Machine learning-based identification of potentially novel non-alcoholic fatty liver disease biomarkers

Shafiha, Roshan; Bahcivanci , Basak; Gkoutos, Georgios; Acharjee, Animesh

DOI:
10.3390/biomedicines9111636

biomedicines

MDPI

Article

Machine Learning-Based Identification of Potentially Novel Non-Alcoholic Fatty Liver Disease Biomarkers

Roshan Shafiha [1], Basak Bahcivanci [1], Georgios V. Gkoutos [1,2,3,4,5,6] and Animesh Acharjee [1,2,3,*]

[1] Centre for Computational Biology, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK; rxx063@student.bham.ac.uk (R.S.); basakbahcivanci@gmail.com (B.B.); g.gkoutos@bham.ac.uk (G.V.G.)
[2] Institute of Translational Medicine, University of Birmingham, Birmingham B15 2TT, UK
[3] NIHR Surgical Reconstruction and Microbiology Research Centre, University Hospital Birmingham, Birmingham B15 2WB, UK
[4] MRC Health Data Research UK (HDR UK), Midlands Site, Birmingham B15 2TT, UK
[5] NIHR Experimental Cancer Medicine Centre, Birmingham B15 2TT, UK
[6] NIHR Biomedical Research Centre, University Hospital Birmingham, Birmingham B15 2TT, UK
* Correspondence: a.acharjee@bham.ac.uk; Tel.: +44-(0)121-414-7012

Abstract: Non-alcoholic fatty liver disease (NAFLD) is a chronic liver disease that presents a great challenge for treatment and prevention.. This study aims to implement a machine learning approach that employs such datasets to identify potential biomarker targets. We developed a pipeline to identify potential biomarkers for NAFLD that includes five major processes, namely, a pre-processing step, a feature selection and a generation of a random forest model and, finally, a downstream feature analysis and a provision of a potential biological interpretation. The pre-processing step includes data normalising and variable extraction accompanied by appropriate annotations. A feature selection based on a differential gene expression analysis is then conducted to identify significant features and then employ them to generate a random forest model whose performance is assessed based on a receiver operating characteristic curve. Next, the features are subjected to a downstream analysis, such as univariate analysis, a pathway enrichment analysis, a network analysis and a generation of correlation plots, boxplots and heatmaps. Once the results are obtained, the biological interpretation and the literature validation is conducted over the identified features and results. We applied this pipeline to transcriptomics and lipidomic datasets and concluded that the C4BPA gene could play a role in the development of NAFLD. The activation of the complement pathway, due to the downregulation of the C4BPA gene, leads to an increase in triglyceride content, which might further render the lipid metabolism. This approach identified the C4BPA gene, an inhibitor of the complement pathway, as a potential biomarker for the development of NAFLD.

Keywords: NAFLD; biomarker; machine learning; transcriptomics; lipidomics

Citation: Shafiha, R.; Bahcivanci, B.; Gkoutos, G.V.; Acharjee, A. Machine Learning-Based Identification of Potentially Novel Non-Alcoholic Fatty Liver Disease Biomarkers. *Biomedicines* **2021**, *9*, 1636. https://doi.org/10.3390/biomedicines9111636

Academic Editors: François R. Jornayvaz and Karim Gariani

Received: 12 October 2021
Accepted: 4 November 2021
Published: 7 November 2021

1. Introduction

Non-alcoholic fatty liver disease (NAFLD) is a form of chronic liver disease that affects 20–30% of the western population and approximately 25% of the global population [1–3] NAFLD is associated with a wide range of diseases, including increased visceral obesity and metabolomic abnormalities, such as insulin resistance, diabetes, hypertension, dyslipidemia, atherosclerosis and systemic micro-inflammation [4–9]. Currently, enhanced by an inactive lifestyle and unhealthy food culture, the spread of NAFLD has increased across countries among different age groups [4,10]. The disease has increased from 15% in 2005 to 25% in 2010 with a subsequent increase in the number of obesity cases [11]. It is also anticipated that there will be an increase in the number of NAFLD cases from 83.1 million (2015) to 100.9 million (2030) [12].

NAFLD consists of a spectrum of hepatic abnormalities ranging from steatosis or non-alcoholic fatty liver to the various levels of necrotic inflammation leading to non-alcoholic steatohepatitis (NASH) [13]. The minority of NAFLD cases progresses to liver disease complications resulting to 4–8% deaths from cirrhosis complications and 1–5% deaths from hepatocellular carcinoma [11]. The initial stages of NAFLD, characterised by complex pathogenesis, include the accumulation of triglycerides in hepatocytes, which might further develop to conditions such as inflammation, fibrosis and cellular death, which are characteristics of NASH [13]. It is considered that there are multiple factors which might lead to NAFLD [13,14]. NAFLD risk factors include an unhealthy diet and an inactive lifestyle and it is expected that the interaction between the genetic characteristics, diet and gut microbiota of an individual play a pivotal role in our understanding of the development and progression of the disease.

Understanding the pathogenesis of NAFLD will cater a better understanding of the pathophysiological processes underlying the disease and will likely highlight potential therapeutic interventions [4]. While the molecular mechanism, involved in the addition of fats in the liver, is not well understood, certain cytokines obtained from inflammation sites, particularly from extrahepatic adipose tissue, have been reported to induce this process [15]. Hepatic de novo lipogenesis is also known to be a unique feature in steatosis [15]. Insulin resistance has also been reported to lead to metabolomic dysregulation in NAFLD that activates and aggravates hepatic steatosis [13]. In total, 20–30% of NAFLD patients with simple steatosis progress to NASH [13].

In this study, we analyse various types of omics datasets, such as transcriptomics and lipidomics, in an effort to gain a better understanding of the NAFLD's underlying pathophysiologic processes. Initially, we analysed transcriptomics data to identify potential gene biomarkers involved in the development of NAFLD and then proceeded to analysing lipidomics data so as to identify potential lipid biomarkers, as well as the pathways which are perturbed by these biomarkers.

2. Materials and Methods

The schematic diagram presents the pipeline developed for the biomarker [16,17] identification using NAFLD-related transcriptomics and lipidomics datasets (Figure 1) .

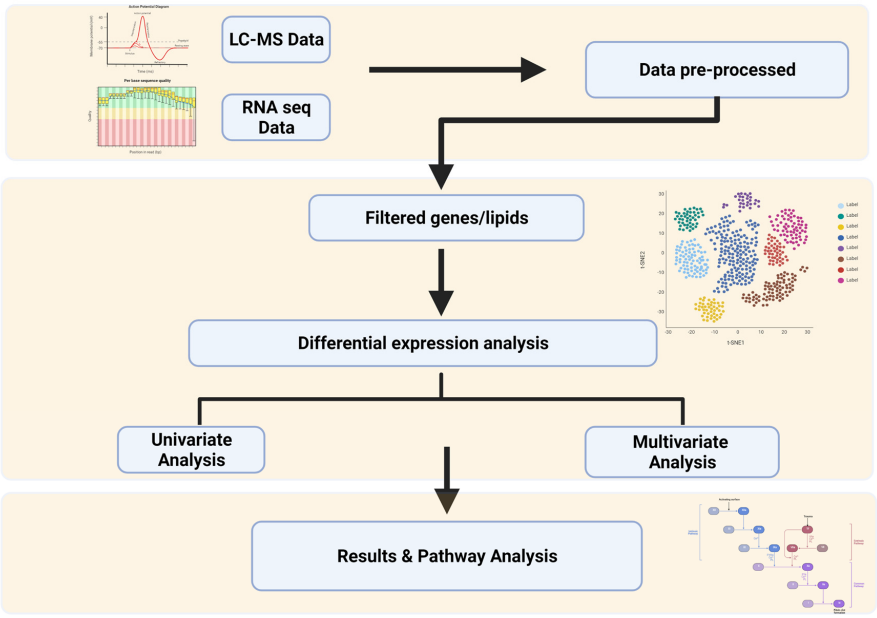


Figure 1. (Created with BioRender.com) NAFLD biomarkers identification study design consisting of pre-processing step followed by differential expression analysis for feature selection and then univariate and multivariate analysis. The final step includes the results interpretation and pathway analysis.

2.1. Transcriptomics

2.1.1. Data Acquisition

The datasets, GSE151158, GSE58979, GSE63067, GSE89632 and GSE33814, employed in this study were downloaded from the Gene Expression Omnibus (GEO) repository on 21 January 2021. In total, these datasets consisted of 146 samples, 81 of which were steatosis-related and 65 were control. The data were split into training, testing and validation sets and were subjected to pre-processing, normalization, data integration, batch-effect correction, PCA analysis, differential gene expression analysis, identification of common significant genes, as well as supervised analysis using random forest and biological interpretation.

Each GEO dataset was downloaded and loaded into R (version 4.0.3) by using the getGEO function in the *GEOquery* package (version 2.58.0) [18]. All datasets, apart from GSE151158, were already normalised. GSE151158 was normalised using the *edgeR* package (version 3.32.1) [19] *cpm* function with a True log parameter.

2.1.2. Derivation of New Transcriptomics Cohort from Multiple GEO Datasets

Due to the number of control and steatosis samples, in each GEO dataset, being low (Table 1), GSE151158, GSE58979, GSE63067 and GSE89632 were integrated to derive a transcriptomics cohort that can be used for the downstream analysis, while GSE33814 was kept for validation. Following the datasets' integration, in the derived cohorts, batch effects were identified using PCA (principle component analysis) plots. The PCs were generated using the function *prcomp* in the *stats* package and the PCA plots were visualized using *ggbiplot* (version 0.55) [20]. The batch correction for the derived cohort was performed based on non-parametric adjustment using *ComBat* [21], where batch effects due to different sequencing platforms were corrected. Following batch correction, PCA was performed to cater their visualisation.

Table 1. Details of the transcriptomics data used in the study.

S.No	GEO	Number of Samples	Number of Features	Platform	Reference
1	GSE89632	Control (n = 24) vs. Steatosis (n = 20)	29,377	Illumina HumanHT-12 WG-DASL V4.0 R2 expression beadchip	[22]
2	GSE151158	Control (n = 21) vs. Steatosis (n =23)	618	NanoString Human Immunology v2 Code Set (NS_Immunology _v2_ C2328+PLS_ Golden_1_ C5164)	[23]
3	GSE58979	Control (n = 0) vs. Steatosis (n = 17)	49,395	Affymetrix Human Gene Expression Array	[24]
4	GSE63067	Control (n = 7) vs. Steatosis (n = 2)	54,675	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	[25]
5	GSE33814	Control (n = 13) vs. Steatosis (n = 19)	48,803	Illumina HumanWG-6 v3.0 expression beadchip	[26]

2.1.3. Differential Gene Expression (DGE) Analysis

The differential gene expression (DGE) analysis was performed using the *lmFit* and *eBayes* functions, available within the *limma* package (version 3.46.0) [27]. The application of the Benjamini–Hochberg (BH) correction method yielded a gene table consisting of the log fold change (logFC) and the adjusted p value. Significant genes with an adjusted p value less than 0.05 were then extracted. This gene list was further filtered to only include genes which were common between the derived cohort and validation set.

2.1.4. Random Forest-Based Predictions

The derived cohort was split into a testing (control, 16; steatosis, 27) and a training set (control, 36; steatosis, 64). The validation set consisted of 13 control and 19 steatosis samples.

The random forest was set for repeated 10-fold cross validation with 5 repeats. The parameters of the random forest were tuned by the *expand.grid* method for the factors *mtry*, (12), *ntree* (55) and *maxnode* (6) using the training data and *train* function from the *caret* package (version 6.0.86) [28]. The model was then tested and validated using the test and validation set.

The accuracy of the model in classifying the steatosis samples for the testing and validation sets was evaluated based on the receiver operating characteristics (ROC) analysis. The area under the curve (AUC), sensitivity and specificity for both datasets were calculated by using the *pROC* package (version 1.17.0.1) [29].

2.1.5. Downstream Data Analysis

The Wilcoxon test was conducted on the training and validation datasets to investigate the genes which were upregulated and downregulated between the control and the steatosis. A further pathway and GO enrichment analysis was performed using *enrichR* (version 3.0) [30]. The correlation plot for the genes in the training dataset was plotted using the package *corrplot* (version 0.84) [31].

2.2. Lipidomics Data Analysis

2.2.1. Data Acquisition

Two cohorts (the Fenland cohort and the Italian cohort) were collected from Sanders et al., for identifying the biomarkers for NAFLD [32]. Both the cohorts consisted of clinical data and lipidomics data.

2.2.2. Data Pre-Processing

The dataset was loaded into R (version 4.0.3) and was separated into clinical data and lipid expression data. Each of the expression data row was annotated according to the lipid names. The dataset was then scaled and the values of the missing features were imputed according to the feature mean.

2.2.3. Differential Lipid Expression Analysis

A differential lipid expression analysis was performed on the Italian cohort using the *lmFit* and *eBayes* functions present in the *limma* package (version 3.46.0) [27]. The sample number difference, for each sample type, resulted in a class imbalance for the Italian cohort where there were 120 samples in steatosis0 and 21 samples in steatosis1. To address this, the steatosis0 samples were separated into 6 different batches, each containing 20 steatosis0 samples, ensuring that the samples in each of the 6 batches were unique and not repeated in other batches. The topTable function in *limma* obtained Benjamini–Hochberg (BH)-corrected *p* values and the logFC change of significant lipids between the steatosis0 and steatosis1 samples. A Volcano plot was plotted using the *ggplot2* (version 3.3.3) [33] for all the 6 different batches to visualize the differentially expressed lipids. The common significant lipids among these 6 different batches were obtained and were further subjected to a differential expression analysis using the 120 steatosis0 and 21 steatosis1 samples to obtain their logFC and *p* values.

2.2.4. Random Forest and ROC Curve Analysis

The first random forest model was formed using the Italian cohort as the training set and the Fenland data as the test set. A stratified k fold cross validation approach was implemented, where the fold value was set to 5. The parameters of the random forest were tuned by using *expand.grid* for the factors *mtry* (12), *ntree* (150) and *maxnode* (6) using the

training data and *train* function from *caret* package (version 6.0.86) [28]. The model was then tested and validated using the test and validation set.

Then, a second random forest model was generated using the Fenland data as the training set and the Italian cohort as the test set. A stratified k fold cross validation approach was implemented, where the fold value was set to 5. The parameters of the random forest were tuned by *expand.grid* for the factors *mtry* (2), *ntree* (55) and *maxnode* (6) using the training data and *train* function from *caret* package (version 6.0.86) [28]. The model was then tested and validated using the test and validation set.

The AUC, sensitivity and specificity for both the models was calculated by using the *pROC* package (version 1.17.0.1) [29].

2.3. Statistical Analysis

A Wilcoxon test was conducted using the *rstatix* package (version 0.7.0) [34]. Since the samples were unpaired, the wilcoxon test paired option was set to false and the confidence level was set to 0.95. The *p* values from this test, for each of the significant features, were extracted and stored in a new data frame. A boxplot was then generated to understand the significant difference between the steatosis0 and steatosis1 samples.

2.3.1. Heatmap-Based Visualization of Significant Lipid Features

Two heatmap were created using the *ComplexHeatmap* (version 2.6.2) [35] package using 7 significant lipids. Heatmap 1 consisted of the 120 steatosis0 samples, while heatmap 2 was formed using the 21 steatosis1 samples. These two heatmaps were further combined to construct the complex heatmap that represented both the steatosis0 and steatosis1 samples variation across the lipids. The row title of the heatmap was set to the lipid names and the columns of the heatmap represents the steatosis0 and steatosis1 sample IDs.

2.3.2. LIPEA-Based Lipid Pathway Enrichment Analysis

The LIPEA lipid pathway enrichment analysis [36] employs the lipid compounds IDs contained in the KEGG Database (Kyto Encyclopedia of Genes and Genomes) and identifies significantly disrupted pathways by applying a Fisher's exact test followed by an over representation analysis (ORA) for each pathway; an output table, consisting of the enriched pathways, the lipids involved in them and their *p* values, is then generated.

2.3.3. Lipid Network Analysis

A network analysis was performed to obtain potential interactions between the various lipid classes. The R *qgraph* package (version 1.6.9) [37] was used to generate the network, using the 7 significant lipids, with nodes, representing lipids, connected to weighted edges resembling the interaction between them. A Benjamini–Hochberg correction was implemented and the significant threshold was set to 0.05.

3. Results

3.1. Feature Selection for Random Forest Model

3.1.1. Gene Signature Identification

By merging GSE151158, GSE58979, GSE63067 and GSE89632, new NAFLD transcriptomics data were derived. Figure 2A depicts the PCA (principle component analysis) for the newly derived cohort before and after batch correction. To identify significant genes, a differential gene expression analysis was conducted over the derived transcriptomics datasets. In the derived transcriptomics cohort training data (GSE151158, GSE58979, GSE63067 and GSE89632), 173 genes were identified as significant (126 were upregulated and 47 were downregulated). Within the validation set (GSE33814), there were 1971 significant genes (772 were upregulated and 1199 were downregulated). Between the training and the validation sets, there were 18 common significant genes (C9, HPRT1, TLR1, B2M, BAX, GAPDH, BTK, PTPN6, SERPING1, ITGAE, IL1RAP, MSR1, TNFRSF14, IL15, CX3CR1, TOLLIP, IFIH1 and C4BPA) forming the group that was used within the training and validation sets.

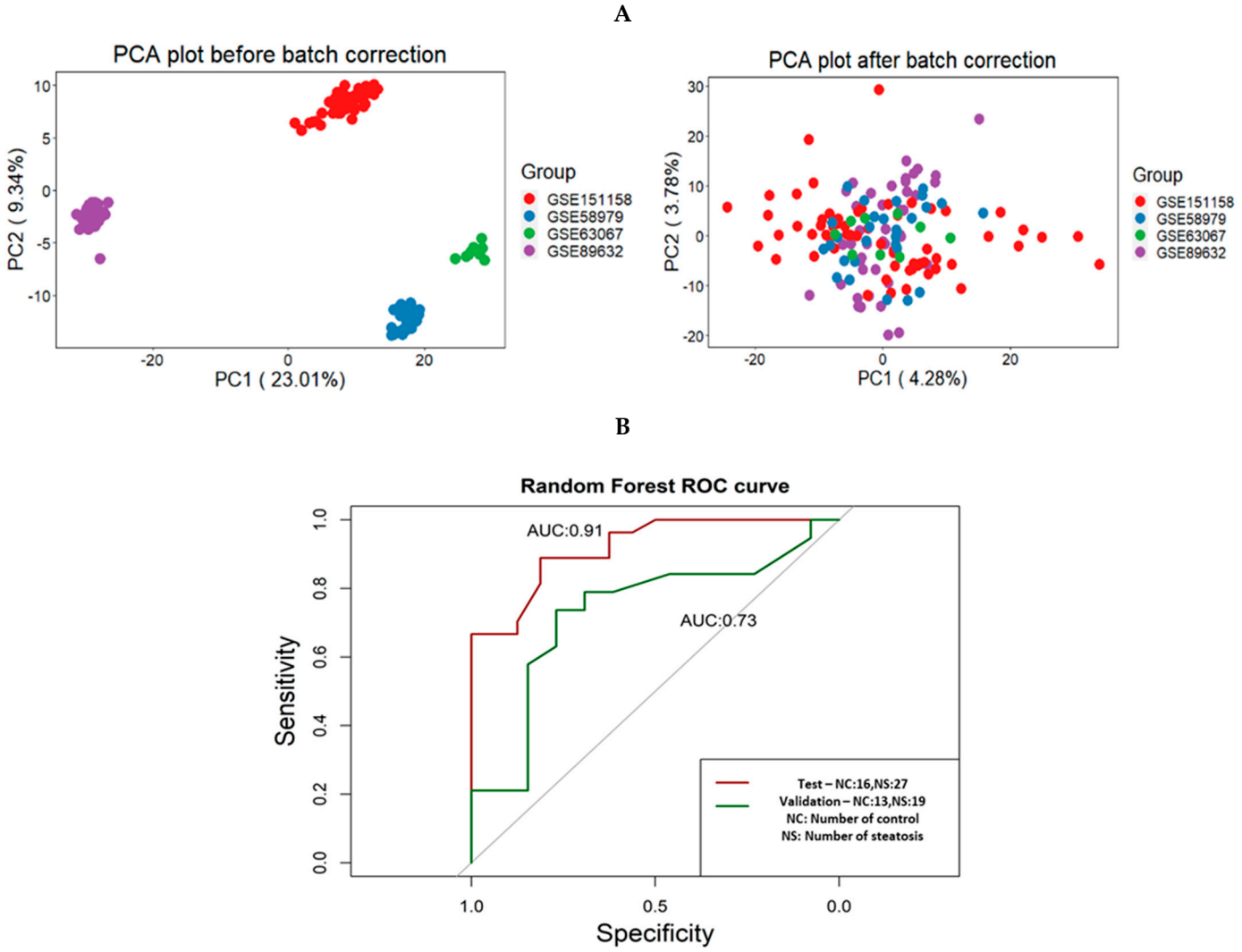


Figure 2. (**A**) A PCA plot that shows how the datasets behaved before and after the batch correction was implemented using *ComBat*. (**B**) The AUC values of the test and validation sets of the random forest model.

3.1.2. Lipid Signature Identification

There were two lipid datasets, the Italian cohort and the Fenland cohort. Within the Italian cohort, 120 steatosis0 and 21 steatosis1 samples were present, indicating a class imbalance. The steatosis0 samples were separated into six batches and a differential expression analysis was conducted six different times using *limma*. A Benjamini–Hochberg (BH) correction was implemented on the differential lipids and the lipids with adjusted *p* value lesser than 0.05 were extracted. The number of significant lipids identified for the different batches were as follows: batch 1, 191; batch 2, 189; batch 3, 171; batch 4, 129; batch 5, 42; batch 6, 57. There were 11 significant lipids which were identified in all six batches (Cholesterol,CE(16:0),DG(34:1),DG(36:2),TG(52:2),TG(52:3),TG(53:2),TG(53:3), TG(53:6),TG(53:7),TG(54:2)) and, of those 11 significant lipids, 2 lipids (Cholesterol, CE(16:0)) were upregulated and the remaining 9 lipids were downregulated (DG(34:1), DG(36:2), TG(52:2),TG(52:3),TG(53:2),TG(53:3),TG(53:6),TG(53:7) and TG(54:2)) in the steatosis0 vs. steatosis1 samples.

3.2. Random Forest Model Performance

3.2.1. Transcriptomic Features Analysis

The area under the receiver operating characteristic curve (AUC) was calculated for the random forest model to be 0.91. The prediction accuracy decreased within the

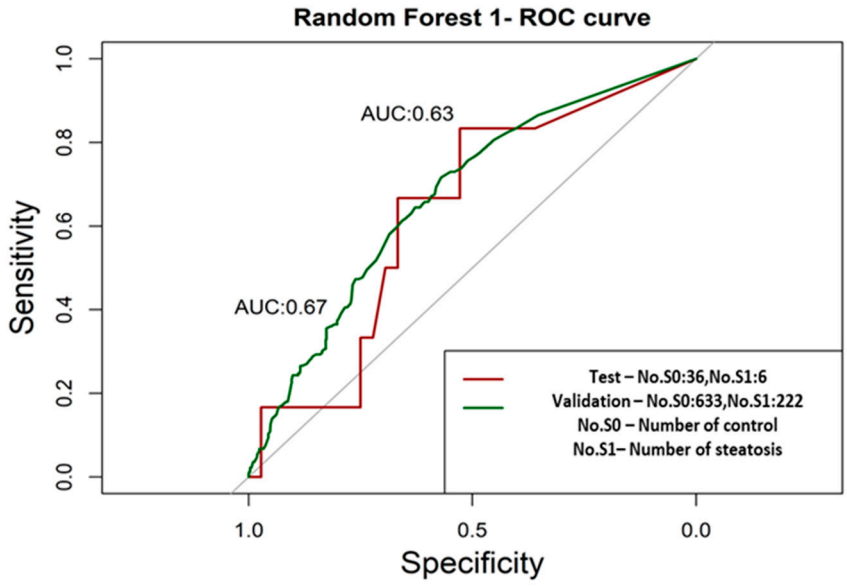
validation data, with an AUC value of 0.73 (Figure 2B). We hypothesise that this prediction accuracy reduction might be a result of the reduced sample number.

3.2.2. Lipidomic Features Analysis

There were seven (Cholesterol, CE(16:0), DG(36:2), TG(52:2), TG(52:3), TG(53:2) and TG(54:2)) lipids common to the and Italian and Fenland cohorts. These lipids were part of the first random forest model. The Italian cohort was split into training and testing sets and the Fenland cohort was kept as the validation set. There were 36 steatosis0 and 6 steatosis1 samples in the test set and 633 steatosis0 and 222 steatosis1 in the validation set.

The resulting AUC value of random forest 1 for the test dataset is 0.63. An accuracy increase was reported for the validation dataset, with an AUC value of 0.67 (Figure 3A).

A



B

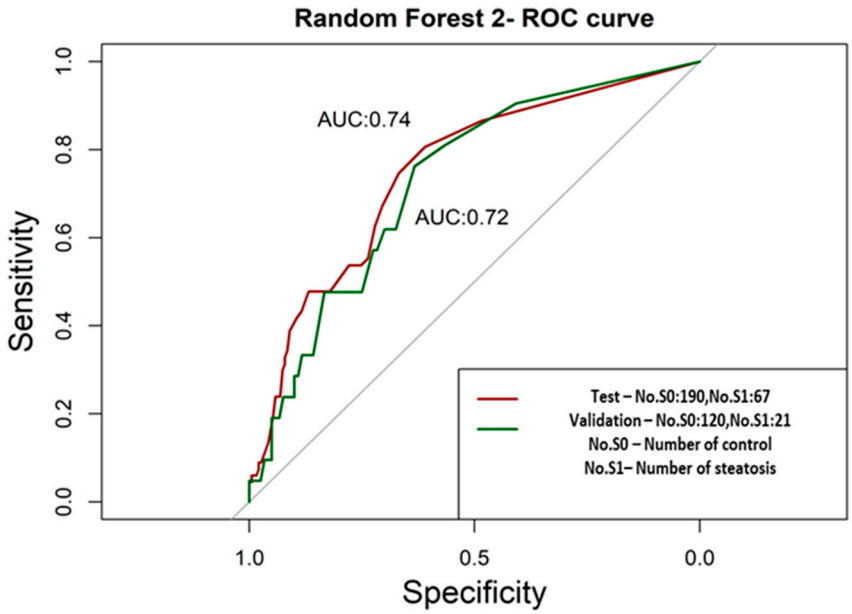


Figure 3. (**A**) Random forest 1 ROC curve that was formed using the Fenland cohort as the validation set. (**B**) Random forest 2 ROC curve that was formed using the Italian cohort as the validation set.

The process was repeated by alternating the Italian cohort as the validation set and the Fenland cohort as the training and testing set. There were 190 steatosis0 and 67 steatosis1 samples in the test set and 120 steatosis0 and 21 steatosis1 in the validation set.

Once the parameters were tuned by using the training set, the second random forest model was validated using the test and validation set. The AUC values for the test and validation datasets are 0.74 and 0.72, respectively (Figure 3B).

3.3. Downstream Analysis

3.3.1. Transcriptomic Feature Study

A pairwise correlation among 18 genes, within the training dataset, was identified within the control and the steatosis samples (Figure 4). The colour determines the sign of the coefficient, where the red colour represents a positive effect and the blue colour indicates a negative one (Figure 4A,B). The intensity of the colour increases proportionally to the magnitude of the correlation coefficient among the genes. When compared to the gene correlation matrix in the steatosis and control samples within the training data, the IFIH1 gene is positively correlated to BTK in the steatosis samples.

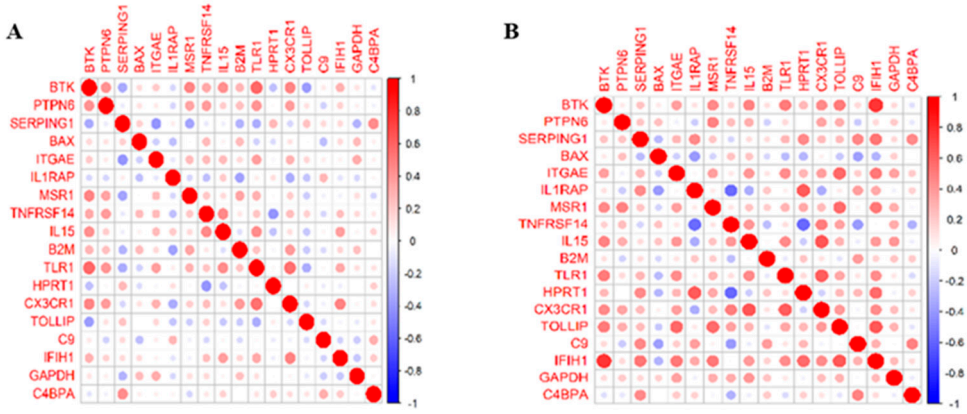


Figure 4. (**A**) A correlation gene matrixof control samples within the training dataset. (**B**) A correlation matrix of steatosis samples within the training dataset.

A pathway enrichment analysis of the transcriptomics data identified the following pathways: complement and coagulation cascades (C9, SERPING1 and C4BPA), cytokine-cytokine receptor interaction (CX3CR1, IL15, TNFRSF14 and IL1RAP), herpes simplex virus 1 infection (IFIH1, BAX, TNFRSF14 and B2M), B cell receptor signaling pathway (BTK and PTPN6), Toll-like receptor signaling pathway (TLR1 and TOLLIP), JAK–STAT pathway (IL15 and PTPN6) and Human immunodeficiency virus 1 infection and primary immunodeficiency pathways (BAX and B2M) (Figure 5). Among the 18 genes within the training and the validation sets, 5 genes (HPRT1, C9, C4BPA, IL1RAP and TNFRSF14) were upregulated in both training and validation sets, whereas the other 13 genes were inconsistent between the training and the validation samples. Five genes, namely, IL1RAP, TOLLIP, HPRT1, C9 and C4BPA, have been previously identified to be in relation with NAFLD or other inflammatory responses [38,39]. A Wilcoxon test on C9 and C4BPA genes revealed gene downregulation within the steatosis samples (Figures 6 and 7).

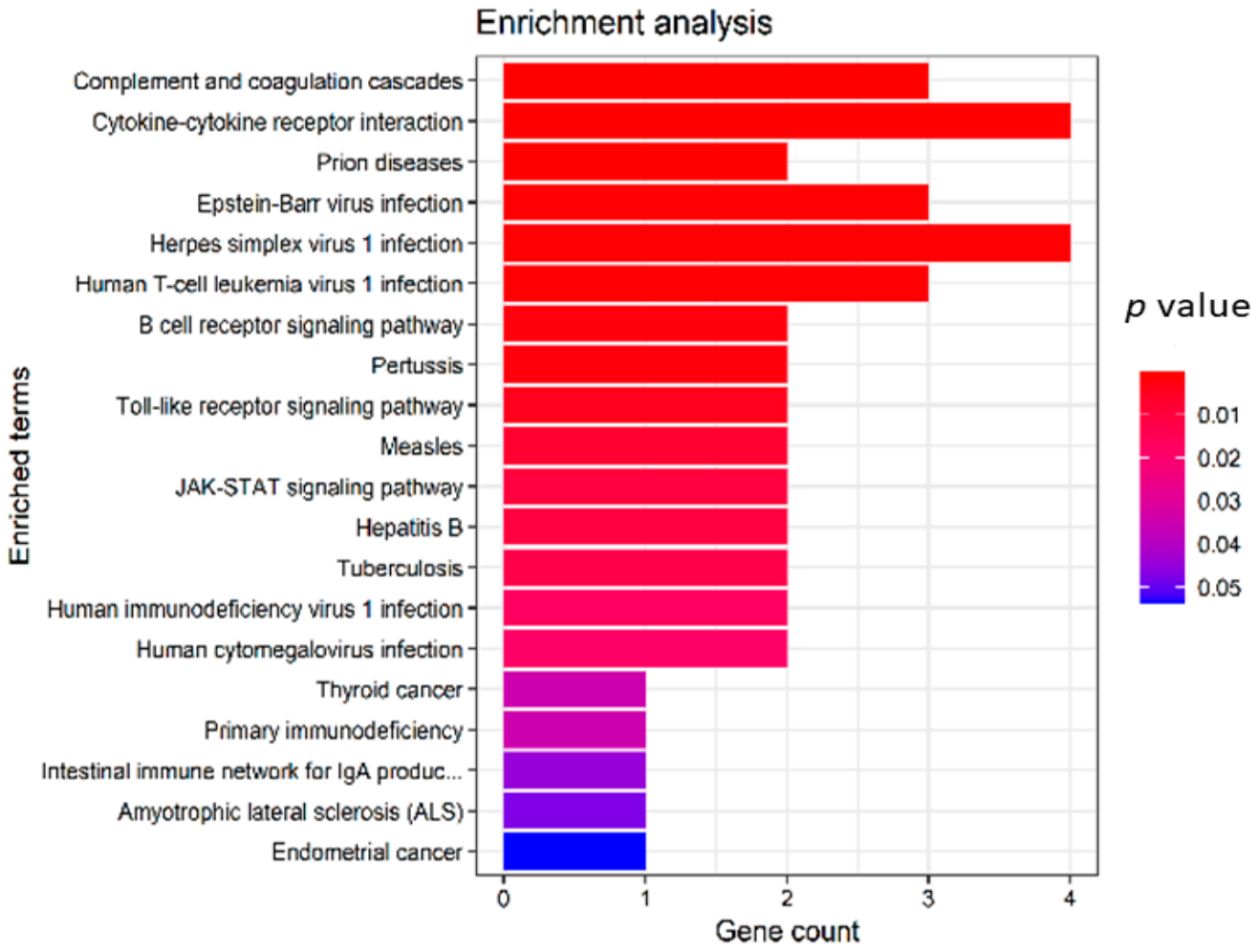


Figure 5. Enrichment analysis of gene signatures. Legend represents the *p* values associated with the pathways.

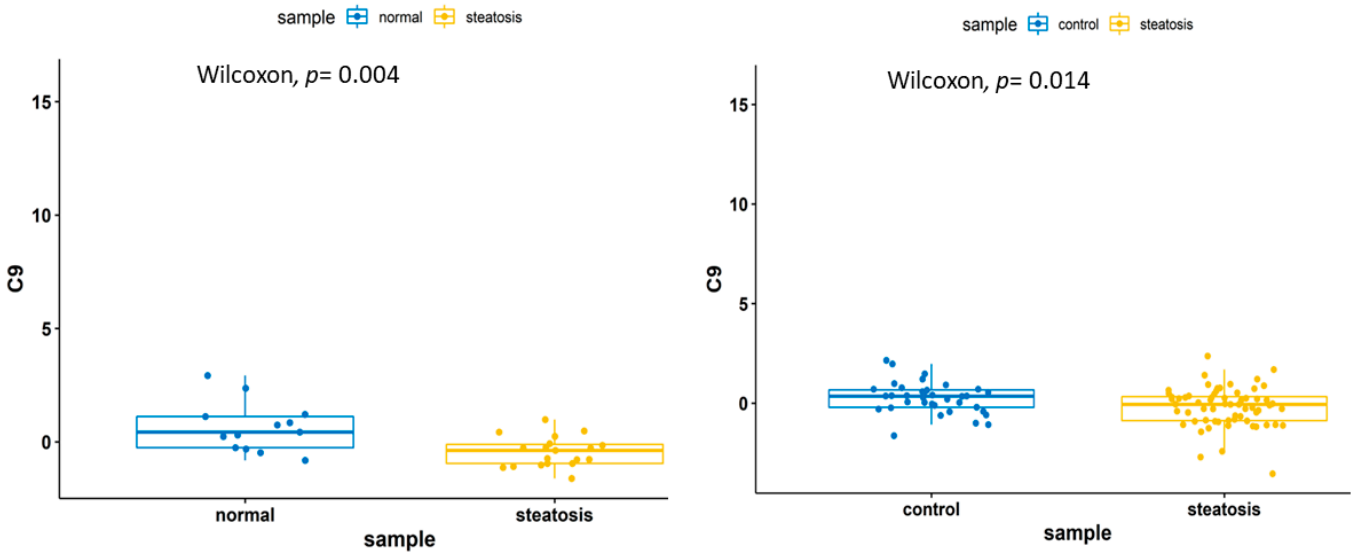


Figure 6. A boxplot denoting the downregulation of the C9 gene in the steatosis samples.

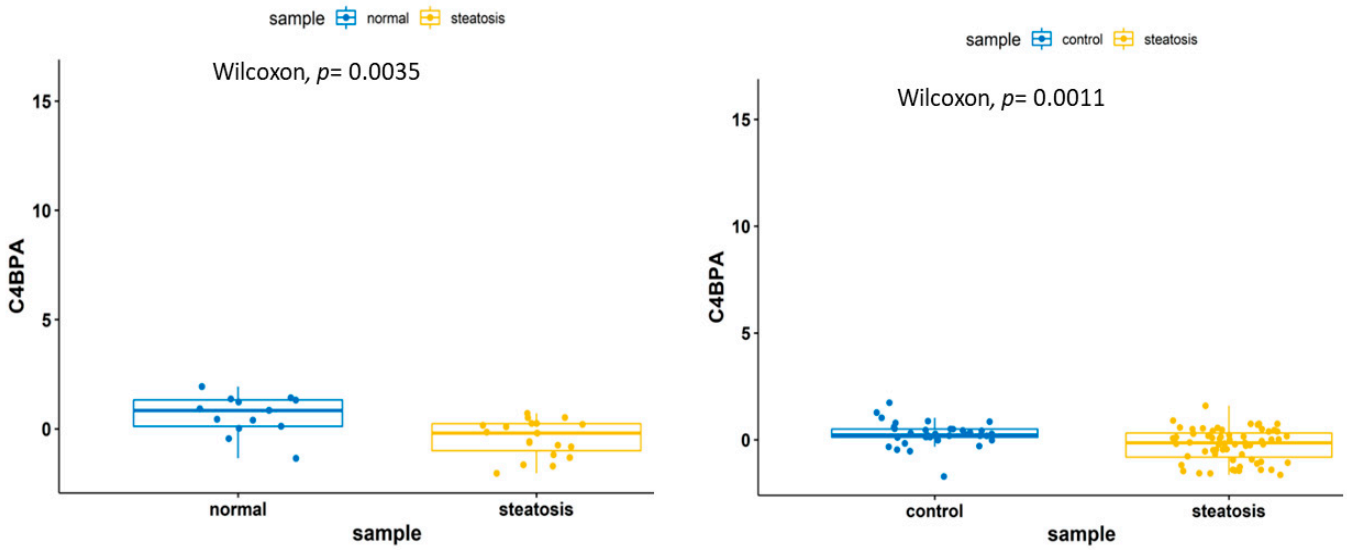


Figure 7. A boxplot showing the downregulation of the C4BPA gene in the steatosis samples.

3.3.2. Lipidomics Feature Study

The seven significant lipids, employed in the random forest model, were subjected to a Wilcoxon nonparametric statistical test and their corresponding *p* values were plotted. Among the seven lipids, TG (52.3) had the highest −log10 (*p* value). Further boxplots were generated based on these lipids, revealing that triglycerides were upregulated in the steatosis1 samples and downregulated in the steatosis0 samples (Figure 8).

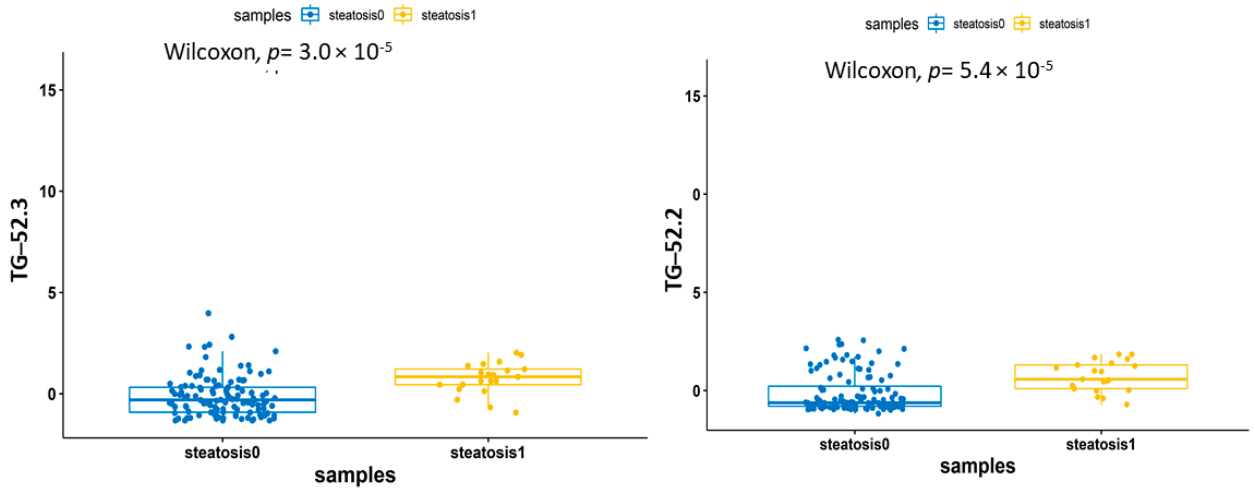


Figure 8. A boxplot denoting the upregulation of different lipid classes in steatosis1 when compared to steatosis0. The *p* value obtained from the Wilxocon test is significant.

A heatmap was constructed for the seven lipids across the steatosis0 and steatosis1 samples. The heatmap colours represent the coefficient signs and, more specifically, the red colour represents a positive effect and the blue colour indicates a negative effect. The majority of the steatosis0 samples exhibit a negative correlation effect on the triglycerides, whereas the steatosis1 samples had a positive correlation, supporting the notion that triglyceride upregulation could indicate NAFLD development (Figure 9). The lipid network shows that most of the triglycerides were positively correlated with one another (Figure 10).

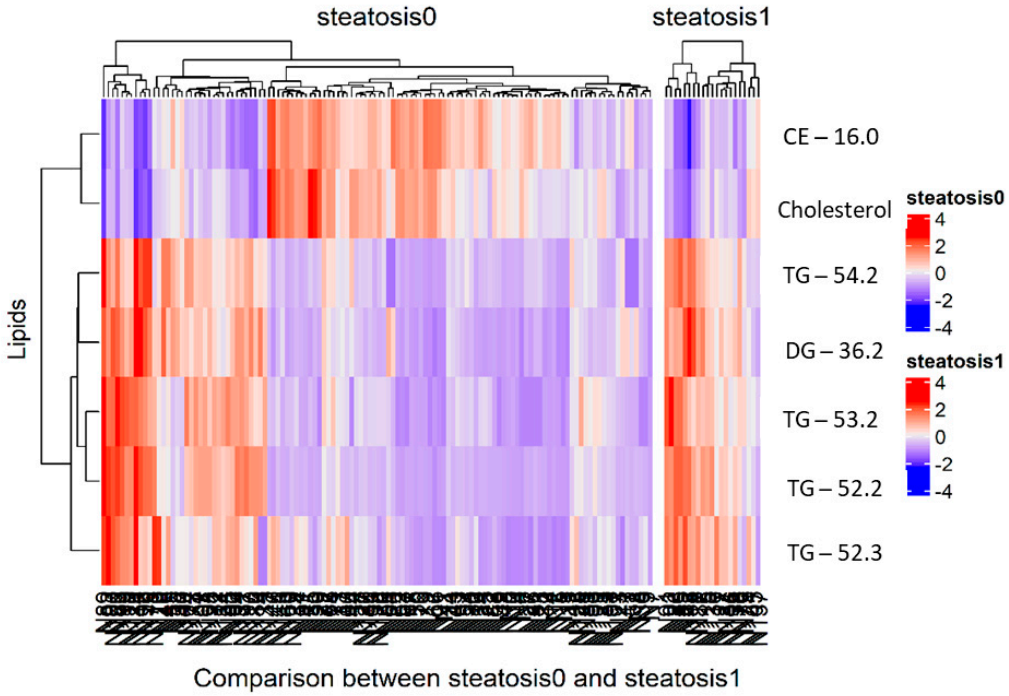


Figure 9. A complex heatmap consisting of steatosis0 and steatosis1 samples. Colours depict whether the lipids are positively or negatively correlated to the sample.

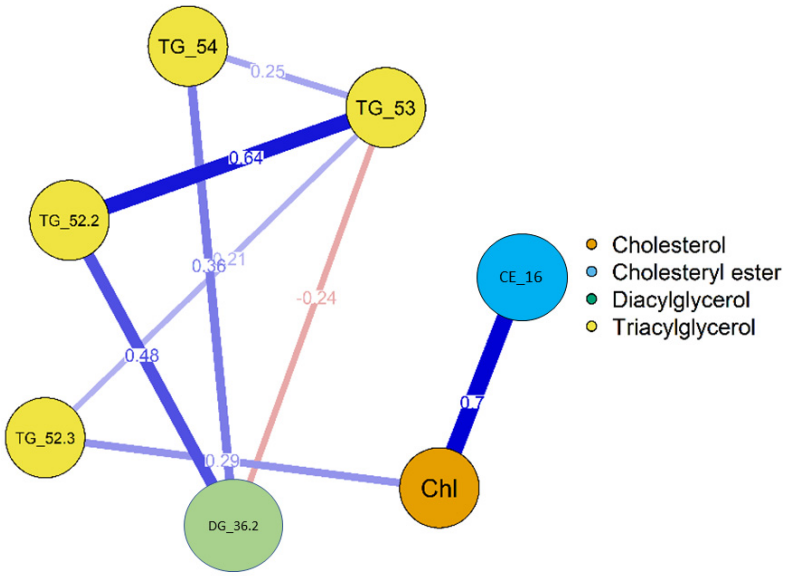


Figure 10. A lipid network showing the interactions among various groups. The blue colour indicates a positive correlation, while the red colour indicates a negative correlation.

A pathway enrichment analysis of the lipids revealed long-term depression, lipolysis in adipocytes, glycerolipid metabolism and insulin resistance as the most significant pathways in which those lipids are involved in (Table 2) (38).

Table 2. Pathway enrichment analysis of the significant lipids and their *p* values.

Pathway Name	Pathway Lipids	Converted Lipids (Number)	Converted Lipids (Percentage)	Converted Lipids (List)	*p*-Value
Long-term depression	3	2	50.00	C00165, C00641	0.0147783
Regulation of lipolysis in adipocytes	6	2	50.00	C00165, C00422	0.0147783
Glycerolipid metabolism	15	2	50.00	C00422, C00641	0.0421456
Insulin resistance	4	2	50.00	C00165, C00422	0.0421456
Fat digestion and absorption	8	2	50.00	C00165, C00422	0.0800387
Rap1 signaling pathway	1	1	25.00	C00165	0.137931
Chemokines signaling pathway	2	1	25.00	C00165	0.137931
Ras signaling pathway	2	1	25.00	C00165	0.137931
MAPK signaling pathway	1	1	25.00	C00165	0.137931
NF-kappa B signaling pathway	1	1	25.00	C00165	0.137931

Abbreviations: NAFLD, non-alcoholic fatty liver disease; ROC, receiver operating characteristics; PCA, principal component analysis; LIPEA, lipid pathway enrichment analysis; NASH, non-alcoholic steatohepatitis; GEO, Gene Expression Omnibus.

4. Discussion

The NAFLD pathogenesis is complex and unhealthy lifestyle trends have substantially increased its health burden over the past few years [11,13]. Omics integrative analytics have been proposed as a promising approach to gain a better understanding of NAFLD's biological underpinnings [40]. In this study, we analysed publicly available transcriptomic datasets to identify potential novel gene biomarkers, as well as the pathways which are perturbed.

The biomarkers identified by our transcriptomics analysis are primarily involved in immune-related pathways. Previous research studies have shown that NAFLD is related to an excessive activation of the immune system [41]. C4BPA, one of the genes identified by our transcriptomics analysis of the NAFLD samples, is primarily involved in immune-related pathways and has been identified as a target by several disease studies [42–44]. Research work has been conducted to study the defense function of C4BP against Influenza A Virus (IAV), an upper respiratory tract infection caused by the Influenza virus under the Orthomyxoviridae family which is known to cause the pandemic [45]. The complement system is safeguarded by various regulatory proteins, C4BP being one such humoral regulator example, to avoid unnecessary inflammation events [46]. Furthermore, C4BP-IgM has also been suggested as a target for the treatment of gonorrhea [47]. C9 and C4BPA have been further identified as key genes involved in the NAFLD development [38]. Moreover, IL1RAP and TOLLIP, involved in cytokine–cytokine interaction and Toll-like receptor signaling pathways, have been reported to play a key role in liver inflammatory diseases [39,48,49].

The complement system pathway is poorly characterised in NAFLD and NASH [50]. It is indicated that the complement system can be activated by three different pathways: the classical, the alternative and the lectin pathways [50]. The gene biomarker identified in our study is present in the complement and coagulation cascade pathway. C4BPA, one of the genes that was identified in this pathway, is also known as C4BP. It is indicated

that C4BP is the main soluble inhibitor of the classical and the lectin pathways [51–53]. If the classical and lectin pathways are activated, they lead to the formation of c3 and c5 convertase and c3 and c5 conversions are central reaction in the complement activation.

The activation of the classical and lectin pathways leads to the apoptosis of hepatocytes which, in turn, renders the lipid metabolism; the improvement in controlling the apoptosis process might help in controlling NASH [54].

It is indicated that there is an increased level of c3 in obese individuals and the action of c3 convertase produces c3a and c3b. c3a has a short half life but is later converted into desArg c3a which has a longer half life [55]. desArg c3a, also known as ASP (acylation-stimulating protein), is involved in the increase in triglycerides in the plasma by causing ASP resistance. Metabolic resistance has also been indicated to be shared between insulin and ASP, where the increase in insulin levels might be caused by obesity [56].

Our differential lipid expression analysis identified 11 significant lipids. The most common class of those lipids was triglycerides. We observed a triglyceride upregulation between the steatosis0 and steatosis1, which is in agreement with recent reports of hypertriglyceridemia being common in NAFLD patients [57–59]. Due to the reason of over-nutrition or insulin resistance, triglyceride concentration within the liver becomes rendered and that might create an increase in the concentration of hepatic triglycerides, which leads to steatosis [58]. It is essential that the triglycerides are exported from the liver in the form of VLDL; if this process is affected, it results in steatosis [60]. Furthermore, the network construction has shown the positive correlation of triglycerides to the other lipid groups, cholesterol and diacylglycerol.

The pathways associated with the identified significant lipids are involved in adipocyte lipolysis. It has been previously reported that elevated body mass causes fat cell lipolysis [61]. This might further cause adipose tissue inflammation, which contributes to insulin resistance. Another function of insulin is to limit lipolysis by inhibiting HSL (hormone-sensitive lipase) [62].

In conclusion, our findings suggest that the downregulation of C4BP results in an activation of the lectin pathway in the complement system triggering the conversion of c3 to c3a and c3b by the action of c3 convertase, thereby increasing the triglyceride levels, as shown in our study. This indicates that C4BP could be a potential biomarker linked to the complement system pathway, that would aid in the treatment of NAFLD.

This study has several limitations. Firstly, a small number of transcriptomics samples were used to train and validate the model. Then, the dataset merging using *ComBat* might have led to the loss of information. Further studies with large sample sizes should be further conducted to validate our findings.

5. Conclusions

We identified C4BPA, which activates the complement and coagulation pathway that renders lipid metabolism, as a potential NAFLD biomarker.

Author Contributions: All authors have made a substantial, direct intellectual contribution to this study. Conceptualization, A.A.; methodology and investigation R.S., B.B. and A.A.; writing—original draft preparation, R.S.; writing—review and editing, R.S., B.B., G.V.G. and A.A.; supervision, G.V.G. and A.A.; project administration, R.S. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by MRC Health Data Research UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health. AA and GVG also acknowledge support from the NIHR Birmingham SRMRC, Nanocommons H2020-EU (731032) and MAESTRIA (Grant agreement ID 965286).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The rmarkdown scripts and the html reports are available at: https://github.com/Roshanshafiha/NAFLD-Multi-Omics-Data-Analysis, accessed on 15 March 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, H.Y.; Baik, S.J.; Lee, H.A.; Lee, B.K.; Lee, H.S.; Kim, T.H.; Yoo, K. Relative fat mass at baseline and its early change may be a predictor of incident nonalcoholic fatty liver disease. *Sci. Rep.* **2020**, *10*, 17491. [CrossRef] [PubMed]
2. Younossi, Z.M. Non-alcoholic fatty liver disease—A global public health perspective. *J. Hepatol.* **2019**, *70*, 531–544. [CrossRef] [PubMed]
3. Younossi, Z.M.; Koenig, A.B.; Abdelatif, D.; Fazel, Y.; Henry, L.; Wymer, M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* **2016**, *64*, 73–84. [CrossRef]
4. Tanaka, N.; Kimura, T.; Fujimori, N.; Nagaya, T.; Komatsu, M.; Tanaka, E. Current status, problems, and perspectives of non-alcoholic fatty liver disease research. *World J. Gastroenterol.* **2019**, *25*, 163–177. [CrossRef]
5. Byrne, C.D.; Targher, G. NAFLD: A multisystem disease. *J. Hepatol.* **2015**, *62* (Suppl. S1), S47–S64. [CrossRef] [PubMed]
6. Zhou, Y.; Llauradó, G.; Orešič, M.; Hyötyläinen, T.; Orho-Melander, M.; Yki-Järvinen, H. Circulating triacylglycerol signatures and insulin sensitivity in NAFLD associated with the E167K variant in TM6SF2. *J. Hepatol.* **2015**, *62*, 657–663. [CrossRef]
7. Lomonaco, R.; Ortiz-Lopez, C.; Orsak, B.; Webb, A.; Hardies, J.; Darland, C.; Finch, J.; Gastaldelli, A.; Harrison, S.; Tio, F.; et al. Effect of adipose tissue insulin resistance on metabolic parameters and liver histology in obese patients with nonalcoholic fatty liver disease. *Hepatology* **2012**, *55*, 1389–1397. [CrossRef] [PubMed]
8. Pagano, G.; Pacini, G.; Musso, G.; Gambino, R.; Mecca, F.; Depetris, N.; Cassader, M.; David, E.; Cavallo-Perin, P.; Rizzetto, M. Nonalcoholic steatohepatitis, insulin resistance, and metabolic syndrome: Further evidence for an etiologic association. *Hepatology* **2002**, *35*, 367–372. [CrossRef] [PubMed]
9. Sanyal, A.J.; Campbell-Sargent, C.; Mirshahi, F.; Rizzo, W.B.; Contos, M.J.; Sterling, R.K.; Luketic, V.A.; Shiffman, M.L.; Clore, J.N. Nonalcoholic steatohepatitis: Association of insulin resistance and mitochondrial abnormalities. *Gastroenterology* **2001**, *120*, 1183–1192. [CrossRef]
10. Mirmiran, P.; Amirhamidi, Z.; Ejtahed, H.S.; Bahadoran, Z.; Azizi, F. Relationship between diet and non-alcoholic fatty liver disease: A review article. *Iran. J. Public Health* **2017**, *46*, 1007–1017. [PubMed]
11. Maurice, J.; Manousou, P. Non-alcoholic fatty liver disease. *Clin. Med.* **2018**, *18*, 245–250. [CrossRef]
12. Estes, C.; Razavi, H.; Loomba, R.; Younossi, Z.; Sanyal, A.J. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology* **2018**, *67*, 123–133. [CrossRef]
13. Yu, J.; Marsh, S.; Hu, J.; Feng, W.; Wu, C. The pathogenesis of nonalcoholic fatty liver disease: Interplay between diet, gut microbiota, and genetic background. *Gastroenterol. Res. Pract.* **2016**, *2016*, 2862173. [CrossRef]
14. Tilg, H.; Adolph, T.E.; Moschen, A.R. Multiple parallel hits hypothesis in nonalcoholic fatty liver disease: Revisited after a decade. *Hepatology* **2021**, *73*, 833–842. [CrossRef]
15. Fabbrini, E.; Sullivan, S.; Klein, S. Obesity and nonalcoholic fatty liver disease: Biochemical, metabolic, and clinical implications. *Hepatology* **2010**, *51*, 679–689. [CrossRef] [PubMed]
16. Davis, S.; Meltzer, P.S. GEOquery: A bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics* **2007**, *23*, 1846–1847. [CrossRef]
17. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef]
18. Aziz, F.; Acharjee, A.; Williams, J.A.; Russ, D.; Bravo-Merodio, L.; Gkoutos, G.V. Biomarker Prioritisation and Power Estimation Using Ensemble Gene Regulatory Network Inference. *Int. J. Mol. Sci.* **2020**, *21*, 7886. [CrossRef] [PubMed]
19. Bravo-Merodio, L.; Acharjee, A.; Russ, D.; Bisht, V.; Williams, J.A.; Tsaprouni, L.G.; Gkoutos, G.V. Translational biomarkers in the era of precision medicine. *Adv. Clin. Chem.* **2021**, *102*, 191–232.
20. Vu, V.Q. Vqv/Ggbiplot: A Biplot Based on Ggplot2. Github. 2015. Available online: http://github.com/vqv/ggbiplot (accessed on 15 March 2021).
21. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef] [PubMed]
22. Arendt, B.M.; Comelli, E.M.; Ma, D.W.; Lou, W.; Teterina, A.; Kim, T.; Fung, S.K.; Wong, D.K.; McGilvray, I.; Fischer, S.E.; et al. Altered hepatic gene expression in nonalcoholic fatty liver disease is associated with lower hepatic n-3 and n-6 polyunsaturated fatty acids. *Hepatology* **2015**, *61*, 1565–1578. [CrossRef] [PubMed]
23. Kriss, M.; Golden-Mason, L.; Kaplan, J.; Mirshahi, F.; Setiawan, V.W.; Sanyal, A.J.; Rosen, H.R. Increased hepatic and circulating chemokine and osteopontin expression occurs early in human NAFLD development. *PLoS ONE* **2020**, *15*, e0236353. [CrossRef]
24. Du Plessis, J.; Van Pelt, J.; Korf, H.; Mathieu, C.; Van der Schueren, B.; Lannoo, M.; Oyen, T.; Topal, B.; Fetter, G.; Nayler, S.; et al. Association of Adipose Tissue Inflammation With Histologic Severity of Nonalcoholic Fatty Liver Disease. *Gastroenterology* **2015**, *149*, 635–648. [CrossRef] [PubMed]

25. Frades, I.; Andreasson, E.; Mato, J.M.; Alexandersson, E.; Matthiesen, R.; Martínez-Chantar, M.L. Integrative genomic signatures of hepatocellular carcinoma derived from nonalcoholic Fatty liver disease. *PLoS ONE* **2015**, *10*, e0124544. [CrossRef]
26. Starmann, J.; Fälth, M.; Spindelböck, W.; Lanz, K.L.; Lackner, C.; Zatloukal, K.; Trauner, M.; Sültmann, H. Gene expression profiling unravels cancer-related hepatic molecular signatures in steatohepatitis but not in steatosis. *PLoS ONE* **2012**, *7*, e46584. [CrossRef]
27. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef]
28. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
29. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77–2105. [CrossRef]
30. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.; Jagodnik, K.M.; Lachmann, A.; et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic. Acids Res.* **2016**, *44(W1)*, W90–W97. [CrossRef]
31. Wei T, S.V. R Package "Corrplot": Visualization of a Correlation Matrix. GitHub. 2017. Available online: https://github.com/taiyun/corrplot (accessed on 15 March 2021).
32. Sanders, F.W.B.; Acharjee, A.; Walker, C.; Marney, L.; Roberts, L.D.; Imamura, F.; Jenkins, B.; Case, J.; Ray, S.; Virtue, S.; et al. Hepatic steatosis risk is partly driven by increased de novo lipogenesis following carbohydrate consumption. *Genome. Biol.* **2018**, *19*, 79. [CrossRef] [PubMed]
33. Wickham, H. *Ggplot2: Elegant Graphics For Data Analysis*; Springer: New York, NY, USA, 2016; ISBN 978-3-319-24277-4. Available online: https://ggplot2.tidyverse.org (accessed on 15 March 2021).
34. Kassambara, A. Rstatix: Pipe-Friendly Framework for Basic Statistical Tests; R Package Version 0.7.0. 2021. Available online: https://CRAN.R-project.org/package=rstatix (accessed on 15 March 2021).
35. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef]
36. Acevedo, A. LIPEA: Lipid Pathway Enrichment Analysis. *bioRxiv* **2018**. [CrossRef]
37. Epskamp, S. Qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Softw.* **2012**, *48*, 1–8. [CrossRef]
38. Wang, R.; Wang, X.; Zhuang, L. Gene expression profiling reveals key genes and pathways related to the development of non-alcoholic fatty liver disease. *Ann. Hepatol.* **2016**, *15*, 190–199. [PubMed]
39. Niederreiter, L.; Tilg, H. Cytokines and fatty liver diseases. *Liver Res.* **2018**, *2*, 14–20. [CrossRef]
40. Tomizawa, M.; Kawanabe, Y.; Shinozaki, F.; Sato, S.; Motoyoshi, Y.; Sugiyama, T.; Yamamoto, S.; Sueishi, M. Triglyceride is strongly associated with nonalcoholic fatty liver disease among markers of hyperlipidemia and diabetes. *Biomed. Rep.* **2014**, *2*, 633–636. [CrossRef] [PubMed]
41. Perakakis, N.; Stefanakis, K.; Mantzoros, C.S. The role of omics in the pathophysiology, diagnosis and treatment of non-alcoholic fatty liver disease. *Metab. Clin. Exp.* **2020**, *111*, 154320. [CrossRef] [PubMed]
42. Kosmalski, M.; Mokros, Ł.; Kuna, P.; Witusik, A.; Pietras, T. Changes in the immune system—The key to diagnostics and therapy of patients with non-alcoholic fatty liver disease. *Cent. Eur. J. Immunol.* **2018**, *43*, 231–239. [CrossRef]
43. Dunkelberger, J.R.; Song, W.C. Complement and its role in innate and adaptive immune responses. *Cell Res.* **2010**, *20*, 34–50. [CrossRef] [PubMed]
44. Luque, A.; Serrano, I.; Ripoll, E.; Malta, C.; Gomà, M.; Blom, A.M.; Grinyó, J.M.; de Córdoba, S.R.; Torras, J.; Aran, J.M. Noncanonical immunomodulatory activity of complement regulator C4BP(β-) limits the development of lupus nephritis. *Kidney Int.* **2020**, *97*, 551–566. [CrossRef]
45. Martin, M.; Gottsäter, A.; Nilsson, P.M.; Mollnes, T.E.; Lindblad, B.; Blom, A.M. Complement activation and plasma levels of C4b-binding protein in critical limb ischemia patients. *J. Vasc. Surg.* **2009**, *50*, 100–106. [CrossRef] [PubMed]
46. Varghese, P.M.; Murugaiah, V.; Beirag, N.; Temperton, N.; Khan, H.A.; Alrokayan, S.H.; Al-Ahdal, M.N.; Nal, B.; Al-Mohanna, F.A.; Sim, R.B.; et al. C4b binding protein acts as an innate immune effector against influenza a virus. *Front. Immunol.* **2021**, *11*, 585361. [CrossRef]
47. Rodriguez de Cordoba, S.; Sanchez-Corral, P.; Rey-Campos, J. Structure of the gene coding for the alpha polypeptide chain of the human complement component C4b-binding protein. *J. Exp. Med.* **1991**, *173*, 1073–1082. [CrossRef] [PubMed]
48. Bettoni, S.; Shaughnessy, J.; Maziarz, K.; Ermert, D.; Gulati, S.; Zheng, B.; Mörgelin, M.; Jacobsson, S.; Riesbeck, K.; Unemo, M.; et al. C4BP-IgM protein as a therapeutic approach to treat Neisseria gonorrhoeae infections. *JCI Insight* **2019**, *4*, e131886. [CrossRef] [PubMed]
49. Chen, K.; Yuan, R.; Zhang, Y.; Geng, S.; Li, L. Tollip deficiency alters atherosclerosis and steatosis by disrupting lipophagy. *J Am. Heart Assoc.* **2017**, *6*, e004078. [CrossRef] [PubMed]
50. Mirea, A.M.; Tack, C.J.; Chavakis, T.; Joosten, L.A.B.; Toonen, E.J.M. IL-1 family cytokine pathways underlying NAFLD: Towards new treatment strategies. *Trends Mol. Med.* **2018**, *24*, 458–471. [CrossRef]
51. Phieler, J.; Garcia-Martin, R.; Lambris, J.D.; Chavakis, T. The role of the complement system in metabolic organs and metabolic diseases. *Semin. Immunol.* **2013**, *25*, 47–53. [CrossRef]
52. Okrój, M.; Blom, A.M. Chapter 24—C4b-binding protein. In *The Complement FactsBook*, 2nd ed.; Barnum, S., Schein, T., Eds.; Academic Press: Cambridge, MA, USA, 2018; pp. 251–259.

53. Moreno-Navarrete, J.M.; Fernández-Real, J.M. The complement system is dysfunctional in metabolic disease: Evidences in plasma and adipose tissue from obese and insulin resistant subjects. *Semin. Cell Dev. Biol.* **2019**, *85*, 164–172. [CrossRef]
54. Rawal, N.; Rajagopalan, R.; Salvi, V.P. Stringent regulation of complement lectin pathway C3/C5 convertase by C4b-binding protein (C4BP). *Mol. Immunol.* **2009**, *46*, 2902–2910. [CrossRef]
55. Rensen, S.S.; Slaats, Y.; Driessen, A.; Peutz-Kootstra, C.J.; Nijhuis, J.; Steffensen, R.; Greve, J.W.; Buurman, W.A. Activation of the complement system in human nonalcoholic fatty liver disease. *Hepatology* **2009**, *50*, 1809–1817. [CrossRef] [PubMed]
56. Reca, R.; Wysoczynski, M.; Yan, J.; Lambris, J.D.; Ratajczak, M.Z. The role of third complement component (C3) in homing of hematopoietic stem/progenitor cells into bone marrow. *Adv. Exp. Med. Biol.* **2006**, *586*, 35–51.
57. Saleh, J.; Wahab, R.A.; Farhan, H.; Al-Amri, I.; Cianflone, K. Plasma levels of acylation-stimulating protein are strongly predicted by waist/hip ratio and correlate with decreased LDL size in men. *ISRN Obes.* **2013**, *2013*, 342802. [CrossRef] [PubMed]
58. Kawano, Y.; Cohen, D.E. Mechanisms of hepatic triglyceride accumulation in non-alcoholic fatty liver disease. *J. Gastroenterol.* **2013**, *48*, 434–441. [CrossRef]
59. Eguchi, Y.; Hyogo, H.; Ono, M.; Mizuta, T.; Ono, N.; Fujimoto, K.; Chayama, K.; Saibara, T. Prevalence and associated metabolic factors of nonalcoholic fatty liver disease in the general population from 2009 to 2010 in Japan: A multicenter large retrospective study. *J. Gastroenterol.* **2012**, *47*, 586–595. [CrossRef] [PubMed]
60. Arvind, A.; Osganian, S.A.; Cohen, D.E.; Corey, K.E. *Lipid and Lipoprotein Metabolism in Liver Disease*; MDText.com, Inc.: Endotext South Dartmouth, MA, USA, 2000.
61. Morigny, P.; Houssier, M.; Mouisel, E.; Langin, D. Adipocyte lipolysis and insulin resistance. *Biochimie* **2016**, *125*, 259–266. [CrossRef] [PubMed]
62. Cignarelli, A.; Genchi, V.A.; Perrini, S.; Natalicchio, A.; Laviola, L.; Giorgino, F. Insulin and insulin receptors in adipose tissue development. *Int. J. Mol. Sci.* **2019**, *20*, 759. [CrossRef]