includes others),CYP2B6
Oxidative Ethanol Degradation III	CT vs AH	4.07E-04	CYP2A6 (includes others),CYP2B6
Nicotine Degradation III	CT vs AH	1.20E-03	CYP2A6 (includes others),CYP2B6
Interferon Signaling	HP vs AC	7.24E-07	IFI6,IFITM1,ISG15
Antioxidant Action of Vitamin C	HP vs AC	1.35E-03	PLA2G2A,SLC2A3
Atherosclerosis Signaling	HP vs AC	1.86E-03	PLA2G2A,S100A8
Role of IL-17A in Psoriasis	HP vs AC	7.24E-03	S100A8
Vitamin-C Transport	HP vs AC	1.17E-02	SLC2A3
Interferon Signaling	HP vs AH	7.76E-06	IFI6,IFITM1,ISG15
PXR/RXR Activation	HP vs AH	4.17E-05	CYP2B6,GSTA2,IGFBP1
Acetone Degradation I (to Methylglyoxal)	HP vs AH	5.37E-04	AKR1B10,CYP2B6
Estrogen Biosynthesis	HP vs AH	9.33E-04	AKR1B10,CYP2B6

GADD45 Signaling	HP vs AH	1.82E-03	GADD45B,GADD45G
Acetone Degradation I (to Methylglyoxal)	HP vs CT	5.37E-04	AKR1B10,CYP2A6 (includes others)
Interferon Signaling	HP vs CT	6.92E-04	IFI6,ISG15
Estrogen Biosynthesis	HP vs CT	9.33E-04	AKR1B10,CYP2A6 (includes others)
SPINK1 General Cancer Pathway	HP vs CT	1.95E-03	MT1M,SPINK1
PXR/RXR Activation	HP vs CT	2.09E-03	CYP2A6 (includes others),IGFBP1
SPINK1 General Cancer Pathway	NF vs AC	6.17E-04	MT1G,MT1M
Senescence Pathway	NF vs AC	6.76E-04	ATF3,GADD45G,SAA2-SAA4
Antioxidant Action of Vitamin C	NF vs AC	1.86E-03	PLA2G2A,SLC2A3
Atherosclerosis Signaling	NF vs AC	2.57E-03	PLA2G2A,S100A8
Role of IL-17A in Psoriasis	NF vs AC	8.51E-03	S100A8
PXR/RXR Activation	NF vs AH	4.90E-07	CYP2A6 (includes others),CYP2B6,GSTA2,IGFBP1
Bupropion Degradation	NF vs AH	2.75E-04	CYP2A6 (includes others),CYP2B6
Oxidative Ethanol Degradation III	NF vs AH	4.07E-04	CYP2A6 (includes others),CYP2B6
Acetone Degradation I (to Methylglyoxal)	NF vs AH	5.01E-04	CYP2A6 (includes others),CYP2B6
Estrogen Biosynthesis	NF vs AH	8.51E-04	CYP2A6 (includes others),CYP2B6
Acetone Degradation I (to Methylglyoxal)	NF vs CT	3.72E-04	AKR1B10,CYP2A6 (includes others)
Estrogen Biosynthesis	NF vs CT	6.31E-04	AKR1B10,CYP2A6 (includes others)
SPINK1 General Cancer Pathway	NF vs CT	1.35E-03	MT1M,SPINK1
PXR/RXR Activation	NF vs CT	1.45E-03	CYP2A6 (includes others),IGFBP1
Antioxidant Action of Vitamin C	NF vs CT	4.07E-03	PLA2G2A,SLC2A3
Interferon Signaling	HP vs NF	2.45E-07	IFI6,IFITM1,ISG15
Systemic Lupus Erythematosus In B Cell Signaling Pathway	HP vs NF	1.15E-02	IGHV3-23,ISG15
Coronavirus Replication Pathway	HP vs NF	1.58E-02	IFITM1
SPINK1 Pancreatic Cancer Pathway	HP vs NF	2.04E-02	SPINK1
SPINK1 General Cancer Pathway	HP vs NF	2.29E-02	SPINK1

* The colors are alternated between blue and white to highlight each pairwise comparison group.

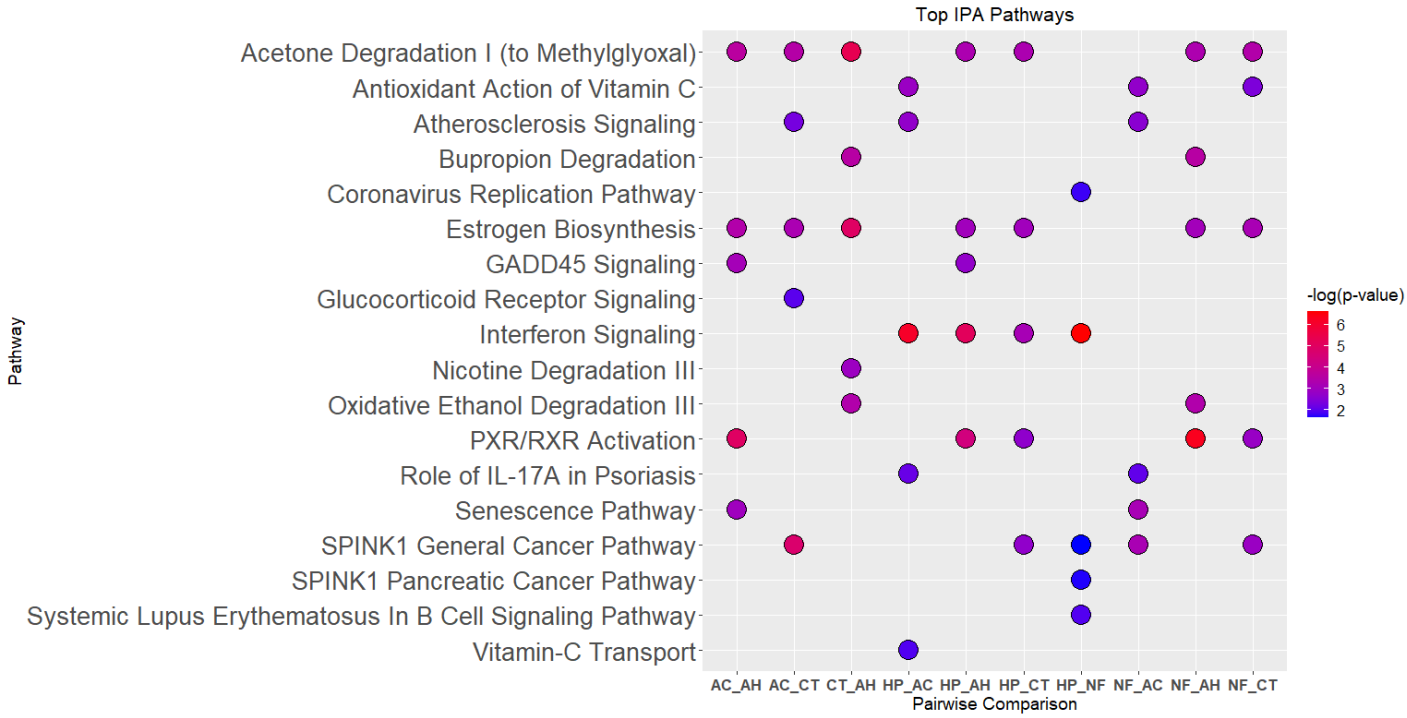

Figure S40: Dot plot of top 5 IPA pathways and their p-value significance for each pairwise comparison of LV 5-Way best gene set. The dots are color-coded by p-value significance, with blue dots representing lower significance and red representing higher significance.

2. *PBMC 5-Way.*

Table S29: Top enriched IPA pathways per pairwise comparison in PBMC 5-Way best gene set.

Ingenuity Canonical Pathways	Pairwise Comparison	-log(p-value)	Molecules
Airway Pathology in Chronic Obstructive Pulmonary Disease	CT vs AH	4.07E-07	CTSG,ELANE,LCN2,MMP8,MPO,ORM1
Iron homeostasis signaling pathway	CT vs AH	2.19E-05	ALAS2,HBD,HBQ1,HP,SLC25A37
Granulocyte Adhesion and Diapedesis	CT vs AH	7.94E-05	CCR2,CXCL5,CXCR1,FPR2,MMP8
Airway Inflammation in Asthma	CT vs AH	7.94E-05	ELANE,RNASE2,RNASE3
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	CT vs AH	5.13E-04	C1QA,C1QB,NLRC4,TLR8
Airway Pathology in Chronic Obstructive Pulmonary Disease	AC vs AH	5.50E-08	CTSG,ELANE,LCN2,MMP8,MPO,ORM1
Airway Inflammation in Asthma	AC vs AH	3.02E-05	ELANE,RNASE2,RNASE3
Iron homeostasis signaling pathway	AC vs AH	1.00E-04	HBD,HBQ1,HP,SLC25A37
Acute Phase Response Signaling	AC vs AH	3.55E-04	C1QA,C1QB,HP,ORM1

IL-8 Signaling	AC vs AH	5.37E-04	AZU1,CXCR1,DEFA1 (includes others),MPO
Iron homeostasis signaling pathway	AC vs CT	1.66E-05	ALAS2,HBD,HBQ1,SLC25A37
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	AC vs CT	2.40E-05	C1QA,C1QB,NLRC4,TLR8
Complement System	AC vs CT	8.13E-04	C1QA,C1QB
TREM1 Signaling	AC vs CT	3.24E-03	NLRC4,TLR8
Pyroptosis Signaling Pathway	AC vs CT	5.01E-03	NLRC4,TLR8
Airway Pathology in Chronic Obstructive Pulmonary Disease	HP vs AH	1.02E-07	CTSG,ELANE,LCN2,MMP8,MPO,ORM1
Iron homeostasis signaling pathway	HP vs AH	7.08E-06	ALAS2,HBD,HBQ1,HP,SLC25A37
Airway Inflammation in Asthma	HP vs AH	3.98E-05	ELANE,RNASE2,RNASE3
Acute Phase Response Signaling	HP vs AH	5.13E-04	C1QA,C1QB,HP,ORM1
Melatonin Degradation III	HP vs AH	2.04E-03	MPO
Iron homeostasis signaling pathway	HP vs CT	3.39E-04	ALAS2,HBD,SLC25A37
Complement System	HP vs CT	6.31E-04	C1QA,C1QB
Tetrapyrrole Biosynthesis II	HP vs CT	5.13E-03	ALAS2
Heme Biosynthesis II	HP vs CT	9.33E-03	ALAS2
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	HP vs CT	9.77E-03	C1QA,C1QB
Tetrapyrrole Biosynthesis II	HP vs AC	2.57E-03	ALAS2
Heme Biosynthesis II	HP vs AC	4.68E-03	ALAS2
Inhibition of Matrix Metalloproteases	HP vs AC	1.95E-02	MMP8
Fcî³ Receptor-mediated Phagocytosis in Macrophages and Monocytes	HP vs AC	4.68E-02	FCGR3A/FCGR3B
Airway Pathology in Chronic Obstructive Pulmonary Disease	HP vs AC	5.75E-02	MMP8
Phagosome Formation	HP vs NF	3.98E-04	CXCR1,FCGR3A/FCGR3B,FFAR2,IGHG4,PLA2G4C
Airway Pathology in Chronic Obstructive Pulmonary Disease	HP vs NF	5.50E-03	LCN2,MMP8
Granulocyte Adhesion and Diapedesis	HP vs NF	1.26E-02	CXCR1,MMP8
Cholesterol Biosynthesis I	HP vs NF	1.29E-02	SC5D
Cholesterol Biosynthesis II (via 24,25-dihydrolanosterol)	HP vs NF	1.29E-02	SC5D
Airway Pathology in Chronic Obstructive Pulmonary Disease	NF vs AH	4.27E-08	CTSG,ELANE,LCN2,MMP8,MPO,ORM1,TNFSF10

Iron homeostasis signaling pathway	NF vs AH	4.68E-05	ALAS2,HBD,HBQ1,HP,SLC25A37
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	NF vs AH	7.24E-05	C1QA,C1QB,NLRC4,TLR8,TNFSF10
Airway Inflammation in Asthma	NF vs AH	1.26E-04	ELANE,RNASE2,RNASE3
Granulocyte Adhesion and Diapedesis	NF vs AH	1.66E-04	CCR2,CXCL5,CXCR1,FPR2,MMP8
IL-15 Signaling	NF vs CT	1.55E-09	IGHG3,IGHG4,IGKV1-12,IGKV1-39,IGLC3,IGLV3-10
B Cell Receptor Signaling	NF vs CT	1.07E-08	IGHG3,IGHG4,IGKV1-12,IGKV1-39,IGLC3,IGLV3-10
Systemic Lupus Erythematosus In B Cell Signaling Pathway	NF vs CT	3.63E-08	IGHG3,IGHG4,IGKV1-12,IGKV1-39,IGLC3,IGLV3-10
Communication between Innate and Adaptive Immune Cells	NF vs CT	4.27E-08	IGHG3,IGHG4,IGKV1-12,IGKV1-39,IGLC3,IGLV3-10
Primary Immunodeficiency Signaling	NF vs CT	2.00E-06	IGHG3,IGHG4,IGLC3
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	NF vs AC	2.04E-07	C1QA,C1QB,NLRC4,TLR8,TNFSF10
Iron homeostasis signaling pathway	NF vs AC	6.46E-06	ALAS2,HBD,HBQ1,SLC25A37
Erythropoietin Signaling Pathway	NF vs AC	5.01E-04	HBD,HBQ1,TNFSF10
Complement System	NF vs AC	5.13E-04	C1QA,C1QB
TREM1 Signaling	NF vs AC	2.04E-03	NLRC4,TLR8

* The colors are alternated between blue and white to highlight each pairwise comparison group.

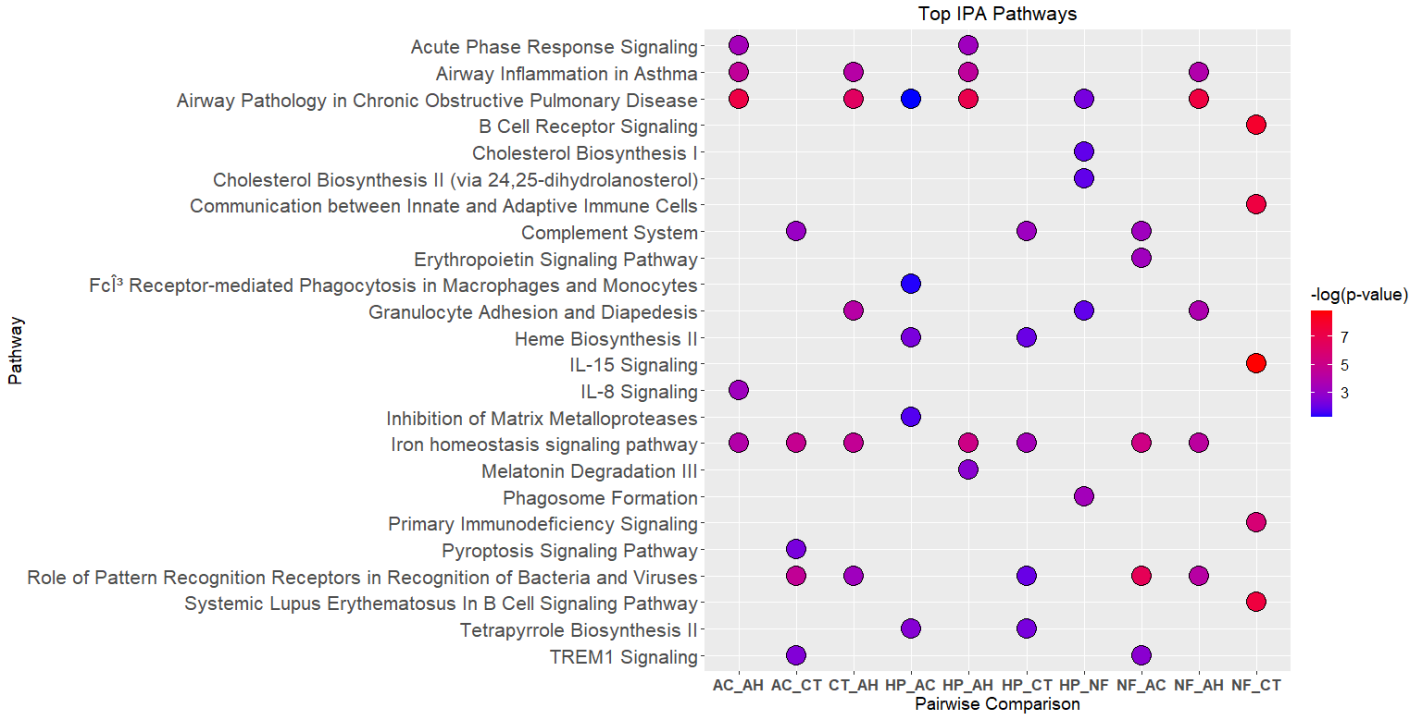

Figure S11: Dot plot of top 5 IPA pathways and their p-value significance for each pairwise comparison of PBMC 5-Way best gene set. The dots are color-coded by p-value significance, with blue dots representing lower significance and red representing higher significance.

GSEAPreranked.

1. *LV 5-Way.*

Table S30: Top enriched GSEA pathways per pairwise comparison in LV 5-Way best gene set.

GSEA Canonical Pathways	Pairwise Comparison	NES	p-value
Biological process involved in interspecies interaction between organisms	AH vs CT	1.374	0.098
Homeostatic process	AH vs CT	1.010	0.464
Regulation of intracellular signal transduction	AH vs CT	0.980	0.478
Pancreas ductal cell	AH vs CT	-1.178	0.255
Pancreas ductal cell	AH vs AC	1.193	0.247
Biological process involved in interspecies interaction between organisms	AH vs AC	0.988	0.473
Homeostatic process	AH vs AC	-1.527	0.047
Regulation of intracellular signal transduction	AH vs AC	-0.681	0.879
Homeostatic process	AH vs NF	-0.848	0.747
Regulation of intracellular signal transduction	AH vs NF	1.273	0.167

Pancreas ductal cell	AH vs NF	1.052	0.418
Biological process involved in interspecies interaction between organisms	AH vs NF	0.491	0.996
Homeostatic process	AH vs HP	-0.938	0.545
Biological process involved in interspecies interaction between organisms	AH vs HP	1.767	0.002
Regulation of intracellular signal transduction	AH vs HP	0.972	0.503
Pancreas ductal cell	AH vs HP	0.912	0.595
Homeostatic process	CT vs AC	-1.529	0.057
Biological process involved in interspecies interaction between organisms	CT vs AC	-1.272	0.182
Regulation of intracellular signal transduction	CT vs AC	-1.105	0.315
Pancreas ductal cell	CT vs AC	1.532	0.063
Pancreas ductal cell	CT vs NF	1.457	0.053
Homeostatic process	CT vs NF	0.851	0.654
Regulation of intracellular signal transduction	CT vs NF	0.539	0.968
Biological process involved in interspecies interaction between organisms	CT vs NF	-1.655	0.009
Pancreas ductal cell	CT vs HP	1.349	0.120
Biological process involved in interspecies interaction between organisms	CT vs HP	1.170	0.287
Homeostatic process	CT vs HP	0.921	0.579
Regulation of intracellular signal transduction	CT vs HP	0.783	0.730
Regulation of intracellular signal transduction	AC vs NF	1.227	0.229
Homeostatic process	AC vs NF	1.147	0.285
Pancreas ductal cell	AC vs NF	0.660	0.876
Biological process involved in interspecies interaction between organisms	AC vs NF	-1.030	0.412
Pancreas ductal cell	AC vs HP	-0.953	0.492
Biological process involved in interspecies interaction between organisms	AC vs HP	1.848	0.001
Homeostatic process	AC vs HP	1.363	0.104
Regulation of intracellular signal transduction	AC vs HP	1.100	0.366
Biological process involved in interspecies interaction between organisms	NF vs HP	1.848	0.001
Homeostatic process	NF vs HP	1.363	0.104
Regulation of intracellular signal transduction	NF vs HP	1.100	0.366
Pancreas ductal cell	NF vs HP	-0.953	0.492

* The colors are alternated between blue and white to highlight each pairwise comparison group. NES is the Normalized Enrichment Score calculated by GSEA.

2. PBMC 5-Way.

Table S31: Top enriched GSEA pathways per pairwise comparison in PBMC 5-Way best gene set.

GSEA Canonical Pathways	Pairwise Comparison	NES	p-value
Neutrophil degranulation	AH vs CT	-1.444	0.051
Defense response to bacterium	AH vs CT	-1.337	0.123
Innate immune system	AH vs CT	-1.279	0.144
Cell cell signaling	AH vs CT	1.319	0.130
Cellular response to oxygen containing compound	AH vs CT	1.197	0.238
Locomotion	AH vs CT	1.075	0.347
Neutrophil degranulation	AH vs AC	-1.829	0
Innate immune system	AH vs AC	-1.733	0.001
Defense response	AH vs AC	-1.692	0.001
Biological process involved in interspecies interaction between organisms	AH vs AC	-1.612	0.007
Response to bacterium	AH vs AC	-1.593	0.004
Response to molecule of bacterial origin	AH vs AC	-1.592	0.005
Lung proliferating macrophage cell	AH vs NF	-1.510	0.054
Pancreas ductal cell	AH vs NF	-1.484	0.062
Lung neutrophil cell	AH vs NF	-1.437	0.085
Innate immune system	AH vs NF	-1.434	0.073
Neutrophil degranulation	AH vs NF	-1.424	0.073
Homeostatic process	AH vs NF	-1.336	0.140
Neutrophil degranulation	AH vs HP	-1.783	0
Defense response	AH vs HP	-1.697	0.001
Defense response to bacterium	AH vs HP	-1.695	0.002
Response to bacterium	AH vs HP	-1.689	0.002
Antimicrobial humoral response	AH vs HP	-1.577	0.007
Innate immune system	AH vs HP	-1.571	0.010
Antimicrobial humoral response	CT vs AC	-2.049	0
Response to lipid	CT vs AC	-1.908	0
Response to molecule of bacterial origin	CT vs AC	-1.785	0.014
Response to bacterium	CT vs AC	-1.734	0.017
Chemical homeostasis	CT vs AC	1.600	0.015
Homeostatic process	CT vs AC	1.507	0.022
Adaptive immune response	CT vs NF	-2.229	0
Immune response	CT vs NF	-1.800	0
Vesicle mediated transport	CT vs NF	-1.532	0.055
Leukocyte mediated immunity	CT vs NF	-1.527	0.045
Chemical homeostasis	CT vs NF	1.443	0.068

Neutrophil degranulation	CT vs NF	1.433	0.058
Defense response to bacterium	CT vs HP	-2.518	0
Response to bacterium	CT vs HP	-2.449	0
Antimicrobial humoral response	CT vs HP	-2.421	0
Antimicrobial peptides	CT vs HP	-2.238	0
Response to molecule of bacterial origin	CT vs HP	-1.934	0.003
Response to lipid	CT vs HP	-1.928	0.004
Antimicrobial humoral response	AC vs NF	2.590	0
Defense response to bacterium	AC vs NF	2.037	0
Response to lipid	AC vs NF	2.009	0
Response to molecule of bacterial origin	AC vs NF	1.973	0.005
Response to bacterium	AC vs NF	1.882	0
Antimicrobial peptides	AC vs NF	1.822	0.013
Phosphorylation	AC vs HP	-1.540	0.067
Antimicrobial peptides	AC vs HP	-1.512	0.045
Positive regulation of molecular function	AC vs HP	-1.421	0.101
Programmed cell death	AC vs HP	-1.419	0.108
Lung neutrophil cell	AC vs HP	1.999	0.004
Adaptive immune response	AC vs HP	1.530	0.038
Lung neutrophil cell	NF vs HP	1.898	0.004
Adaptive immune response	NF vs HP	1.564	0.037
Antimicrobial humoral response	NF vs HP	-2.647	0
Antimicrobial peptides	NF vs HP	-2.472	0
Defense response to bacterium	NF vs HP	-2.323	0
Neutrophil degranulation	NF vs HP	-2.140	0

* The colors are alternated between blue and white to highlight each pairwise comparison group. NES is the Normalized Enrichment Score calculated by GSEA.

Blood Transcription Module analysis (BloodGen3Module).

1. PBMC 5-Way.

Differential blood transcription module analysis was performed on our best gene set from the 5-way PBMC dataset, using BloodGen3Module software. Thirty-nine modules were identified as shown in Table S32. The cells of the table are color-coded according to the direction and expression level of the module for each comparison group. The AH group demonstrated upregulation in most modules. Differential blood transcription module analysis is useful for adding annotation to PBMC gene expression research (14, 15).

Table S32: Blood Transcription Module (BTM) response by pairwise comparison of PBMC 5-Way best gene set.

Module name	AH_CT	AH_AC	AH_NF	AH_HP	AC_CT	AC_NF	AC_HP	NF_CT	HP_CT	HP_NF	Module title	Top GOTERM BP
M13.18											B cells	RNA processing
M12.8											B cells	B cell activation
M16.107											Oxidative stress	N/A
M9.1											Cytotoxic lymphocytes	cellular defense response
M14.27											Protein synthesis	negative regulation of catalytic activity
M16.49											Inflammation	negative regulation of cell proliferation
M12.2											Monocytes	defense response
M13.11											TBD	response to calcium ion
M15.127											Interferon	immune response
M15.58											Monocytes	molting cycle process
M16.16											TBD	negative regulation of catalytic activity
M16.27											TBD	hemopoiesis
M16.37											TBD	regulation of cellular localization
M14.50											Inflammation	inflammatory response
M16.80											Cytokines/chemokines	epidermal growth factor receptor signaling pathway
M16.67											TBD	immune response
M16.44											Protein synthesis	immune response
M16.1											TBD	regulation of leukocyte migration
M15.66											TBD	retinoic acid metabolic process
M16.102											TBD	protein kinase cascade
M13.12											Inflammation	response to wounding
M13.16											Cytokines/chemokines	glucan catabolic process
M15.26											Neutrophils	leukocyte activation
M13.22											Neutrophils	response to bacterium
M15.37											Inflammation	apoptosis
M15.84											Cytokines/chemokines	protein kinase cascade
M12.10											Inflammation	regulation of cell morphogenesis
M15.109											Inflammation	inflammatory response
M16.82											Gene transcription	response to organic substance
M9.2											Erythrocytes	erythrocyte differentiation
M12.11											Erythrocytes	hexose metabolic process
M13.30											Erythrocytes	oxygen transport
M10.4											Neutrophil activation	defense response to bacterium
M16.96											Erythrocytes	defense response
M16.11											Protein synthesis	intracellular transport
M16.3											T cells	lymphocyte activation
M16.8											TBD	protein transport
M15.6											Cell cycle	modification-dependent macromolecule catabolic proce
M16.30											Complement	immune response

* Cells in shades of red are upregulated for the condition listed first, and shades of green if downregulated for the condition listed first.

e. Misclassified Sample Analysis:

As part of our analysis, we also examined whether there were any samples that proved to be particularly difficult to classify within our data. For the misclassified sample analysis, we examined only a fraction of our 36 configurations. Specifically, we examined the following 6 configurations: (LR + DE Feature Selection) x (Intersection/Union) x (2.5/3.0/3.5 Threshold). If a given sample was misclassified across all feature sizes in each of the 6 configurations, it was labeled as frequently misclassified. For example, if the logistic regression algorithm could never correctly classify the sample, regardless of feature size, and how the features were selected and filtered, then it was labeled as frequently misclassified. Table S30 summarizes the frequently misclassified samples found in our dataset.

Table S33: Frequently misclassified samples in each dataset.

Dataset	Frequently Misclassified Samples
LV 2-Way	None.
LV 3-Way	Two AC samples. Both misclassified as AH.
LV 5-Way	Three AC samples. Two misclassified as AH, one as NF.
PBMC 5-Way	3 AC, 1 NF, and 1 HP samples. All AC samples were primarily misclassified as AH. The NF sample was misclassified as AC. The HP sample was mostly misclassified as CT.

While the clinical data was lacking for most liver samples, the clinical data was available for all our PBMC samples. Therefore, we were able to examine whether the frequently misclassified PBMC samples were unusual in any way, based upon clinical parameters. Specifically, we examined the BMI, MELD, and DF scores. All of the frequently misclassified AC samples were notable for having some of the highest MELD and DF scores for their

condition. This suggests that severity may play a role in the way that the AC and AH conditions were being distinguished within the PBMC 5-Way dataset. The frequently misclassified NF and HP samples did not possess any unusual or outlying clinical parameters. Therefore, we could only speculate as to the reasons behind their frequent misclassification.

f. AH PBMC-LV Analysis:

Both liver and PBMC samples were collected from 19 alcohol-associated hepatitis (AH) participants. We performed differential expression analysis of 19 AH liver samples against 8 CT liver samples, and of 19 AH PBMC samples against 20 CT PBMC samples. We filtered the results using the following cutoffs: FPKM > 1, Q-Value < 0.05, and log2(FC) > 1. We then identified genes that were similarly upregulated and downregulated within both tissues as compared to CT samples from the same tissue type. As shown in Table S34, there were 37 genes that were upregulated in AH compared to CT within both tissues, and 3 genes that were downregulated in AH compared to CT within both tissues.

Table S34: Genes that were similarly upregulated and downregulated within both PBMC and LV tissues for AH vs CT comparison.

Downregulated	IFITM1, IGFBP4, MFAP3L.
Upregulated	ADAM9, AIF1, ANXA3, APOBEC3A, BLVRA, C3AR1, CPVL, CSF2RA, CTSS, FCN1, FGR, IFI44L, LGALS3, LGALS9, LILRB4, LY96, MILR1, MMP14, MNDA, MS4A4A, NCF1, NCF2, OSCAR, PECAM1, PILRA, PTAFR, SECTM1, SIRPB1, SLC11A1, SLC7A7, SNCA, ST14, TESC, TIMP2, TNFRSF21, TNFSF13B, VCAN.

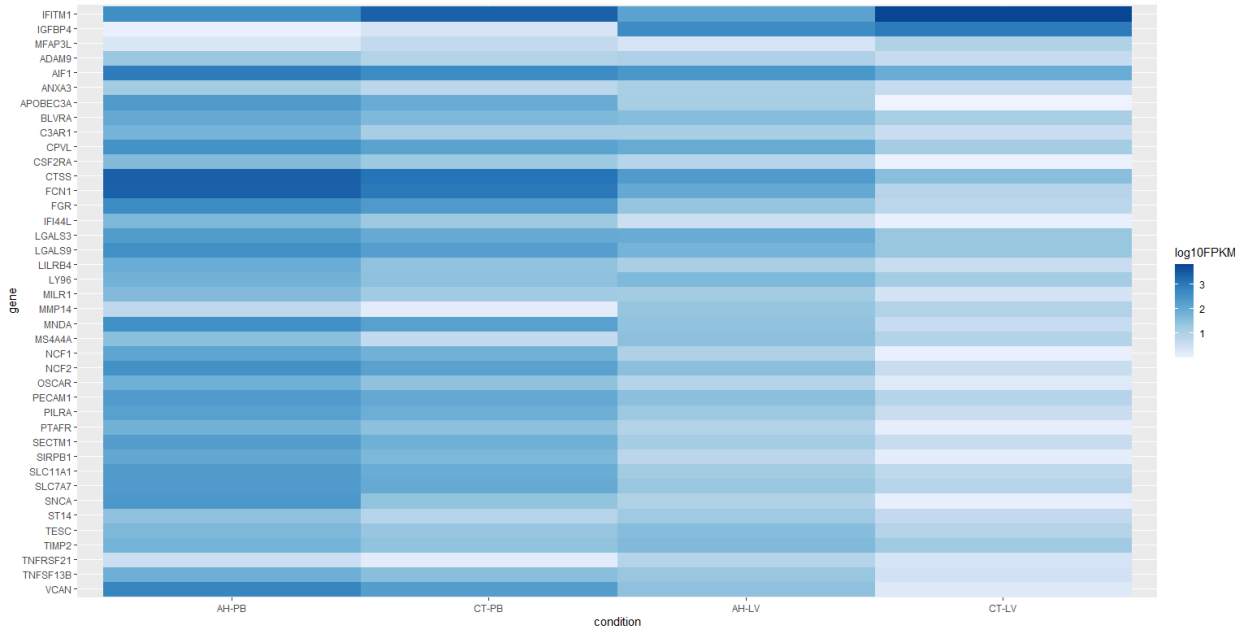

Figure S12: Heatmap of genes that were similarly upregulated and downregulated within both PBMC and LV tissues for AH vs CT comparison.

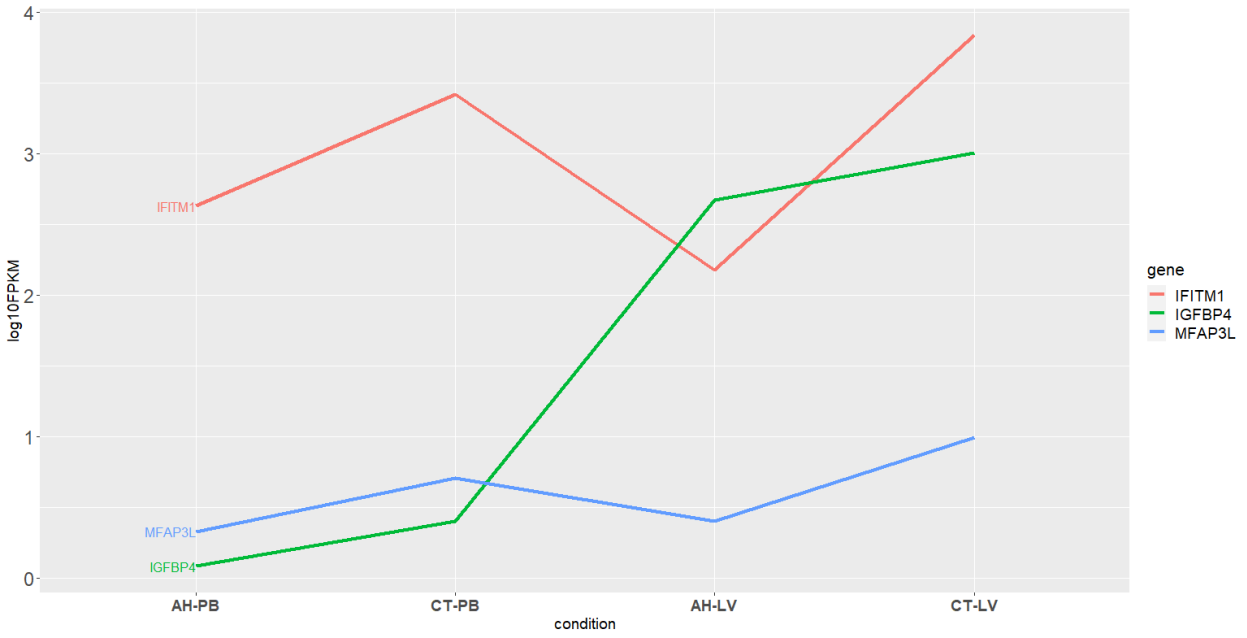

Figure S13: Line plot of 3 genes that were downregulated in AH vs. CT, within both PBMC and liver tissues.

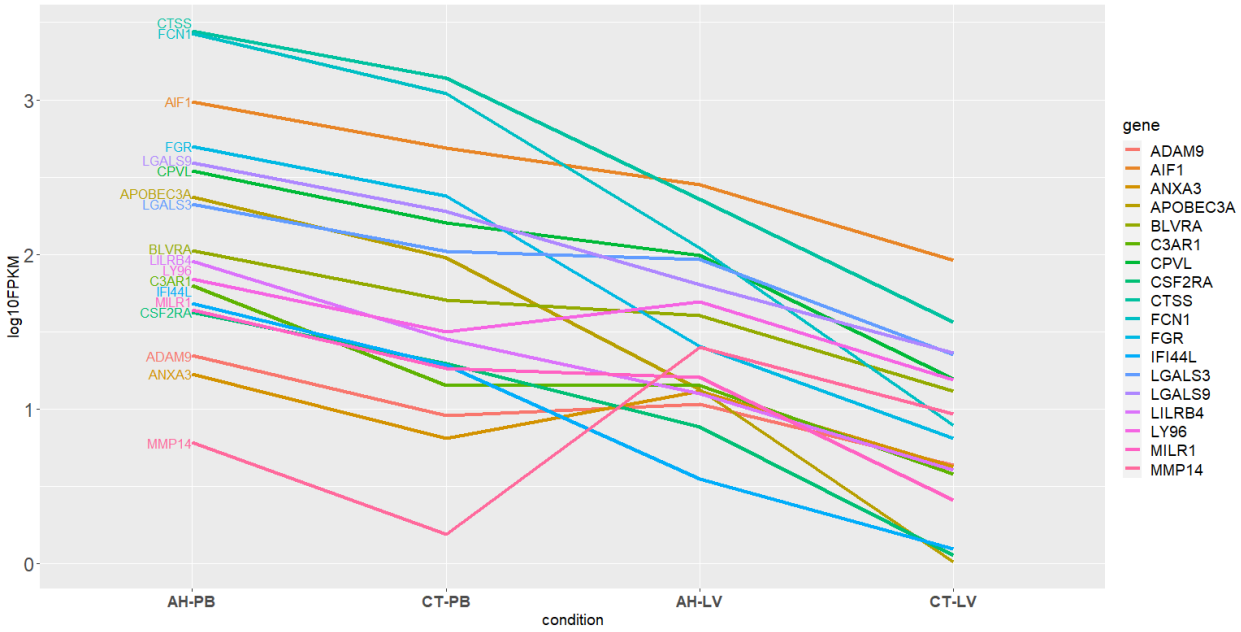

Figure S14a: Line plot of 18 genes that were upregulated in AH vs. CT within both PBMC and liver tissues.

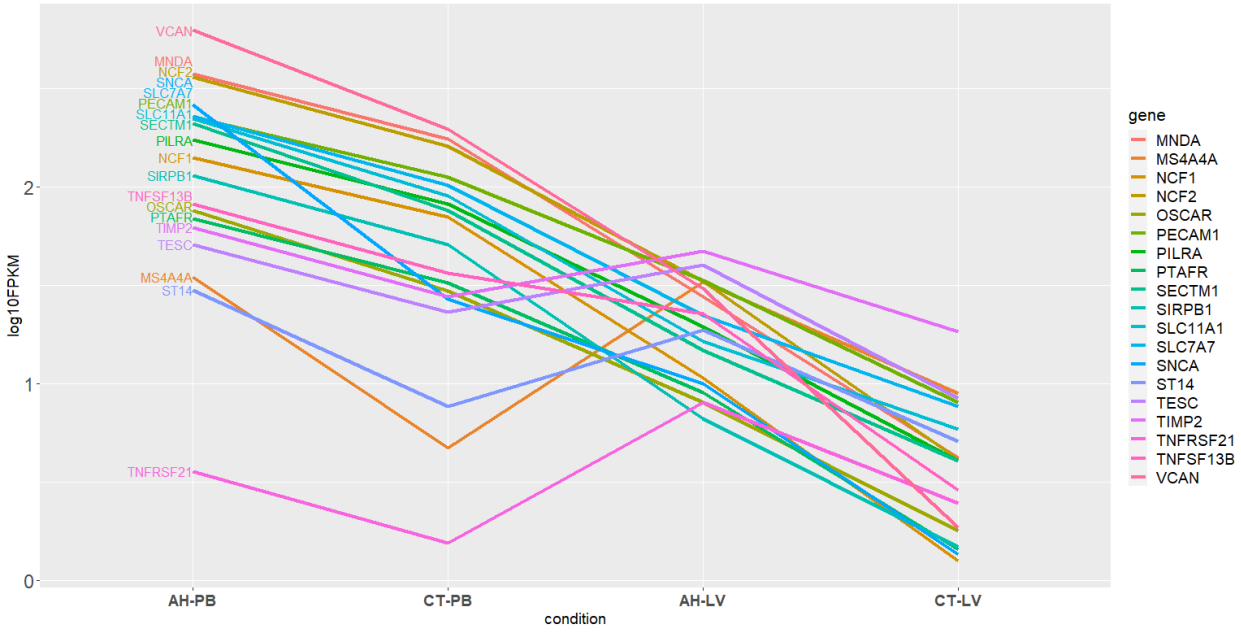

Figure S14b: Line plot of remaining 19 genes that were upregulated in AH vs. CT within both PBMC and liver tissues.

The 40 genes that were similarly up or down regulated in both tissues are visualized by a heatmap (Fig. S12) and line plots (Figs. S13 and S14). Fig. S14 was split into two line plots to improve the readability of the individual lines. Additionally, we examined which of these 40 genes were also present in PBMC and LV 5-Way best gene sets, and found that there were 3 genes that matched exactly to our best gene sets: ANXA3, IFITM1, and IFI44L. There were also several genes belonging to the same gene families within both tissues (e.g., matrix metalloproteinase: MMP7, MMP8, MMP14; iron homeostasis: SLC25A37, SLC11A1; and Tumor Necrosis Factor: TNFS10, TNFRSF21, TNFSF13B). These genes are present in several of the key pathways that are altered during alcohol-associated hepatitis. Because these genes show similar expression directionality within both liver tissue and PBMCs, they may potentially serve as effective biomarkers for AH.

APPENDIX A: CHAPTER 3 SUPPLEMENTAL REFERENCES

1. Massey V, Parrish A, Argemi J, Moreno M, Mello A, García-Rocha M, et al. Integrated Multiomics Reveals Glucose Use Reprogramming and Identifies a Novel Hexokinase in Alcoholic Hepatitis. Gastroenterology. 2021;160(5):1725-1740.

2. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley DR, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 2013;14.

3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods 2012;9:357-U354.

4. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 2019;37:907-915.

5. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.

6. Kampf C, Mardinoglu A, Fagerberg L, Hallstrom BM, Edlund K, Lundberg E, et al. The human liver-specific proteome defined by transcriptomics and antibody-based profiling. Faseb Journal 2014;28:2901-2914.

7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114-2120.

8. 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;7:562-578.

9. Hart SN, Therneau TM, Zhang YJ, Poland GA, Kocher JP. Calculating Sample Size Estimates for RNA Sequencing Data. Journal of Computational Biology 2013;20:970-978.

10. 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;14.

11. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30(4):523–30.

12. Mootha V, Lindgren C, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 2003;34:267–273.

13. Rinchai D, Roelands J, Toufiq M, Hendrickx W, Altman MC, Bedognetti D, et al. BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R. *Bioinformatics*. 2021;37(16):2382–2389.

14. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nature Immunology. 2014;15:195–204.

15. Sharma S, Baweja S, Maras JS, Shasthry SM, Moreau R, Sarin SK. Differential blood transcriptome modules predict response to corticosteroid therapy in alcoholic hepatitis. JHEP Reports. 2021;3(3).

APPENDIX B: CHAPTER 4 SUPPLEMENTAL

1. SUPPLEMENTARY METHODS

Sections **a-j** below briefly describe the collection and processing of the samples that were RNA sequenced in the current study. For full methods regarding RNAseq data please refer to our previous publication (1).

The study was approved by the Department of Veterans Affairs VA Long Beach Healthcare Systems Institutional Review Board (IRB# 1254), by the Human Subjects Committee, Los Angeles Biomedical Research Institute (Project No. 20607-0), University of Southern California Health Sciences Campus Institutional Review Board (Project # HS-13-00815), and by the University of California, Irvine Institutional Review Board, HS #2016-3064. All participants signed written consents prior to providing biospecimens.

Liver tissue and PBMC RNAseq data is being deposited in dbGaP (1).

For information about the independent RNA-seq liver tissue dataset used for external validation, please refer to GSE142530 (2).

Liver tissue proteomic test and validation data can be found in MassIVE repository (assession number MSV000089168).

PBMC proteomic data is pending deposition in MassIVE.

a. Inclusion and Exclusion Criteria (RNAseq and Proteomics):

Alcohol-associated Liver Disease (AH, AC) Donors:

Common Inclusion Criteria: History of chronic alcohol consumption sufficient to cause liver damage. Generally, this is considered to be >40 g/day for women and >60 g/day for men, for many years.

Common Exclusion Criteria: Liver disease significantly caused by hemochromatosis, autoimmune liver disease, Wilson disease, NAFLD, hepatitis C, or hepatitis B.

Specific to Alcohol-Associated Hepatitis Donors (AH):

Inclusion Criteria: A clinical diagnosis of possible alcoholic hepatitis. Serum total bilirubin >3 mg/dL.

Specific to Alcohol-Associated Liver Cirrhosis Donor (AC):

Inclusion Criteria: This group contained both abstinent and recently drinking alcohol associated cirrhosis. Inclusion Criteria for Abstinent donors: Abstinent (consumption of less than one standard drink*/week) during the 6 months prior to enrollment. Inclusion Criteria for Recently drinking donors: Heavy alcohol use until recently (stopped/reduced alcohol use within past 60 days). For the current study, both groups were combined into a single group for analysis.

Healthy Donors:

Inclusion Criteria: AUDIT-C scores of <4 for men and <3 for women (signifying no alcohol misuse). Abstinent (consumption of less than one standard drink*/week) during the 6 months prior to enrollment.

Exclusion Criteria: Clinical history or laboratory evidence of liver disease including alcoholic liver disease, NAFLD, hemochromatosis, alcoholic hepatitis, autoimmune liver disease, Wilson disease, hepatitis C, or hepatitis B. BMI>32. Any of the following laboratory abnormalities

within 90 days prior to signing the consent. - Creatinine: >1.5 mg/dL; - Hemoglobin: <12 g/dL; Total bilirubin: >1.5 mg/dL; - AST: >40 IU/mL; - ALT: >40 IU/mL.

b. Sample Processing and Liquid Chromatography Mass Spectrometry (Proteomic):

The methods used to process liver tissue samples and perform liquid chromatography-tandem mass spectrometry are described in (3). The methods used to process PBMC samples are nearly identical to the methods used to process plasma samples in the follow publication (4).

c. Reference Genome (RNAseq):

We used hg38 (GRCh38 assembly) human reference genome, downloaded from the UCSC Genome Browser. ChrM was not included in the assembly.

d. Gene Annotations (RNAseq):

We used Ensembl release 91 (Dec 2017) annotataion.

e. Short-read alignment to reference genome and transcriptome (RNAseq):

We used STAR 2.6.0 (5) aligner with default settings (STARCQ).

f. Sample Sequencing (RNAseq):

RNA was isolated from the cell pellets and liver tissue according to total RNA extraction kit instructions (Qiagen RNAeasy kit). Total RNA was monitored for quality control using the Agilent Bioanalyzer Nano RNA chip and Nanodrop absorbance ratios for 260/280nm and 260/230nm. Library construction was performed according to the Illumina TruSeq mRNA stranded protocol.

All samples included in this study were RNA sequenced on an Illumina platform by the Genomics High-Throughput Facility (GHTF) at the University of California, Irvine (UCI), except for one healthy liver sample for which the sequencing data was directly downloaded from the European Bioinformatics Institute (EBI) ArrayExpress database (accession number E-

MTAB-1733) (6). The number of paired or single reads per sample was approximately 140M before filtering and decontamination.

g. Read Trimming & Quality Filters (RNAseq):

The sequencing reads in each dataset were first filtered to remove low quality reads. On average, 9.62% of the original reads were discarded during this step and 15.43% of the paired reads were orphaned. The mean PHRED quality score of the remaining reads was approximately 40.

h. Sample Decontamination (RNAseq):

The remaining quality-filtered and trimmed reads for each dataset were then further filtered to remove possible contaminants in each sample such as PhiX control reads or bacterial contamination. In addition, both the human mitochondrial genome and ribosomal DNA/RNA sequences were treated as contaminants during this step due to highly variable quantities of these reads in the various datasets generated during the experiment, ranging from a few percent of the reads in most cases to about 80% of the reads for some highly contaminated samples. On average, approximately 115M paired and single reads were left per sample and used for the gene expression analysis described in the next sections.

i. Normalized counts before and after application of log transformation (RNAseq and Proteomics):

RNAseq counts were transformed using ln(1+count) formula. The proteomic counts did not exhibit same properties as RNAseq counts and were not log transformed.

j. Alignment Pipeline Selection (RNAseq):

Based on preliminary analysis we decided to use hg38 Ensembl Starcq pipeline throughout rest of the study.

k. Nested Cross-Validation Setup (RNAseq and Proteomics):

We utilized nested cross-validation to attain the estimates of classification performance for various feature selection (FS) strategies, classifiers, and feature sizes within our data. The best feature (gene) sets selected for each of the datasets were then validated in the independent test set. The nested cross-validation was implemented in the standard configuration with k = 5 in both the inner and outer loops. The outer loop was used for model evaluation (i.e., classification performance), while the inner loop was used for model selection (i.e., hyper-parameter tuning). The feature selection was done within both inner and outer loops. That is FS was done for each training set in inner and outer loops. This means that effectively there were 30 training sets (25 in inner loop, 5 in outer loop) as part of a single nested cross-validation execution. Feature selection occurred for each of these training sets.

Since one of our classification strategies relied on differential expression as computed by Cuffdiff (7), the feature selection process within nested cross validation was time consuming. A single Cuffdiff analysis could require anywhere from 30 minutes to 5 hours depending on the number of samples. In order to keep runtime reasonable, all folds were pre-defined, and only a single splitting of samples into folds (for both inner and outer loops) was used within each dataset. Typically, multiple repeated data splits of samples to folds are desired to obtain best estimate of classifier's performance. However, due to Cuffdiff's large runtime performing multiple data splits proved to be prohibitive.

The proteomic differential expression was computed using INFERNORDN, otherwise the same nested cross validation procedure was applied to both gene and protein expression data (8).

l. Feature Selection Strategies (RNAseq and Proteomics):

Based on preliminary analysis we have identified filter feature selection to be best suited for small sample size RNAseq data. The two filter feature selection methods we selected are: differential expression and information gain.

For proteomic data we immediately settled on using filter feature selection in the form of differential expression. The RNAseq and proteomic data are similar in sample (~10s) and feature sizes (~10,000s). Therefore, we assumed that filter feature selection would be the best approach in both types of data.

m. Differential Expression (DE) Feature Selection (RNAseq and Proteomics):

RNAseq:

For every training set all pairwise comparisons were filtered by normalized FPKM (> 1.0) and q-values (< 0.05). All of the genes belonging to each pairwise comparison were then sorted by absolute log2(fold change) value, and the top gene for each pairwise comparison was taken. If that gene was not already in the top genes list, the gene was added to the list. The algorithm continued to cycle through the pairwise comparisons until the desired number of genes was reached. This procedure was used for all the datasets. The best features for each training set were then stored in text files.

Other DE feature selection approaches were implemented and tested by us as well. However, we found that pairwise DE selection was best performer since other DE feature selection approaches, we tested were too easily biased by the most strongly differentially expressed pairwise comparisons.

Proteomics:

The INFERNORDN was used to generate fold changes and q-values for proteomic counts. The results were filtered by q-value (< 0.05). Additionally, depending on imputation threshold entries that were missing data for too many samples were filtered out. The pairwise DE selection described above was used for proteins as well.

n. Information Gain (IG) Feature selection (RNAseq):

For every training set, the genes within normalized RNA-seq counts were ranked using the scikit-learn's mutual_info_classif function.

o. Imputation (Proteomics):

We used median and replacement with zero imputation strategies. Median: replace missing values using the median along each column (feature, in this case protein). Zero: replace all missing values with zeros.

We only imputed values for proteins that were missing data for 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 < threshold % of total samples were missing data. Threshold of 0% means no imputation took place and all proteins with missing values were removed.

p. Feature Sizes (RNAseq and Proteomics):

RNAseq:

We refer to the number of features selected during filter feature selection as "feature size". The feature sizes used with DE & IG feature selection were: 2, 3, 4, 5, 10, 15, 20, 25, and, 50 for LV 2-Way dataset and 10, 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 for the other three datasets. The feature sizes denote the number of features selected within each training set. We found during preliminary testing that we required at least 5-10 features per training set to attain reasonable classification performance and that we generally did not see benefit in using

more than 500 features per training set. The maximum feature size was also influenced by our power size calculation (that is number of significantly differentially expressed genes within our datasets).

Proteomics:

The feature sizes for proteomic data were largely based on our findings when dealing with RNAseq data. The following feature sizes were selected: 15, 25, 35, 50, 60, 70, 80, 90, 100, 150, and 200.

q. Performance Metrics (RNAseq and Proteomics):

Several different ML performance metrics were evaluated for use in this project including overall accuracy, per-class accuracy, balanced accuracy, confusion matrices, Matthews Correlation Coefficient (MCC), and F1-score. Balanced accuracy, MCC, and F1-score attempt to account for class sizes when evaluating performance, while the confusion matrices provide information about both class sizes and also per-class accuracies. Therefore, we chiefly reported our classification performance in the form of confusion matrices.

Given that our sample sizes for proteomic data were largely the same as for RNAseq data we continued to use accuracies and confusion matrices as chief means of evaluating classification performance.

r. Machine Learning Classifiers (RNAseq and Proteomics):

For RNAseq classification we decided to use logistic regression (LR), k nearest neighbors (kNN), and support vector machine (SVM) classifiers based on preliminary analysis. For proteomic analysis we used logistic regression classifier as it proved to be the fastest, best performing, and easiest to interpret during preliminary analyses.

s. Sample Size Calculation (RNAseq and Proteomic):

We expected there to be <= 450 significantly differentially expressed genes (SDEGs) in our RNAseq data based on preliminary power size calculation. In proteomic data we primarily relied on q-value as output by INFERNORDN to establish significance of differentially expressed proteins.

t. Enrichr Libraries (RNAseq):

The genes selected during feature selection were computationally evaluated using gene enrichment analysis via Enrichr (9) with pathway, tissue, and disease Enrichr libraries listed below. Custom code was written using regular expressions to match: a) immune system pathways; b) cell types that comprise blood and liver tissues; c) diseases included the conditions within this study (AH, AC) along with several other liver and blood disorders.

In order to attain the top three Enrichr hit tables (Tables S3 and S5) we performed the following steps. Enrichr hits for the best gene sets, after matching using the regular expressions, were sorted by adjusted p-value with a cutoff of 0.05. We removed entries with redundant term names or genes. We then displayed up to three top entries for each category: pathway, tissue, disease.

Enrichr Libraries used:

Pathways: 'BioPlanet_2019', 'WikiPathways_2019_Human', 'KEGG_2019_Human', 'GO_Biological_Process_2018'.

Tissues: 'ARCHS4_Tissues', 'Human_Gene_Atlas'.

Diseases: 'Disease_Perturbations_from_GEO_up', 'Disease_Perturbations_from_GEO_down'.

u. AGOTOOL Libraries (Proteomics):

The proteins selected during feature selection were computationally evaluated using protein enrichment analysis via AGOTOOL (10) with pathway, tissue, and disease AGOTOOL libraries listed below. Custom code was written using regular expressions to match: a) immune system pathways; b) cell types that comprise blood and liver tissues; c) diseases included the conditions within this study (AH, AC) along with several other liver and blood disorders.

In order to attain the top three AGOTOOL hit tables (Tables S4 and S6) we performed the following steps. AGOTOOL hits for the best protein sets, after matching using the regular expressions, were sorted by adjusted p-value with a cutoff of 0.05. We removed entries with redundant term names or proteins. We then displayed up to three top entries for each category: pathway, tissue, disease.

AGOTOOL Libraries used:

Pathways: ‘GO biological process’, ‘KEGG’, ‘WikiPathways’.

Tissues: ‘Brenda Tissue Ontology’.

Diseases: ‘Disease Ontology’.

v. Regular Expression (Regex) Patterns for Enrichr Libraries (RNAseq and Proteomics):

The regular expression (regex) patterns used for filtering the results returned by Enrichr and AGOTOOL are listed below.

Disease Regex:

'hepa|liver|cirrhosis|NAFLD|liver fibrosis|NASH|steatohepatitis|HCV|alcohol|sepsis|septic shock|hypercholesterolemia|hyperlipidemia|obesity'

Tissue Regex:

'Blood|Macrophage|Erythro|Platelet|Basophil|Neutrophil|Eosinophil|Cytokine|Tumor Necrosis Factor|Monocyte|Lymphocyte|Granulocyte|Dendritic|Megakaryocyte|T Cell|B Cell|NK Cell|Toll-like receptor|Fc receptor|Liver|Hepatocyte|Stellate|Kupffer|Sinusoidal Endothelial Cells|CD34+|Natural Killer Cell|PBMC|Tcell|Bcell|lymphoblast|CD8+|CD19+|CD4+|CD71+|Omentum'

Pathway Regex:

'Interferon|Immun|Interleukin|Prolactin|Complement|Chemokine|Oncostatin M|Rejection|Inflamma|IL1|IL-|selenium|osteopontin|circulation|coagulation|clotting|biosynthesis|degradation|cholesterol|lipid|TNF|steroid|metal ion|heme|metallo|CXCR|LDL|Phagocytosis|metabolism|TYROBP|AP-1|'

Additionally, the pathway regex included all of the disease and tissue terms.

w. Impact of Outlier Gene (Feature) Removal – Variance, Intersection, and Union Filtering (RNAseq and Proteomics):

RNAseq:

RNA-seq serves as a proxy for the level of gene expression in a biological sample. One challenge with interpretation of RNA-seq output, however, involves expression of non-coding genes that were presumed to be removed via poly(A)-selection. It is also common to observe genes with aberrant expression that poorly distinguish between the study conditions, thereby hindering classification performance.

Based on our observations and explanation above, we developed three strategies for removing undesirable genes: Variance, Intersection, and Union filtering. Variance filtering was implemented by removing genes in which the RNA-seq counts for at least one sample were

further than a standard deviation multiplied by the threshold from the mean in any of the conditions (AH, CT, etc.). Throughout the study, we used three threshold values: 2.5, 3.0, and 3.5. Lower thresholds resulted in more genes being eliminated, while higher thresholds resulted in less genes being eliminated. The filtered-out genes were not used in the subsequent feature selection process. The Union filter built upon the Variance filter by removing all genes that were either highly variant (as defined above) *or* non-coding as determined by ENSEMBL database's gene "biotype" column. The Intersection filter was similar to the Union filter, except that only the genes that were both highly variant *and* non-coding were removed. In addition to improving the odds of successful classification, the outlier feature filtering was also found to improve in silico biological validation of identified gene signatures, since protein coding genes are more extensively annotated than non-coding ones. These three filters also removed all genes whose counts were mostly zeroes across all samples.

Proteomics:

The concept of coding and non-coding did not apply to proteins. In case of proteomics we simply used variance filter with standard deviation thresholds of 2.5 and 3.0.

x. Summary of Methods (RNAseq and Proteomics):

RNAseq:

Table S1: The RNAseq methods used within the study.

Methods	Feature Selection	Outlier Feature Removal	ML Classifiers
Final Configuration	Filter (DE, IG).	Intersection and Union filtering. (Standard deviation thresholds: 2.5, 3.0, 3.5)	LR, kNN, SVM.

The final analysis included the following method configurations for each of the datasets: 2 feature selection strategies (DE, IG), 2 outlier feature removal strategies (Intersection, Union) each paired with three different thresholds (2.5, 3.0, 3.5), and 3 ML classifiers (LR, kNN, SVM). This resulted in a total of 36 configurations. For each configuration there was also a range of possible feature sizes as described in the feature size section above. The nested cross-validation ML metrics were recorded for each of these configurations, for each feature size.

Proteomics:

For proteomics data the settings were further narrowed down to 1 feature selection strategy, 1 outlier feature removal strategy, and 1 classifier (Table S2).

Table S2: The proteomic methods used within the study.

Methods	Feature Selection	Outlier Feature Removal	ML Classifiers
Final Configuration	Filter (DE).	Variance (Standard deviation thresholds: 2.5, 3.0)	LR.

y. Candidate Gene and Protein Sets (RNAseq and Proteomics):

RNAseq:

Since one of the overarching goals of this study was to identify characteristic gene expression signatures to diagnose liver disease using liver tissue and PBMC RNA-seq data, the next step of our pipeline involved selecting the best gene sets for our datasets. Within nested cross-validation, feature selection was performed for every training set in both inner and outer loops, resulting in 30 total gene sets (5 in outer, 25 in inner) for each feature size. The gene sets selected in the inner loops are not relevant, since the inner loop was only used for hyper-

parameter tuning. Therefore, we developed a method of merging the gene sets produced for each of the outer loop training sets. The strategy used was as follows: if a given gene appeared in N out of the 5 (k = 5 in outer loop) gene sets it was added to the merged gene set. After examining the results, we determined that N = 4 and N = 5 yielded our best results. The candidate gene sets were analyzed using Enrichr to establish their biological relevancies. The classification accuracy attained from the associated instance of the nested cross-validation of each candidate gene set was also examined.

Proteomics:

Identical methods were used to generate candidate protein sets. The only difference is that N = 3, N=4, and N=5 were used for proteomic data.

z. Best Gene Set Selection (RNAseq):

From the large collection of candidate gene sets attained by running the 36 different method configurations for each dataset across multiple feature sizes, we used the following strategy to select a single best candidate gene set for each of the four datasets. This process involved the evaluation of a combination of candidate gene set's size, classification performance, and biological relevancy metrics. The algorithm for picking best gene sets is described below.

1) The candidate gene set size was restricted between 5 (genes per pairwise comparison) to 100 total genes, if possible. Gene set sizes of between 100 and 200 were also considered, if suitable performance was not observed in candidate gene sets below 100 genes. The LV 3-Way dataset contains 3 pairwise comparisons. Therefore, the candidate gene set sizes, using the range guidelines above for each dataset, are as follows: 15-100 genes for LV 3-Way. The gene set size guidelines were developed to minimize the chance of either under- or overfitting.

2) Biological relevancy as indicated by Enrichr was prioritized slightly higher than the classification accuracy. That is, gene sets with highest number of pathway, tissue, and disease hits were examined in detail first. Gene sets were only considered if they included at least 10 pathway hits, 1 tissue hit, and 3 disease hits. The tissue, pathway, and disease hits were examined to verify that they were appropriate and relevant to the disease groups.
3) Total and per-class classification accuracies were considered after the in silico biological relevancy. In general, only gene sets within 10% of the best recorded performance (for a given dataset) were considered.

Once a single gene set that best satisfied all 3 criteria was selected, it was used to generate the heatmaps, confusion matrices, and pathway analysis. The liver tissue gene sets selected from our data set were evaluated with the independent validation dataset.

aa. Best Protein Set Selection (Proteomics):

This was identical to best gene set selection, except one additional criterial was added to the best gene set algorithm. The protein sets generated by configurations with least imputation were preferred.

ab. Codebase (RNAseq and Proteomics):

Github: https://github.com/staslist/AH-Project The repository contains the code used to perform the analysis. Directories and sample names have been removed from the codebase.

2. SUPPLEMENTAL RESULTS

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

RNAseq:

Test:

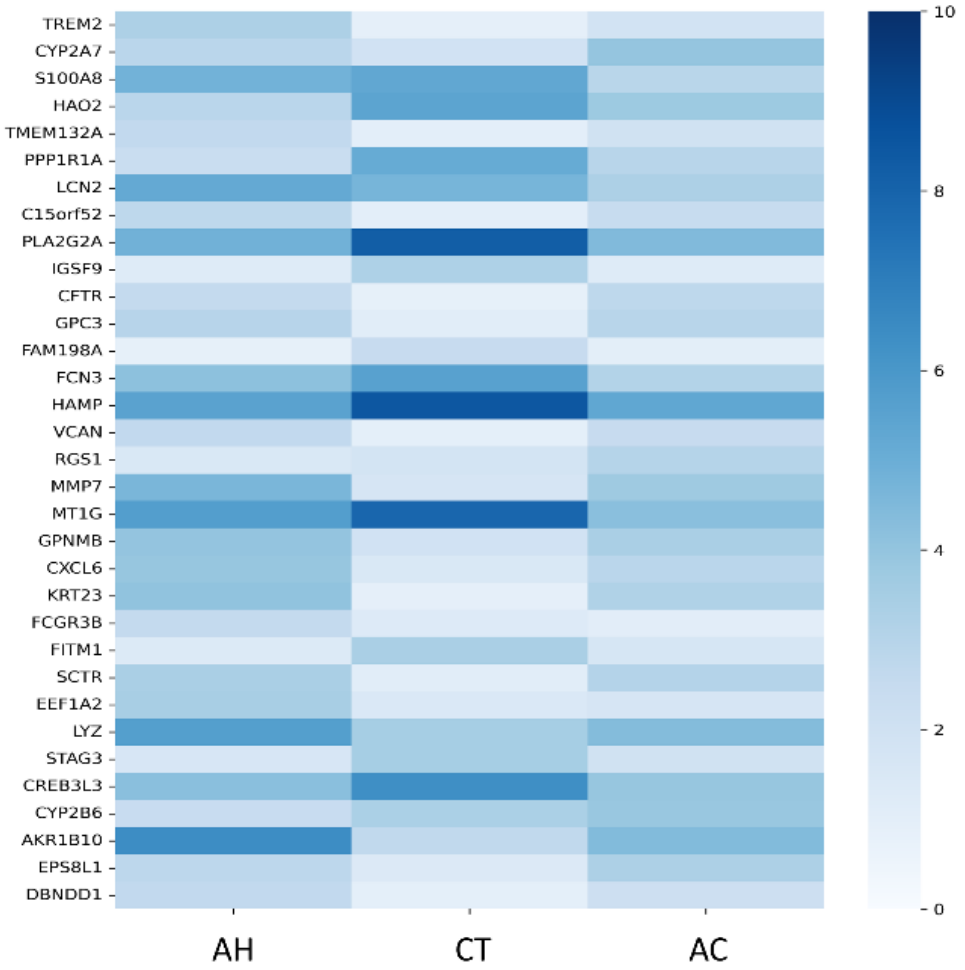

Figure S5: Heatmap of RNAseq counts for Liver 3-Way Full dataset averaged per condition.

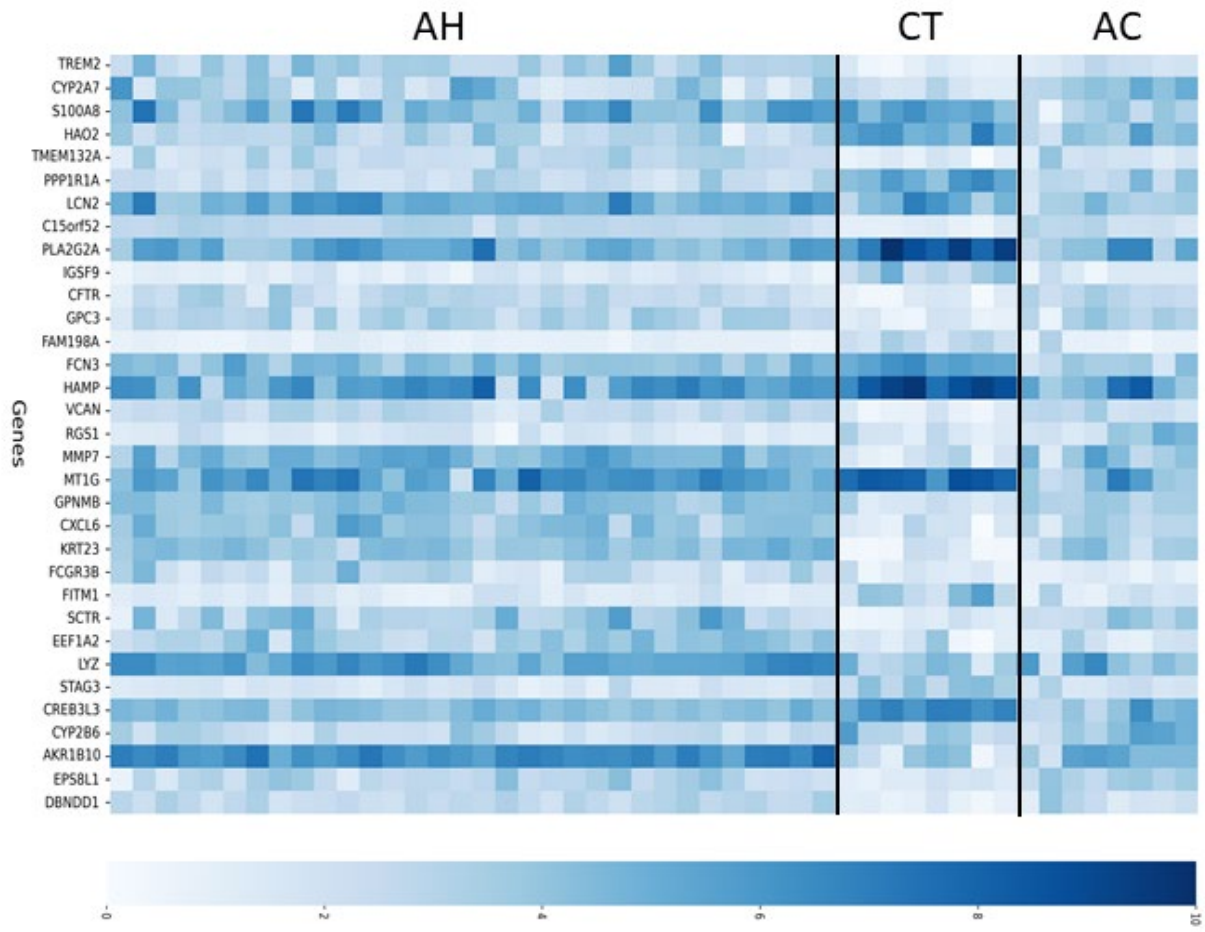

Figure S6: Heatmap of RNAseq counts for Liver 3-Way dataset.

Validation:

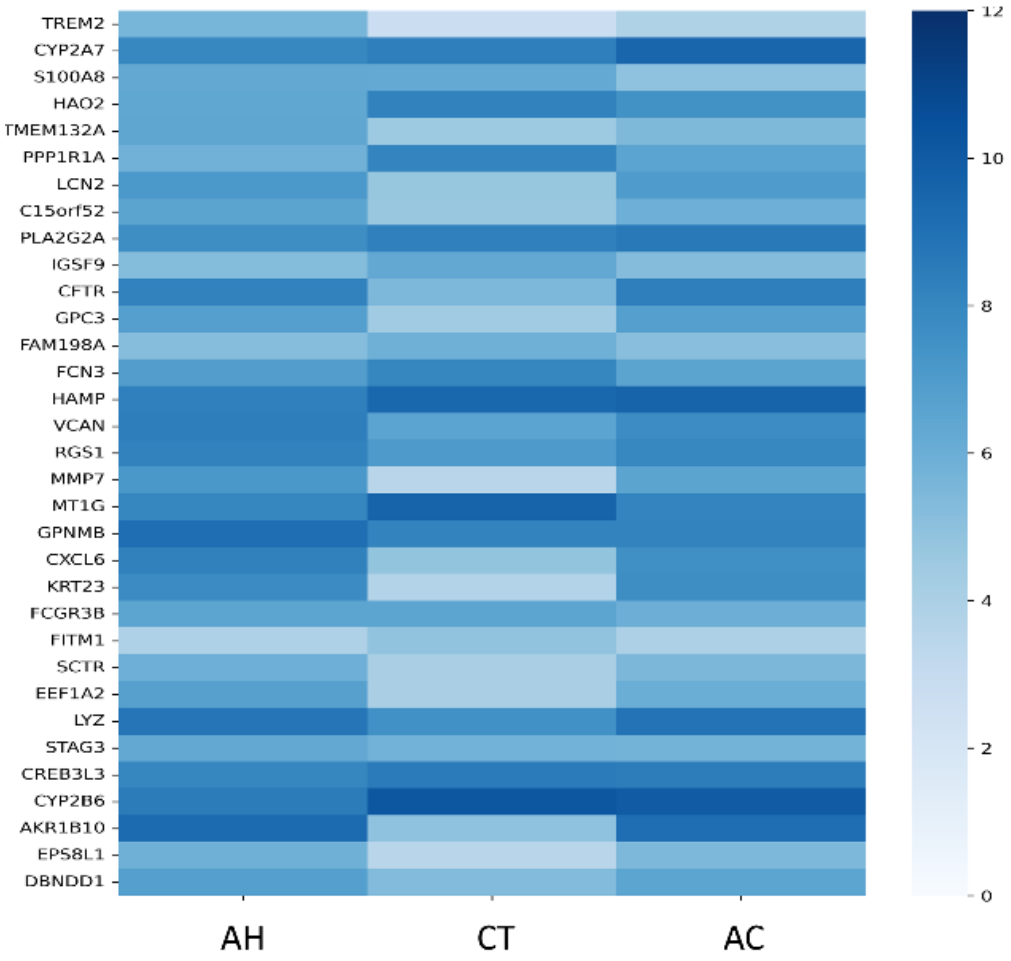

Figure S7: Heatmap of RNAseq counts for independent liver validation dataset averaged per condition.

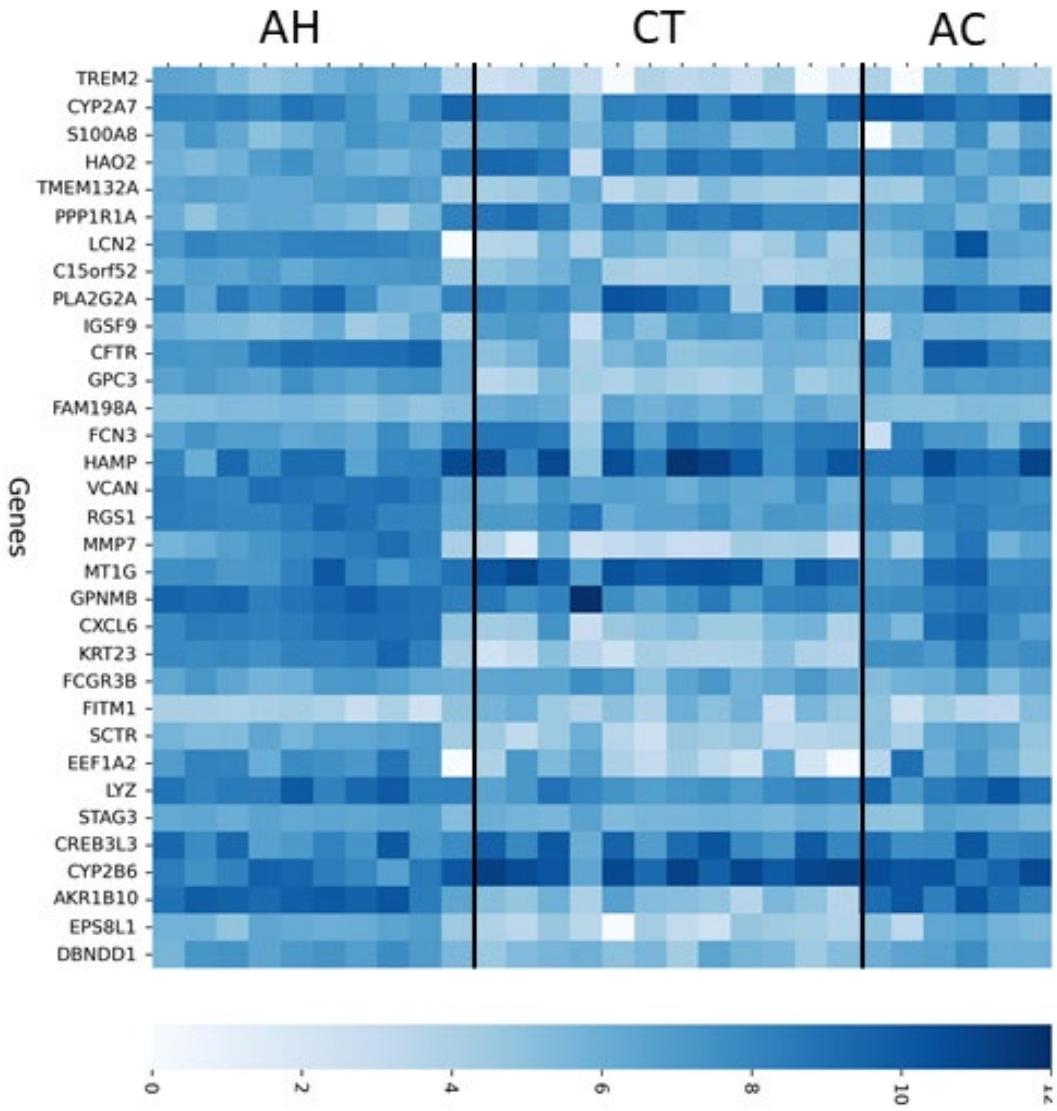

Figure S8: Heatmap of RNAseq counts for independent liver validation dataset.

Proteomics:

Test:

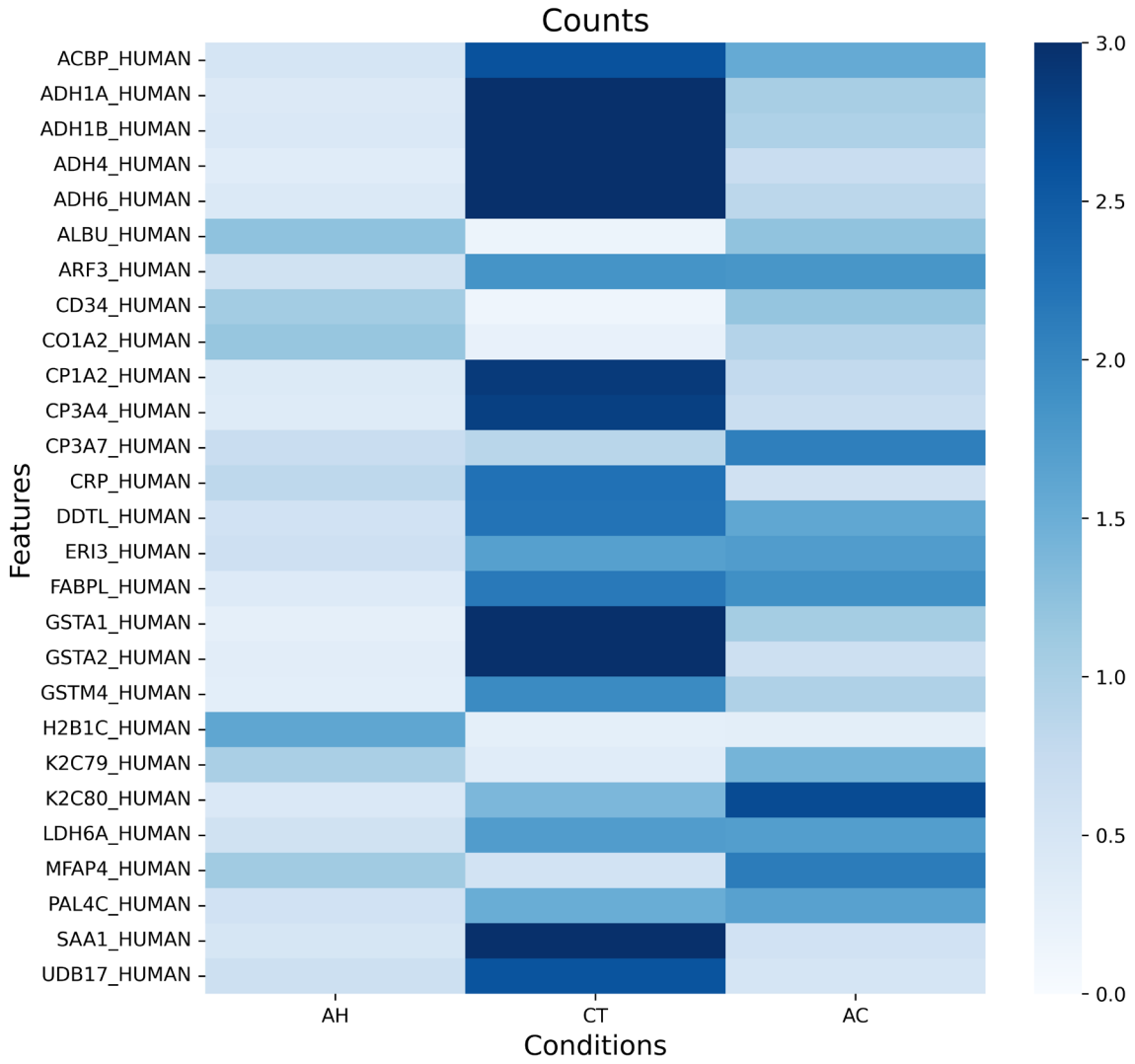

Figure S9: Heatmap of proteomic counts for Liver 3-Way Full dataset averaged per condition.

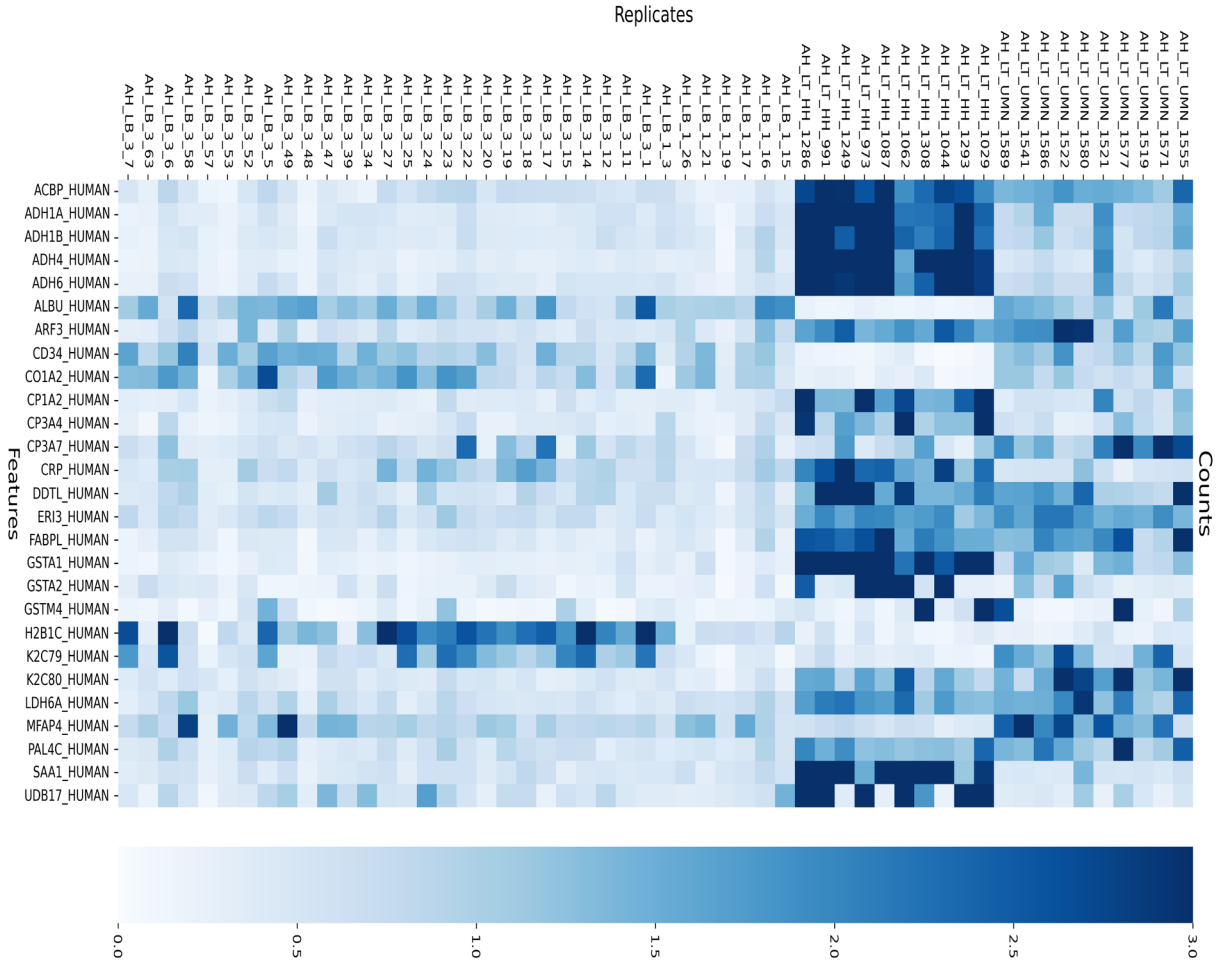

Figure S10: Heatmap of proteomic counts for Liver 3-Way Full dataset.

Validation:

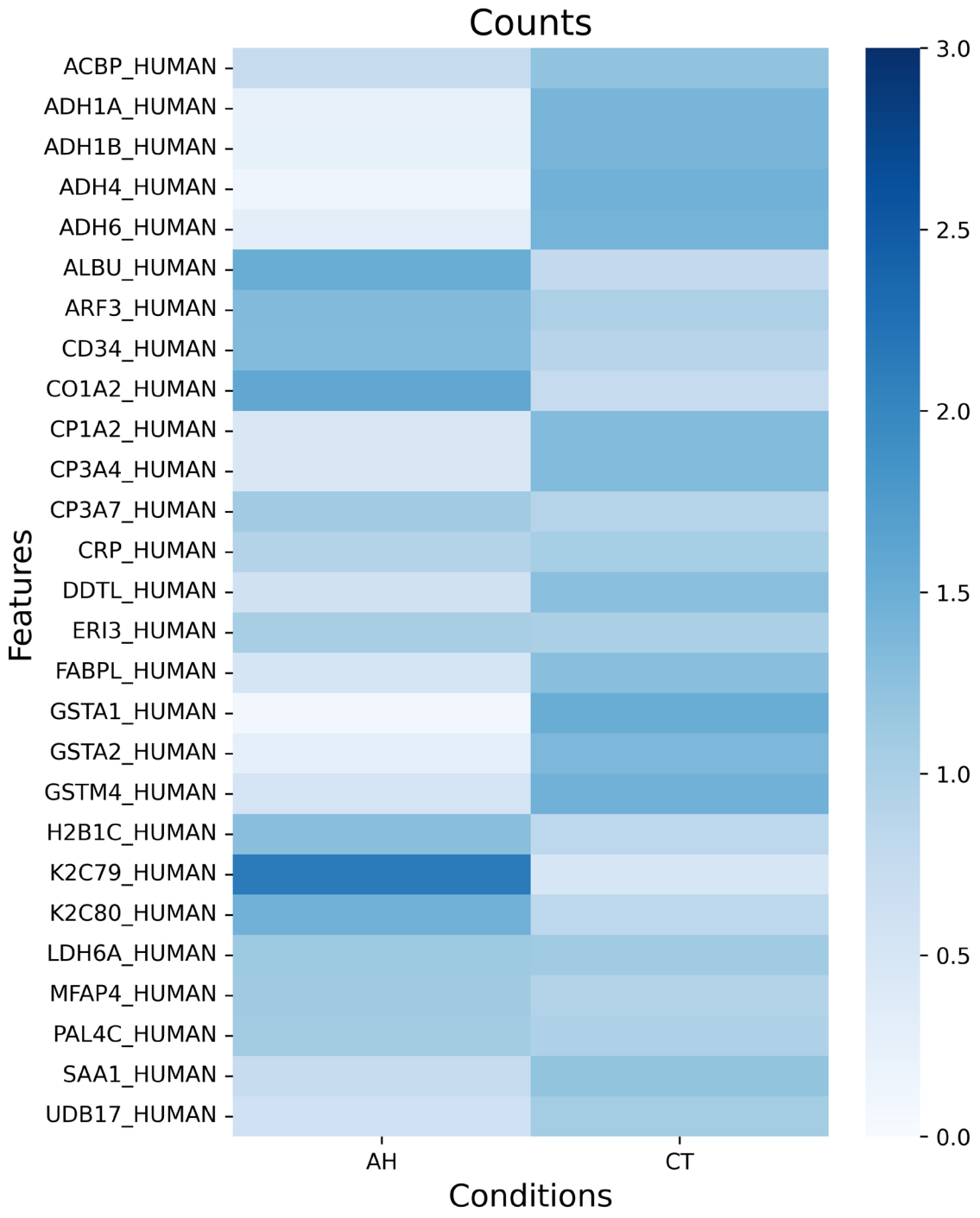

Figure S11: Heatmap of proteomic counts for independent liver validation dataset averaged per condition.

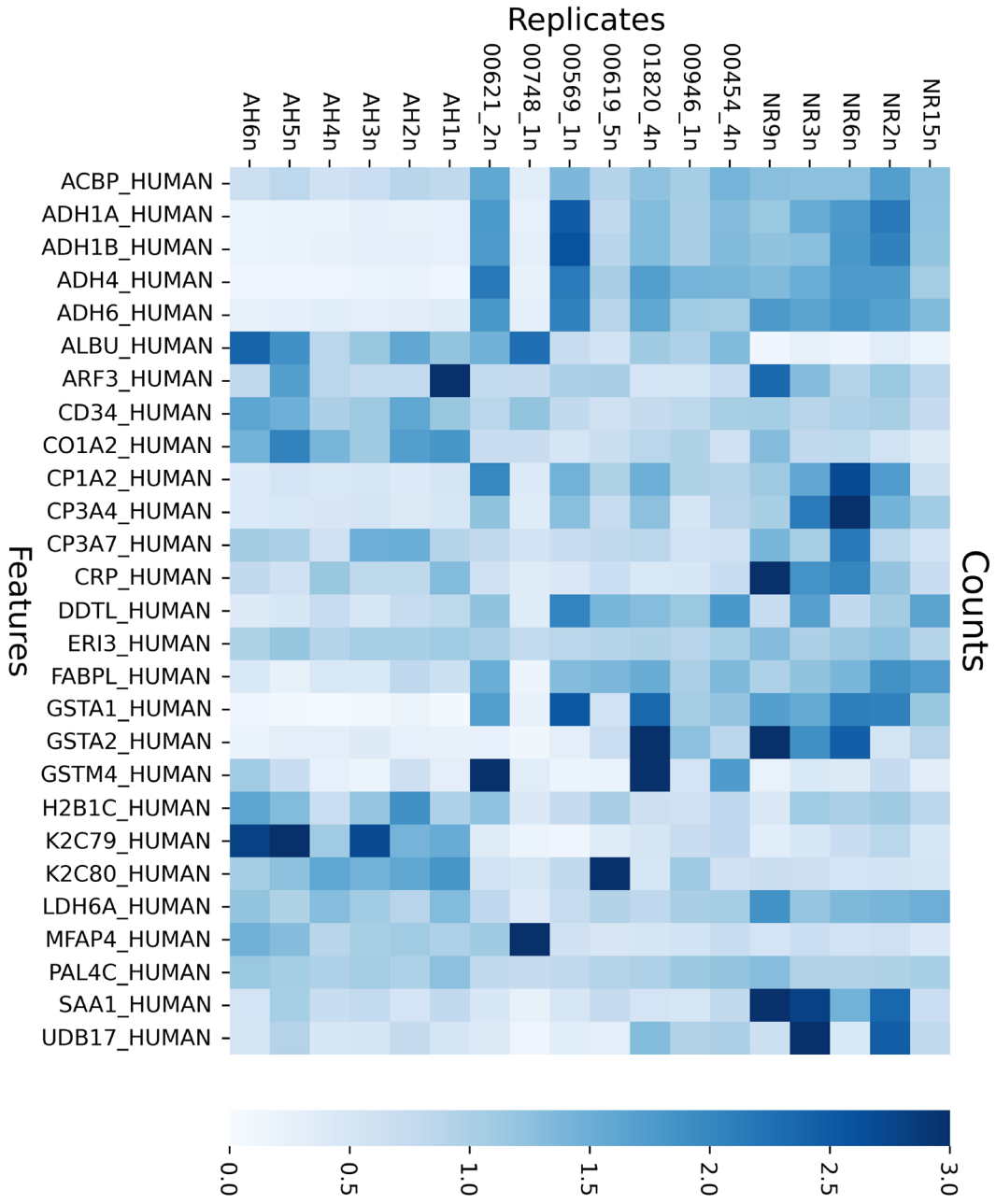

Figure S12: Heatmap of proteomic counts for independent liver validation dataset.

Enrichr:

Table S3: Top Enrichr hits for Liver 3-Way Full dataset.

Pathway		
Term	Adjusted P-Value	Genes
Oncostatin M	1.81e-02	CXCL6;AKR1B10;LCN2;HAMP;S100A8
IL-17 signaling pathway	2.20e-02	CXCL6;LCN2;S100A8

Endogenous Toll-like receptor signaling	2.59e-02	VCAN;S100A8
Tissue		
HEPATOCYTE	2.53e-07	FCN3;PLA2G2A;SCTR;FITM1;KRT23;TREM2;IGSF9;FAM198A;DBNDD1;CYP2A7;AKR1B10;CYP2B6;PPP1R1A;CREB3L3;LCN2;GPC3;MT1G;HAO2
LIVER (BULK TISSUE)	4.02e-05	FCN3;PLA2G2A;SCTR;FITM1;IGSF9;FAM198A;CYP2A7;AKR1B10;CYP2B6;CREB3L3;GPC3;MT1G;HAMP;HAO2;CFTR
OMENTUM	1.06e-04	CXCL6;FCN3;MMP7;PLA2G2A;TREM2;IGSF9;FAM198A;PPP1R1A;GPNMB;RGS1;GPC3;MT1G;EPS8L1;S100A8
Disease		
Alcoholic Hepatitis human	8.50e-09	CXCL6;VCAN;MMP7;AKR1B10;GPNMB;PLA2G2A;EEF1A2;LCN2;KRT23;TREM2
hepatocellular carcinoma human	4.53e-05	CYP2A7;CXCL6;FCN3;MMP7;PPP1R1A;MT1G;HAMP;S100A8
Carcinoma, Hepatocellular human	9.65e-05	FCN3;CYP2B6;PPP1R1A;MT1G;HAMP;HAO2;S100A8

AGOTOOL:

Table S4: Top AGOTOOL hits for Liver 3-Way Full dataset.

Pathway		
Term	Adjusted P-Value	Proteins
Drug metabolism - cytochrome P450	1.29e-05	ADH1A_HUMAN;ADH1B_HUMAN;ADH4_HUMAN;ADH6_HUMAN;CP1A2_HUMAN;CP3A4_HUMAN;GSTA1_HUMAN;GSTA2_HUMAN;GSTM4_HUMAN;UDB17_HUMAN
Drug metabolism - other enzymes	1.29e-05	CP3A4_HUMAN;GSTA1_HUMAN;GSTA2_HUMAN;GSTM4_HUMAN;UDB17_HUMAN
Steroid hormone biosynthesis	5.98e-05	CP1A2_HUMAN;CP3A4_HUMAN;CP3A7_HUMAN;UDB17_HUMAN
Tissue		
Liver	1.08e-03	ACBP_HUMAN;ADH1A_HUMAN;ADH1B_HUMAN;ADH4_HUMAN;ADH6_HUMAN;ALBU_HUMAN;CO1A2_HUMAN;CP1A2_HUMAN;CP3A4_HUMAN;CRP_HUMAN;FABPL_HUMAN;GSTA1_HUMAN;GSTA2_HUMAN;SAA1_HUMAN;UDB17_HUMAN
Venous blood	9.38e-03	ALBU_HUMAN;CRP_HUMAN
Hepatocyte	2.47e-02	ALBU_HUMAN;CP3A4_HUMAN
Disease		
Alcohol dependence	2.36e-02	ADH1B_HUMAN;ADH4_HUMAN
Alcohol use disorder	4.54e-02	ADH1B_HUMAN;ADH4_HUMAN

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

RNAseq:

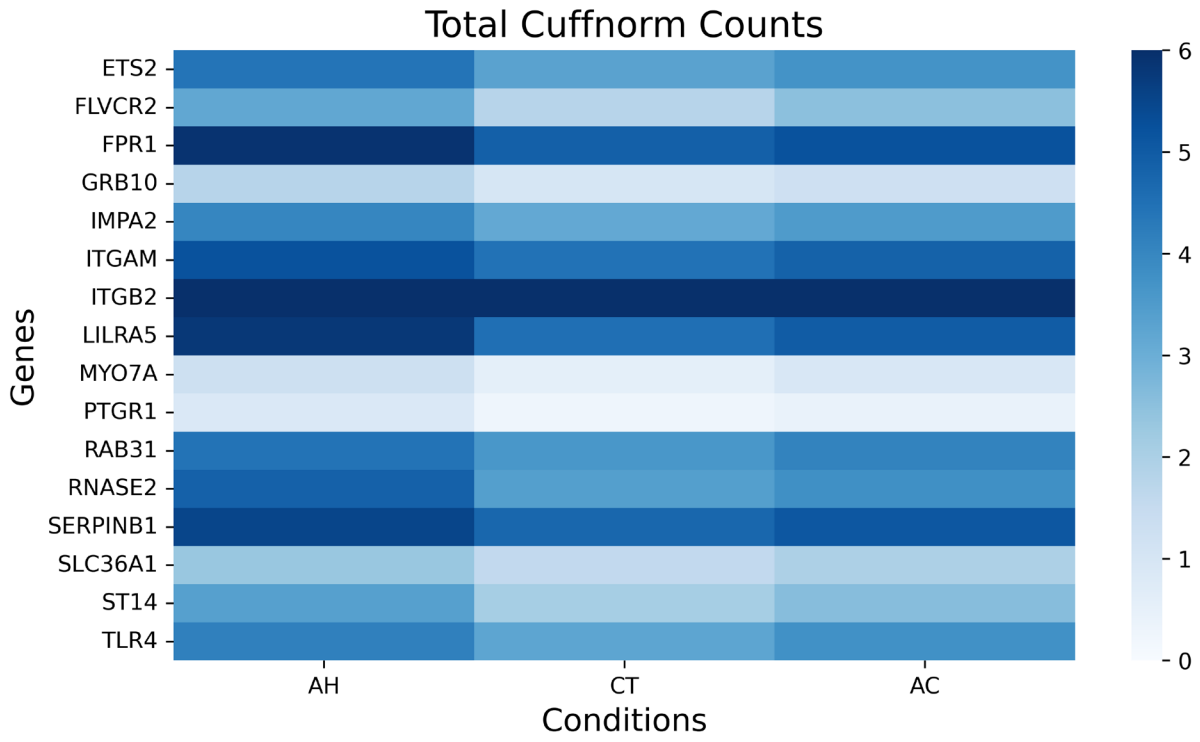

Figure S13: Heatmap of RNAseq counts for PBMC 3-Way Full dataset averaged per condition.

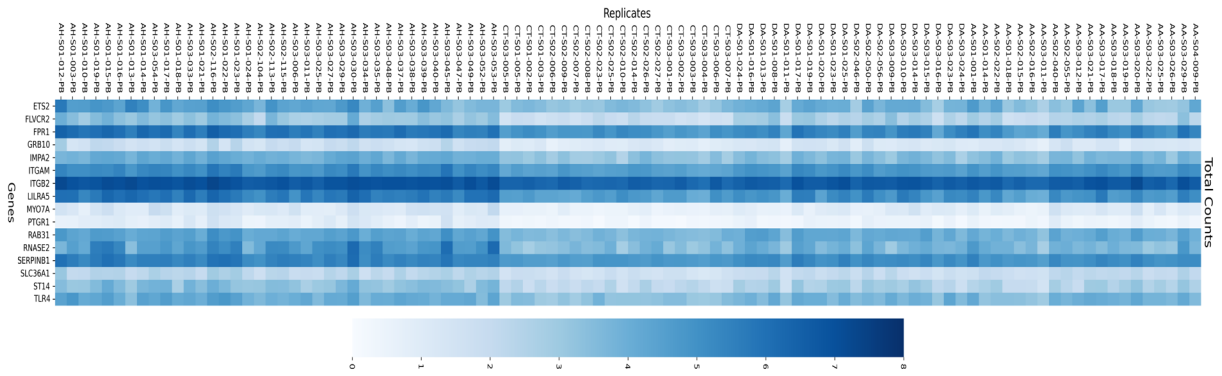

Figure S14: Heatmap of RNAseq counts for PBMC 3-Way Full dataset.

Proteomics:

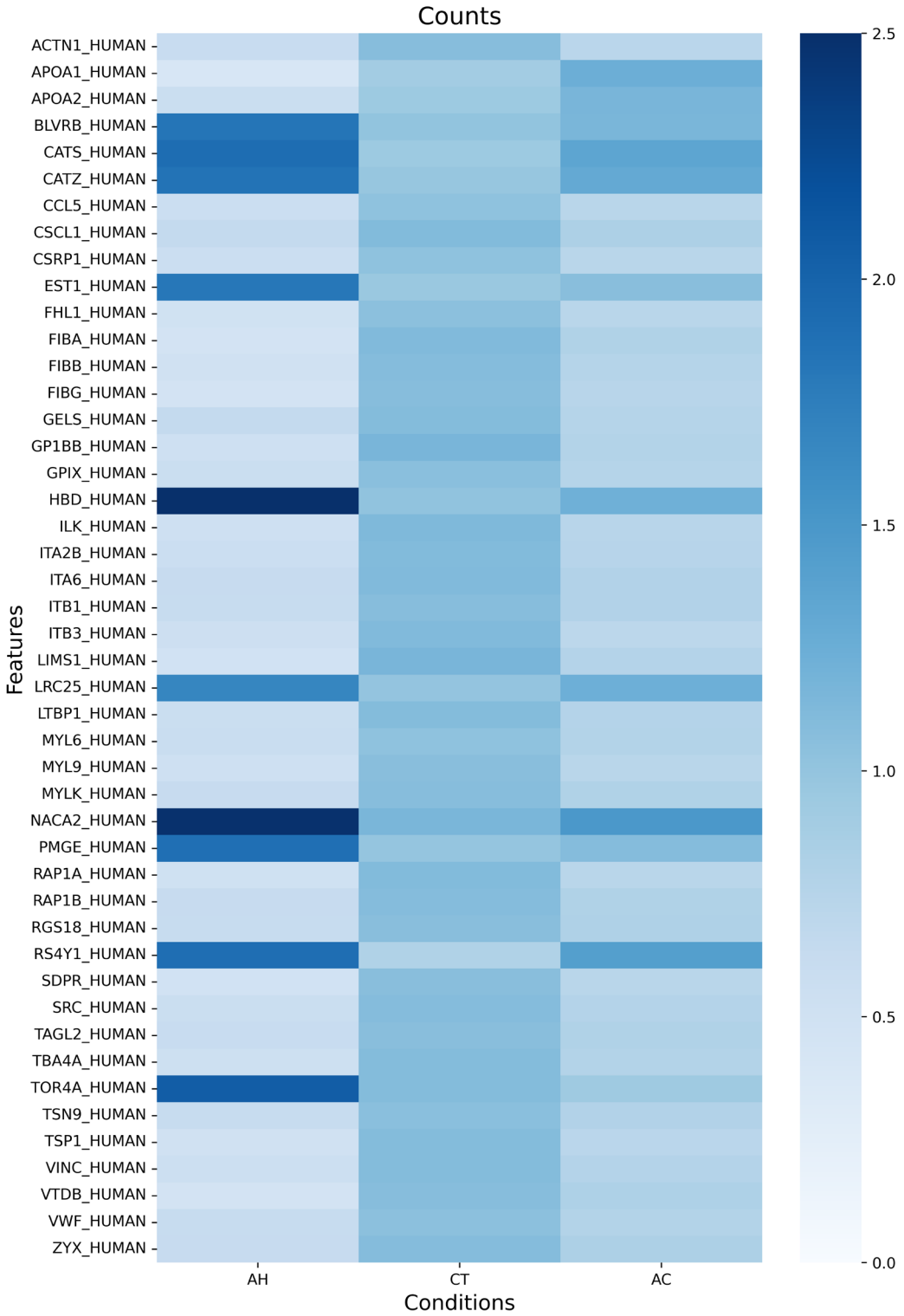

Figure S15: Heatmap of proteomic counts for PBMC 3-Way dataset averaged per condition.

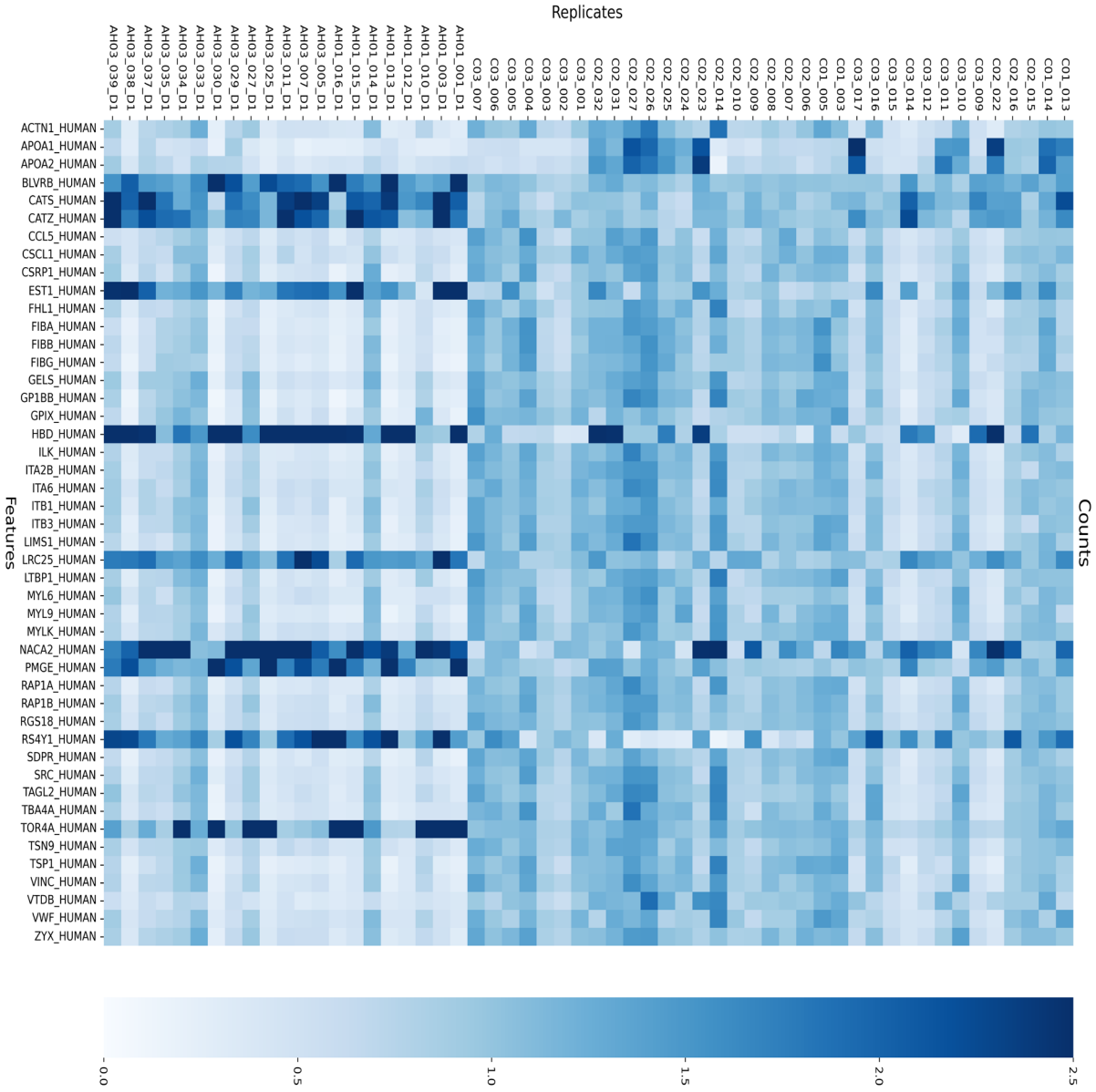

Figure S16: Heatmap of proteomic counts for PBMC 3-Way Full dataset.

Enrichr:

Table S5: Top Enrichr hits for PBMC 3-Way Full dataset.

Pathway		
Term	Adjusted P-Value	Genes
toll-like receptor 4 signaling pathway	5.12e-05	ITGAM;ITGB2;TLR4
neutrophil degranulation	8.87e-05	SERPINB1;ITGAM;RAB31;ITGB2;FPR1;RNASE2
Interleukin-2 signaling pathway	3.71e-03	ITGAM;RAB31;ITGB2;TLR4;ETS2

Tissue		
MACROPHAGE	9.46e-06	SLC36A1;ST14;ITGAM;RAB31;ITGB2;FPR1;FLVCR2;MYO7A;RNASE2;TLR4;LILRA5
PERIPHERAL BLOOD	3.78e-04	SLC36A1;ST14;ITGAM;ITGB2;FPR1;FLVCR2;RNASE2;TLR4;LILRA5
CD14+ Monocytes	1.80e-03	SERPINB1;RAB31;FPR1;LILRA5
Disease		
Septic Shock human	4.14e-07	SERPINB1;RAB31;FPR1;GRB10;RNASE2;TLR4;ETS2;LILRA5
familial combined hyperlipidemia human	1.26e-05	IMPA2;ITGB2;FPR1;RNASE2;TLR4;ETS2
familial hypercholesterolemia human	2.45e-03	ITGAM;ITGB2;FPR1;RNASE2

AGOTOOL:

Table S6: Top AGOTOOL hits for PBMC 3-Way Full dataset.

Pathway		
Term	Adjusted P-Value	Proteins
Platelet activation	4.39e-05	FIBA_HUMAN;FIBB_HUMAN;FIBG_HUMAN;GP1BB_HUMAN;GPIX_HUMAN;ITA2B_HUMAN;ITB1_HUMAN;ITB3_HUMAN;MYLK_HUMAN;RAP1A_HUMAN;RAP1B_HUMAN;SRC_HUMAN;VWF_HUMAN
Complement system	1.77e-04	APOA1_HUMAN;FIBA_HUMAN;FIBB_HUMAN;FIBG_HUMAN;ITA2B_HUMAN;ITB3_HUMAN;TSP1_HUMAN
Blood clotting cascade	1.77e-04	FIBA_HUMAN;FIBB_HUMAN;FIBG_HUMAN;VWF_HUMAN
Tissue		
Blood	2.88e-04	ACTN1_HUMAN;APOA1_HUMAN;BLVRB_HUMAN;CATS_HUMAN;CATZ_HUMAN;CCL5_HUMAN;EST1_HUMAN;FHL1_HUMAN;FIBA_HUMAN;FIBB_HUMAN;FIBG_HUMAN;GELS_HUMAN;GP1BB_HUMAN;GPIX_HUMAN;HBD_HUMAN;ITA2B_HUMAN;ITA6_HUMAN;ITB3_HUMAN;LIMS1_HUMAN;LRC25_HUMAN;LTBP1_HUMAN;MYL6_HUMAN;MYLK_HUMAN;RAP1A_HUMAN;RAP1B_HUMAN;RGS18_HUMAN;SRC_HUMAN;TAGL2_HUMAN;TBA4A_HUMAN;TSP1_HUMAN;VINC_HUMAN;VWF_HUMAN;ZYX_HUMAN
Blood platelet	2.88e-04	ACTN1_HUMAN;APOA1_HUMAN;FIBA_HUMAN;FIBB_HUMAN;FIBG_HUMAN;GELS_HUMAN;GP1BB_HUMAN;GPIX_HUMAN;ITA2B_HUMAN;ITA6_HUMAN;ITB3_HUMAN;LIMS1_HUMAN;LTBP1_HUMAN;MYLK_HUMAN;RAP1A_HUMAN;RAP1B_HUMAN;RGS18_HUMAN;SRC_HUMAN;TAGL2_HUMAN;TBA4A_HUMAN;TSP1_HUMAN;VINC_HUMAN;VWF_HUMAN;ZYX_HUMAN
Blood plasma	2.88e-04	ACTN1_HUMAN;APOA1_HUMAN;FHL1_HUMAN;FIBA_HUMAN;FIBB_HUMAN;FIBG_HUMAN;GELS_HUMAN;GP1

		BB_HUMAN;GPIX_HUMAN;ITA2B_HUMAN;ITA6_HUMAN;ITB3_HUMAN;LIMS1_HUMAN;LTBP1_HUMAN;MYLK_HUMAN;RAP1A_HUMAN;RAP1B_HUMAN;RGS18_HUMAN;SRC_HUMAN;TAGL2_HUMAN;TBA4A_HUMAN;TSP1_HUMAN;VINC_HUMAN;VWF_HUMAN;ZYX_HUMAN
Disease		
NA	NA	NA
NA	NA	NA
NA	NA	NA

Intersection Analysis of LV 3-Way Matched Balanced Integrated (AH vs Healthy vs AC)

There are 1304 DEGs and 2957 DEPs. The overlap between the two is 409 elements.

The best gene set consists of 59 DEGs, while best protein set consists of 27 DEPs.

Assume overlap signifies the 409 overlapping DEGs/DEPs.

Let us calculate the probability of there being no elements in common between 59 random DEGs and 27 random DEPS.

Framing: assume we first picked 59 DEGs, then replaced them, then picked 27 DEPs. What is the probability of there being 0 elements in common between 59 DEGs and 27 DEPS?

$$P(of\ 0\ best\ genes\ in\ overlap) * (2716/2716)^{27} + P(of\ 1\ best\ gene\ in\ overlap) * (2715/2716)^{27} + \ldots + P(of\ 59\ best\ genes\ in\ overlap) * (2657/2716)^{27}$$

$$P(of\ 0\ best\ genes\ in\ overlap) = (895\ /\ 1304)^{59} * (409\ /\ 1304)^{0} * C(59,0)$$

$$P(of\ 1\ best\ gene\ in\ overlap) = (895\ /\ 1304)^{58} * (409\ /\ 1304)^{1} * C(59,1)$$

$$P(of\ m\ best\ gene\ in\ overlap) = (895\ /\ 1304)^{(59-m)} * (409\ /\ 1304)^{m} * C(59,m)$$

What is the probability of n element(s) in common?

$$P(of\ 0\ best\ genes\ in\ overlap) * (2716/2716)^{(27-n)} * (0/2716)^{n} * C(27,n) + P(of\ 1\ best\ gene\ in\ overlap) * ((2716-1)/2716)^{(27-n)} * (1/2716)^{n} * C(27,n) + \ldots + P(of\ 59\ best\ genes\ in\ overlap) * ((2716-59)/2716)^{(27-n)} * (59/2716)^{n} * C(27,n)$$

The calculation above was written in Python.

Probability of 0 elements in common between best genes and proteins ≈ 83.2%

Probability of 1 elements in common between best genes and proteins ≈ 15.3%

Probability of 2 elements in common between best genes and proteins ≈ 1.4%

Assume probability of ≥ 3 elements in common is negligible.

Expected value ≈ 0*0.832 + 1*0.153 + 0.014*2 ≈ 0.181

What is the expected number of matches between 59 random DEGs and 27 random DEPs? The answer is 0.181. The probability of there being ≥ 1 element in common between 59 random DEGs and 27 random DEPs is 16.8%.

Intersection Analysis of PBMC 3-Way Matched Balanced Integrated (AH vs Healthy vs AC)

Calculation was done using identical approach to the one in section above, except the total number of DEGs was 971, the total number of DEPs 986, the overlap between the two is 103, DEGs picked was 16, and DEPs picked was 24.

Probability of 0 elements in common between best genes and proteins ≈ 95.9%

Probability of 1 elements in common between best genes and proteins ≈ 3.9%

Probability of 2 elements in common between best genes and proteins ≈ 0.12%

Assume probability of ≥ 3 overlap is negligible.

Expected value ≈ 0*0.959 + 1*0.039 + 0.0012*2 ≈ 0.0414

What is the expected number of matches between 16 random DEGs and 24 random DEPs? The answer is 0.0414. The probability of there being ≥ 1 element in overlap between 16 random DEGs and 24 random DEPs is 4%.

APPENDIX B: CHAPTER 4 SUPPLEMENTAL REFERENCES

1. Listopad S, Magnan C, Asghar A, Stolz A, Tayek JA, Liu Z, Morgan, T.R., and Norden-Krichmar, T.M. Differentiating between liver diseases by applying multiclass machine learning approaches to transcriptomics of liver tissue or blood based samples. JHEP Reports. 2022;4(10).

2. Massey V, Parrish A, Argemi J, Moreno M, Mello A, García-Rocha M, et al. Integrated Multiomics Reveals Glucose Use Reprogramming and Identifies a Novel Hexokinase in Alcoholic Hepatitis. Gastroenterology. 2021;160(5):1725-1740.

3. Hardesty J, Day L, Warner J, Warner D, Gritsenko M, Asghar A, Stolz A, et al. Hepatic protein and phosphoprotein signatures of alcohol-associated cirrhosis and hepatitis. The American Journal of Pathology. 2022;192(7):1066-1082.

4. Argemi J, Kedia K, Gritsenko M, Clemente-Sanchez A, Asghar A, Herranz J, et al. Integrated transcriptomic and proteomic analysis identifies plasma biomarkers of hepatocellular failure in alcohol-associated hepatitis. The American Journal of Pathology. 2022.

5. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.

6. Kampf C, Mardinoglu A, Fagerberg L, Hallstrom BM, Edlund K, Lundberg E, et al. The human liver-specific proteome defined by transcriptomics and antibody-based profiling. Faseb Journal 2014;28:2901-2914.

7. 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;7:562-578.

8. 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; 24(13):1556-8.

9. 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;14.

10. 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;12(11):1003-4.