

UNIVERSITY OF TURIN



Department of Medical Sciences

PhD in Biomedical Sciences and Oncology

Omics Approaches in Non-Alcoholic Fatty Liver Disease
(NAFLD)

Tutor
Prof. Giuseppe Matullo

PhD coordinator
Prof. Emilio Hirsch

Candidate
Dr. Chiara Catalano

XXXV cycle

Academic years 2019/2023

Science-disciplinary sector: MED/03

Contents

1	Introduction	1
1.1	Pathogenesis of NAFLD	2
1.2	Progression to NASH	4
1.3	Omics Approach	4
1.4	Omics in NAFLD	5
1.5	Aim of study	7
2	Whole exome sequencing to identify genes predisposing to the progression of non-alcoholic fatty liver (NAFL)	8
2.1	Introduction	8
2.2	Methods	9
2.2.1	Cohort of study	9
2.2.2	DNA isolation and library preparation	9
2.2.3	Whole Exome Sequencing	10
2.2.4	Sequencing data analysis	10
2.2.5	Functional enrichment analysis	10
2.2.6	Validation cohort	11
2.3	Results	11
2.3.1	Study cohort characteristics	11
2.3.2	Whole Exome sequencing (WES) analysis	13
2.3.3	Variants analysis	13
2.3.4	Gene analysis	14
2.3.5	Functional enrichment analysis	18
2.3.6	Functional enrichment analysis in the validation cohort	26
2.4	Discussion	29
3	Mirnome profile of circulating extracellular vesicles (EVs) across the histological spectrum of non-alcoholic fatty liver disease (NAFLD)	35
3.1	Introduction	35
3.2	Methods	36
3.2.1	Cohort of study	36
3.2.2	Isolation of serum extracellular vesicles	36
3.2.3	Nanoparticle Tracking Analysis	36
3.2.4	Super-resolution microscopy	36
3.2.5	Library preparation and next-generation sequencing	37
3.2.6	Sequencing data analysis	37
3.2.7	MiRNA targets enrichment analysis	37

3.2.8	Computational and Statistical Analysis	37
3.3	Results	38
3.3.1	Study cohort characteristics	38
3.3.2	EVs isolation	39
3.3.3	Sequencing Data	40
3.3.4	MiRNA expression analysis	41
3.3.5	Target genes and Enrichment analysis	44
3.4	Discussion	53
4	Untargeted analysis to identify metabolites related to non-alcoholic fatty liver disease (NAFLD) and non-alcoholic steatohepatitis (NASH) condition	55
4.1	Introduction	55
4.2	Methods	56
4.2.1	Cohort of study	56
4.2.2	Untargeted metabolomics analysis	56
4.2.3	Statistical analysis	57
4.3	Results	58
4.3.1	Cross-Validation analysis	58
4.3.2	Enrichment analysis	60
4.4	Discussion	64
5	Conclusions	67
	Aknowledgments	70
	Bibliography	71
	List of publications	83

Chapter 1

Introduction

Non-alcoholic fatty liver disease (NAFLD) is currently defined as the accumulation of excessive fat in the liver in the absence of excessive alcohol consumption (less than 20 g per day) or other liver disease aetiologies. [1]

It is the most common cause of liver disease worldwide, in most of the cases it is of primary origin and its aetiology is not yet completely understood, even if it is closely linked to type II diabetes (T2DM), obesity, insulin resistance (IR), arterial hypertension and dyslipidemia. Among the secondary causes of NAFLD, there are relevant causes, such as nutritional (e.g., malnutrition, rapid weight loss), metabolic (e.g., abetalipoproteinemia, lipodystrophy), and drug-induced (e.g., glucocorticoids) as well as other conditions (e.g., jejunal diverticulitis with bacterial overgrowth, inflammatory bowel disease, occupational exposure). [2]

A recent meta-analysis estimated that about one-third of the world population has NAFLD. These rates have increased from 25% to 38% over the past three decades [3]. According to mathematical studies, the number of patients with NAFLD and NASH will continue to increase over the next years worldwide. In Europe, a study demonstrated that from 2016 to 2030, NAFLD cases could potentially increase from 13.9 million to 16.0 million in France; 15.2 million to 17.42 million in Italy; 18.4 million to 20.9 million in Germany; 14.0 million to 16.9 million in the United Kingdom (UK) and 10.5 million to 12.6 million in Spain. By 2030, the estimated NAFLD prevalence was forecasted to be highest in Italy (29.5%), followed by Spain (27.6%), Germany (26.4%), the UK (24.7%), and France (23.6%) [4].

NAFLD is a broad range of disease states (Figure 1.1), and most people who received NAFLD diagnosis have simple steatosis with a low risk of progression and a low risk of liver-related mortality or morbidity.

The NAFLD is defined as the percentage of liver fat greater than or equal to 5% when evaluated by liver biopsy, representing the gold standard in diagnostic tests or magnetic resonance imaging (NMR). Generally, 10%-20% of patients can develop the progressive form: non-alcoholic steatohepatitis (NASH), characterized by the concomitant presence of steatosis, lobular inflammation, and ballooning of hepatocytes, which, in the presence of advanced fibrosis, can lead to cirrhosis and its complications including liver failure and hepatocellular carcinoma (HCC) [5]. Despite the important clinical impact, most people living with NAFLD are undiagnosed also because they do not show symptoms [6].

Also, disease course prediction is uncertain, and no treatments are licensed for this condition. A NAFLD diagnosis is often discovered due to abnormal liver function tests such

as incidental findings of hepatic steatosis on radiologic abdominal findings or aminotransferases (ALT and AST) [7].

Several studies demonstrated that NAFLD is linked to different pathologies. It is associated with an increased incidence of cardiovascular (CV) morbidity and mortality [8], with a higher risk for cerebrovascular events [9], and numerous evidence suggest that NAFLD ultimately represents an independent risk factor for chronic kidney disease (CKD) [10]. To date, the natural history of the disease still needs to be fully understood. Still, it is known that the development of NAFLD is mainly associated with lipid accumulation, oxidative stress, ER stress, and lipotoxicity [11].

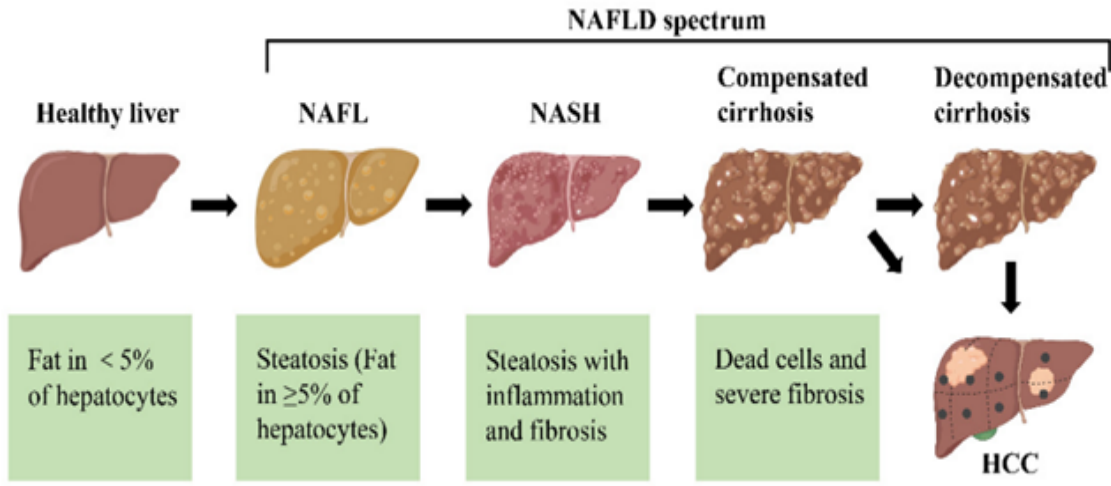


Figure 1.1: The different non-alcoholic fatty liver disease (NAFLD) stages.[12]

1.1 Pathogenesis of NAFLD

The pathogenesis of NAFLD is heterogeneous, already evident by the fact that NAFLD comprises a clinical spectrum from simple steatosis to cirrhosis, the end stage of liver disease. However, it is not clear why some people with simple steatosis develop NASH while others do not undergo the same evolution, just as it is not clear why only some people with NASH develop advanced liver disease. In addition, little is known regarding the temporal development of fatty liver disease.

Much literature reports that NAFLD and its progression are attributed to the so-called 'two-hits hypothesis' theory [13], which suggests that the first step would consist in the accumulation of triglycerides inside the hepatocytes, a phenomenon mediated by the IR, which would bring to a second step responsible of disease progression (e.g. mitochondrial dysfunction, the action of inflammatory cytokines/adipokines and oxidative stress).

However, most recent evidence has suggested more complex mechanisms, denominated as the 'multiple parallel hits hypothesis' theory, suggesting that different elements act together rather than in series [14]. According to that theory, IR, mitochondrial dysfunction, endoplasmic reticulum stress, microbiota, chronic inflammation, genetic and epigenetic factors, and dysfunction of adipose tissue represent, altogether, synchronic causes of both NAFLD development and progression. As already mentioned, IR has a key role in NAFLD

development, as it leads to an increase in hepatic lipogenesis and a suppression of adipose tissue lipolysis, with a consequent elevated flow of fatty acids in the liver [15].

Fat accumulates within the hepatocytes mainly as triglycerides deriving from the esterification of glycerol and free fatty acids (FFAs) [16]. The hepatic accumulation of triglycerides does not act as a hepatotoxic event but rather as a defense mechanism that balances the excess FFAs in the plasma. Other bioactive intermediates, like ceramides and diacylglycerol (DAG), can induce lipotoxicity, leading to inflammation, necrosis, and liver fibrosis. NAFLD progresses to NASH when the mechanisms defending hepatocytes from lipotoxicity are removed. This causes necrosis and secondary repair phenomena mediated by hepatic stellate cells, with the deposition of scar collagen tissue the development and progression of fibrosis [17].

Significantly, NAFLD itself contributes to the development of IR, leading to the establishment of a vicious circle. In NAFLD patients, genetic and environmental aspects can interfere with the insulin post-receptor signal cascade through the phosphorylation of serine residues on insulin [18], the activation of the nuclear factor kappa-light-chain-enhancer of activated B cells (NF- κ B), and the Suppressor Of Cytokine Signaling-3 (SOCS-3) [19]. Therefore, the relationship between IR and NAFLD seems bidirectional.

Regarding visceral fat, it may also be an independent risk factor for NAFLD development. Indeed, visceral fat releases increased FFAs and triglycerides into the circulation and can cause inhibition of glucose uptake, glucose oxidation, and glycogen synthesis via several mechanisms. Increased FFAs to the liver via the portal vein also act as ligands for Toll-like receptor (TLR) 4, inducing cytokine production, thereby leading to inflammation pathways associated with NAFLD. [20] Also, structural and functional alterations in the mitochondria contribute to the pathogenesis of NAFLD [21]. In fact, if the mitochondrial and peroxisomal functions are not able to handle the increase in lipid flow, respiratory oxidation collapses, thus leading to impairment in lipid homeostasis, generation of toxic metabolites, and overproduction of reactive oxygen species (ROS) [22]. These molecules cause oxidative stress, participate in hepatic necro-inflammatory processes, and worsen mitochondrial damage. Indeed, IR, obesity, and TNF- α levels resulted to be directly associated with mitochondrial dysfunction [23]. In addition, ROS, along with oxidized LDL particles, can activate Kupffer and hepatic stellate cells, thus resulting in the addition of collagen and secondary liver fibrosis [24].

Increasing evidence proposes that gut microbiota may be involved in both pathogenesis and progression of NAFLD, promoting inflammation and favoring the mechanisms of liver damage progression [25] [26].

Overall, genetic and epigenetic factors also resulted in the progression of NAFLD. Genetic variants can affect several processes as the FFAs flow into the liver, response to endotoxins, oxidative stress, production, and activity of cytokines. Therefore, they seem crucial both in the development and progression of NAFLD [27]. In addition to the role of genetic variants, some studies have been focusing on the potential impact of epigenetics on the development of NAFLD. For example, non-coding microRNAs (miRNAs) were investigated as key components that regulate epigenetic gene expression mechanisms. Changes in miRNA expression and circulating levels have been related to NAFLD and NASH pathogenesis. Different miRNAs are expressed in individuals with NASH and are associated with lipid and glucose metabolism alterations [28].

1.2 Progression to NASH

As already described above, NASH is the progressive form of simple steatosis, in which a specific injury pattern is characterized by ballooning hepatocellular injury combined with macrovesicular steatosis and inflammation. Fibrosis and Mallory-Denk bodies (MDBs) are also often seen but not required for the diagnosis [29].

Compared to NAFL, NASH is a specific pattern of liver injury that may be recognized even if present with other liver diseases. The gold standard for the NASH diagnosis remains the liver biopsy; however, the features of steatohepatitis are not present in equivalent proportions in every biopsy, and no single feature by itself is diagnostic, making the diagnosis difficult at times. Recently, it was demonstrated that among NAFLD patients who underwent a random liver biopsy, the prevalence of NASH was 6.67%. In addition, the prevalence of NASH among NAFLD patients with a clinical indication for a liver biopsy was 59.10%. Given these estimates, the average prevalence of NASH was estimated to be between 1.5% and 6.45% [30]. Since liver biopsy is not practicable in the general population, no studies can accurately assess the incidence or prevalence of NASH. Unfortunately, plasma biomarkers for the NASH diagnosis have low specificity, between 78% and 82%, and sensitivity, ranging from 62% to 66% [31].

While simple steatosis can remain for years with a benign or significant course improvement, on the other hand, a 10-year prognosis of cirrhosis-associated NASH results very severe, with the need for transplant in 50% of cases, death associated with hepatic causes in 20% and development of HCC in 7% [32].

1.3 Omics Approach

Adding 'omics' to a molecular term implicates a complete, or global, assessment of a set of molecules [33]. There are various branches of omics sciences; the more known are proteomics, transcriptomics, genomics, metabolomics, lipidomics, and epigenomics, which correspond to global analyses of proteins, RNA, genes, metabolites, lipids, and methylated DNA or modified histone proteins in chromosomes, respectively [34]. As a rapidly developing and ever-growing field, omics has empowered us to uncover the intricate molecular mechanism underlying different phenotypic manifestations of disordered traits in an overwhelming and systematic manner at high accuracy. However, the complexity of cellular behavior and its decision-making system may persistently drive the establishment of novel omics and associated techniques. Sequencing and mass spectrometry (MS) are basic experimental tools availing in investigating the omics of a given biological system. While sequencing-based approaches are feasible for studies on genome, transcriptome, their epitomes, and interactomes involving DNA/RNA, MS-based techniques can be used to interrogate proteome, metabolome, and interactomes that do not involve DNA/RNA [35].

Such high-throughput assays generate large amounts of data, which can be used for modeling and other forms of analysis but also present data storage challenges. Multiple databases have been constructed to house data from omics studies and to make the data publicly available [36].

In recent years, different omics approaches have been applied to study different complex disorders, including cardiovascular diseases, different types of cancer, and Alzheimer's disease, with the aim to detect the early onset of the pathologies, novel biomarker molecules,

or molecular signatures with potential value in clinical practice [37] [38] [39] [40].

Each type of omics data, on its own, usually provides a list of differences associated with the disease. These data can be helpful both as markers of the disease process and to give understanding as to which biological pathways or mechanisms are different between the disease and control groups or between different stages of the same pathology. However, analysis of only a singular data type is limited to correlations, mostly reflecting reactive mechanisms rather than causative ones. Integrating different omics data types is often used to explain potential causative changes that lead to disease and identify mechanisms of action, pathways, biomarkers, and other relationships present in pathological processes [41].

The multi-omics approach provides an integrated perspective to power discovery across multiple levels of biology. This combines genomic data with data from other modalities, like transcriptomics, epigenetics, and proteomics, to measure gene activation, gene expression, and protein levels. Multi-omics profiling studies enable a more comprehensive understanding of molecular changes supporting normal development, cellular response, and disease. Using integrative omics approaches, it is possible to better link genotype to phenotype and fuel the discovery of novel drug targets and biomarkers.

Importantly, -omics technologies permit the systematic use and combination of multiple data sets, creating more dimensionality in the interpretation of the results that can lead to the development of new conceptual approaches and treatment strategies [42].

Nonetheless, significant challenges remain in the omics sciences, especially concerning data complexity, data management, and the integration of data from omics studies with data from other sources, such as clinical data gathered during routine physician visits.

1.4 Omics in NAFLD

Knowledge in the field of liver diseases, especially NAFLD, has benefitted in the last years from the rapid development of high-throughput technologies, including genomics, transcriptomics, proteomics, and metabolomics. Indeed, due to the several limitations of the histological diagnosis, significant research and clinical efforts are currently directed toward the search for solid biomarkers aimed to predict the disease prognosis and severity [43]. Certain omics have been widely used so far, either for identifying high-risk patients (genomics), for evaluating the effect of treatments in key pathophysiological mechanisms (transcriptomics), or for developing non-invasive diagnostic tools (metabolomics/lipidomics). Genomics was the first and one of the most extensively investigated omics to identify the genetic susceptibility underlying NAFLD development and progression. The Genome Wide Association Studies (GWAS) have significantly contributed to enhancing the knowledge of the genetic component of NAFLD, providing robust evidence about the causal relationship of certain common SNPs with the development and rapid progression of NAFLD. Among others, the rs738409 in the PNPLA3 gene explains about 5.3% of the total variance in NAFLD [44] [45] [46]. Furthermore, several Whole-Exome sequencing (WES) studies were conducted to investigate novel variants conferring susceptibility to NAFLD [47] [48] [49].

In the last ten years, researchers have tried to incorporate genetic markers into non-invasive tests that discriminate between NAFLD stages. Despite some limitations, there have been some interesting attempts with encouraging results [50] [51]. Importantly, genomics is not a dynamic process; thus, they cannot reflect the impact of environmental factors on the

pathogenesis and progression of NAFLD. However, Multi-gene models are promising tools for patient risk stratification.

Also, epigenomics was investigated in NAFLD. The few studies performed so far have shown that the extensive hepatic epigenetic modifications observed in NAFLD are causally related to the disease; however, it emerged as an attractive target for therapeutic intervention. Indeed, changes in DNA methylation could be used as a biomarker that allows monitoring, for instance, the effectiveness of pharmacotherapy [52]. It is interesting to note that proteomics is not widely used in NAFLD research. The proteins and signaling pathways associated with the disease's pathophysiology have been almost exclusively identified by approaches based on data from mechanistic animals, observational humans, and, in some cases, large epidemiological studies. This is probably linked to methodological challenges and technical limitations related to proteomics. In the future, the expected main contribution of proteomics will not be in the direct development of non-invasive diagnostic scores but rather in identifying novel molecules and/or pathways with important roles in the pathophysiology of the disease and with diagnostic and therapeutic potentials [53].

On the other hand, Metabolomic and Lipidomics emerged among the most investigated omics in NAFLD. They have provided important information about the pathophysiology and course of the disease, and changes in metabolome and lipidome have a major role in mechanisms of NAFLD development and progression [54] [55] [56].

Thus, most drug candidates aim to restore the metabolic and lipid profile abnormalities observed in NAFLD [57] [58].

These abnormalities are observed in the liver and often in peripheral blood, which has led to the extensive use of metabolomics/lipidomics to develop non-invasive tests for diagnosing and staging NAFLD. Despite several limitations correlated to the techniques, among the omics, metabolomics/lipidomics are still considered to have the best potential for leading to the development of important tools for diagnosis and staging of the disease. Overall, omics approaches have contributed to the creation of models for diagnosing and staging the disease and have identified potential therapeutic targets. The power of these technologies creates significant opportunities for addressing unmet clinical needs, such as the non-invasive diagnosis and treatment of NAFLD.

1.5 Aim of study

The present thesis integrates multiple aims consistent with a 4-years PhD work.

Assuming that omics approaches are contributing to the major evolution observed in health research in the last years, this work was focused on measuring multiple omics data types on a single set of NAFLD patients, composed of histologically characterized individuals, exploiting the potential of technologies, such as Next Generations Sequencing (NGS) and mass spectrometry (MS). The main aim is to provide deep insight into biological mechanisms and networks, allowing for the development of more powerful patient stratification strategies for a significant step forward in personalized medicine.

Specifically, in Chapter 2, the Whole exome Sequencing (WES) was performed to identify the different genetic landscapes highlighted the various stages of the diseases, NAFL and NASH. Chapter 3 aimed to use the NGS approach to perform a comprehensive miRNome profile of miRNAs contained in circulating EVs, aiming at identifying specific miRNAs associated with the presence of NASH and significant fibrosis.

In Chapter 4, a comprehensive untargeted metabolomics analysis was performed to identify metabolites differentially expressed between NAFL and NASH patients.

Finally, a brief discussion of the main results is provided.

Chapter 2

Whole exome sequencing to identify genes predisposing to the progression of non-alcoholic fatty liver (NAFL)

2.1 Introduction

NAFLD is a complex disease with multiple factors involved in the pathogenesis that include both environmental and genetic modifiers. Although the understanding of the genetic underpinnings of NAFLD has advanced, known risk variants explain only a tiny portion of heritability, suggesting the existence of 'missing heritability' [59].

Today, several familial clustering studies involving monozygotic twins and inter-ethnic differences point towards the role of genetics in the pathogenesis of NAFLD [60].

Major evidence on the genetic basis of NAFLD comes from studies showing the ethnic differences in the prevalence and severity of NAFLD. In a study including cryptogenic cirrhosis or NASH-related cirrhosis, it was reported that majority of cirrhosis among Hispanic and African American patients was 3.1 times higher and 3.9 times lower than European American patients, despite similar prevalence of type 2 diabetes mellitus among these groups [61]. It is now known that particular genetic variants such as the *PNPLA3* (patatin-like phospholipase domain-containing 3) rs738409 [62], *TM6SF2* (transmembrane 6 superfamily member 2) rs58542926 [63], *HSD17B13* (hydroxysteroid 17-beta dehydrogenase 13) rs6834314 [64], *MBOAT7* (membrane-bound O-acyltransferase domain containing 7) rs641738 and *GCKR* (glucokinase regulatory protein) rs1260326 may disturb specific pathways in hepatic lipid metabolism and are considered key determinants in the susceptibility to progress to non-alcoholic steatohepatitis and cirrhosis among NAFLD individuals [65] [66].

Multiple Genome Wide Association Studies (GWAS) and Epigenome Wide Association Studies (EWAS) have been conducted to find genetic and DNA methylation associations with specific features of NAFLD, such as hepatic steatosis, fibrosis, and liver inflammation, as well as the full spectrum of disease. Some variants have been associated with the full NAFLD spectrum, while others correlate only with certain phenotypes [67].

In contrast to GWAS, in which the majority of variants identified are located in non-coding

genomic sequences distant from protein-coding genes, Whole Exome Sequencing (WES) analyses almost entirely use variants in the exonic regions of the genome, which can make causal gene identification easier as the variants likely alter the sequence of the encoded protein. However, to date, the findings of NAFLD WES are still limited.

This work aims to use the WES analysis to analyze a cohort of 154 very well-histologically defined NAFLD patients to identify the different genetic landscapes highlighted by the different stages of the disease.

2.2 Methods

2.2.1 Cohort of study

The cohort of the following study included 154 patients with biopsy-proven NAFLD consecutively enrolled at the Liver Unit (prof. Bugianesi) of the Department of Medical Sciences, University of Torino, Turin, Italy, in the period between June 2010 and May 2020.

The inclusion criteria were:

- people whose age was equal to or greater than 18 years.
- diagnosis of NAFLD confirmed by liver biopsy.
- availability of both serum and DNA samples collected at the time of liver biopsy.

Other etiologies of liver disease, including viral, autoimmune, cholestatic, genetic, alcoholic, and drug-induced, were not considered. At the time of the liver biopsy, a complete medical history and physical examination were undertaken, and anthropometric data were collected too. All patients underwent serological screening for hepatitis B surface antigen, antibodies to hepatitis B core antigen, and antibodies to hepatitis C virus. As part of the clinical workup, standard laboratory data have been collected, including full blood count, liver biochemistry, and metabolic profile. Among all participants, 108 of them (70.13% of the cohort) received a histological diagnosis of NASH due to the concomitant presence of liver steatosis greater than 5%, lobular inflammation, and hepatocyte ballooning.

All patients signed a written informed consent for the collection of personal data and for the use of blood samples for research purposes. This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Ethics Committee of A.O.U. Città della Salute e della Scienza di Torino (CEI/522, 17 Nov 2015).

2.2.2 DNA isolation and library preparation

For each patient, DNA was isolated from a whole blood sample using the EZ1 DSP DNA Blood Kit (Qiagen; Hilden, Germany).

Before libraries preparation, all DNA samples have been checked in terms of quality and quantity. The absorbance ratio at 260 nm and 280 nm in a specific range ($1.7 < 260/280 < 2$) was used to assess the purity of DNA using a NanoDrop Spectrophotometer (Thermo Fisher Scientific; Waltham, Massachusetts, USA).

The samples concentration was determined using Qubit dsDNA Broad Range (BR) Assay Kits, designed for use with Qubit Fluorometer (Thermo Fisher Scientific; Waltham, Massachusetts, USA).

For the library construction, DNA was prepared for Illumina paired-end sequencing according to the KAPA HyperCap Workflow v3.0 (Roche Sequencing Solutions, Inc.; Pleasanton, USA). This approach includes enzymatic fragmentation (KAPA HyperPlus Kit), amplification, and purification steps (KAPA UDI Primer Mixes and KAPA HyperPure Beads). Unique barcodes were added to the DNA fragments to facilitate multiplexed exome capture and sequencing. Equal amounts of sample were pooled before overnight exome capture, utilising KAPA HyperExome Probes technology (Roche Sequencing Solutions, Inc.; Pleasanton, USA). Afterward, the captured DNA was PCR-amplified and purified.

To assure the final best result, quality and quantity were checked respectively with the High Sensitivity DNA assay kit on the 2100 Bioanalyzer System (Agilent Technologies; Santa Clara, California, USA) and with the Qubit dsDNA High Sensitivity (HS) Assay Kit on the Qubit Fluorometer (Thermo Fisher Scientific; Waltham, Massachusetts, USA).

2.2.3 Whole Exome Sequencing

The multiplexed samples were pooled and then sequenced using 150-base-pair paired-end reads with two 8-base-pair index reads on the Illumina Novaseq 6000 platform using SP (the first 16 samples) or S1 (for all other samples) flow cells (Illumina, San Diego, California, USA). The sequencing was performed at the Italian Institute for Genomic Medicine (IIGM, c/o IRCCS hospital Candiolo (TO) - Italy).

2.2.4 Sequencing data analysis

The generated data were subsequently demultiplexed and changed into FastQ files using the bcl2fastq conversion software (Illumina, San Diego, California, USA).

An average alignment coverage of 150X and more than 97% of bases in the target bed with coverage equal to or greater than 20 were obtained.

The alignment of the sequence reads on the GRCh38/hg38 genome assembly, indexing of the reference genome, variant calling, and annotation were performed with the closed JuliaOmix™ pipeline (GenomeUp; Rome, Italy). All Single Nucleotide Variants (SNVs) annotated as upstream_gene_variant, downstream_gene_variant, intron_variant, and intergenic_variant have been removed. Only high-quality SNVs, identified as those that passed internal quality control points, with read depth > 20 and annotated on canonical transcripts, were used for the downstream analysis.

Variants were classified as pathogenic, likely pathogenic, uncertain significance, likely benign, or benign variants using the recent recommendations of the American College of Medical Genetics (ACMG) [68].

2.2.5 Functional enrichment analysis

Starting from variants analysis, different genes resulted to be mutated respectively only in NASH or NAFL patients or in both groups. The gene lists were used separately to conduct the functional enrichment analysis.

The Kyoto Encyclopaedia of Genes and Genomes (KEGG), the Gene Ontology (GO) Cellular component, and the Jensen Disease databases, implemented in ShinyGO (v0.77) [69], were used as reference for the pathway enrichment analyses.

A value of FDR<0.05 was regarded as statistically significant.

2.2.6 Validation cohort

Exome sequencing data of 79 individuals, 14 (17.2%) NAFL and 65 NASH (82.8%) patients were obtained from the Division of Digestive and Liver Diseases, Department of Medicine of the Columbia University Irving Medical Center, New York, NY 10032, USA. Detailed information about samples collection, sequencing process, and variant annotation are reported in previous publications [70]. Only a list of variants, including the ACMG annotations and information about frequency and variant's function, was available and used to confirm results from the functional analysis performed on the discovery cohort. Written informed consent was obtained from each patient, and the study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a priori approval by CUIMC Institutional Review Board.

2.3 Results

2.3.1 Study cohort characteristics

A total of 154 patients with biopsy-proven NAFLD were included in the present study. Among them, 108 individuals (70.1%) received a histological diagnosis of NASH according to the concomitant presence of liver steatosis $>5\%$, lobular inflammation, and hepatocyte ballooning. The demographic, clinical, and biochemical characteristics of both groups are reported in Table 2.1. Median age was 48 ± 13 years in NASH patients and 46 ± 12 years in NAFL ones. In the study cohort, most patients were males (68.2%). Among them, (64.8%) received the histological diagnosis of NASH, while (35.2%) were in the NAFL group. Only 49 patients (31.8%) were females, among them (81.6%) with a NASH diagnosis. Significant liver fibrosis (defined as Fibrosis score ≥ 2) was observed in the NASH group (50.0%) compared to the (39.1%) observed in the NAFL patients. In both groups, most of the patients (62%) in NASH and (69.6%) in NAFL were not affected by type 2 diabetes mellitus.

Subjects(n=154)	NASH (%)	NAFL (%)	P-value
	108 (70.1%)	46 (29.9%)	
Gender			
M	68 (44.2%)	37 (24%)	0.04*
F	40 (26%)	9 (5.8%)	ref
Mean Age [years, mean±sd]	48±13	46±12	0.53
BMI [mean±sd] (kg/m ²)	29.5±4.4	28.0±5.0	0.05*
Fibrosis_score			
0	33 (30.6%)	14 (30.4%)	ref
1	21 (19.4%)	14 (30.4%)	0.36
2	14 (13%)	11 (23.9%)	0.3
3	22 (20.4%)	3 (6.5%)	0.14
4	18 (16.7%)	4 (8.7%)	0.39
Type 2 Diabetes			
yes	41(38%)	14(30.4%)	0.46
no	67(62%)	32(69.6%)	ref
GLUCOSE [mean±sd] (mg/dl)	102.8±29.4	99.15±15.34	0.79
INSULIN [mean±sd] (uUI/ml)	14.7±12.5	12.79±14.41	0.23
TGC [mean±sd] (mg/dl)	140.9±88.4	137±85.86	0.41
CHOL tot [mean±sd] (mg/dl)	189.2±47.1	181.57±57.28	0.26
CHOL HDL [mean±sd] (mg/dl)	45.5±16.1	42.46±17.67	0.26
AST [mean±sd] (U/l)	38.6±26.1	36.22±28.48	0.25
ALT [mean±sd] (U/l)	57.1±33	61.04±49.30	0.91
ALBUMIN [mean±sd] (g/dl)	4.2±0.9	4.32±0.72	0.31
PLT [mean±sd] (x10 ⁹ /l)	244.6±206.4	226.15±83.14	0.58
γGT [mean±sd] (U/l)	93.5±107.4	87.67±137.93	0.29
BIL Tot [mean±sd] (mg/dl)	0.8±0.5	0.72±0.36	0.78
Hb [mean±sd] (g/dl)	14.7±1.5	14.85±1.23	0.65
PRO-C3 [mean±sd] (ng/ml)	14±11.7	10.18±4.44	0.2
PRO-C6 [mean±sd] (ng/ml)	8.1±4.2	6.95±3.01	0.13
sCD163 [mean±sd] (ng/ml)	542.3±195.6	493.61±161.41	0.11

Table 2.1: Summary of the demographic, clinical, and biochemical characteristics of the 154 individuals included in the study cohort. Fisher's exact Test, and unpaired two-samples Wilcoxon test were used. P-value ≤ 0.5 are indicated with *.

2.3.2 Whole Exome sequencing (WES) analysis

The WES analysis was performed for all the 154 samples. The quality of raw data obtained from the Illumina Novaseq 6000 system was evaluated and reported in Table 2.2 for each run. Importantly, the Run 1 included 16 samples, and it was performed using the SP Illumina flowcell; the other runs (Run 2, Run 3, and Run 4) were performed using the S1 Illumina flowcells, including 48 samples in each run. For each patient, the $82.9 \pm 5.3\%$ (Max 92.03%; Min 79.18%) of bases in the target region obtained a coverage higher than 100X. A total of $83.1 \pm 1.5\%$ (Max 85.47%; Min 81.52%) of aligned reads was obtained, and the average alignment coverage was 159 ± 23.2 X (Max 199.21X; Min 144.95X). An average of $14.5 \pm 1.6\%$ (Max 16.95%; Min 12.32%) duplicates marked reads were obtained.

	RUN1*	RUN2	RUN3	RUN4	Mean Value	Standard Deviation	Max value	Min value
% of bases in target_bed with coverage (100x:inf)	92.03	79.76	79.18	80.76	82.9	5.3	92.03	79.18
% Aligned reads	85.47	83.23	81.52	82.1	83.1	1.5	85.47	81.52
Average alignment coverage	199.21	144.95	145.2	146.7	159.0	23.2	199.21	144.95
% Number of duplicate marked reads	16.95	12.32	14.22	14.32	14.5	1.6	16.95	12.32

Table 2.2: Summary of the raw quality data obtained from the WES analysis of 154 NAFLD patients. *Run 1 includes 16 samples (SP Illumina Flowcell); Run 2, Run 3, and Run 4 include 48 samples, respectively (S1 Illumina Flowcell).

2.3.3 Variants analysis

Only High-quality SNVs that passed all quality filters reported above were analyzed. A total of 90877 SNVs were identified in the entire cohort, of which 39678 (43.66%) were uncommon variants (minor allele frequency (MAF) < 1% across all ancestries). As reported in Table 2.3, in NAFL patients (42 individuals), approximately 12986 (24.46%) uncommon variants were identified considering the ACMG guidelines, including 22 (0.04%) pathogenic variants, 288 (0.54%) likely pathogenic variants and 32130 (60.54%) variants with Uncertain Significance (VUS). Of NASH patients (108 individuals), 28991 (37.66%) resulted in uncommon variants, of which 62 (0.08%) reported as pathogenic variants, 513 (0.67%) as likely pathogenic variants, and 50737 (65.81%) as VUS according ACMG criteria.

	Cohort (154)	NAFL PATIENTS (42)	NASH PATIENTS (108)
Number of variants	90877 (100%)	53093	77097
Rare Variants MAF< 0.01	39678 (43.66%)	12986 (24.46%)	28991 (37.60%)
ACMG Classification of rare variants (MAF<1%)			
<i>Pathogenic</i>		22 (0.04%)	62 (0.08%)
<i>Likely Pathogenic</i>		288 (0.54%)	513 (0.67%)
<i>VUS</i>		32130 (60.52%)	50737 (65.81%)
<i>Other</i>		20653 (38.90%)	25785 (33.44%)

Table 2.3: Number of variants discovered in exome sequencing data from 154 NAFLD patients included in the discovery cohort.

2.3.4 Gene analysis

In the entire study cohort, approximately 460 genes reported at least one uncommon (MAF<1%) pathogenic or likely pathogenic variants, among which 102 (22.2%) only in NAFL patients, 262 (57%) only in NASH patients and 96 genes in both groups (Table 2.4). A total of 117 genes were selected to report more than one uncommon pathogenic or likely pathogenic variants, in particular, 37 genes only among NASH patients, 8 genes only among NAFL patients, and 72 genes were in common between the two groups. In 5 genes among the 37 genes exclusively mutated in NASH patients, more than two different uncommon pathogenic or likely pathogenic variants were found: *FLNB* (rs138220431; rs151259375; rs138034708; rs143831841); *ADGRV1* (rs200945405; rs201386977; rs61744480); *COL6A3* (rs769072338; rs369169235; rs150907698); *SBDS* (rs113993993; rs147652512; rs373730800) and *SLC7A9* (rs79389353; rs147344717; rs1266495325) (Table 2.5 and Table 2.6).

In addition, in 9 genes, including *FLNB*, *CEL*, *CYP24A1*, *GHR*, *GNPTAB*, *KARS1*, *SLC7A9*, *C9*, and *UROCI*, the same mutation was observed in more than one NASH patient belonging to the study cohort. In particular, the variant rs143831841 in the *FLNB* gene and the rs756449511 in the *CEL* gene were identified respectively in 4 and 3 NASH patients.

Among the 8 genes exclusively mutated in NAFL patients, with more than one uncommon pathogenic or likely pathogenic variants, the *APOB* gene reported three different variants and only the rs33995463 in the *LRRK2* gene was observed in more than one NAFL patient. (Table 2.7). Regarding the uncommon VUS, 9448 genes resulted in carrying at least one uncommon VUS mutation, and among them, 863 (9.1%) genes reported this type of mutation only in NAFL patients, 3557 (37.6%) only in NASH and 5028 (53.2%) were in common between the two groups.

	Total number of genes (%)	ONLY IN NAFL PATIENTS (%)	ONLY IN NASH PATIENTS (%)	COMMON (%)
Gene with P-LP mutations MAF<1%	460 (100%)	102 (22.2%)	262 (57%)	96 (20.9%)
Gene with VUS mutations MAF<1%	9448 (100%)	863 (9.1%)	3557 (37.6%)	5028 (53.2%)
Gene with more than 1 P-LP mutations MAF<1%	117 (100%)	8 (6.8%)	37 (31.6%)	72 (61.5%)

Table 2.4: Summary of genes resulted to be mutated respectively in NASH, NAFL, and both groups.

Gene	Variant ID	Number of NASH patients with the variant	MAF in general population	Gene Definition
FLNB	rs138220431	2	0.005	This gene encodes a member of the filamin family. The encoded protein interacts with glycoprotein I b alpha as part of the process to repair vascular injuries.
	rs151259375	1	0.002	
	rs138034708	1	0.001	
	rs143831841	4	0.003	
CEI	rs756449511	3	0.001	The protein encoded by this gene is a glycoprotein secreted from the pancreas into the digestive tract and from the lactating mammary gland into human milk. The physiological role of this protein is in cholesterol and lipid-soluble vitamin ester hydrolysis and absorption.
	rs763199261	2	0.0004	
CYP24A1	rs114368325	2	0.002	This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.
	rs35873579	1	0.004	
GHR	rs121909362	2	0.008	This gene encodes a member of the type I cytokine receptor family, which is a transmembrane receptor for growth hormone.
	rs143814221	1	0.002	
GNPTAB	rs34002892	2	0.001	This gene encodes two of three subunit types of the membrane-bound enzyme N-acetylglucosamine-1-phosphotransferase. These recognition markers are essential for appropriate trafficking of lysosomal enzymes. There are two specific forms of GNPTAB-related disorders, mucopolidosis II and mucopolidosis III alpha/beta.
	rs35333334	1	-	
KARS1	rs77573084	2	0.004	Aminoacyl-tRNA synthetases are a class of enzymes that charge tRNAs with their cognate amino acids. Lysyl-tRNA synthetase is a homodimer localized to the cytoplasm which belongs to the class II family of tRNA synthetases. It has been shown to be a target of autoantibodies in the human autoimmune diseases.
	rs144274136	1	0.001	
SLC7A9	rs79389953	2	0.008	This gene encodes a protein that belongs to a family of light subunits of amino acid transporters. This protein plays a role in the high-affinity and sodium-independent transport of cystine and neutral and dibasic amino acids, and appears to function in the reabsorption of cystine in the kidney tubule.
	rs147344717	1	0.005	
	rs1266495325	1	0.0002	
C9	rs34000044	2	0.002	This gene encodes the final component of the complement system. It participates in the formation of the Membrane Attack Complex (MAC). The MAC assembles on bacterial membranes to form a pore, permitting disruption of bacterial membrane organization.
	rs771605119	1	0.00007	
UROCL	rs200128917	2	0.001	This gene encodes an enzyme involved in the second step of histidine catabolism, metabolizing urocanic acid to form iminoglutamic acid.
	rs145069129	1	0.0001	
ACADS	rs539219309	1	0.0008	This gene encodes a tetrameric mitochondrial flavoprotein, which is a member of the acyl-CoA dehydrogenase family. This enzyme catalyzes the initial step of the mitochondrial fatty acid beta-oxidation pathway.
	-	1	-	
ADGRV1	rs200945405	1	0.003	This gene encodes a member of the G-protein coupled receptor superfamily. The encoded protein contains a 7-transmembrane receptor domain, binds calcium and is expressed in the central nervous system. Mutations in this gene are associated with Usher syndrome 2 and familial febrile seizures.
	-	1	-	
	rs201386977	1	0.006	
	rs61744480	1	0.004	
ANOS	rs137854523	1	0.003	This gene encodes a member of the anoctamin family of transmembrane proteins. The encoded protein is likely a calcium activated chloride channel. Mutations in this gene have been associated with gnathodiaphyseal dysplasia.
	rs1564930625	1	-	
BBS4	rs1375331013	1	0.00001	This gene is a member of the Bardet-Biedl syndrome (BBS) gene family. Bardet-Biedl syndrome is an autosomal recessive disorder characterized by severe pigmentary retinopathy, obesity, polydactyly, renal malformation and cognitive disability.
	rs775710800	1	0.00005	
BRIP1	rs28903098	1	0.0005	The protein encoded by this gene is a member of the RecQ/DEAH helicase family and interacts with the BRCT repeats of breast cancer, type 1 (BRCA1). The bound complex is important in the normal double-strand break repair function of breast cancer, type 1 (BRCA1).
	rs786203717	1	0.0002	
CACNA1S	rs780390034	1	0.0002	This gene encodes one of the five subunits of the slowly inactivating L-type voltage-dependent calcium channel in skeletal muscle cells. Mutations in this gene have been associated with hypokalemic periodic paralysis, thyrotoxic periodic paralysis and malignant hyperthermia susceptibility.
	rs760246258	1	0.0003	
COL6A3	rs769072338	1	0.0002	This gene encodes the alpha-3 chain, one of the three alpha chains of type VI collagen, a beaded filament collagen found in most connective tissues.
	rs369169235	1	0.0003	
	rs150907698	1	0.002	
CRAT	rs779910980	1	0.0003	This gene encodes carnitine O-acetyltransferase, a member of the carnitine acyltransferase family and a key metabolic pathway enzyme which plays an important role in energy homeostasis and fat metabolism. This enzyme catalyzes the reversible transfer of acyl groups from an acyl-CoA thioester to carnitine and regulates the ratio of acyl-CoA/CoA.
	-	1	-	
ECE1	rs117012322	1	0.006	This gene encodes a member of the M13 family of endopeptidases. Members of this family are zinc-containing type II integral-membrane proteins that are important regulators of neuropeptide and peptide hormone activity. Mutations in this gene are associated with autosomal recessive distal arthrogryposis, type 5D.
	rs564717204	1	0.008	

Table 2.5: Summary of genes reported more than one uncommon pathogenic or likely pathogenic variants only in NASH patients. The table reports gene name; variant ID; Maximum Allele Frequency (MAF) in the general population, and a gene description from the Reference Sequence (RefSeq) database.

Gene	Variant ID	Number of NASH patients with the variant	MAF in general population	Gene Definition
ERBB3	rs779924065	1	0.0002	This gene encodes a member of the epidermal growth factor receptor (EGFR) family of receptor tyrosine kinases. Amplification of this gene and/or overexpression of its protein have been reported in numerous cancers, including prostate, bladder, and breast tumors.
	rs770291856	1	0.000009	
IHH	rs150661368	1	0.0005	The protein encoded by this gene specifically plays a role in bone growth and differentiation. Mutations in this gene are the cause of brachydactyly type A1, which is characterized by shortening or malformation of the fingers and toes. Mutations in this gene are also the cause of acrocapitofemoral dysplasia.
	rs140093604	1	0.004	
IRF2BPL	-	1	-	This gene encodes a transcription factor that may play a role in regulating female reproductive function.
KCND3	rs201340369	1	0.001	This gene encodes a member of the potassium channel, voltage-gated, shal-related subfamily, members of which form voltage-activated A-type potassium ion channels and are prominent in the repolarization phase of the action potential.
	rs14908060	1	0.001	
KIF21A	rs781202644	1	0.0005	This gene encodes a member of the KIF4 subfamily of kinesin-like motor proteins. The encoded protein is characterized by an N-terminal motor domain a coiled-coil stalk domain and a C-terminal WD-40 repeat domain. This protein may be involved in microtubule dependent transport. Mutations in this gene are the cause of congenital fibrosis of extraocular muscles-1.
	rs145318959	1	0.0005	
KRT13	rs141432690	1	0.007	The protein encoded by this gene is a member of the keratin gene family. The keratins are intermediate filament proteins responsible for the structural integrity of epithelial cells and are subdivided into cytokeratins and hair keratins.
	rs148399399	1	0.002	
KRT71	rs763992766	1	0.00006	Keratins are intermediate filament proteins responsible for the structural integrity of epithelial cells and are subdivided into epithelial keratins and hair keratins. This gene encodes a protein that is expressed in the inner root sheath of hair follicles.
	rs141914500	1	0.004	
MCCC1	rs148616219	1	0.003	This gene encodes the large subunit of 3-methylcrotonyl-CoA carboxylase. This enzyme functions as a heterodimer and catalyzes the carboxylation of 3-methylcrotonyl-CoA to form 3-methylglutaconyl-CoA. Mutations in this gene are associated with 3-Methylcrotonylglycinuria, an autosomal recessive disorder of leucine catabolism.
	rs772395858	1	0.00004	
MMP21	-	1	-	This gene encodes a member of the matrix metalloproteinase family. Proteins in this family are involved in the breakdown of extracellular matrix for both normal physiological processes, such as embryonic development, reproduction, and tissue remodeling, and disease processes, such as asthma and tumor metastasis.
	rs886041273	1	0.002	
MSMO1	rs139920569	1	0.002	The protein is localized to the endoplasmic reticulum membrane and is believed to function in cholesterol biosynthesis.
MYL2	rs730880950	1	0.00004	The encoded protein plays a role in embryonic heart muscle structure and function, while phosphorylation of the encoded protein is involved in cardiac myosin cycling kinetics, torsion and function in adults. Mutations in this gene are associated with hypertrophic cardiomyopathy 10 and infant-onset myopathy.
	-	1	-	
PHLPP1	-	1	-	The encoded protein promotes apoptosis by dephosphorylating and inactivating the serine/threonine kinase Akt, and functions as a tumor suppressor in multiple types of cancer. Increased expression of this gene may also play a role in obesity and type 2 diabetes by interfering with Akt-mediated insulin signaling.
	-	1	-	
PKD1L1	rs1247239831	1	0.000009	This gene encodes a member of the polycystin protein family containing 11 transmembrane domains, a receptor for egg jelly (REJ) domain, and a polycystin-1, lipoxigenase, alpha-toxin (PLAT) domain.
	rs1259250158	1	0.000009	
PMS2	rs587780046	1	0.0004	The protein encoded by this gene is a key component of the mismatch repair system that functions to correct DNA mismatches and small insertions and deletions that can occur during DNA replication and homologous recombination.
	rs374704824	1	0.002	
PROC	rs759557871	1	0.00009	This gene encodes a vitamin K-dependent plasma glycoprotein. The encoded protein is cleaved to its activated form by the thrombin-thrombomodulin complex. This activated form contains a serine protease domain and functions in degradation of the activated forms of coagulation factors V and VIII. Mutations in this gene have been associated with thrombophilia due to protein C deficiency.
	rs1321566264	1	0.0002	
SBD5	rs113993993	1	0.001	This gene encodes a highly conserved protein that plays an essential role in ribosome biogenesis. The encoded protein interacts with elongation factor-like GTPase 1 to disassociate eukaryotic initiation factor 6 from the late cytoplasmic pre-60S ribosomal subunit allowing assembly of the 80S subunit. Mutations within this gene are associated with the autosomal recessive disorder Shwachman-Bodian-Diamond syndrome.
	rs147652512	1	0.002	
	rs373730800	1	0.001	
SC5D	rs146322594	1	0.00004	This gene encodes an enzyme of cholesterol biosynthesis. The encoded protein catalyzes the conversion of lathosterol into 7-dehydrocholesterol. Mutations in this gene have been associated with lathosterolosis.
	-	1	-	
SLC10A2	rs141838179	1	0.002	This gene encodes a sodium/bile acid cotransporter. This transporter is the primary mechanism for uptake of intestinal bile acids by apical cells in the distal ileum. Bile acids are the catabolic product of cholesterol metabolism, so this protein is also critical for cholesterol homeostasis. Mutations in this gene cause primary bile acid malabsorption (PBAM); mutations in this gene may also be associated with other diseases of the liver and intestines, such as familial hypertriglyceridemia (FHTG).
	-	1	-	
USH2A	rs80338902	1	0.003	This gene encodes a protein that contains laminin EGF motifs, a pentaxin domain, and many fibronectin type III motifs. The protein is found in the basement membrane, and may be important in development and homeostasis of the inner ear and retina. Mutations within this gene have been associated with Usher syndrome type IIa and retinitis pigmentosa.
	rs1553313793	1	-	

Table 2.6: Summary of genes reported more than one uncommon pathogenic or likely pathogenic variants only in NASH patients. The table reports gene name; variant ID; Maximum Allele Frequency (MAF) in the general population, and a gene description from the Reference Sequence (RefSeq) database.

Gene	Variant ID	Number of NAFL patients	MAF in general population	Gene Definition
AMBN	rs773197079	1	0.00004	This gene encodes the nonamelogenin enamel matrix protein ameloblastin. The encoded protein may be important in enamel matrix formation and mineralization.
	.	1	.	This gene is located in the calcium-binding phosphoprotein gene cluster on chromosome 4. Mutations in this gene may be associated with dentinogenesis imperfect and autosomal dominant amyogenesis imperfect.
APOB	rs142066904	1	0.00003	This gene product is the main apolipoprotein of chylomicrons and low density lipoproteins (LDL), and is the ligand for the LDL receptor. Mutations in this gene or its regulatory region cause hypobetalipoproteinemia, normotriglyceridemic hypobetalipoproteinemia and hypercholesterolemia due to ligand-defective apoB, diseases affecting plasma cholesterol and apoB levels.
	.	1	.	
	.	1	.	
GATA5	rs200383755	1	0.005	The protein encoded by this gene is a transcription factor that contains two GATA-type zinc fingers. The encoded protein is known to bind to hepatocyte nuclear factor-1alpha (HNF-1alpha), and this interaction is essential for cooperative activation of the intestinal lactase-phlorizin hydrolase promoter.
	.	1	.	
LRRK2	rs33995463	3	0.003	This gene is a member of the leucine-rich repeat kinase family and encodes a protein with an ankyrin repeat region, a leucine-rich repeat (LRR) domain, a kinase domain, a DFG-like motif, a RAS domain, a GTPase domain, a MLK-like domain, and a WD40 domain. The protein is present largely in the cytoplasm but also associates with the mitochondrial outer membrane. Mutations in this gene have been associated with Parkinson disease-8.
	rs34637584	1	0.009	
MED15	rs773058463	1	0.0006	The protein encoded by this gene is a subunit of the multiprotein complexes PC2 and ARC/DRIP and may function as a transcriptional coactivator in RNA polymerase II transcription. Diseases associated with MED15 include Hepatocellular Carcinoma and DiGeorge Syndrome.
	.	1	.	
MYO18B	.	1	.	The protein encoded by this gene may regulate muscle-specific genes when in the nucleus and may influence intracellular trafficking when in the cytoplasm. The encoded protein functions as a homodimer and may interact with F-actin.
	.	1	.	Mutations in this gene are associated with lung cancer.
OTOGL	rs200797401	1	0.004	This gene is expressed in the inner ear of vertebrates with the highest level of expression seen at the embryonic stage and lowest in adult. Knockdown studies in zebrafish suggest that this gene is essential for normal inner ear function. Mutations in this gene are associated with autosomal recessive deafness.
	rs777911261	1	0.00005	
TRPC6	.	1	.	The protein encoded by this gene forms a receptor-activated calcium channel in the cell membrane. The channel is activated by diacylglycerol and is thought to be under the control of a phosphatidylinositol second messenger system. Defects in this gene are a cause of focal segmental glomerulosclerosis 2.
	rs1276794439	1	.	

Table 2.7: Summary of genes reported more than one uncommon pathogenic or likely pathogenic variants only in NAFL patients. The table reports gene name; variant ID; Maximum Allele Frequency (MAF) in the general population, and a gene description from the Reference Sequence (RefSeq) database.

2.3.5 Functional enrichment analysis

Functional enrichment analysis has been performed to investigate the different pathways in which the genes, resulting from being mutated with at least one uncommon pathogenic or likely pathogenic variant, were involved and to explore if the genes were involved in relevant molecular networks related to the progression of NAFLD. Genes exclusively mutated in NASH (262), in NAFL (102) patients, and in both groups (96) were separately analyzed. The three lists of genes were investigated through ShinyGO v0.77 online tool using The Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology Cellular component (GO-CC), and Disease Jensen databases.

The 262 genes mutated only in NASH patients showed 27 pathways statistically significantly enriched ($FDR \leq 0.05$) in KEGG database (Table 2.8). Among them, several key pathways such as metabolic pathways (50 genes; $FDR=8.86 \times 10^{-10}$), lysosome (10 genes;

FDR= 1.32×10^{-4}), bile secretion (8 genes; FDR= 2.94×10^{-4}), and cholesterol metabolism (7 genes; FDR= 1.32×10^{-4}) were selected as the most relevant.

In the Gene Ontology Cellular component database, 64 pathways (FDR < 0.05 and with at least 3 genes involved) resulted in being statistically significantly enriched with the selected genes. Importantly, as reported in Table 2.9 and Table 2.10, 6 of them were correlated with the Cilium component (Cilium with 28 genes and FDR= 2.35×10^{-6} ; Ciliary plasm with 10 genes and FDR= 2.27×10^{-4} ; non-motile cilium with 9 genes and FDR= 2.21×10^{-3} ; Ciliary basal body with 8 genes and FDR= 7.93×10^{-3} ; 9+0 non-motile cilium with 7 genes and FDR= 9.31×10^{-3} and Ciliary membrane with 5 genes and FDR= 1.58×10^{-2}) and 39 genes are reported to be involved at least in more than one pathway. In the Jensen disease database, 17 pathways were enriched with at least 7 respective target genes with statistically significant relevance (FDR ≤ 0.05). All data are reported in Table 2.11, indicating that the Liver cancer pathway was the top one with 26 target genes involved and FDR= 2.47×10^{-7} .

Pathway Enrichment Analysis in KEGG DATABASE	Number of Genes identified	Pathway Genes	Enrichment FDR
Metabolic pathways	50	1527	8.86E-10
Lysosome	10	128	1.32E-04
Biosynthesis of cofactors	9	155	1.37E-03
Drug metabolism	8	79	1.58E-04
Bile secretion	8	89	2.94E-04
Focal adhesion	8	200	2.12E-02
Cholesterol metabolism	7	50	1.32E-04
Cardiac muscle contraction	7	87	1.36E-03
Hypertrophic cardiomyopathy	7	90	1.37E-03
Dilated cardiomyopathy	7	96	1.90E-03
Adrenergic signaling in cardiomyocytes	7	150	2.12E-02
Porphyrin and chlorophyll metabolism	6	43	2.94E-04
ABC transporters	6	45	3.31E-04
Complement and coagulation cascades	6	84	6.20E-03
Carbon metabolism	6	115	2.12E-02
Alanine, aspartate and glutamate metabolism	5	37	1.36E-03
Arrhythmogenic right ventricular cardiomyopathy	5	77	2.12E-02
Peroxisome	5	82	2.32E-02
Pentose and glucuronate interconversions	4	34	8.76E-03
Fanconi anemia pathway	4	54	2.93E-02
Steroid hormone biosynthesis	4	61	4.05E-02
Steroid biosynthesis	3	20	2.07E-02
Histidine metabolism	3	22	2.12E-02
Mismatch repair	3	23	2.24E-02
Folate biosynthesis	3	26	2.93E-02
Ascorbate and aldarate metabolism	3	30	3.92E-02
Glyoxylate and dicarboxylate metabolism	3	30	3.92E-02

Table 2.8: Summary of the relevant pathways, obtained from the KEGG database, in which the 262 genes with uncommon P-LP variants only in NASH patients are involved and the respective enrichment rate. All pathways are enriched with respective target genes with statistically significant relevance ($FDR \leq 0.05$).

Pathway Enrichment Analysis in GO_Cellular Component	Number of Genes identified	Pathway Genes	Enrichment FDR
Intrinsic component of plasma membrane	39	1965	4.64E-03
Mitochondrion	38	1816	2.24E-03
Integral component of plasma membrane	38	1881	4.24E-03
Plasma membrane region	34	1323	2.28E-04
Supramolecular complex	34	1455	8.25E-04
Supramolecular fiber	30	1102	2.28E-04
Supramolecular polymer	30	1110	2.38E-04
Membrane protein complex	30	1415	7.28E-03
Microtubule cytoskeleton	29	1391	1.01E-02
Cilium	28	696	2.35E-06
Organelle envelope	27	1352	1.88E-02
Envelope	27	1352	1.88E-02
Neuron projection	27	1444	3.60E-02
Microtubule organizing center	21	868	9.31E-03
Mitochondrial matrix	20	512	2.27E-04
Plasma membrane protein complex	19	712	6.09E-03
Mitochondrial membrane	19	822	1.82E-02
Mitochondrial envelope	19	872	2.86E-02
Apical part of cell	18	453	2.27E-04
Apical plasma membrane	17	385	2.27E-04
Transporter complex	17	442	3.22E-04
Cell projection membrane	16	368	2.27E-04
Transmembrane transporter complex	16	427	6.80E-04
External encapsulating structure	16	601	1.23E-02
Extracellular matrix	16	600	1.23E-02
Organelle inner membrane	16	627	1.58E-02
Centrosome	16	658	2.20E-02
Cell body	15	614	2.67E-02
Mitochondrial inner membrane	14	558	2.81E-02
Myofibril	13	244	2.27E-04
Contractile fiber	13	253	2.28E-04
Neuronal cell body	13	535	4.47E-02

Table 2.9: Summary of the relevant pathways, obtained from the GO-CC database, in which the 262 genes with uncommon P-LP variants only in NASH patients are involved and the respective enrichment rate. All pathways are enriched with respective at least 3 target genes, with statistically significant relevance ($FDR \leq 0.05$).

Pathway Enrichment Analysis in GO_Cellular Component	Number of Genes identified	Pathway Genes	Enrichment FDR
Sarcomere	12	221	2.48E-04
Plasma membrane bounded cell projection cytoplasm	12	230	3.28E-04
Cytoplasmic region	12	275	1.40E-03
Collagen-containing extracellular matrix	12	446	2.98E-02
Cluster of actin-based cell projections	11	171	2.27E-04
Ion channel complex	11	324	1.22E-02
Endoplasmic reticulum lumen	11	328	1.23E-02
Axoneme	10	134	2.27E-04
Ciliary plasm	10	136	2.27E-04
Cation channel complex	10	238	5.67E-03
Non-motile cilium	9	167	2.21E-03
Brush border	8	116	1.13E-03
I band	8	148	4.51E-03
Ciliary basal body	8	167	7.93E-03
Vacuolar lumen	8	205	1.82E-02
9+0 non-motile cilium	7	133	9.31E-03
Z disc	7	136	1.01E-02
Sarcolemma	7	143	1.23E-02
Primary lysosome	7	170	2.20E-02
Azurophil granule	7	170	2.20E-02
Brush border membrane	6	59	1.13E-03
Lysosomal lumen	6	114	1.58E-02
Myosin filament	5	24	2.28E-04
Myosin complex	5	59	7.27E-03
Ciliary membrane	5	78	1.58E-02
Golgi-associated vesicle	5	98	3.20E-02
Muscle myosin complex	4	16	6.80E-04
Myosin II complex	4	22	1.98E-03
DNA repair complex	4	40	1.18E-02
Golgi-associated vesicle membrane	4	55	2.37E-02
Mismatch repair complex	3	11	3.96E-03
ATPase dependent transmembrane transport complex	3	26	2.20E-02

Table 2.10: Summary of the relevant pathways, obtained from the GO-CC database, in which the 262 genes with uncommon P-LP variants only in NASH patients are involved and the respective enrichment rate. All pathways are enriched with respective at least 3 target genes, with statistically significant relevance ($FDR \leq 0.05$).

Pathway Enrichment Analysis in Jensen Disease	Number of Genes identified	Pathway Genes	Enrichment FDR
Liver cancer	26	568	2.47E-07
Skin cancer	25	439	1.57E-08
Melanoma	24	487	2.47E-07
Breast cancer	20	413	5.00E-06
Endometrial cancer	13	284	6.57E-04
Hypertension	12	283	2.17E-03
Coronary artery disease	10	200	2.17E-03
Fundus dystrophy	9	74	1.28E-05
Retinitis pigmentosa	9	78	1.53E-05
Large intestine cancer	9	115	2.81E-04
Distal arthrogryposis	8	44	3.82E-06
DOID:2627	8	85	2.80E-04
Polydactyly	8	89	3.04E-04
Blindness	7	78	8.01E-04
Immune system cancer	7	128	9.46E-03
Pancreatic cancer	7	130	1.01E-02
Lymphoid leukemia	7	144	1.54E-02

Table 2.11: Summary of the relevant pathways, obtained from the Jensen Disease database, in which the 262 genes with uncommon P-LP variants only in NASH patients are involved and the respective enrichment rate. All pathways are enriched with respective at least 7 target genes, with statistically significant relevance ($FDR \leq 0.05$).

Using the same approach, the 102 genes that presented uncommon pathogenic or likely pathogenic mutations only in NAFL patients and the 96 genes with the same type of mutations in both groups of patients were respectively analyzed. Tables 2.12 and 2.13 summarise the results obtained among the three different databases in both cases. Regarding the KEGG database, the metabolic pathway emerged to be relevant both in the enrichment analysis with the NAFL genes (18 genes and $FDR = 7.91 \times 10^{-3}$) and in common genes (19 genes and $FDR = 1.7 \times 10^{-3}$). In the last analysis, the lysosome (5 genes and $FDR = 8.99 \times 10^{-3}$), insulin resistance (4 genes and $FDR = 1.49 \times 10^{-2}$), and cholesterol metabolism (3 genes and $FDR = 1.49 \times 10^{-2}$) pathways were selected. In Gene Ontology Cellular Component and Jensen Diseases databases, no significant enrichments of interest were observed when genes mutated only in NAFL patients were used. On the other hand, when the 96 genes found to have at least one uncommon P-LP mutation in both groups were used, some pathways correlated to the lipoproteins, such as the plasma lipoprotein particle (3 genes and $FDR = 2.03 \times 10^{-2}$) and the protein-lipid complex (3 genes and $FDR = 2.23 \times 10^{-2}$), and to other metabolic disorders, including liver cancer (8 genes and $FDR = 2.25 \times 10^{-2}$), were selected among the key pathways involved.

Pathway enrichment analysis in KEGG database	Pathway	nGenes	Pathway Genes	Enrichment FDR
	Homologous recombination	4	41	4.77E-03
	Metabolic pathways	18	1527	7.91E-03
	Vitamin digestion and absorption	3	24	7.91E-03
	Peroxisome	4	82	1.79E-02
	TGF-beta signaling pathway	4	93	2.31E-02
Pathway enrichment analysis in GO_CC database	Pathway	nGenes	Pathway Genes	Enrichment FDR
	Neuronal cell body	11	535	9.40E-03
	Cell body	11	614	1.63E-02
Pathway enrichment analysis in Jensen Disease database	Pathway	nGenes	Pathway Genes	Enrichment FDR
	Intellectual disability	7	287	1.97E-02
	Breast cancer	7	413	3.66E-02
	Fundus dystrophy	4	74	1.97E-02
	Polydactyly	4	89	2.48E-02
	Diarrhea	4	147	4.96E-02
	Vitamin B12 deficiency	3	17	1.65E-02
	Leber congenital amaurosis	3	23	1.97E-02
	Methylmalonic acidemia	3	29	1.97E-02
	Cone dystrophy	3	52	3.43E-02
	Ovarian cancer	3	51	3.43E-02
	Congenital heart disease	3	61	3.66E-02

Table 2.12: Summary of the relevant pathways obtained from the KEGG, GOCC, and Jensen Disease databases, in which the 102 genes with uncommon P-LP variants only in NAFL patients are involved and the respective enrichment rate. All pathways are enriched with respective at least 3 target genes, with statistically significant relevance ($FDR \leq 0.05$).

Pathway Enrichment analysis in KEGG database		Number of genes Identified	Pathway Genes	Enrichment FDR
	Metabolic pathways	19	1527	1.79E-03
	PI3K-Akt signalling pathway	7	354	1.34E-02
	Lysosome	5	128	8.99E-03
	ECM-receptor interaction	4	88	1.15E-02
	Glucagon signaling pathway	4	107	1.49E-02
	Insulin resistance	4	108	1.49E-02
	AMPK signaling pathway	4	120	1.83E-02
	Insulin signaling pathway	4	137	2.54E-02
	Glycosaminoglycan degradation	3	19	4.45E-03
	Citrate cycle (TCA cycle)	3	30	9.03E-03
	Starch and sucrose metabolism	3	35	1.15E-02
	Valine, leucine and isoleucine degradation	3	48	1.49E-02
	Cholesterol metabolism	3	50	1.49E-02
	PPAR signalling pathway	3	75	3.50E-02
Pathway Enrichment analysis in GO_CC database		Number of genes Identified	Pathway Genes	Enrichment FDR
	Integral component of plasma membrane	17	1881	3.02E-02
	Intrinsic component of plasma membrane	17	1965	4.12E-02
	Mitochondrion	16	1816	4.44E-02
	Plasma membrane region	14	1323	2.23E-02
	Vacuole	11	973	3.64E-02
	External encapsulating structure	10	601	1.03E-02
	Extracellular matrix	10	600	1.03E-02
	Receptor complex	9	422	8.49E-03
	Mitochondrial matrix	9	512	1.26E-02
	Endoplasmic reticulum lumen	8	328	8.49E-03
	Transmembrane transporter complex	8	427	1.60E-02
	Collagen-containing extracellular matrix	8	446	1.71E-02
	Transporter complex	8	442	1.71E-02
	Apical part of cell	8	453	1.72E-02
	Apical plasma membrane	7	385	2.23E-02
	Cluster of actin-based cell projections	6	171	8.49E-03
	Brush border membrane	4	59	8.58E-03
	Brush border	4	116	2.66E-02
	Plasma lipoprotein particle	3	44	2.03E-02
	Lipoprotein particle	3	44	2.03E-02
	Protein-lipid complex	3	47	2.23E-02
Pathway Enrichment analysis in Jensen Disease database		Number of genes Identified	Pathway Genes	Enrichment FDR
	Melanoma	14	487	3.77E-06
	Skin cancer	12	439	3.87E-05
	Breast cancer	10	413	3.43E-04
	Liver cancer	8	568	2.25E-02
	Cardiovascular system disease	7	233	1.93E-03
	Lung cancer	7	381	1.31E-02
	Hypoglycemia	6	73	5.72E-05
	Coronary artery disease	6	200	4.64E-03
	Type 2 diabetes mellitus	5	252	3.10E-02
	Hyperuricemia	4	30	3.43E-04
	Nonsyndromic deafness	4	82	7.97E-03
	Sensorineural hearing loss	4	101	1.29E-02

Table 2.13: Summary of the relevant pathways obtained from the KEGG, GOCC, and Jensen Disease databases, in which the 96 genes with uncommon P-LP variants in both groups of patients are involved and the respective enrichment rate. All pathways are enriched with respective at least 3 target genes, with statistically significant relevance ($FDR \leq 0.05$).

2.3.6 Functional enrichment analysis in the validation cohort

To confirm the result from the functional enrichment analysis, the same approach was used on a different cohort of patients. Data of 79 individuals, 14 (17.2%) NAFL and 65 NASH (82.8%) patients were received from the Division of Digestive and Liver Diseases, Department of Medicine of the Columbia University Irving Medical Center, New York, NY 10032, USA. A total of 68815 variants were identified in the entire cohort, of which 32720 were uncommon variants with $MAF < 1\%$. As reported in Table 2.14, in NAFL patients (16 individuals), approximately 7876 (11.5%) uncommon variants were identified that considering the ACMG guidelines included 6 (0.1%) likely pathogenic variants; 4340 (55.1%) VUS and 3530 (44.8%) of variants with other annotations. In NASH patients (65 individuals), 28935 (42%) resulted to be uncommon variants, of which 1 (0.003%) reported as a pathogenic variant, 47 (0.2%) as likely pathogenic variants, 17674 (61.1%) as VUS and 11214 (38.8%) as other types according to ACMG criteria. Regarding the mutated genes, as shown in the same table, approximately 44 genes reported at least one uncommon ($MAF < 1\%$) pathogenic or likely pathogenic variants, of which 5 (11.4%) only in NAFL patients, 38 (86.4%) only in NASH patients and only 1 gene (2.2%) in both groups.

Considering the uncommon VUS, 10549 genes resulted to carry on at least one uncommon VUS mutation, and among them, 638 (6%) genes reported this type of mutation only in NAFL patients, 7318 (69.4%) only in NASH and 2593 (24.6%) in individuals affected by one of both conditions.

Regarding genes, 38 were identified as exclusively mutated in NASH patients (4 of them were also mutated only in the NASH patients of the discovery cohort), 5 genes only in NAFL, and one gene resulted to be mutated in both groups of patients (as in the discovery cohort).

Due to the small number of genes that are reported to be mutated only in NAFL patients or in both groups, only the 38 genes exclusively mutated in NASH patients were analyzed through ShinyGO v0.77 online tool using the same databases KEGG, Gene Ontology Cellular component (GO-CC) and Disease Jensen databases to confirm the functional enrichment analysis results obtained in the discovery cohort.

<i>Validation Cohort</i>	Cohort (79)	NAFL PATIENTS (16)	NASH PATIENTS (65)	
Number of variants	68815	30609	62569	
Variants MAF< 0.01	32720	7876	28935	
ACMG Classification of uncommon variants (MAF<1%)				
<i>Pathogenic</i>		0	1(0.003%)	
<i>Likely Pathogenic</i>		6 (0.1%)	47 (0.2%)	
<i>VUS</i>		4340 (55.1%)	17674 (61.1%)	
<i>Other</i>		3530 (44.8%)	11214 (38.8%)	
Gene with uncommon (MAF<1%) mutations				
	Total number of genes (%)	ONLY IN NAFL PATIENTS (%)	ONLY IN NASH PATIENTS (%)	COMMON (%)
P-LP mutations	44 (100%)	5 (11.4%)	38 (86.4%)	1 (2.2%)
Vus mutations	10549 (100%)	638 (6%)	7318 (69.4%)	2593 (24.6%)

Table 2.14: Number of variants discovered in exome sequencing data from 79 NAFLD patients included in the validation cohort from the Division of Digestive and Liver Diseases, Department of Medicine of the Columbia University Irving Medical Center, New York, NY 10032, USA.

As shown in Table 2.15, 10 pathways were statistically significantly enriched with the target genes ($FDR \leq 0.05$) in KEGG database. Among them, the metabolic pathways (11 genes; $FDR=8.42 \times 10^{-4}$) and the lysosome one (3 genes; $FDR=2.55 \times 10^{-2}$) were selected as top enriched pathways, confirming the results obtained in the cohort of the study.

Interestingly, most of the rest of the significant enriched pathways were involved in mechanisms correlated with fatty acid metabolism.

In the Gene Ontology Cellular component database, 3 pathways ($FDR<0.05$) resulted to be statistically significantly enriched with the selected genes, and the Lysosomal Lumen pathway (3 genes; $FDR=4.33 \times 10^{-2}$) was selected as one of the tops involved as in the discovery cohort, confirming the crucial role of the lysosome pathway in the progression of the disease.

Considering the Disease Jensen database, 15 pathways showed an $FDR<0.05$, although 4 of them were present in the analysis of the two cohorts; none of them are significantly related to the pathology under study but are related to more generic conditions.

Furthermore, although some pathways that emerged from this analysis were not also observed in the discovery cohort, they could still be interestingly linked to complications from NAFL/NASH.

	Pathway	Number of Genes identified	Pathway Genes	Enrichment FDR
Pathway Enrichment Analysis in KEGG DATABASE	Metabolic pathways	11	1527	8.42E-04
	Glycosaminoglycan degradation	2	19	1.48E-02
	Lysosome	3	128	2.55E-02
	Fatty acid degradation	2	43	2.67E-02
	Cysteine and methionine metabolism	2	50	2.67E-02
	Glutathione metabolism	2	57	2.67E-02
	Amino sugar and nucleotide sugar metabolism	2	49	2.67E-02
	Fatty acid metabolism	2	57	2.67E-02
	Ovarian steroidogenesis	2	51	2.67E-02
	MicroRNAs in cancer	3	161	2.67E-02
Pathway Enrichment Analysis in GO_Cellular Component DATABASE	Vacuolar lumen	4	205	4.12E-02
	Centriolar satellite	3	121	4.33E-02
	Lysosomal lumen	3	114	4.33E-02
Pathway Enrichment Analysis in Jensen Disease	Carnitine palmitoyltransferase II deficiency	2	12	5.69E-03
	Glucosephosphate dehydrogenase deficiency	2	14	5.69E-03
	Lichen planus	2	14	5.69E-03
	Systemic primary carnitine deficiency disease	2	13	5.69E-03
	Tyrosinemia	2	11	5.69E-03
	Multiple acyl-CoA dehydrogenase deficiency	2	23	1.31E-02
	Immune system cancer	3	128	1.52E-02
	Congenital hemolytic anemia	2	29	1.52E-02
	Methylmalonic acidemia	2	29	1.52E-02
	Lymphoid leukemia	3	144	1.92E-02
	Congenital disorder of glycosylation	2	40	2.16E-02
	Melanoma	4	487	3.95E-02
	Blindness	2	78	3.95E-02
	Cardiomyopathy	2	80	3.95E-02
	Hypoglycemia	2	73	3.95E-02

Table 2.15: Summary of the relevant pathways obtained from the KEGG, GOCC, and Jensen Disease databases, in which the 38 genes with uncommon P-LP variants in NASH patients belonging to the validation cohort are involved and the respective enrichment rate. All pathways are enriched with statistically significant relevance ($FDR \leq 0.05$).

2.4 Discussion

The aim of this study is to identify new possible genes or pathways that can have a key role in the different stages of fatty liver disease from NAFL to NASH.

DNA extracted from blood samples of 154 very well histologically characterized adults were analyzed using the whole exome sequencing method. Due to the large number of data obtained, in order to prioritize information, the analysis was focused on uncommon mutations because it showed a MAF minor than 1%. Among them, a total of 39,678 variants were selected after a very stringent filtering process.

The ACMG criteria were used to classify variants, and as expected, a different distribution between NAFL and NASH patients was immediately noticed. Indeed, despite the unbalancing of the two groups (108 NASH vs. 46 NAFL individuals), it was possible to show that the pathogenic or likely pathogenic variants identified in the NASH group were more frequent in percentage than those identified in the NAFL group (0.75% and 0.58% respectively). The same trend was observed concerning variants known to be with an uncertain significance (VUS), 65.81% in NASH patients and 60.52% in NAFL ones. These data are in line with the hypothesis that patients with an aggressive form of the disease can have an increased genetic susceptibility due to a higher prevalence of damaging mutations. To verify if damaging variants were enriched in specific genes and pathways with a crucial role in the progression of the disease, the analysis was targeted at the genes carrying uncommon pathogenic or likely pathogenic variants. With the same aim, genes were divided into those exclusively mutated in NASH or NAFL patients and analyzed separately.

As a first approach, all genes that reported more than one pathogenic or likely pathogenic uncommon variants were individually analyzed. Among the genes emerged to be exclusively mutated only in NASH patients, 37 genes reported more than one damaging mutation (the entire list is reported in Table 2.5 and 2.6). Several genes are of particular interest, such as the *FLNB* gene, which provides instructions for making the protein called filamin B. This protein helps to build the network of protein filaments (cytoskeleton) that provides structure to cells and allows them to change shape and to move [71]. The protein expression is highly abundant in endothelial cells. It is known that it is part of the process to repair vascular injuries. In the cohort of study, 4 pathogenic or likely pathogenic variants (rs138220431; rs151259375; rs138034708; rs143831841) were identified on this gene in about 8 NASH patients. However, its link with the pathology remains still unclear. From research carried out in the literature, this gene does not appear to be related to NAFLD. This gene resulted to be involved in different autosomal dominant conditions as Atelosteogenesis type I, Atelosteogenesis type III, Boomerang dysplasia, Larsen syndrome and Spondylocarpotarsal synostosis syndrome. Mutations recognize to be involved in these conditions, such as rs80356519, rs121908897, resulted to have a stop gained functional consequence. Others, such as the rs121908896, involves residue highly evolutionarily conserved among vertebrate filamins. Compared to them, those identified in NASH patients are all heterozygous missense variants localized in different domains.

CEL and *CYP24A1* are the other two interesting genes highlighted in this analysis. In the first one, the rs756449511 and rs763199261 were identified respectively in three and two NASH patients.

Carboxyl-ester lipase (*CEL*) is secreted into the pancreatic juice and helps to the hydrolysis of dietary lipids in the duodenum; indeed, the physiological role of this protein is in cholesterol and lipid-soluble vitamin ester hydrolysis and absorption. The *CEL* gene

harbours a variable number of tandem repeats (VNTR) region in exon 11 that emerged can be associated with alcoholic liver cirrhosis (ALC) [72]. This data can be important because, despite the two forms of liver pathology being recognized as two distinct conditions, both diseases are shown to have an overlap in the pathophysiology and share genetic and epigenetic factors. Indeed, frequently, the two forms were observed to coexist [73].

Regarding the *CYP24A1* gene, encoding the cytochrome P450 family 24 subfamily A member 1, or the enzyme 24-hydroxylase, which catabolizes 25(OH)D and 1,25(OH)2D into inactive carboxylic acids [74], it is important to note that in a recent study focused on the association of vitamin-D-related genetic variations with NAFLD, it was selected as a gene associated with presence, severity, or response to treatment of NAFLD. More in detail, the rs3787557, rs6068816, rs2296241, rs2248359 resulted to be associated with the presence of the disease; while the rs3787557, rs6068816, rs6097809, rs6127119, rs2248359, rs3787554, rs4809960, rs6022999 resulted to be related with its severity [75]. In the present study, two novel variants not yet described to be associated with NAFLD, rs114368325, and rs35873579, were identified respectively in two and one NASH individuals.

Also, the *C9* (Complement 9) gene, part of the membrane attack complex (MAC), which plays a crucial role in the both innate and adaptive immune response by building pores in the plasma membrane of target cells, appeared to be interesting. In fact, here, rs34000044 and rs771605119 were found, but this gene was already observed with a reduced expression in NASH patients without fibrosis in recent work on hepatic gene-expression patterns associated with different liver injuries [76].

Among the 37 genes individuated, there is also the *COL6A3* (CollagenVI-alpha3 chain) gene reported rs769072338, rs369169235 and rs150907698. This gene encodes the alpha-3 chain, one of the three alpha chains of type VI collagen, a beaded filament collagen present in most connective tissues. It is already known to be a biomarker for both hepatic fibrosis and poor prognosis of hepatocellular carcinoma (HCC) [77].

In addition, among them, many genes, including *GNPTAB*, *BBS4*, *CACNA1S*, *ERBB3*, *MMP21*, *MYL2*, *PMS2*, *SBDS*, and *SC5D* for which evidence of their implication in the NAFLD pathophysiology are not available. However, they demonstrated associations with several phenotypes related to liver damage, such as elevated hepatic transaminase, hepatomegaly, hepatocellular carcinoma, intrahepatic cholestasis, or impairment of bile flow due to obstruction in the small bile ducts within the liver [78].

Moreover, some of them, including *PMS2*, which provides instructions for making a protein that has an essential role in repairing DNA, *BBS4*, a member of the Bardet-Biedl syndrome (BBS) gene family, although in the literature they are not directly related to the pathogenesis of NAFLD, they are however clearly involved in several mechanisms known to be involved in the progression of NAFLD.

On the other hand, considering genes exclusively mutated in NAFL patients, with more than one uncommon pathogenic or likely pathogenic variants (list reported in Table 2.7), only the *APOB* gene, which reported three different variants, and the *LRRK2* gene with the rs33995463 observed in more than one NAFL patient, emerged as of peculiar importance. *APOB* (Apolipoprotein B) gene product is the main apolipoprotein of chylomicrons and low-density lipoproteins (LDL), and it is the ligand for the LDL receptor. Mutations in this gene will cause hypobetalipoproteinaemia and hypercholesterolemia due to ligand-defective apoB, diseases affecting plasma cholesterol and apoB levels. It was already demonstrated that mutations in the *APOB* gene are quite common among NAFLD patients with very low LDL-Cholesterol and may be associated with increased amino-

transferase levels and steatosis severity [79]. Regarding *LRRK2*, a gene member of the leucine-rich repeat kinase family that encodes a protein present largely in the cytoplasm but also correlates with the mitochondrial outer membrane, it is known its association with Parkinson’s disease. However, studies on mice models were performed to investigate the role of *LRRK2* in lipid metabolism, and results suggested that this gene plays a role in the regulation of β -oxidation. Moreover, they demonstrated that the downregulation of *LRRK2* in the livers of NAFLD mice may result in the suppression of β -oxidation, which may result in inflammation [80].

For all these genes, in this study we found heterozygous mutations with missense or frameshift functional consequence not yet described in other disorders involving the same genes. In addition to this, compared to the mutations already found in these genes and that are correlated with disorders with an autosomal dominant Inheritance, they were located in different domains. This could explain the lack of overlap of the NAFL or NASH phenotype and the other conditions caused by mutations identified in the same genes.

As the second step, a functional enrichment analysis was performed to investigate the different pathways in which the genes that reported at least one uncommon pathogenic or likely pathogenic variants were involved. Genes exclusively mutated in NASH (262), in NAFL (102) patients, and in both groups (96) were separately investigated using three different databases (KEGG, Gene Ontology Cellular component (GO-CC), and Disease Jensen databases) and the significant pathways were selected according to the FDR value ($FDR < 0.05$). Differences between the two groups of patients immediately emerged from this analysis. In fact, many of the pathways that emerged from the enrichment analysis obtained using the mutated genes only in NASH patients, in addition to being more numerous, probably due to the higher number of genes identified, are clearly correlated to aspects involved in the pathology. On the other hand, using the mutated genes exclusively in NAFL individuals, few and mostly generic pathways were highlighted.

According to the KEGG database, the metabolic pathway was observed as one of the most important using all three gene lists.

Furthermore, the number of genes resulted to be involved in this pathway increases considering the three different groups. Indeed, a total of 18 genes involved in this pathway were identified considering genes mutated only in NAFL patients, 19 genes when analyzed, were mutated in both groups and 50 genes using only genes reported damaging uncommon mutations exclusively in NASH individuals.

This is consistent with the fact that NAFLD is linked with an imbalanced metabolic network that involves lipids, glucose, amino acids, bile acids and iron. Indeed, it is well known that lipid metabolism is the core metabolic process and directly leads to lipid deposits. However, glucose is associated with lipids by insulin, which is abundant but inadequately processed in NAFLD.

Additionally, amino acids simultaneously regulate lipid and glucose metabolism. Bile acid is associated with lipids and glucose through FXRs and choline. Furthermore, iron metabolism is a unique pathway and iron accumulation induces ROS to participate in inflammation and oxidative stress, which also increases lipid accumulation by adding fatty acids and cholesterol [81].

Another important pathway that emerged in the KEGG database is the lysosome. It was selected as a key pathway both using genes exclusively mutated in NASH and those mutated in both groups, but not when the genes mutated only in NAFL were used.

As a degradation organelle, the lysosomes receive, digest, and distribute lipids to provide

cells with building blocks and energy. As a signaling hub, the lysosome senses lipid signals and regulates gene expression to control lipid biogenesis and turnover. A lot of evidence is already present in literature that correlates this organelle with the pathology.

Indeed, it was demonstrated that an overabundance of cellular nutrients in NAFLD leads to a hyperactivation of *mTOR* that increases lipogenesis, inhibits autophagy, and blocks lysosome biogenesis. All these mechanisms can promote hepatic steatosis and lipotoxicity. Restoring normal lysosomal function through, for example, the inhibition of *mTOR* was proposed as one possible strategy in the NAFLD treatment [82].

It was also shown that a reduced activity of lysosomal acid lipase (LAL), a hydrolase that has a key role in intracellular cholesterol trafficking, promotes increased lysosomal cholesterol ester storage. Importantly, it was found that its activity progressively reduces with liver disease severity by 10% in patients with NAFL, by 20% in NASH and by 50% in cryptogenic cirrhosis (CC) [83].

Regarding more specifically NASH, in the last few years, several studies have highlighted the role that lysosome and correlated components can play in the NAFLD progression. Autophagy, for example, was recently proposed as another cellular process that enhances NASH progression. This process involves the delivery and degradation of cytoplasmic materials by lysosomes and plays a prominent role in cell survival, differentiation, development, and homeostasis by controlling cytoplasmic physiology through energy balance and the removal of damaged organelles, misfolded proteins and lipid droplets [84].

In another recent work aimed to investigate the association between NASH and lysosome components, it was demonstrated that the expression of the lysosomal-associated protein transmembrane 5 (*LAPTM5*) was significantly down-regulated in the livers of both human NASH subjects and mouse NASH models. They showed that depletion of *LAPTM5* in hepatocytes significantly exacerbated hepatic steatosis, inflammation, and fibrosis in high-fat and high-cholesterol (HFHC) diet-induced mouse NASH models, whereas *LAPTM5* over-expression in hepatocytes substantially delayed and mitigated the foregoing pathological changes [85].

Together, the data are in line with what was observed in the study cohort, suggesting the lysosome pathways as significantly involved in the progression of NASH disease. In addition, this evidence was also verified in the validation cohort composed of 79 individuals, 14 NAFL and 65 NASH patients, received from the Division of Digestive and Liver Diseases, Department of Medicine of Columbia University. Indeed, using the 38 genes in which at least one uncommon pathogenic or likely pathogenic variants only in NASH was found, the lysosome pathway was selected among the most significantly enriched. Furthermore, this pathway emerged with a key role also in the enrichment analysis performed on the other database (Gene Ontology- Cellular Component).

Considering the results obtained from the enrichment analysis conducted on the Gene Ontology- Cellular Component database, while using the exclusively mutated genes in the NAFL patients, nothing relevant was found on the contrary using the exclusively mutated genes in the NASH subjects, among the several enriched pathways, the cilium component emerged as extremely relevant. Indeed, six cilium-related pathways were selected among the most significantly enriched. This correlation can be of particular interest from different points of view. It can involve cholangiocytes, epithelial cells that line the biliary tree. Their major physiological function lies in modifying the hepatic canalicular bile secreted by hepatocytes via absorptive and secretory processes, while being transported along the biliary tree through transmembrane molecules present on the apical or basolateral membranes

[86]. These cells present primary cilia extending from their apical plasma membranes into the ductal lumen, being ideally positioned in order to detect changes in bile flow, composition and osmolality [87]. The most important and well-studied function of cholangiocytes is the release of bicarbonate into bile as induced by secretin, in a mechanism dependent on the increment of the levels of intracellular cyclic AMP (cAMP) and activation of protein kinase A (PKA). Defects in the structure but also in the function of these organelles lead to cholangiociliopathies [88]. Recent works have begun to unravel the impact of cholangiocytes on NAFLD and NASH progression. For instance, NAFLD patients with cholestasis showed more advanced histological impairments, including cholangitis, advanced fibrosis, and cirrhosis when compared with age- and sex-matched NAFLD cohort [89]. Moreover, a study conducted on mice models focused on the secretin (SCT) and secretin receptor (SR), which are critically involved in functions of the cholangiocytes pathway and represent one of the most studied messengers/receptors in cholestatic liver injury, emerged that SR-/- mice show less liver steatosis with high-fat diet comparing with wild-type mice, suggesting that the SCT/SR axis may play a key role in NASH [90]. In addition, several studies revealed that NASH shares common aetiology and phenotype with cholangiopathies, including cholestasis, bridging fibrosis, ductular reaction, hepatic inflammation and biliary senescence [91]. Together, this data strongly supports results from the enrichment analysis in which the cilium component emerged as clearly involved in the more aggressive form of the disease, with 6 different pathways enriched and a total of 28 genes involved in these pathways that were found to report uncommon damaging mutations only among NASH patients. The cilium component did not emerge from the enrichment analysis conducted with the same approach on the validation cohort, probably due to the small size of the cohort. Considering the last database used for the enrichment analysis, the Jensen Disease Database, it was observed that the first pathway selected as the most significantly enriched was Liver cancer when the exclusively mutated in NASH genes were used. This pathway also emerged when genes mutated in both groups were used, but with a small number of genes involved, however not when genes exclusively mutated in NAFL individuals were analyzed. This is in line with the expectations because about 3-15% of patients with NASH progress to cirrhosis and about 4 -27% of NASH with cirrhosis patients transform to HCC [92]. It is known, too, that HCC can develop de novo in NASH patients in the absence of cirrhosis. Although the mechanism is not fully understood, the malignant transformation of hepatic adenomas into carcinomas was hypothesized as the underlying probable cause [93].

In summary, this work is one of the first in which the analysis of the entire exome, using the NGS approach, is used to investigate the genetic background of two very well-histologically characterized groups of different NAFLD stage patients.

Here, this analysis allowed to identify novel potential risk alleles associated with the different stages of the pathology. Due to the large number of identified variants, at least the most relevant could be further investigated by functional validation in molecular assays and animal models to confirm the role and to understand the mechanisms involved, which can then be the targets of drug development.

In addition, starting from the genetic landscape, several pathways already known to give a contribution to the NAFLD progression emerged as mostly correlated with NASH stage, suggesting the importance of the genetic contribution to the progression of the disease. In summary, the robustness of these results is supported by the fact that some pathways were confirmed in a second cohort genetically very different from the discovery one, representing

important information useful for patients' stratification and potentially also to perform targeted interventions on patients at high risk of progression due to a higher genetic predisposition.

The limited number of individuals recruited in both the discovery and validation cohorts, the imbalance between the two groups taken into analysis and the lack of a control cohort composed of healthy subjects represent limitations for this study. For this reason, replicating the data in larger and heterogeneous patient cohorts is a necessary step together with the development of functional studies.

Chapter 3

Mirnome profile of circulating extracellular vesicles (EVs) across the histological spectrum of non-alcoholic fatty liver disease (NAFLD)

3.1 Introduction

MicroRNAs (miRNAs) are short non-coding RNAs (19 to 24 nucleotides long) responsible for post-transcriptional regulation of gene expression [94]. Individual miRNAs have been observed to regulate the expression of several genes, while multiple miRNAs can regulate the same gene. Furthermore, miRNAs are released in extracellular space, for example, in vesicles, and are measurable in various biological fluids, such as plasma, serum and urine. MiRNAs have cell-specific expression profiles and play crucial roles in maintaining metabolic processes, including lipid and glucose metabolism [95].

In addition, miRNAs have great potential as non-invasive biomarkers for diagnosis, prognosis and therapeutic targets of liver diseases [96].

Most circulating miRNAs are released in the bloodstream packaged in extracellular vesicles (EVs), such as exosomes and microvesicles, small lipid bilayer membranes (40-150 nm) released by all cell types and possess a relevant role in cell-to-cell communication, particularly during pathological conditions. EVs are membranous, differ in size and composition, are very heterogeneous, and are able to transfer biologically active molecules and genetic material to target cells. The active secretion of EVs-miRNAs can contribute to intercellular communication, promoting inflammation, angiogenesis, and fibrosis [97].

As a matter of fact, miRNAs have been recognized as important epigenetic factors involved in the pathogenesis of NAFLD and liver fibrosis [98].

Despite the considerable amount of available literature, results are still conflicting, and most studies failed to be replicated; the different characteristics of populations studied, the various experimental conditions, and the diverse methods adopted [99], are all factors that contributed to the lack of consensus, hampering the use of miRNAs in clinical practice.

In addition, the most commonly used technique to measure circulating miRNAs is qPCR,

with its inherent limitations due to the lack of unequivocally accepted normalization strategies [100].

Next-generation sequencing (NGS) permits the analysis of both microRNA expression and sequence, allowing for elucidation of a comprehensive landscape of miRNAs in a given tissue and sample type [101].

Although EVs-miRNA has advantages as biomarkers, only a few studies have profiled miRNA expression in EVs [102]. EV-derived miRNAs present higher specificity and stability than circulating miRNAs, they are protected from RNases degradation and they reflect the pathological state of the cells and tissues of origin. Thus, these characteristics make miRNAs carried by EVs ideal biomarkers for disease diagnosis [103].

In the present study, the aim was to use the NGS approach to perform a comprehensive miRNome profile of circulating EVs in a cohort of NAFLD patients, aiming at identifying specific miRNAs associated with the presence of NASH and significant ($F \geq 2$) fibrosis.

3.2 Methods

3.2.1 Cohort of study

A total of 228 patients with biopsy-proven NAFLD were included in the present study. All participants belonging to the discovery cohort described in Chapter 2. This study includes the 154 individuals for which the exome sequencing analysis was performed in addition to 74 individuals, part of the same cohort, but not the DNA was available.

3.2.2 Isolation of serum extracellular vesicles

EVs were isolated from 200 μ l of serum using ExoQuick precipitation solution (System Biosciences, USA) according to the manufacturer's instructions. Briefly, 200 μ l of serum was processed with 50.4 μ l ExoQuick solution and stored at 4°C overnight. The EVs pellets were resuspended with 200 μ l of nuclease-free water and RNA was immediately isolated from the solution. Total RNA was extracted with miRNeasy serum/plasma kit (Qiagen, Germany) using the QIAcube extractor (Qiagen, Germany) according to the manufacturer's instructions. RNA concentration was determined for all samples with Qubit 2.0 Fluorometer with miRNA assay kit (ThermoFisher, USA).

3.2.3 Nanoparticle Tracking Analysis

EV concentration was measured by Nanosight NS300 (Malvern Instruments Ltd., Malvern, UK) equipped with a 488 nm laser module that utilizes Brownian motion and refraction index. Samples were diluted 1:200 in physiologic solution filtered with 100nm pore size. For each sample, 3 videos of 30 s at camera level 15 and threshold 5 were captured. All samples were characterized with NTA 3.2 Analytical software.

3.2.4 Super-resolution microscopy

Super-resolution microscopy acquisition of EVs isolated from serum was obtained using Nanoimager S Mark II microscope from ONI (Oxford Nanoimaging, Oxford, UK) equipped with a 100x, 1.4NA oil immersion objective, an XYZ closed-loop piezo 736 stage, and triple

emission channels split at 640, 488 and 555 nm. EV profiler Kit (ONI) was used to perform the experiments following the manufacturer’s protocol. The Kit contains fluorescent antibodies anti CD9-488, CD63-568, and CD81-647 and the imaging buffer. Images were taken in dSTORM mode acquired sequentially in total reflection fluorescence (TIRF) mode. Single-molecule data was filtered using NimOS software (v.1.18.3, ONI). Data has been processed with the Collaborative Discovery (CODI) online analysis platform www.alto.codi.bio from ONI and the drift correction pipeline version 0.2.3 was used [104].

3.2.5 Library preparation and next-generation sequencing

Small noncoding RNA libraries were constructed starting from 6 ul of total RNA, using NEBNext Multiplex Small RNA Library Prep set for Illumina (New England Biolabs, Inc., USA). The cDNA libraries were purified with the Qiagen PCR Purification kit following the modifications indicated by the manufacturer’s protocol. Six pools of cDNA barcoded samples were prepared, and each pool was finally size-selected miRNAs in a 6% PolyAcrylamide Gel (ThermoFisher, USA). Fragments with an insert of 140 nucleotides corresponding to miRNAs were cut out and purified with Qiagen Gel Extraction MiniElute columns (Qiagen, Germany) following the modifications indicated in NEBNext Multiplex Small RNA Library Prep Protocol. A BioAnalyzer 2100 system (Agilent Technologies, USA) was used to check the size of each cDNA library fragment. Single-end sequencing (75 nt) was performed on the NextSeq550 platform (Illumina, USA).

3.2.6 Sequencing data analysis

The data from small RNA sequencing were computationally processed and analyzed. Briefly, adapter removal was performed using Cutadapt [105], which clipped the reads that passed the quality control from the adapter sequences. Reads longer than 14 nucleotides were mapped to a miRBase v.22 reference using BWA algorithm (v0.7.17) with the default settings, while reads shorter than 14 nucleotides were discarded. Only alignments without internal mismatches or indels were considered, and those with the highest quality were used to assign each read to a unique miRNA.

3.2.7 MiRNA targets enrichment analysis

The validated target genes of selected deregulated miRNAs were retrieved with Multi-MiR Bioconductor’s package from Mirecords, Mirtarbase, and Tarbase databases [106]. The analysis was performed separately for up-regulated and down-regulated miRNAs, and genes that were validated in at least two of the three databases were selected. Gene Ontology-Biological Process enrichment analysis was performed using ShinyGO v0.741, showing the top 50 significant terms (p-value $\leq$ 0.05 for FDR) [107]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed using PathfindR [108], setting a minimum of five genes in each term.

3.2.8 Computational and Statistical Analysis

MiRNAs were quantified for each sample and then merged into a single count matrix, setting missing miRNAs to zero. MiRNAs whose counts were less than 20 in more than half of the samples were discarded. Differential expression analysis was performed with the DESeq2 Bioconductor’s package (version 1.22.2) [109]. Statistical analyses were performed

using R software [110]. For all the analyses, a p-value <0.05 was considered statistically significant.

3.3 Results

3.3.1 Study cohort characteristics

A total of 228 patients with biopsy-proven NAFLD were included in the present study. The demographic, clinical, and biochemical characteristics are reported in Table 3.1. Considering the study cohort, a median age of 48.5 (37.5-58.0) years was observed. Eighty-two individuals were males, while 146 were females, representing the 64% of the entire cohort. Among the patients, 165 (72.5%) of them received a histological diagnosis of NASH according to the concomitant presence of liver steatosis $>5\%$, lobular inflammation, and hepatocyte ballooning. A significant liver fibrosis (defined as $F\geq 2$) was observed in 94 (41.2%) patients.

Variables	Total, n = 228
Age (years), median (IQR)	48.5 (37.5–58.0)
Gender (M/F)	82/146
BMI (kg/m²), median (IQR)	29.1 (25.9–32.9)
Waist (cm), median (IQR)	99 (98–108)
T2DM, n (%)	76 (33.3%)
Hypertension, n (%)	91 (39.9%)
NASH, n (%)	165 (72.4%)
Liver fibrosis, n (%)	
F0	65 (28.5%)
F1	69 (30.3%)
F2	29 (12.7%)
F3	33 (14.5%)
F4	32 (14.0%)
ALT (U/L), median (IQR)	45 (30–70)
AST (U/L), median (IQR)	30 (22–42)
γGT (U/L), median (IQR)	45 (26–92)
Platelets (x10⁹/L), median (IQR)	227 (190–270)
Albumin (g/dL), median (IQR)	4.4 (4.1–4.6)
Total bilirubin (mg/dL), median (IQR)	0.7 (0.5–0.9)
Fasting glucose (mg/dL), median (IQR)	94 (86–110)
Fasting insulin (mU/L), median (IQR)	13.9 (9.4–23.9)
Total cholesterol (mg/dL), median (IQR)	184 (161–211)
HDL cholesterol (mg/dL), median (IQR)	46 (39–55)
Triglycerides (mg/dL), median (IQR)	124 (90–171)

Table 3.1: Summary of the demographic, clinical, and biochemical characteristics of the 228 individuals included in the study cohort. Abbreviations: alanine aminotransferase (ALT), aspartate aminotransferase (AST), body mass index (BMI), gamma-glutamyl transferase (γGT), high-density lipoprotein (HDL), interquartile range (IQR), number (n), non-alcoholic steatohepatitis (NASH), type 2 diabetes mellitus (T2DM).

3.3.2 EVs isolation

The presence and integrity of EVs isolated from serum were confirmed by Super-resolution microscopy and by NanoSight analyses (Figure 3.1). Super-resolution microscopy was performed on 3 representative samples confirming EV size in the range of 100/200 nm and the expression of the classical exosomal markers, tetraspanins (CD63, CD9 and CD81). Results showed that EVs were single, double or triple positive for tetraspanins expression, with the majority of them double positive for CD63 and CD9, and single positive for CD9.

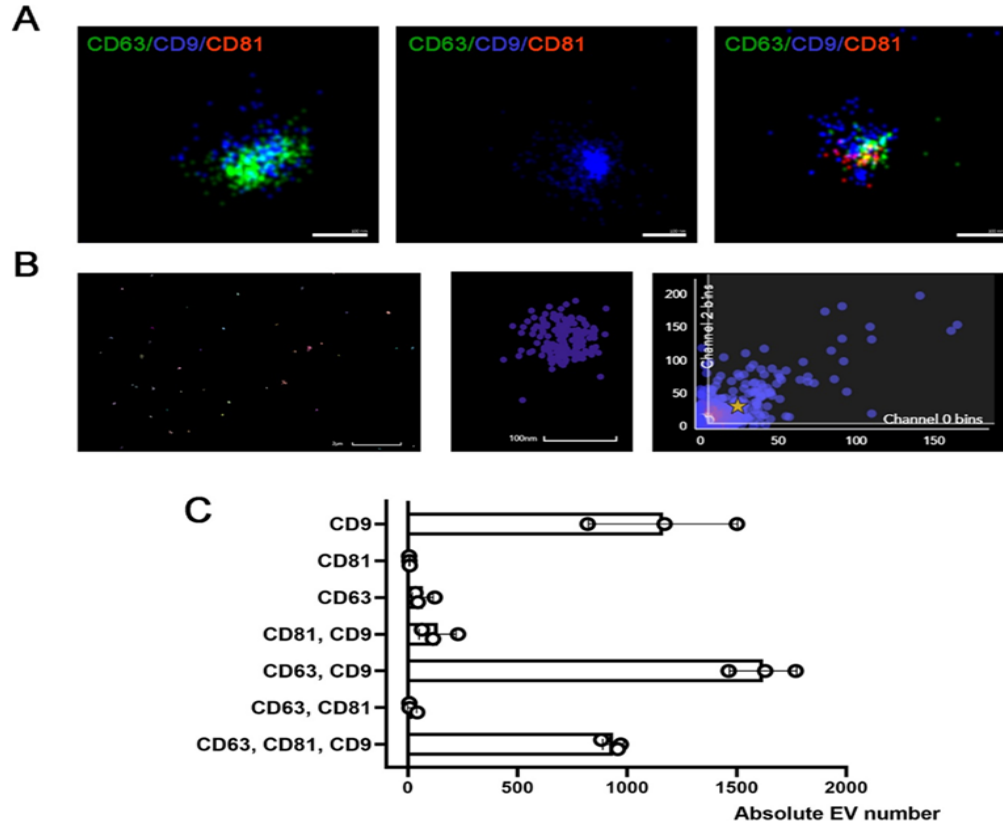


Figure 3.1: A. Representative super-resolution microscopy images of EVs showing double (CD63 and CD9), single (CD9) and triple (CD63, CD9 and CD81) expression of CD63 (green), CD81 (red), CD9 (blue). The scale bars are below each EV image (100nm). B. Representative clustering strategy of EV analysis showing a large field of view (left panel), a selected cluster and a graph of CD63/CD9 cluster distribution (right panel). C. Clustering analysis of super-resolution microscopy images showing the single, double, and triple positive EV fractions expressing the tetraspanins. The analyses were performed in 3 EV preparations using CODI software; the graph shows the mean $\pm$ SD of a cumulative analysis of 3 fields for each preparation.

3.3.3 Sequencing Data

Data analysis obtained from Massive Parallel Sequencing is reported in Table 3.2. Sequencing generated per sample an average of 18 million reads, ranging from 0.2 to 59 million reads. The mean ratio of raw reads/clean reads (reads cleared from adapters and sequences ≥ 14 bp) is equal to 90%. On average, 4% of the reads obtained from miRNA-seq aligned on miRNA sequences (miRBase v22). Overall, NGS analysis led to the identification of 132 miRNAs.

	Raw reads	Clean reads	Clean_reads/Raw_reads	Aligned reads	Well aligned/Raw reads
Mean Value	18062780	16567770	0.9	8417431	0.040
Min Value	160707	116013	0.6	44978	0.001
Max Value	59781058	58036335	1.0	31189188	0.325

Table 3.2: Sequencing Data. The table shows mean, Minimum and Maximum values of the reads after NGS.

3.3.4 MiRNA expression analysis

A differential expression analysis was performed to find miRNAs specifically associated with NASH or NAFL conditions and with the presence of significant fibrosis (F01 vs. F234). Following adjustment for library, age, gender, BMI, and T2DM, 23 and 31 miRNAs were observed to be differentially expressed according to the diagnosis of NASH and the presence of significant fibrosis, respectively. Tables 3.3 and 3.4 listed the deregulated signals. The corresponding Volcano plots of miRNAs changing significantly according to NASH and fibrosis are reported in Figure 3.2.

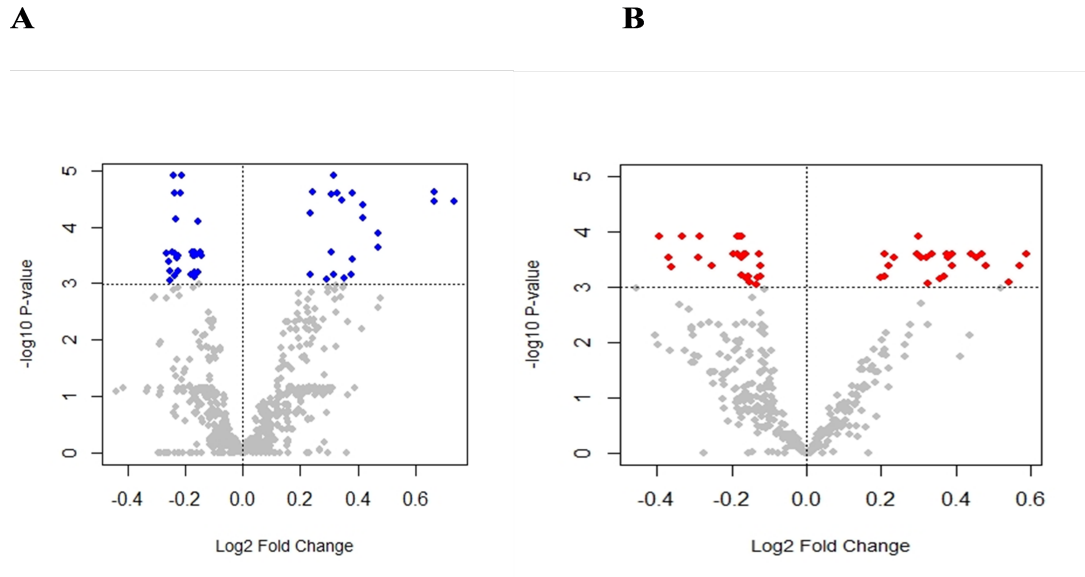


Figure 3.2: Volcano plots depicting differentially expressed miRNAs according to the diagnosis of NASH (A), and the presence of significant fibrosis (B). Scattered points represent miRNAs, x-axis is the log2 Fold change, and y-axis represents $-\log_{10}$ of adjusted p-value. Blue and red dots indicate significantly deregulated miRNAs in NASH and fibrosis analysis, respectively.

miRNAs	log2 FC	log2 FC SE	p value	Adjusted p value*
miR-30a-5p	0.32	0.09	<0.001	0.029
miR-130b-5p	-0.37	0.11	0.001	0.029
miR-143-3p	0.38	0.11	<0.001	0.029
let-7i-5p	-0.17	0.05	0.001	0.029
miR-1301-3p	-0.29	0.09	0.001	0.029
miR-193a-5p	0.46	0.14	0.001	0.029
miR-148a-3p	0.31	0.09	0.001	0.029
miR-340-3p	-0.25	0.08	0.002	0.034
miR-27b-3p	0.22	0.07	0.002	0.034
miR-100-5p	0.57	0.19	0.002	0.034
miR-375-3p	0.48	0.15	0.002	0.034
miR-423-5p	-0.36	0.12	0.002	0.035
let-7d-5p	-0.2	0.06	0.003	0.04
miR-151a-3p	-0.15	0.05	0.004	0.041
miR-215-5p	0.37	0.13	0.004	0.041
miR-29a-3p	0.21	0.07	0.003	0.041
miR-128-3p	-0.12	0.04	0.004	0.041
let-7f-5p	-0.16	0.06	0.005	0.043
miR-192-5p	0.36	0.13	0.005	0.043
miR-122-5p	0.54	0.19	0.005	0.046
miR-26a-5p	-0.15	0.05	0.006	0.046
miR-338-5p	0.33	0.12	0.006	0.047
miR-103a-3p	-0.13	0.05	0.006	0.047

Table 3.3: Summary of the 23 differentially expressed miRNAs in serum EVs comparing NASH and NAFL patients. *Adjusted for library, age, gender, BMI, and T2DM. Abbreviations: fold change (FC), miRNA (microRNA), standard error (SE).

miRNAs	log2 FC	log2 FC SE	p value	Adjusted p value*
miR-221-3p	-0.24	0.06	<0.001	0.01
miR-378c	0.38	0.1	<0.001	0.01
miR-378a-3p	0.33	0.09	<0.001	0.01
miR-146a-5p	-0.21	0.06	<0.001	0.01
miR-30a-5p	0.31	0.09	<0.001	0.01
miR-122-5p	0.67	0.19	<0.001	0.012
miR-1246	0.73	0.21	<0.001	0.012
miR-27b-3p	0.24	0.07	0.001	0.014
miR-99a-5p	0.42	0.12	0.001	0.016
miR-28-5p	-0.23	0.07	0.001	0.016
miR-375-3p	0.47	0.15	0.002	0.026
miR-30c-5p	-0.18	0.06	0.003	0.029
miR-103a-3p	-0.15	0.05	0.002	0.029
miR-191-5p	-0.15	0.05	0.003	0.029
let-7d-5p	-0.17	0.06	0.003	0.029
miR-199a-5p	-0.25	0.08	0.002	0.029
let-7b-3p	0.31	0.1	0.003	0.029
let-7f-5p	-0.17	0.05	0.003	0.029
miR-374b-5p	-0.27	0.09	0.003	0.029
miR-199a-3p	-0.23	0.08	0.004	0.032
miR-101-3p	0.38	0.13	0.004	0.032
miR-151a-5p	-0.26	0.09	0.004	0.034
miR-340-5p	-0.15	0.06	0.006	0.041
miR-584-5p	-0.17	0.06	0.005	0.041
miR-185-5p	0.31	0.12	0.007	0.043
miR-532-5p	0.24	0.09	0.007	0.043
miR-4443	0.38	0.14	0.007	0.043
miR-98-5p	-0.18	0.07	0.007	0.043
miR-1301-3p	-0.24	0.09	0.007	0.043
miR-486-3p	0.35	0.13	0.008	0.046
miR-320c	0.29	0.11	0.008	0.047

Table 3.4: Summary of the 31 differentially expressed miRNAs in serum EVs comparing F01and F234 patients. *Adjusted for library, age, gender, BMI, and T2DM. Abbreviations: fold change (FC), miRNA (microRNA), standard error (SE).

As reported in Table 3.3, considering a log2 fold change (FC) ≥ 0.5 , in subjects with a diagnosis of NASH, miR-122-5p (log2 FC = 0.54, adjusted p = 0.046) and miR-100-5p (log2 FC = 0.57, adjusted p = 0.034) were the most significantly up-regulated miRNAs. Regarding the fibrosis stage, the most deregulated miRNAs were miR-1246 (log2 FC = 0.67, adjusted p = 0.012) and again miR-122-5p (log2 FC = 0.67, adjusted p = 0.012),

both up-regulated in patients with $F \geq 2$ (Table 3.4).

Overall, among the two lists of significantly deregulated miRNAs, eight of them were present in both lists of miRNAs (Figure 3.3). Importantly, miR-27b-3p, miR-30a-5p, miR-122-5p, and miR-375-3p resulted consistently up-regulated, while miR-103a-3p, miR-1301-3p, let-7d-5p, and let-7f-5p were significantly down-regulated.

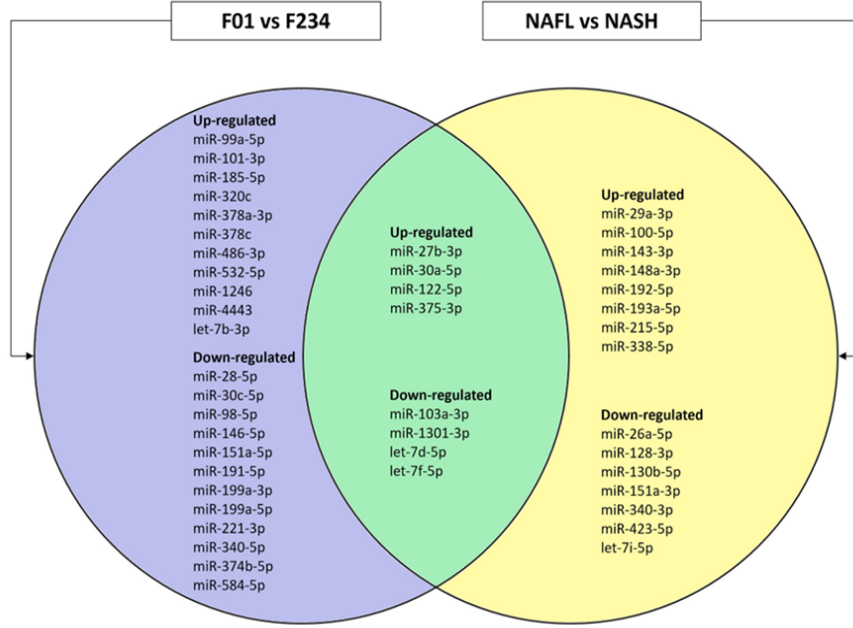


Figure 3.3: Venn diagram depicting deregulated miRNAs according to liver fibrosis and NASH diagnosis and their intersection.

3.3.5 Target genes and Enrichment analysis

Functional enrichment analysis was performed to explore if the significantly deregulated miRNAs were involved in relevant molecular networks related to the pathogenesis of NAFLD or fibrosis progression.

The multiMiR Bioconductor's package was used to retrieve validated target genes of the identified deregulated miRNAs associated with both the severity of liver fibrosis and NASH. A total of 562 and 1011 experimentally validated target genes for the down- and up-regulated miRNAs respectively were identified.

Concerning down-regulated miRNAs (i.e., miR-103a-3p, miR-1301-3p, let-7d-5p, and let-7f-5p), the results of Gene Ontology enrichment analysis of their target genes are summarized in Figure 3.4. This analysis identified biological processes belonging to the regulation of cellular metabolic and biosynthetic processes, cell cycle, cell death, and genes involved in transcriptional regulation as the most pathways involved. In addition, the analysis of the KEGG database, as reported in Table 3.5 and 3.6, revealed several intriguing terms related to Hippo and p53 signaling, as well as to cancers, including hepatocellular carcinoma (HCC).

On the other hand, when the target genes of the up-regulated miRNAs were used for the Gene Ontology enrichment analysis, they resulted to be mainly involved in the reg-

ulation of cellular metabolic processes, cell cycle, cell death, apoptotic processes, and protein transport (Figure 3.5). In this case, considering the KEGG pathway analysis, the genes were significantly grouped in pathways related to proteoglycans in cancer, adherens junction, protein processing in endoplasmic reticulum, mitogen-activated protein kinases (MAPK) signaling, mitophagy, spliceosome, non-alcoholic fatty liver disease, lipid, and atherosclerosis (Table 3.7).

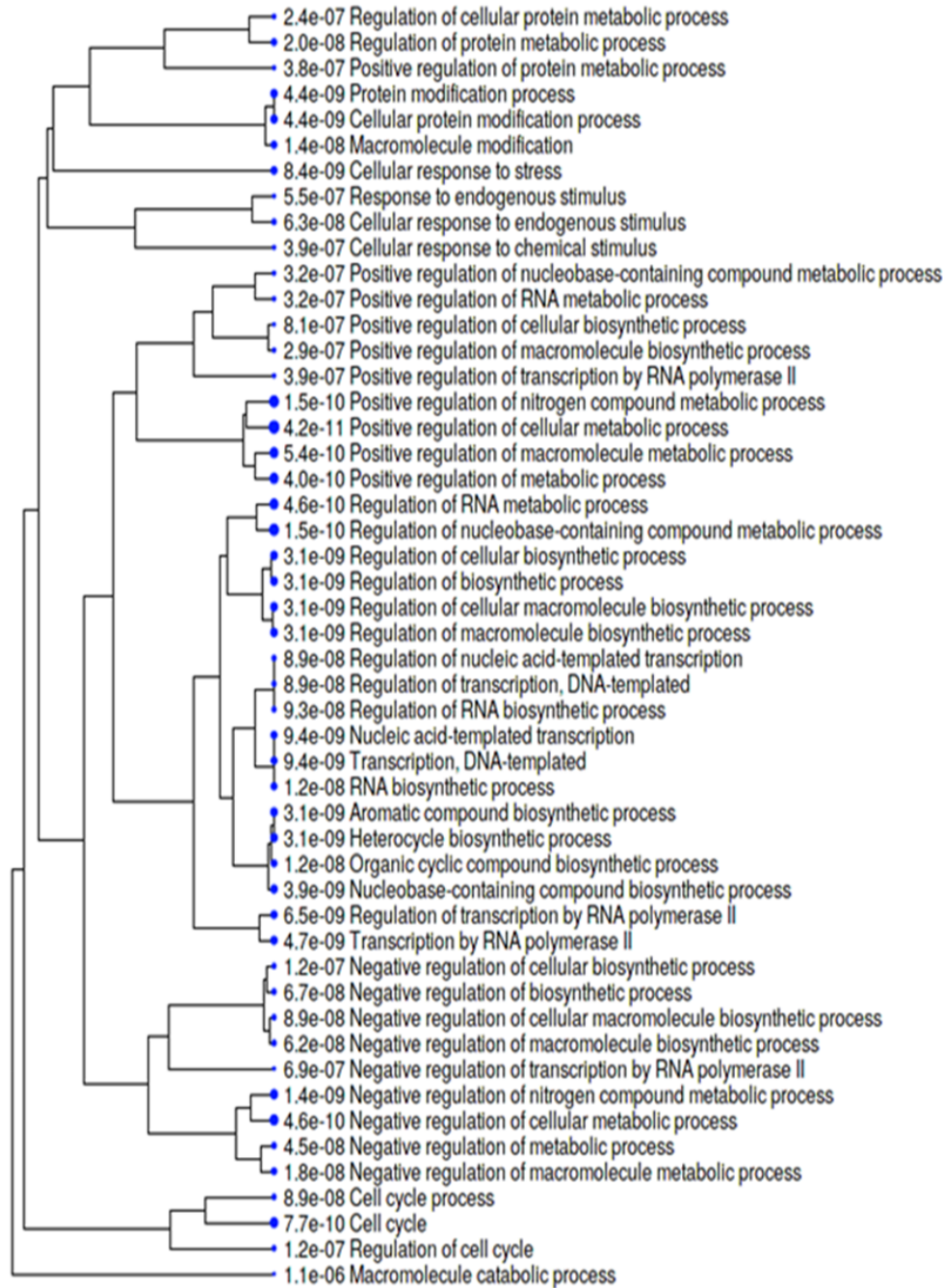


Figure 3.4: Gene Ontology terms Tree of down-regulated miRNAs target genes. Pathways with many shared genes are clustered together. Bigger dots indicate more significant p values.

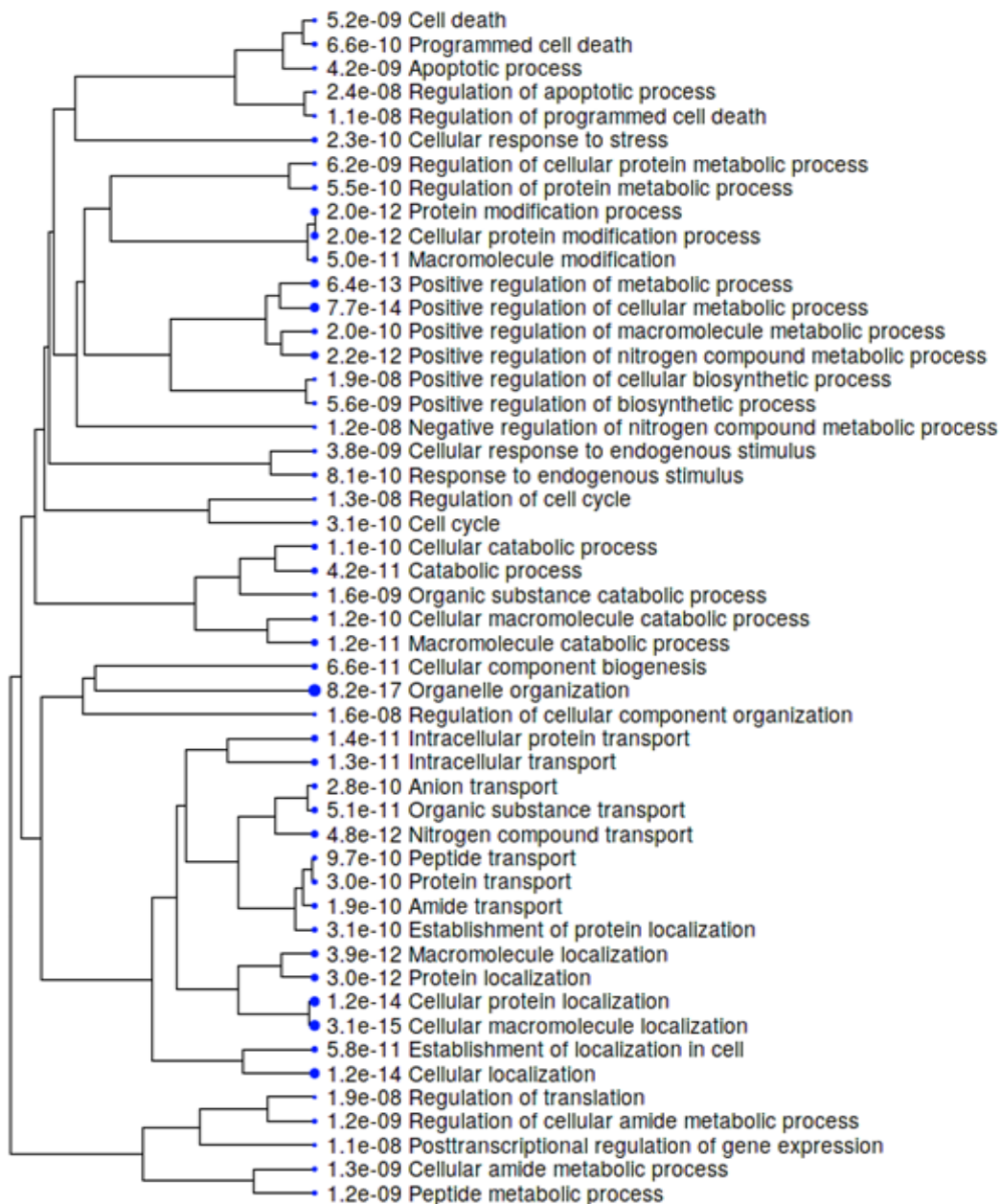


Figure 3.5: Gene Ontology terms Tree of up-regulated miRNAs target genes. Pathways with many shared genes are clustered together. Bigger dots indicate more significant p values.

Term description	Fold enrichment	Occurrence	Lowest p value	Highest p value	miRNAs target genes
Cell cycle	4.4	10	2.90E-15	2.10E-08	CCND1, CDK6, HDAC2, MYC, CDKN1A, CDK1, YWHAZ, YWHAH, WEE1, ESPL1, SMC1A, CREBBP, CDC25A, MCM7
p53 signaling pathway	5.95	10	6.90E-09	6.10E-08	MDM4, CDKN1A, CCND1, CDK6, CDK1, TNFRSF10B, PMAIP1, EI24, THBS1, RRM2, PTEN
Hippo signaling pathway	3.1	10	1.00E-08	1.20E-04	AMOT, SAV1, LATS1, TGFBR1, WNT3A, FZD6, DVL3, YWHAZ, YWHAH, AXIN2, MYC, CCND1
Cushing syndrome	3.29	10	2.70E-08	1.50E-04	ADCY9, KMT2A, KMT2D, CDK6, CCND1, AHR, CDKN1A, NR4A1, SP1, WNT3A, FZD6, DVL3, AXIN2
Breast cancer	3.47	10	5.00E-08	2.30E-06	NCOA3, SP1, CCND1, MYC, FGF2, IGF1R, PTEN, CDKN1A, WNT3A, FZD6, DVL3, AXIN2, CDK6
Glioma	3.64	10	5.00E-08	4.10E-06	PDGFB, IGF1R, CALM1, PTEN, CDKN1A, CCND1, CDK6
Transcriptional misregulation in cancer	2.72	10	7.20E-08	1.00E-06	HDAC2, MYC, KMT2A, CCNT1, CCNT2, RUNX2, HMGA2, IGF1R, CDKN1A, ELK4, MYCN, SP1
Hepatocellular carcinoma	2.78	10	9.00E-08	1.30E-05	IGF1R, PTEN, MYC, CDKN1A, CDK6, CCND1, TGFBR1, WNT3A, FZD6, DVL3, AXIN2, SMARCC1
Cellular senescence	3.27	10	1.00E-07	7.40E-07	TGFBR1, CDK6, CCND1, CDKN1A, HLA-B, PTEN, LIN52, MYC, CDK1, RAD9A, CDC25A, CALM1, VDAC1
Viral carcinogenesis	2.36	10	1.10E-07	2.60E-05	YWHAZ, YWHAH, CDKN1A, CDK1, CREBBP, CCND1, CDK6, HDAC2, PMAIP1, HLA-B
MAPK signaling pathway	1.87	10	2.80E-07	7.30E-04	GNG12, FGF2, PDGFB, IGF1R, ELK4, MYC, TGFBR1, TAB2, STK4, MAP3K7, MAP2K7, CRKL, DUSP1, NR4A1
Hepatitis B	1.94	10	2.80E-07	9.90E-06	TGFBR1, CDKN1A, MYC, CREBBP, YWHAZ, TAB2, MAP3K7, MAP2K7
Nucleocytoplasmic transport	2.62	10	4.20E-07	1.60E-03	NUP155, NUP50, KPNA1, KPNA5, IPO9, XPO7, THOC5
FoxO signaling pathway	3.35	10	6.20E-07	3.60E-05	TGFBR1, PRKAB2, CREBBP, STK4, IGF1R, PTEN, CCND1, CDKN1A, BNIP3, GABARAPL1, SOD2
Human T-cell leukemia virus 1 infection	1.96	10	6.50E-07	1.70E-04	MYC, HLA-B, VDAC1, ESPL1, CCND1, CDKN1A, PTEN, TGFBR1, ADCY9, ELK4, CREBBP
Non-small cell lung cancer	3.25	10	8.30E-07	2.90E-05	CDK6, CCND1, STK4, CDKN1A, KIF5B, EML4
Bladder cancer	4.75	10	1.20E-06	4.30E-06	MYC, DAPK1, CDKN1A, CCND1, THBS1
Prostate cancer	2.41	10	1.20E-06	9.90E-05	PDGFB, IGF1R, PTEN, CDKN1A, CREBBP, CCND1
Melanoma	3.79	10	1.30E-06	3.70E-05	FGF2, PDGFB, IGF1R, CCND1, PTEN, CDKN1A, CDK6
Chronic myeloid leukemia	3.59	10	1.40E-06	4.10E-06	CRKL, MYC, CDKN1A, CCND1, CDK6, TGFBR1, HDAC2
Epstein-Barr virus infection	1.75	10	1.40E-06	4.20E-06	TAB2, MAP3K7, HLA-B, HDAC2, MYC, CDKN1A, MAP2K7, CCND1, CDK6
Chemical carcinogenesis - receptor activation	1.7	10	1.50E-06	3.60E-03	CDC25A, CCND1, MYC, FGF2, ADCY9, KLF4, AHR, KPNA1, KPNA5
Hepatitis C	1.77	10	1.60E-06	1.90E-05	SOC3, YWHAZ, YWHAH, CDKN1A, CDK6, CCND1, MYC
Adherens junction	4.13	10	2.30E-06	5.50E-05	WASL, VCL, CSNK2A1, IGF1R, TGFBR1, CREBBP, MAP3K7
Thyroid hormone signaling pathway	2.25	10	2.40E-06	4.40E-05	HDAC2, NCOA3, CREBBP, MED13, CCND1, MYC, ATP2A2
Small cell lung cancer	2.54	10	3.40E-06	9.80E-05	CDKN1A, MYC, CDK6, CCND1, ITGA2, PTEN
Kaposi sarcoma-associated herpesvirus infection	2.42	10	3.50E-06	1.00E-04	CREBBP, CDKN1A, CDK6, HLA-B, GNG5, GNG12, MAP2K7, CALM1, FGF2, PDGFB, CCND1, MYC
HIF-1 signaling pathway	2.91	10	5.10E-06	1.20E-03	IL6R, IFNGR2, IGF1R, CREBBP, TFRC, EDN1, ALDOA, CDKN1A
Pathogenic Escherichia coli infection	2.01	10	5.20E-06	4.50E-03	ARF6, ACTR2, MYO5A, MYH9, WASL, TUBB4A, MAP3K7, TAB2, TNFRSF10B, ABCF2
Human cytomegalovirus infection	2.1	10	6.10E-06	4.80E-04	SP1, GNG5, GNG12, CALM1, ADCY9, CCND1, MYC, IL6R, CRKL, HLA-B, CDKN1A, CDK6
Osteoclast differentiation	1.89	10	7.40E-06	2.10E-03	IFNGR2, TGFBR1, MAP3K7, TAB2, MAP2K7, SOCS3
Endometrial cancer	3.36	10	8.50E-06	1.50E-04	PTEN, AXIN2, MYC, CCND1, CDKN1A
TGF-beta signaling pathway	2.63	10	1.20E-05	1.20E-05	THBS1, TGFBR1, ACVR2B, CREBBP, SP1, MYC
Endocytosis	2.43	10	1.50E-05	2.80E-04	ARF6, AP2A1, TGFBR1, TFRC, IGF1R, WASL, ACTR2, CAPZA2, VPS4A, HLA-B, RABEP1, RAB10, KIF5B, RAB11FIP4, GBF1
Autophagy - animal	3.64	10	1.50E-05	1.50E-05	IGF1R, PTEN, BNIP3, ATG9A, SMCR8, RAGC, CAMKK2, MAP3K7, DAPK1, MTMR3, MTMR4, GABARAPL1, LAMP2
Human immunodeficiency virus 1 infection	2.1	10	1.60E-05	3.60E-05	CRKL, CALM1, GNG5, GNG12, TAB2, MAP3K7, MAP2K7, HLA-B, CUL4A, WEE1, CDK1
GABAergic synapse	2.27	6	1.60E-05	3.90E-02	GABARAPL1, NSF, GNG5, GNG12, ADCY9
Colorectal cancer	2.72	10	1.60E-05	1.10E-04	AXIN2, MYC, CCND1, TGFBR1, PMAIP1, CDKN1A

Table 3.5: Analysis of KEGG pathway enrichment for target genes of down-regulated miRNAs. From left to right, the table reports the description of the enriched term, the fold enrichment value for the enriched term, the number of iterations that the given term was found to be enriched over all iterations (occurrence), the lowest adjusted-p value of the given term over all iterations, the highest adjusted-p value of the given term over all iterations, and the list of miRNAs target genes.

Pancreatic cancer	2.56	10	2.60E-05	4.60E-05	CDK6, CCND1, CDKN1A, TGFB1, RAD51
Thyroid cancer	3.16	10	3.40E-05	2.80E-04	CDKN1A, MYC, CCND1
ErbB signaling pathway	2.35	10	3.60E-05	4.20E-03	CRKL, ABL2, MAP2K7, MYC, CDKN1A
Focal adhesion	1.97	10	4.90E-05	1.60E-03	COL6A1, THBS1, ITGA2, PDGFB, IGF1R, ZYX, VCL, PTEN, CRKL, CCND1
Salmonella infection	2.21	10	6.40E-05	1.50E-04	DYNC1L2, TUBB4A, ARL8A, ARL8B, KIF5B, ACTR2, ARF6, WASL, MAP3K7, TAB2, KPN1, MAP2K7, MYC, TNFRSF10B
Relaxin signaling pathway	2.15	9	6.60E-05	5.00E-02	GNG5, GNG12, EDN1, TGFB1, COL3A1, ADCY9, MAP2K7
Wnt signaling pathway	2.87	10	6.70E-05	7.20E-05	WNT3A, ZNRF3, FZD6, DVL3, CSNK2A1, AXIN2, TLE4, CREBBP, MAP3K7, MYC, CCND1, ROR2
Apelin signaling pathway	2.31	10	7.40E-05	3.20E-02	ADCY9, GNG5, GNG12, CALM1, CCND1, PRKAB2, MEFD2, TGFB1
Shigellosis	1.98	10	8.40E-05	6.70E-04	CRKL, ACTR2, VCL, WASL, ARF6, VDAC1, BNP3, MAP3K7, TAB2, RAGC, GLMN
Oocyte meiosis	2.95	10	1.30E-04	8.10E-04	IGF1R, ADCY9, CREB3, YWHAZ, YWHAH, CDK1, ESPL1, SMC1A, CALM1
Base excision repair	2.36	1	1.40E-04	1.40E-04	POLL, POLD3
Hippo signaling pathway - multiple species	2.78	10	1.50E-04	4.80E-03	SAV1, LATS1
DNA replication	3.25	10	1.60E-04	1.70E-02	DNA2, POLD3, MCM7
Gastric cancer	2.35	10	1.70E-04	1.20E-03	CDKN1A, WNT3A, FZD6, DVL3, AXIN2, MYC, CCND1, TGFB1, FG2
JAK-STAT signaling pathway	2.98	10	1.70E-04	6.10E-04	IL13, PDGFB, IL6R, IFNGR2, IFNLR1, MPL, SOCS3, SOCS7, MYC, CCND1, CDKN1A, CREBBP
Proteoglycans in cancer	1.92	10	1.70E-04	2.30E-03	CCND1, IGF1R, CDKN1A, MYC, THBS1, ITGA2, FZD6, RDX, FGF2, WNT3A
TNF signaling pathway	1.76	10	1.80E-04	7.80E-03	MAP3K7, TAB2, MAP2K7, SOCS3, EDN1
Parathyroid hormone synthesis, secretion and action	2.23	10	1.80E-04	2.00E-03	ADCY9, SP1, PDE4D, MEFD2, CDKN1A, RUNX2
Protein processing in endoplasmic reticulum	1.92	10	1.80E-04	1.00E-02	RPN1, EDEM1, EDEM3, SVIP, DNAJA1, YOD1, RAD23B, MAP2K7
Fluid shear stress and atherosclerosis	1.98	10	1.80E-04	2.90E-03	CALM1, DUSP1, MAP3K7, MAP2K7, PDGFB, EDN1, ACVR2B
Signaling pathways regulating pluripotency of stem cells	4.21	10	2.20E-04	2.20E-04	KLF4, MYC, ACVR2B, WNT3A, FZD6, DVL3, AXIN2, FGF2, IGF1R, SKI, SMARCA1, RIF1, PCGF3, HAND1, ZFX3
Regulation of actin cytoskeleton	1.82	10	2.40E-04	8.70E-03	FGF2, PDGFB, ITGA2, GNG12, CRKL, MYH9, WASL, ACTR2, RDX, VCL
NF-kappa B signaling pathway	1.5	10	2.80E-04	9.80E-04	MAP3K7, TAB2, ERK1, CSNK2A1
Measles	1.41	10	2.80E-04	5.10E-04	CSNK2A1, MAP3K7, TAB2, CCND1, CDK6
Nucleotide excision repair	2.54	9	2.90E-04	2.20E-02	CUL4A, RAD23B, POLD3
Notch signaling pathway	2.83	10	3.30E-04	2.70E-03	DVL3, CREBBP, TLE4, HDAC2
Bacterial invasion of epithelial cells	2.26	10	4.10E-04	1.20E-02	CRKL, WASL, ACTR2, VCL
Tight junction	1.69	10	4.20E-04	7.00E-03	AMOT, CCND1, RDX, MAP2K7, ACTR2, PRKAB2, MYH9
RNA polymerase	2.69	5	5.50E-04	9.80E-04	POLR2D, POLR3D
Lipid and atherosclerosis	1.27	10	6.30E-04	2.80E-03	SOD2, MAP2K7, TAB2, MAP3K7, MIB1, CALM1, TNFRSF10B
Morphine addiction	1.81	2	7.00E-04	7.00E-04	ADCY9, GNG5, GNG12, PDE4D
Circadian rhythm	3.9	10	7.20E-04	8.80E-04	PER1, CRY2, PRKAB2
Renal cell carcinoma	2.4	10	7.50E-04	2.20E-03	CREBBP, PDGFB, CRKL, CDKN1A
Rap1 signaling pathway	1.52	10	7.50E-04	2.90E-02	CALM1, ADCY9, FGF2, PDGFB, IGF1R, CRKL, SIPA1L2, THBS1
Acute myeloid leukemia	1.16	10	7.60E-04	2.00E-02	MYC, CCND1
Central carbon metabolism in cancer	1.13	10	7.90E-04	1.50E-02	MYC, PTEN
Chemical carcinogenesis - reactive oxygen species	1.14	9	8.00E-04	4.50E-02	SOD2, VDAC1, AHR, MAP2K7, PTEN, ABL2
AMPK signaling pathway	2.64	10	8.30E-04	4.00E-03	MAP3K7, CAB39, CAMKK2, ADIPOR2, PRKAB2, CCND1, RAB10, IGF1R
Circadian entrainment	2.05	3	8.70E-04	3.50E-02	CALM1, GNG5, GNG12, ADCY9, PER1
Ras signaling pathway	1.73	10	9.40E-04	5.30E-03	FGF2, PDGFB, IGF1R, GNG5, GNG12, CALM1, STK4, SHOC2, ABL2, ARF6
Alcoholism	1.84	10	9.60E-04	2.00E-02	HDAC2, GNG5, GNG12, CALM1, CAMKK2
Influenza A	2.06	10	1.00E-03	9.70E-03	CREBBP, SOCS3, IFNGR2, CDK6, TNFRSF10B, VDAC1, HNRNPUL1, KPN1, KPN5

Table 3.6: Analysis of KEGG pathway enrichment for target genes of down-regulated miRNAs. From left to right, the table reports the description of the enriched term, the fold enrichment value for the enriched term, the number of iterations that the given term was found to be enriched over all iterations (occurrence), the lowest adjusted-p value of the given term over all iterations, the highest adjusted-p value of the given term over all iterations, and the list of miRNAs target genes.

Term description	Fold enrichment	occurrence	Lowest p Value	Highest p value	miRNAs target genes
Ubiquitin mediated proteolysis	3.47	10	1.80E-18	3.40E-10	UBA1, SAE1, UBE2D1, UBE2D3, UBE2H, UBE2L3, UBE2K, UBE3C, SMURF2, ITCH, TRIP12, NEDD4L, UBR5, STUB1, PRPF19, CUL1, SKP2, VHL, SOCS3, SOCS3, CDC20, ANAPC5
Cell cycle	2.96	10	1.10E-12	3.70E-09	CDK4, RBL1, CCNE2, CUL1, SKP2, CCNA2, DBF4, YWHAZ, YWHA, WEE1, ANAPC5, CDC20, BUB3, TP53, PCNA, ORC2, ORC6
Protein processing in endoplasmic reticulum	3.07	10	5.10E-12	4.00E-08	SEC62, HSPA5, DNAAK3, CALR, MAN1A2, SEC23A, SEC24B, SEC24A, ERP29, SSR3, DERL1, DNAA1A, DNAB1, BAG2, RAD23B, NFE2L2, MAPK8, BAX, CAPN2, STUB1, UBE2D1, UBE2D3, CUL1
Proteoglycans in cancer	2.34	10	6.70E-11	6.20E-10	CD44, VAV3, NRAS, MAPK1, RAC1, PAK1, GAB1, PIK3CD, SLC9A1, PPP1CC, PPP1R12A, IGF1R, EGFR, CASP3, CTNNB1, MET, THBS1, TP53, ITGA2, FRS2, WNT5A, CTSL
Viral carcinogenesis	3.28	10	8.10E-11	3.20E-08	NRAS, MAPK1, CREB1, YWHAZ, YWHA, JAK1, CASP3, PIK3CD, TP53, CCNE2, BAX, TBLP1, GTF2H1, GTF2H2, GTF2E2, GTF2B, CCNA2, SKP2, RBL1, JUN, PKM, RAC1, CDK4, CDC20, PMAIP1
Adherens junction	3.6	10	1.20E-10	8.80E-07	RAC1, WASF1, CTNND1, CTNNB1, IGF1R, MET, EGFR, MAPK1, SNAI1, TGFBR1, NLK
E protein B and virus infection	2.15	10	2.00E-10	4.10E-08	HLA-DQA1, IFNA1, IFNAR2, JAK1, BAX, CASP3, TABP, CALR, TP53, PIK3CD, MAPK8, MAP2K3, JUN, CD44, VIM, RAC1, SKP2, CDK4, CCNA2, CCNE2
MAPK signaling pathway	1.86	10	2.00E-10	6.50E-09	GNIG12, PPP3CA, PPP3CB, PPP3R1, RASGRP3, NFI, BDNF, EGFR, IGF1R, MET, NRAS, MAPK1, MKN2, STMN1, PLA2G4A, TGFBR1, RAC1, CASP3, PAK1, TAOX1, MAP2K3, MAPK8, JUN, TP53, NLK
Spliceosome	1.84	10	4.00E-10	4.40E-06	DHX8, PRPF40A, SFB3, SNRNFP27, PRPF8, PRPF19, HNRNPA1, HNRNPA12, HNRNPM, HNRNPU, SRSF1, SRSF7
Mitophagy	3.71	10	5.00E-10	6.50E-10	MAPK8, JUN, PINK1, GABARAPL2, MAP1LC3B, BECN1, BCL2L1, BNIP3, NRAS, TP53, BNIP3L, SP1
Hepatitis B	2.28	10	5.10E-10	1.70E-06	TGFBR1, MAPK8, YWHAZ, PIK3CD, BAX, CASP3, TP53, CCNA2, CCNE2, JAK1, NRAS, JUN, CREB1, PCNA, MAPK1, MAP2K3, IFNA1
Pancreatic cancer	3.41	10	6.40E-10	2.30E-08	PIK3CD, RAC1, BCL2L1, MAPK1, MAPK8, EGFR, JAK1, CDK4, TP53, BAX, TGFBR1, BRCA2
Shigellosis	2.4	10	6.90E-10	6.50E-09	CD44, RAC1, WASF1, ARPC5, EGFR, PIK3CD, ARF6, CAST, CAPN2, TP53, BCL2L1, HK1, BAX, BNIP3, PPID, MAPK8, JUN, UBE2D1, UBE2D3, CUL1, MAPK1, CBX3, BECN1, MAP1LC3B
p53 signaling pathway	3.9	10	3.30E-09	1.30E-04	MDM4, TP53, CDK4, CCNE2, TNFRSF10B, BAX, PMAIP1, BCL2L1, CASP3, THBS1, SERPINB5, RRM2, CCNG1
Hepatocellular carcinoma	2.57	10	4.70E-09	1.20E-08	EGFR, IGF1R, SHC1, NRAS, MAPK1, PIK3CD, BCL2L1, TP53, BAX, CDK4, TGFBR1, WNT5A, CTNNB1, NFE2L2, HMOX1, GSTM3, MGST2, MET, GAB1, PBRM1
Kaposi sarcoma-associated herpesvirus infection	2.8	10	5.40E-09	5.40E-09	IFNA1, IFNAR2, JAK1, ZFP36, BAX, CASP3, TP53, CDK4, BECN1, MAP1LC3B, GNG2, GNG5, GNG12, NRAS, MAPK1, JUN, PIK3CD, RAC1, MAPK8, CALM3, PPP3CA, PPP3CB, PPP3R1, CTNNB1, CREB1
Colorectal cancer	3.52	10	5.80E-09	4.20E-08	CTNNB1, PIK3CD, MAPK1, JUN, RAC1, MAPK8, CASP3, TGFBR1, MLH1, BAX, PMAIP1, TP53, EGFR, NRAS
Lipid and atherosclerosis	2.63	10	6.50E-09	6.50E-09	LDLR, OLR1, PIK3CD, BAX, BCL2L1, MAPK8, SOD2, RAC1, JUN, MAP2K3, NRAS, MAPK1, MIB1, IFNA1, PPARG, VAV3, HSPA5, NFE2L2, CALM3, PPP3CA, PPP3CB, PPP3R1, TNFRSF10B, CASP3, CASP7, TP53
Human cytomegalovirus infection	2.42	10	7.40E-09	7.40E-09	PIK3CD, EGFR, NRAS, MAPK1, SP1, RAC1, IFNA1, GNG2, GNG5, GNG12, CALM3, PPP3CA, PPP3CB, PPP3R1, GNAI2, GNAI3, CTNNB1, JAK1, CREB1, TABP, CALR, TP53, CDK4, BAX, CASP3
Human T cell leukemia virus 1 infection	2.57	10	1.20E-08	8.10E-07	TRRAP, NRAS, MAPK1, XPOL, CALR, PPP3CA, PPP3CB, PPP3R1, JAK1, BUB3, ANAPC5, CDC20, CDK4, CCNA2, CCNE2, TP53, PIK3CD, MAPK8, JUN, ZFP36, BCL2L1, TGFBR1, HLA-DQA1, TBLP1, CREB1, BAX
AGE-RAGE signaling pathway in diabetic complications	2.18	10	1.30E-08	2.60E-05	TGFBR1, MAPK1, JUN, MAPK8, RAC1, NRAS, PIK3CD, BAX, CASP3, CDK4
Central carbon metabolism in cancer	3.76	10	1.40E-08	1.30E-06	PKM, TP53, SLC1A5, MET, RET, EGFR, NRAS, MAPK1, HK1, IDH1, PIK3CD, SLC7A5
Choline metabolism in cancer	2.5	10	1.60E-08	1.60E-06	EGFR, NRAS, MAPK1, MAPK8, PLA2G4A, PIK3CD, RAC1, WASF1, SP1, SLC44A1, JUN
Rap1 signaling pathway	1.69	10	2.10E-08	5.80E-06	CALM3, ADORA2B, RASGRP3, EGFR, IGF1R, MET, CTNNB1, GNAI2, THBS1, RAC1, VAV3, CTNND1, MAPK1, MAP2K3, PIK3CD, NRAS
TNF signaling pathway	1.95	10	2.40E-08	6.10E-04	MAPK8, JUN, ITCH, MAP2K3, MAPK1, CREB1, CASP7, CASP3, SOCS3, PIK3CD
Breast cancer	2.37	10	2.80E-08	5.80E-08	NCOA3, JUN, SP1, IGF1R, EGFR, SHC1, NRAS, MAPK1, PIK3CD, NOTCH2, WNT5A, CTNNB1, TP53, BAX, CDK4, BRCA2
TGF-beta signaling pathway	2.91	10	2.80E-08	4.50E-04	THBS1, FBN1, NODAL, BMP2, TGFBR1, SMURF2, ID4, RBL1, SP1, CUL1, MAPK1, PPP2R1B
Ras signaling pathway	2.11	10	3.20E-08	6.90E-08	BDNF, EGFR, IGF1R, MET, GAB1, SHC1, RASGRP3, GNG2, GNG5, GNG12, CALM3, NRAS, NF1, RAC1, PAK1, PIK3CD, BCL2L1, MAPK1, PLA2G4A, MAPK8, ABL2, ARF6
Gastric cancer	2.18	10	3.20E-08	1.70E-05	TP53, BAX, WNT5A, CTNNB1, EGFR, SHC1, NRAS, MAPK1, PIK3CD, MLH1, CCNE2, TGFBR1, JUN, MET, GAB1
Human immunodeficiency virus 1 infection	2.76	10	4.00E-08	4.20E-08	CUL1, GNAI2, PIK3CD, JUN, NRAS, MAPK1, CALM3, PPP3CA, PPP3CB, PPP3R1, GNG2, GNG5, GNG12, RAC1, PAK1, BCL2L1, IFNA1, MAPK8, SAMHD1, MAP2K3, BAX, CASP3, TABP, CALR, AP1G1, WEE1
Relaxin signaling pathway	2.38	10	4.10E-08	7.40E-05	GNAI2, GNG2, GNG5, GNG12, CREB1, PIK3CD, MAPK1, MMP13, TGFBR1, MAPK8, JUN, NRAS, SHC1, EGFR
Epithelial cell signaling in Helicobacter pylori infection	3.09	10	4.20E-08	3.00E-06	MET, EGFR, ADAM10, RAC1, PAK1, MAPK8, JUN, CASP3, ATP6V1B2, ATP6V1C1
Osteoclast differentiation	2.61	10	4.90E-08	6.90E-05	MAPK1, PIK3CD, TGFBR1, MAPK8, JUN, RAC1, PPP3CA, PPP3CB, PPP3R1, CREB1, PPARG, IFNAR2, JAK1, SOCS1, SOCS3
Chemical carcinogenesis - reactive oxygen species	1.78	10	4.90E-08	1.70E-05	GSTM3, MGST2, NDUFB3, SOD2, PPF, CAT, EGFR, MET, NRAS, MAPK1, MAPK8, JUN, PIK3CD, RAC1, ABL2, NFE2L2, HMOX1
Non-small cell lung cancer	3.3	10	6.40E-08	7.00E-06	CDK4, PIK3CD, EGFR, MET, NRAS, MAPK1, TP53, BAX, KIF5B, RET, EML4
Hepatitis C	2.53	10	7.10E-08	7.10E-08	LDLR, IFNA1, SOCS3, IFNAR2, JAK1, PPP2R1B, PPP2R2A, EGFR, NRAS, YWHAZ, YWHA, MAPK1, PIK3CD, CTNNB1, TP53, CDK4, CASP3, BAX
Prostate cancer	2.01	10	7.10E-08	7.10E-08	CCNE2, EGFR, IGF1R, PIK3CD, TP53, CREB1, CTNNB1, NRAS, MAPK1
Cellular senescence	2.93	10	7.10E-08	7.10E-08	TGFBR1, CDK4, RBL1, PIK3CD, CCNE2, NRAS, CCNA2, MYBL2, LINS4, TP53, PPP1CC, MAPK1, MAP2K3, ZFP36L1, ZFP36L2, CAPN2, CALM3, PPP3CA, PPP3CB, PPP3R1, PPID
Th17 cell differentiation	2	10	7.80E-08	1.80E-03	TGFBR1, JAK1, HLA-DQA1, PPP3CA, PPP3CB, PPP3R1, MAPK1, MAPK8, JUN, RUNX1

Table 3.7: Analysis of KEGG pathway enrichment for target genes of up-regulated miRNAs. From left to right, the table reports the description of the enriched term, the fold enrichment value for the enriched term, the number of iterations that the given term was found to be enriched over all iterations (occurrence), the lowest adjusted-p value of the given term over all iterations, the highest adjusted-p value of the given term over all iterations, and the list of miRNAs target genes.

Neurotrophin signaling pathway	2.75	10	8.30E-08	4.90E-07	BDNF, NRAS, MAPK1, KIDINS220, FRS2, SHC1, GAB1, PIK3CD, CALM3, RAC1, MAPK8, JUN, TP53, BAX, YWHAZ
mRNA surveillance pathway	2.76	10	1.40E-07	3.90E-05	NXT1, TARDBP, NUDT21, CPSF1, CPSF3, CPSF4, WDR82, PPP1CC, PABPC1, PPP2R1B, PPP2R2A, PPP2R3C
GnRH signaling pathway	1.86	10	1.40E-07	5.00E-05	CALM3, MAP2K3, PLA2G4A, MAPK8, JUN, EGFR, NRAS, MAPK1
Estrogen signaling pathway	2.21	10	1.60E-07	1.00E-06	NCOA3, KRT10, KRT14, KRT18, JUN, SP1, CREB1, EGFR, SHC1, NRAS, MAPK1, PIK3CD, GNAI2, CALM3
ErbB signalling pathway	2.6	10	1.60E-07	2.10E-06	EGFR, ABL2, PAK1, MAPK8, JUN, SHC1, NRAS, MAPK1, GAB1, PIK3CD
Focal adhesion	2.29	10	2.00E-07	4.10E-07	LAMC1, THBS1, THBS2, ITGA2, ITGA6, ITGB4, IGFBP1, EGFR, MET, PPP1CC, PPP1R12A, CAPN2, CTNNB1, PIK3CD, VAV3, RAC1, PAK1, MAPK8, JUN, SHC1, MAPK1
Bacterial invasion of epithelial cells	2.82	10	2.00E-07	8.40E-04	CTNNB1, MET, GAB1, PIK3CD, RAC1, WASF1, ARPC5, SHC1, CD2AP
Melanoma	2.7	10	2.00E-07	3.60E-07	MET, IGFBP1, EGFR, NRAS, MAPK1, CDK4, PIK3CD, TP53, BAX
FoxO signalling pathway	2.7	10	2.20E-07	7.20E-07	TGFB1, PRKAA1, NLK, SKP2, MAPK8, IGFBP1, PIK3CD, EGFR, NRAS, MAPK1, PRMT1, PLK2, BNIP3, GABARAPL2, CAT, SOD2
T cell receptor signaling pathway	2.35	10	2.20E-07	2.00E-06	VAV3, PAK1, PPP3CA, PPP3CB, PPP3R1, NRAS, MAPK1, JUN, MAPK8, PIK3CD, CDK4
Glioma	2.88	10	2.30E-07	4.20E-07	EGFR, IGFBP1, CALM3, SHC1, NRAS, MAPK1, PIK3CD, TP53, CDK4, BAX
Thyroid cancer	4.09	10	2.50E-07	3.20E-05	RET, NRAS, MAPK1, PPARG, TP53, BAX, CTNNB1
Chronic myeloid leukemia	2.84	10	2.90E-07	8.20E-06	PIK3CD, BCL2L1, TP53, NRAS, MAPK1, SHC1, CDK4, BAX, TGFB1, RUNX1
Bladder cancer	3.16	10	3.60E-07	1.90E-05	NRAS, MAPK1, TP53, CDK4, THBS1, EGFR
Toxoplasmosis	2.51	10	3.90E-07	6.30E-04	JAK1, HLA-DQA1, SOCS1, MAPK1, MAPK8, MAP2K3, BCL2L1, LAMC1, ITGA6, LDLR, CASP3, GNAI2, PPIF
Transcriptional misregulation in cancer	2.14	10	4.40E-07	2.80E-06	RUNX1, PER2, JUP, CCNA2, CCNT1, CCNT2, RUNX2, HOTA10, IGFBP1, BCL2L1, PPARG, TP53, SP1, PAX7, PAX3, MET, BAX
Small cell lung cancer	2.58	10	5.70E-07	1.90E-05	TP53, BAX, CASP3, CDK4, SKP2, CCNE2, LAMC1, ITGA2, ITGA6, PIK3CD, BCL2L1
Apoptosis	2.4	10	9.60E-07	9.70E-05	TNFRSF10B, CASP3, CASP7, BAX, LMNB2, CAPN2, CTL, BCL2L1, MAPK8, JUN, TP53, PMAIP1, PIK3CD, NRAS, MAPK1
Renal cell carcinoma	2.99	10	9.70E-07	1.00E-06	VHL, MET, GAB1, PIK3CD, NRAS, MAPK1, JUN, RAC1, PAK1
Measles	2.04	10	1.00E-06	1.00E-04	IFNA1, MAPK8, JUN, IFNAR2, JAK1, ADAR, TP53, BAX, BCL2L1, CASP3, CCNE2, CDK4, PIK3CD
Regulation of actin cytoskeleton	2.02	10	1.20E-06	3.60E-05	EGFR, ITGA2, ITGA6, ITGB4, GNAI3, GNGL2, NRAS, PIK3CD, VAV3, MAPK1, RAC1, PAK1, PPP1CC, PPP1R12A, SLC9A1, PIP4K2A, NCKAP1, WASF1, ARPC5, SPATA13
Apoptosis - multiple species	5.04	10	1.20E-06	2.20E-04	CASP7, CASP3, PMAIP1, BCL2L1, BAX, MAPK8, BECN1
Salmonella infection	2.45	10	1.20E-06	1.60E-05	VPS33A, DYNC1H1, DYNC1U1, DYNC2U1, DYNLL2, DCTN4, DCTN5, TUBB3, FYCO1, KIF5B, RAC1, NCKAP1, ARPC5, PAK1, PIK3CD, ARF6, KPN1A1, KPN1A4, MAPK1, MAP2K3, MAPK8, JUN, CTNNB1, TNFRSF10B, BAX, CASP3, CASP7, RHOB
Endometrial cancer	2.98	10	1.40E-06	3.40E-06	EGFR, PIK3CD, NRAS, MAPK1, MLH1, CTNNB1, TP53, BAX
Pertussis	1.99	10	1.50E-06	4.60E-03	CALM3, CASP3, CASP7, MAPK8, MAPK1, JUN, GNAI2
Natural killer cell mediated cytotoxicity	2.4	10	1.50E-06	6.50E-04	VAV3, RAC1, PAK1, MAPK1, PIK3CD, SHC1, NRAS, PPP3CA, PPP3CB, PPP3R1, IFNA1, IFNAR2, TNFRSF10B, CASP3
Arrhythmogenic right ventricular cardiomyopathy	2.56	10	1.90E-06	4.10E-04	ITGA2, ITGA6, ITGB4, DAG1, CTNNB1, ATP2A2, CDH2, JUP, DSG2
Insulin signaling pathway	2.23	10	2.10E-06	7.50E-06	PIK3CD, PPP1CC, PRKAA1, CALM3, FASN, HK1, PRKAA1, SHC1, NRAS, MAPK1, MNK2, SOCS1, SOCS3, MAPK8
AMPK signalling pathway	2.75	10	2.40E-06	4.80E-05	CAB39, PRKAA1, PRKRB2, CREB1, CCNA2, EEF2, FASN, SCD, PPP2R1B, PPP2R2A, PPP2R3C, RAB14, IGFBP1, PIK3CD, PPARG
IL-17 signalling pathway	1.9	10	2.50E-06	8.50E-03	CASP3, JUN, ANAPC5, MAPK8, MAPK1, USP25, SRFBP1, MMP13
Vasopressin-regulated water reabsorption	3.44	10	2.70E-06	5.20E-03	CREB1, DYNC1H1, DYNC1U1, DYNC2U1, DYNLL2, DCTN4, DCTN5
Influenza A	2.42	10	2.70E-06	3.20E-04	DNAJB1, DNAJC3, IFNA1, IFNAR2, JAK1, SOCS3, ADAR, HLA-DQA1, PIK3CD, CDK4, TNFRSF10B, BAX, CASP3, CPSF4, NXT1, MAPK1, XPO1, KPN1A1, KPN1A6
HIF-1 signaling pathway	1.82	10	2.80E-06	2.80E-06	EGFR, IGFBP1, MAPK1, MNK2, PIK3CD, VHL, HMOK1, HK1, ALDOA
PD-L1 expression and PD-1 checkpoint pathway in cancer	2.91	10	2.80E-06	1.70E-05	EGFR, NRAS, MAPK1, JUN, EML4, PIK3CD, JAK1, BATF2, MAP2K3, PPP3CA, PPP3CB, PPP3R1
NOD-like receptor signaling pathway	1.48	10	3.00E-06	3.50E-03	MAPK1, MAPK8, JUN, GABARAPL2, MAP1LC3B, IFNA1, BCL2L1, DHX33, TXNIP, YWHAZ, IFNAR2, JAK1
Prolactin signaling pathway	2.47	10	3.90E-06	2.70E-04	SHC1, PIK3CD, NRAS, MAPK1, MAPK8, SOCS1, SOCS3, SOCS6
C-type lectin receptor signaling pathway	2.14	10	4.70E-06	2.80E-05	NRAS, CALM3, PPP3CA, PPP3CB, PPP3R1, MAPK1, MAPK8, JUN, PAK1, PIK3CD
Oocyte meiosis	2.91	10	4.70E-06	4.70E-06	IGFBP1, CPEB4, MAPK1, YWHAZ, YWHAZ, PPP1CC, CUL1, ANAPC5, CDK20, CCNE2, PPP2R3C, PPP2R1B, CALM3, PPP3CA, PPP3CB, PPP3R1
Chagas disease	1.91	10	4.80E-06	3.60E-03	MAPK8, MAPK1, JUN, PPP2R1B, PPP2R2A, CALR, GNAI2, PIK3CD, TGFB1
Pathogenic Escherichia coli infection	2.34	10	4.80E-06	8.10E-06	RAC1, ARF6, NCKAP1, WASF1, ARPC5, MYO10, GNAI3, TUBB3, NCL, MAPK1, MAPK8, JUN, TNFRSF10B, TM6IM6, BAX, CASP3, CASP7, PAK1, SEC24B, SEC24A, TMED10
Non-alcoholic fatty liver disease	1.85	10	4.80E-06	2.40E-03	SOCS3, PIK3CD, MLNIP, PRKAA1, PPARG, RAC1, MAPK8, ITCH, JUN, BAX, CASP3, CASP7, NDUFB3

Table 3.8: Analysis of KEGG pathway enrichment for target genes of up-regulated miRNAs. From left to right, the table reports the description of the enriched term, the fold enrichment value for the enriched term, the number of iterations that the given term was found to be enriched over all iterations (occurrence), the lowest adjusted-p value of the given term over all iterations, the highest adjusted-p value of the given term over all iterations, and the list of miRNAs target genes.

Toll-like receptor signalling pathway	1.66	10	5.00E-06	4.80E-04	RAC1, PIK3CD, MAPK1, MAP2K3, MAPK8, JUN, IFNA1, IFNAR2
Thyroid hormone signalling pathway	2.32	10	5.00E-06	5.00E-06	NRAS, MAPK1, TP53, NCOA3, MED13, MED14, ATP2A2, CTNNB1, NOTCH2, SLCA1, PIK3CD, TBC1D4, PFKFB2
Parathyroid hormone synthesis, secretion and action	1.85	10	5.10E-06	6.80E-04	GNAI2, SP1, MAPK1, MEK2, CREB1, RUNX2, MMP13, GNAI3, EGFR
Phospholipase D signalling pathway	1.34	10	5.40E-06	1.20E-03	EGFR, GAB1, SHC1, NRAS, PIK3CD, ARF6, PLA2G4A, MAPK1, GNAI3
Nucleocytoplasmic transport	2.29	10	5.40E-06	5.00E-05	NUP133, KPN1, KPN4, KPN4, TNPO1, TNPO3, XPO1, XPO7, XPO7, EEF1A1, NXT1
Fc epsilon RI signalling pathway	2.34	10	5.80E-06	5.80E-06	PIK3CD, RAC1, MAP2K3, MAPK8, VAV3, NRAS, MAPK1, PLA2G4A
cAMP signalling pathway	1.63	10	6.60E-06	6.60E-06	EDNRA, GNAI2, CALM3, MAPK8, VAV3, RAC1, PAK1, PIK3CD, MAPK1, PPP1CC, CREB1, BDNF, JUN, PPP1R12A, ATP2A2, SLCA1
Autophagy - animal	2.18	10	7.60E-06	8.00E-06	IGF1R, PIK3CD, NRAS, MAPK1, BNIP3, PRKAA1, MAPK8, BCL2L1, BECN1, MTMR3, GABARAPL2, MAP1LC3B, LAMP1, CTSL
Fluid shear stress and atherosclerosis	2.19	10	7.80E-06	7.80E-06	CALM3, NFE2L2, HMOX1, GSTM3, MGST2, CTNNB1, PRKAA1, PIK3CD, MAPK8, JUN, CTSL, BMPR2, RAC1, TP53
Cushing syndrome	1.68	10	8.80E-06	4.90E-05	CREB1, RBBP5, CDK4, GNAI2, CCNE2, MAPK1, EGFR, SP1, NCEH1, LDLR, WNT5A, CTNNB1
Growth hormone synthesis, secretion and action	1.82	10	8.90E-06	5.90E-05	GNAI2, CREB1, SHC1, NRAS, MAPK1, PIK3CD, SOCS1, SOCS3, MAP2K3, MAPK8
Axon guidance	1.94	10	1.00E-05	3.00E-04	PPP3CA, PPP3CB, PPP3R1, RAC1, ABLIM1, PAK1, NRAS, MAPK1, GNAI2, PLXNA1, SEMA6A, SEMA6D, MET, WNT5A, PIK3CD, BMPR2
Sphingolipid signalling pathway	2.77	10	1.00E-05	5.50E-05	SPTLC1, SPTLC3, BAX, MAPK8, PPP2R1B, PPP2R2A, PPP2R5C, TP53, PIK3CD, GNAI2, MAPK1, SGPL1, NRAS, RAC1, GNAI3
Coronavirus disease - COVID-19	1.13	10	1.30E-05	5.10E-03	EGFR, JAK1, RPS15A, RPS24, RPL19, PIK3CD, MAPK1, MAPK8, JUN, IFNA1, IFNAR2, ADAR
Yersinia infection	1.44	10	1.60E-05	1.20E-03	RAC1, ARF6, ARPC5, MAPK1, MAP2K3, MAPK8, JUN, PIK3CD, VAV3
Long-term depression	2.61	10	1.70E-05	6.00E-03	PPP2R1B, NRAS, MAPK1, GNAI2, GNAI3, PLA2G4A, IGF1R
Longevity regulating pathway	2.43	10	1.70E-05	1.70E-05	IGF1R, NRAS, PIK3CD, CREB1, SOD2, CAT, PRKAA1, BAX, TP53, PPARG
Wnt signalling pathway	1.59	10	1.70E-05	1.80E-05	WNT5A, CTNNB1, NLK, JUN, TP53, TBL1XR1, CUL1, RAC1, MAPK8, PPP3CA, PPP3CB, PPP3R1
Signaling pathways regulating pluripotency of stem cells	2.18	10	1.90E-05	1.10E-04	UFR, JAK1, SOX2, MAPK1, PIK3CD, NODAL, BMPR2, ID4, WNT5A, CTNNB1, NRAS, IGF1R, RIF1, ZFXD3
Tight junction	1.61	10	2.40E-05	5.60E-04	RAC1, PPP2R1B, PPP2R2A, NEDD4L, MAPK8, JUN, CDK4, YBX3, PCNA, RUNX1, ARPC5, PRKAA1
B cell receptor signalling pathway	2.74	10	2.60E-05	1.80E-04	VAV3, RAC1, PPP3CA, PPP3CB, PPP3R1, RASGRP3, NRAS, MAPK1, JUN, PIK3CD
Oxytocin signalling pathway	2.02	10	3.10E-05	3.50E-04	NRAS, MAPK1, PLA2G4A, JUN, EEF2, CALM3, PPP3CA, PPP3CB, PPP3R1, PRKAA1, PPP1CC, PPP1R12A, GNAI2, EGFR
Insulin resistance	2.04	10	3.50E-05	3.50E-05	PPP1CC, PIK3CD, SOCS3, MAPK8, TBC1D4, PRKAA1, PTPA, PPARGC1B, MLXIP, CREB1
Diabetic cardiomyopathy	1.27	10	3.60E-05	3.60E-05	ATP2A2, SP1, RAC1, MAPK8, PTPA, PPP1CC, PIK3CD, TBC1D4, NDUFB3, PPIF, TGFBR1
cGMP PKG signalling pathway	1.88	10	3.70E-05	3.70E-05	EDNRA, GNAI3, PPP3CA, PPP3CB, PPP3R1, MEK2, ATP2A2, PPP1R12A, PPP1CC, CALM3, GNAI2, PPIF, CREB1, MAPK1
Leishmaniasis	1.4	10	4.10E-05	5.10E-03	MAPK1, JUN, JAK1, HLA-DQA1, EEF1A1
Parkinson disease	1.48	10	4.30E-05	1.40E-04	GNAI2, UBA1, UBE2L3, CALM3, HSPA5, SLC11A2, NDUFB3, PINK1, PPIF, CASP3, BCL2L1, BAX, KIF5B, TUBB3, TP53, NFE2L2, MAPK8
Ribosome biogenesis in eukaryotes	1.99	10	4.40E-05	4.40E-03	UTP15, WDR75, UTP6, UTP14A, BMS1, RPP40, GNL3L, XPO1, NXT1
Type II diabetes mellitus	3.29	10	4.60E-05	1.70E-03	PIK3CD, MAPK1, SOCS1, SOCS3, MAPK8, HK1, PKM
Nucleotide excision repair	2.35	10	5.00E-05	9.40E-03	RAD23B, GTF2H1, GTF2H2, RPA2, PCNA
Gap junction	1.26	10	5.10E-05	2.90E-03	GNAI2, EGFR, NRAS, MAPK1, TUBB3
Tuberculosis	1.57	10	5.40E-05	3.00E-03	CASP3, BAX, JAK1, CREB1, HLA-DQA1, MAPK1, MAPK8, IFNA1, CALM3, LAMP1, PPP3CA, PPP3CB, PPP3R1
VEGF signalling pathway	2.93	10	5.90E-05	6.90E-05	NRAS, MAPK1, PLA2G4A, PPP3CA, PPP3CB, PPP3R1, PIK3CD, RAC1
Chemical carcinogenesis - receptor activation	1.36	10	6.10E-05	3.90E-04	PIK3CD, NRAS, MAPK1, JUN, CREB1, GNAI2, EGFR, NOTCH2, KPN1, KPN4, KPN4, GSTM3, MGST2
Th1 and Th2 cell differentiation	2.14	10	7.50E-05	8.20E-03	JAK1, HLA-DQA1, PPP3CA, PPP3CB, PPP3R1, MAPK1, MAPK8, JUN, NOTCH2
Alcoholism	2.45	10	8.60E-05	3.20E-04	GNAI2, GNG2, GNG5, GNG12, ADR2B, CREB1, BDNF, SHC1, NRAS, MAPK1, CALM3, PPP1CC
Dopaminergic synapse	2.29	10	8.80E-05	8.80E-05	CALM3, PPP3CA, PPP3CB, CREB1, MAPK8, PPP1CC, GNAI2, GNG2, GNG5, GNG12, PPP2R1B, PPP2R2A, PPP2R5C, KIF5B
Chemokine signalling pathway	1.26	10	9.70E-05	3.70E-03	GNAI2, SHC1, NRAS, MAPK1, PIK3CD, VAV3, RAC1, PAK1, GNG2, GNG5, GNG12
Hippo signalling pathway	1.43	10	1.10E-04	1.10E-04	PPP2R1B, PPP2R2A, PPP1CC, TGFBR1, BMPR2, WNT5A, YWHAE, YWHAQ, CTNNB1, SOX2
Ferroptosis	3.69	10	1.20E-04	8.70E-03	SLC7A11, GCLC, ACSL4, TP53, SLC11A2, MAP1LC3B, HMOX1
Protein export	4.91	10	1.20E-04	2.50E-04	SEC62, HSPA5, SPIC3, SEC11C, SEC11A
Phagosome	2.28	10	1.30E-04	1.60E-03	VAMP3, STX7, HLA-DQA1, ATP6V1B2, ATP6V1C1, DYNC1H1, DYNC1U1, TUBB3, LAMP1, CTSL, CALR, ITGA2, THBS1, THBS2, OLRI, RAC1
SNARE interactions in vesicular transport	3.28	10	1.70E-04	8.80E-04	STX7, STX16, USE1, VAMP3, YKT6
Acute myeloid leukemia	2.26	10	2.00E-04	1.50E-03	PIK3CD, NRAS, MAPK1, RUNX1, PER2, CCNA2, JUP
Long-term potentiation	2.26	10	2.40E-04	1.30E-03	PPP1CC, PPP3CA, PPP3CB, PPP3R1, CALM3, NRAS, MAPK1
Prion disease	1.35	10	2.50E-04	1.50E-03	LAMC1, BAX, HSPA5, PPP3CA, PPP3CB, PPP3R1, PPIF, CASP3, NDUFB3, PIK3CD, RAC1, MAPK8, MAPK1, CREB1, KIF5B, TUBB3
Necroptosis	1.46	10	3.10E-04	3.40E-02	PPID, MAPK8, PLA2G4A, CAPN2, TNFRSF10B, IFNA1, IFNAR2, JAK1, BAX
Apelin signalling pathway	2.24	10	3.30E-04	1.20E-03	GNAI2, GNG2, GNG5, GNG12, CALM3, SLCA1, NRAS, MAPK1, BECN1, MAP1LC3B, PRKAA1, GNAI3, MEK2, TGFBR1

Table 3.9: Analysis of KEGG pathway enrichment for target genes of up-regulated miRNAs. From left to right, the table reports the description of the enriched term, the fold enrichment value for the enriched term, the number of iterations that the given term was found to be enriched over all iterations (occurrence), the lowest adjusted-p value of the given term over all iterations, the highest adjusted-p value of the given term over all iterations, and the list of miRNAs target genes.

3.4 Discussion

The aim of this study was to investigate the miRNome profile of circulating EVs obtained from patients at different NAFLD stages using high-throughput sequencing. First, the presence of EVs was confirmed with a super-resolution microscopy assay showing the classical exosomal markers (CD9, CD63, and CD81).

Using the NGS approach, eight miRNAs (miR-27b-3p, miR-30a-5p, miR-375-3p, miR-122-5p, miR-103a-3p, miR-1301-3p, let-7d-5p, and let-7f-5p) resulted in being significantly altered according to both the NAFLD stage (NAFL or NASH diagnosis) and the presence of significant fibrosis.

This miRNAs panel includes signals observed to be up-regulated (miR-27b-3p, miR-30a-5p, miR-122-5p, and miR-375-3p) and others found to be down-regulated (miR-103a-3p, miR-1301-3p, let-7d-5p, and let-7f-5p) in serum EVs of patients with significant fibrosis and in those with NASH.

Among the up-regulated miRNAs, the miR-122-5p is of particular interest. It is a liver-specific miRNA accounting for more than 70% of the entire hepatic miRNAs pool [111]. This signal is known to be involved in the regulation of cholesterol metabolism [112], and showed different behavior in chronic viral infections, favoring HCV replication [113], and inhibiting HBV transcription [114]; furthermore, miR-122-5p acts as a tumor suppressor against HCC [115].

Given its unique characteristics, miR-122-5p has been widely studied in chronic liver diseases both as a diagnostic/prognostic biomarker [116] [117], as well as a possible therapeutic target [118] [119].

Regarding the NAFLD, pre-clinical studies showed promising results on the possible exploitation of miR-122-5p as a therapeutic target. Indeed, several *in vitro* and in mouse model studies demonstrated that circulating FFAs enhanced miR-122 promoter activity via retinoid-related orphan receptor alpha ($ROR\alpha$), increasing miR-122 expression and secretion from the liver; in turn, miR-122 promoted β -oxidation by inhibiting triglycerides synthesis both in the liver and in the adipose tissue [120]. Subsequently, it has been shown that targeting the miR-122 promoter through a $ROR\alpha$ agonist led to increased liver miR-122 expression and concomitant reduction of liver steatosis, inflammation, and fibrosis, together with the improvement of IR in the mouse model of NASH [121].

In agreement with these findings, in this study, it was observed that miR-122-5p levels in serum EVs were up-regulated in patients with significant fibrosis and in those with NASH; miR-122-5p might qualify as a valuable candidate for the development of target therapies for the cure of NAFLD.

Besides the regulation of hepatic lipid metabolism, it has been reported that ectopic over-expression of miR-122 is also involved in hepatic differentiation of adipose tissue-derived stem cells (ADSCs) [122]. A recent *in vitro* study showed that the combination of miR-122 overexpression and let-7f silencing led to a more rapid and spontaneous differentiation of ADSCs toward functional hepatocyte-like cells [122]. Indeed, let-7f targets the hepatocyte nuclear factor 4 alpha ($HNF4\alpha$), a critical transcription factor involved in the regulation of several genes responsible for liver development [123]; in turn, $HNF4\alpha$ enhances the effect of miR-122 on hepatic differentiation of ADSCs [122]. In agreement with these results, this study highlighted a parallel down-regulation of miRNAs belonging to let-7 family, namely let-7d-5p and let-7f-5p, in patients with more advanced liver disease; in particular, let-7f-5p showed a strong inverse correlation with adipose tissue IR and SFAs.

Taken together, these results suggest an intertwined connection of miR-122 and let-7 in the counter-regulation of lipid derangements to promote the repair of liver damage in patients with severe NAFLD.

Also, the miR-375-3p resulted to be up-regulated in significant liver fibrosis and NASH. Previous studies showed that miR-375 was highly expressed in pancreatic islets of humans and mice, where it was involved in the regulation of glucose-mediated insulin secretion by β cells [124]. Moreover, it has been shown that miR-375 inhibition leads to increased expression of adiponectin, reduction of lipid accumulation, and proinflammatory cytokine expression in palmitic acid-induced HepG2 cells [125].

Consistently, a recent study showed that circulating levels of miR-375 increased stepwise in patients with NAFL as compared to healthy controls and further increased in patients with NASH [126].

It is important to note that the focus of this study was to investigate the expression of miRNAs packaged in EVs, mainly exosomes, thus restricting the analysis to those miRNAs actively released into circulation. On the other hand, the exclusion of miRNAs passively released into the bloodstream following cell death allows to obtain an unbiased insight into the pathophysiology of liver damage and confirms a possible mechanistic role of miR-122 in the setting of deranged lipid metabolism in NAFLD with significant liver damage.

Considering the pathway analysis for target genes of up-regulated miRNAs, it identified several genes involved in the MAPK signaling pathway. In the liver, increased FFA levels can generate reactive oxygen species that contribute to the activation of MAPKs, which in turn can impair hepatic gluconeogenesis and FFA synthesis [127].

Regarding down-regulated miRNAs, target genes pointed out various signaling pathways involved in the regulation of liver homeostasis, suggesting that the down-regulated miRNAs may act as tumor suppressor genes. For instance, data from this study showed that the target genes belong to Hippo signaling, a key regulator in liver metabolism and cell proliferation [128]. Indeed, the metabolic activities of hepatocytes, including glycolysis, bile synthesis, glutamine-genesis, and ammonia clearance, are strictly linked to the Hippo signaling pathway. Its alteration has been associated with metabolic disorders, such as T2DM, NAFLD, cardiovascular diseases, and cancer.

In conclusion, in a cohort of well-characterized patients with NAFLD, a comprehensive profile of miRNAs contained within EVs provided further evidence of the involvement of several miRNAs in the pathophysiological mechanisms underlying liver damage in NAFLD. In particular, miR-122-5p, let-7f-5p, and miR-375-3p emerged to be valuable candidates for the development of miRNA-targeting drugs for the treatment of NASH and liver fibrosis. These preliminary data must be confirmed in a larger cohort of patients to confirm their key role in the pathology.

Chapter 4

Untargeted analysis to identify metabolites related to non-alcoholic fatty liver disease (NAFLD) and non-alcoholic steatohepatitis (NASH) condition

4.1 Introduction

In recent years, numerous studies regarding pathophysiology, genetics, transcriptomics, metabolomics and lipidomics have improved the scientific understanding of NAFLD [129] [130].

Metabolomics and lipidomics approaches have provided insightful evidence of altered metabolic pathways in NAFLD and NASH, mainly involving alterations in several aspects of amino acid, bile acid and lipid metabolism, including circulating fatty acids, branched-chain amino acids (BCAAs), phospholipids, triglycerides (TGs) [131] [132] [133] [134].

These approaches have a key role also in clinical practice. Indeed, although liver histology is currently the reference standard for the assessment of NASH progression and treatment efficacy, there is still a great demand to optimize and validate other biomarkers that predict NASH progression and/or response to treatment because repeated liver biopsies are difficult to perform in patients with NASH due to several limitations (health risks, interpretations and cost) [135].

In addition, compared to traditional biomarkers, metabolomics can more accurately reflect the current metabolic state of the body, allowing the profile of patients' disease progression by potential biomarkers, not just general parameters [136].

Regarding metabolomics, this method aims to assess metabolism and identify downstream functions of genes and proteins. The chemical and biological diversity of the metabolome is enormous [137]. For this reason, several analytical strategies, with an extensive array of techniques, to cover a wide range of metabolites while ensuring data output quality are required. Metabolomics can be split into targeted metabolomics and untargeted metabolomics based on the research objectives. The first one involves the multiple analyses

of known metabolites, including absolute quantitative and qualitative analysis of the test substance using standard substances. On the other hand, untargeted metabolomics aims to obtain as much information on substances as possible. Overall, targeted metabolomics has higher accuracy and sensitivity than untargeted metabolomics, but the latter has a larger coverage of substances.

In recent years, metabolomics approaches have been widely applied both in clinical studies with the aim of understanding the changes of metabolites and metabolic pathways before and after the onset of the disease and in basic research using animal models [138].

A variety of possible NAFLD-related biomarkers were identified; however, the clinical universality of these biomarkers remains to be further verified.

Here, with the aim of identifying metabolites differentially expressed between NAFL and NASH patients, a comprehensive untargeted metabolomics analysis was performed in a cohort of well-histologically characterized NAFLD patients.

4.2 Methods

4.2.1 Cohort of study

This study was conducted on the 154 patients already described in Chapter 2.

4.2.2 Untargeted metabolomics analysis

Metabolomic analyses of serum samples were prepared as previously reported here [139]. Briefly, serum metabolites were obtained with protein precipitation with cold acetonitrile/isopropanol/water, followed by derivatization with methoxamine and BSTFA. Small molecules were analyzed using a LECO Pegasus BT 4D GCXGC/TOFMS instrument (Leco Corp., St. Josef, MI, USA) equipped with a LECO dual-stage quad jet thermal modulator. The GC part of the instrument was an Agilent 7890 gas chromatograph (Agilent Technologies, Palo Alto, CA, USA) equipped with a split/splitless injector. The first dimension column was a 30 m Rxi-5Sil (Restek Corp., Bellefonte, PA, USA) MS capillary column with an internal diameter of 0.25 mm and a stationary phase film thickness of 0.25 μ m, and the second dimension chromatographic column was a 2 m Rxi-17Sil MS (Restek Corp., Bellefonte, PA, USA) with a diameter of 0.25 mm and a film thickness of 0.25 μ m. High-purity helium (99.9999%) was used as the carrier gas with a flow rate of 1.4 mL/min. 1 μ L of sample was injected in splitless mode at 250 °C. The temperature program was as follows: the initial temperature was 100 °C for 2 min, then ramped 20°C/min up to 330 °C and then held at this value for 2 min. The secondary column was maintained at +5°C relative to the GC oven temperature of the first column. The programming rate was the same for the two columns. Electron impact ionization was applied (70 eV). The ion source temperature was set at 250 °C; the mass range was 25 to 550 m/z with an extraction frequency of 32 kHz. The acquisition rates were 200 spectra/s. The modulation period for the bi-dimensional analysis was 4 s for the entire run. The modulator temperature offset was set at +15 °C relative to the secondary oven temperature, while the transfer line was set at 280 °C. The chromatograms were acquired in total ion current mode. Peaks with signal-to-noise (S/N) values lower than 500.0 were rejected. ChromaTOF version 5.31 was used for the raw data processing. Mass spectral assignment was performed by matching with NIST MS Search 2.3 libraries and the FiehnLib. An in-house library of standards was also used for small molecule identification.

4.2.3 Statistical analysis

All statistical analyses were conducted using the open-source software Rv4.1.0. Cross-validation analysis was conducted by mixing and splitting the dataset randomly each time for a total of 10 cycles; each run consisted of a generalized linear model analysis done on 80% of the whole sample size, with the option of maintaining comparable proportions of NAFL and NASH patients.

Subsequent analysis was focused on those signals that showed a nominal p-value ≤ 0.05 and a Fold Change (FC) higher than 1.3 or lower than 0.769 in at least 8 out of 10 runs. All the analyses were adjusted for sex, age and BMI. Multiple regression models were done with glm R function and barplots with ggplot2 R function. MetaboAnalyst 5.0 software [140] was used for enrichment analysis.

4.3 Results

From the analysis of serum samples from 154 NAFLD patients (46 NAFL and 108 NASH), a total of 380 small molecules were identified. A differential expression analysis on small molecule abundances was performed, using a model adjusted for gender, age and BMI, to investigate particular metabolites correlated to the NAFL or NASH condition. As is observed in the Volcano plot in Figure 4.1, although all metabolites showed an FDR value >0.05 , at least 34 small molecules reported a nominal p-value <0.05 .

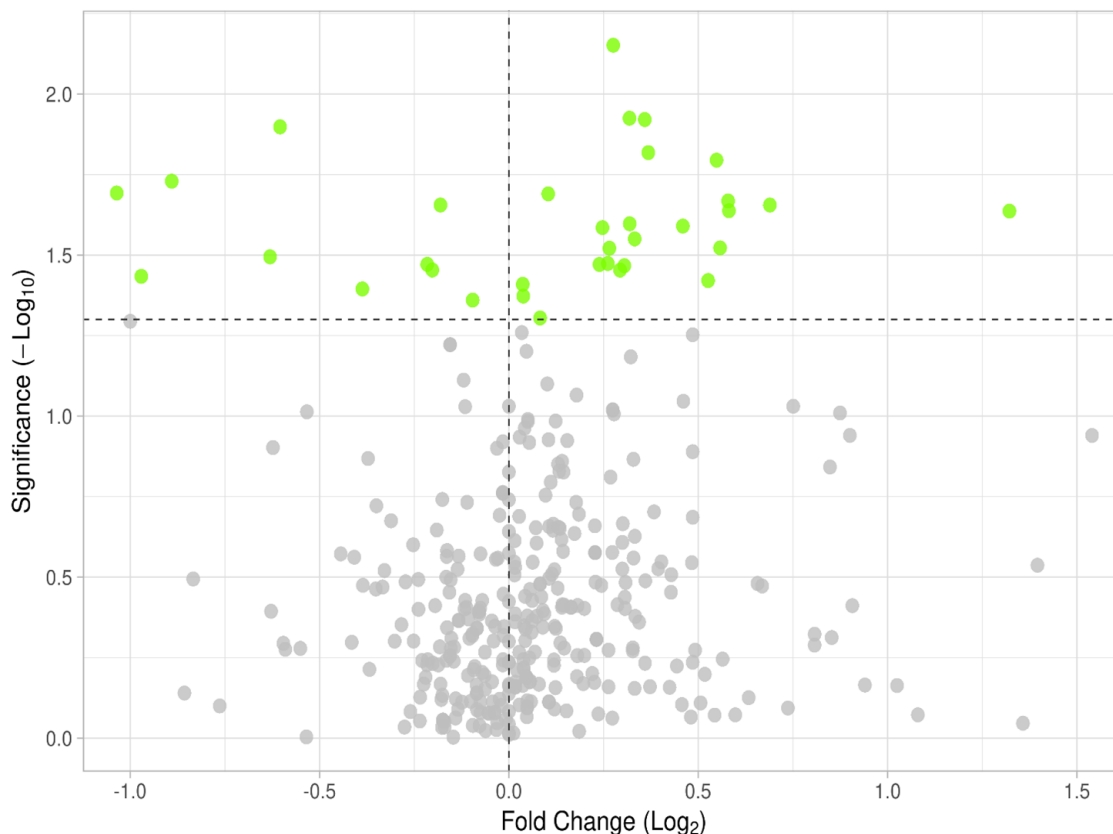


Figure 4.1: Volcano plot depicting differentially expressed metabolites according to the diagnosis of NASH. Scattered points represent metabolites, x-axis is the log₂ Fold change, and y-axis represents -log₁₀ of adjusted p-value. Green dots indicate significantly deregulated metabolites in NASH analysis.

4.3.1 Cross-Validation analysis

A cross-validation analysis was performed to provide a more robust evaluation of the model. This approach works by splitting the dataset into multiple training and test sets and running the evaluation multiple times. In this study, the initial dataset composed of data from 154 NAFLD patients was split 10 times, selecting 80% of the sample each time as training and 20% as test sets, maintaining the proportions between the two categories (NAFL and NASH patients) and used for a differential expression analysis using a model adjusted for gender, age and BMI. As reported in Figure 4.2, a total of 8 small molecules

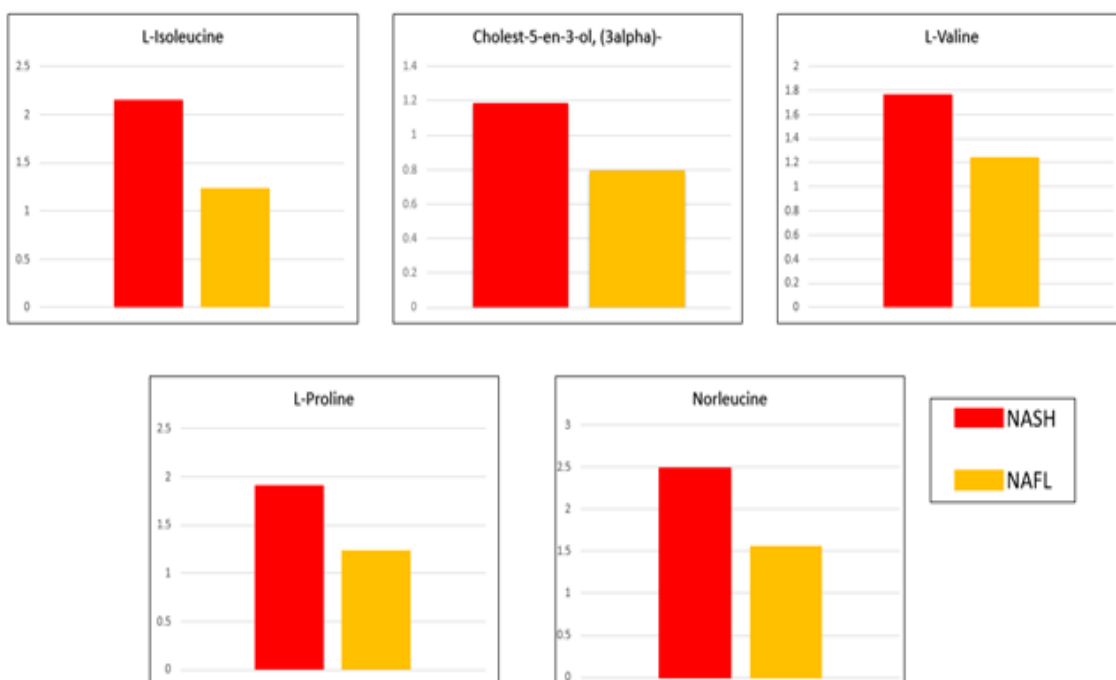


Figure 4.3: The Barplots show the 5 metabolites with higher abundance in NASH patients compared to the NAFL Patients with a $p\text{-value} < 0.05$ and $1.3 < FC < 0.769$ in the cross-validation analysis. The average concentration observed was reported in the graph.

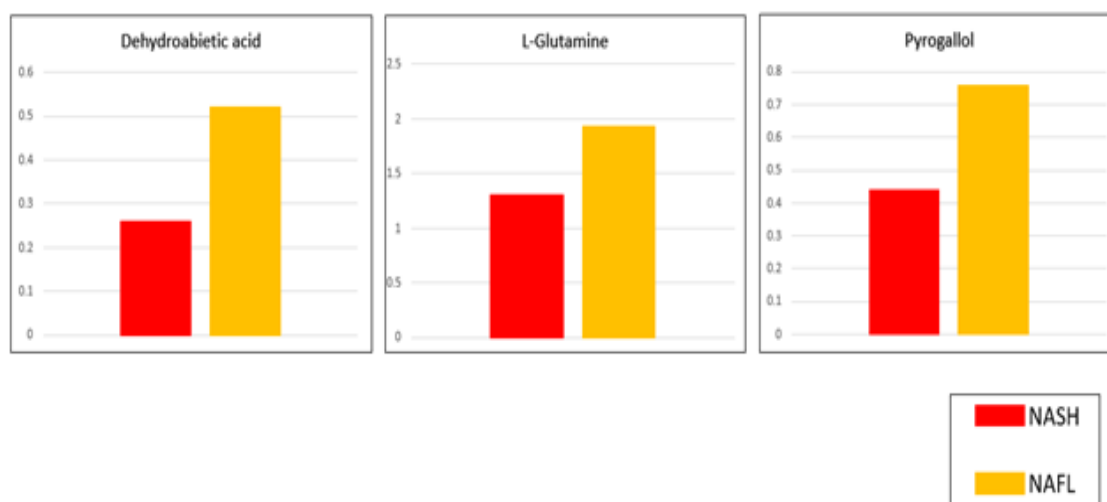


Figure 4.4: The barplots show the 3 metabolites with higher abundance in NAFL patients compared to the NASH Patients with a $p\text{-value} < 0.05$ and $1.3 < FC < 0.769$ in the cross-validation analysis. The average concentration observed was reported in the graph.

4.3.2 Enrichment analysis

Using the MetaboAnalyst 5.0 tool [141] the 8 small molecules obtained from the cross-validation analysis were used for metabolites set enrichment analysis (MSEA) and path-

ways analysis.

Regarding the MSEA, a Over Representation analysis (ORA) enrichment algorithm and a metabolite set library including 99 metabolite sets based on normal human metabolic pathways were used. As reported in Figure 4.5, 2 metabolite sets were selected among the top involved: the Valine, Leucine and Isoleucine Degradation set (p-value= 0.02) and the Phenylacetate Metabolism set (p-value=0.04).

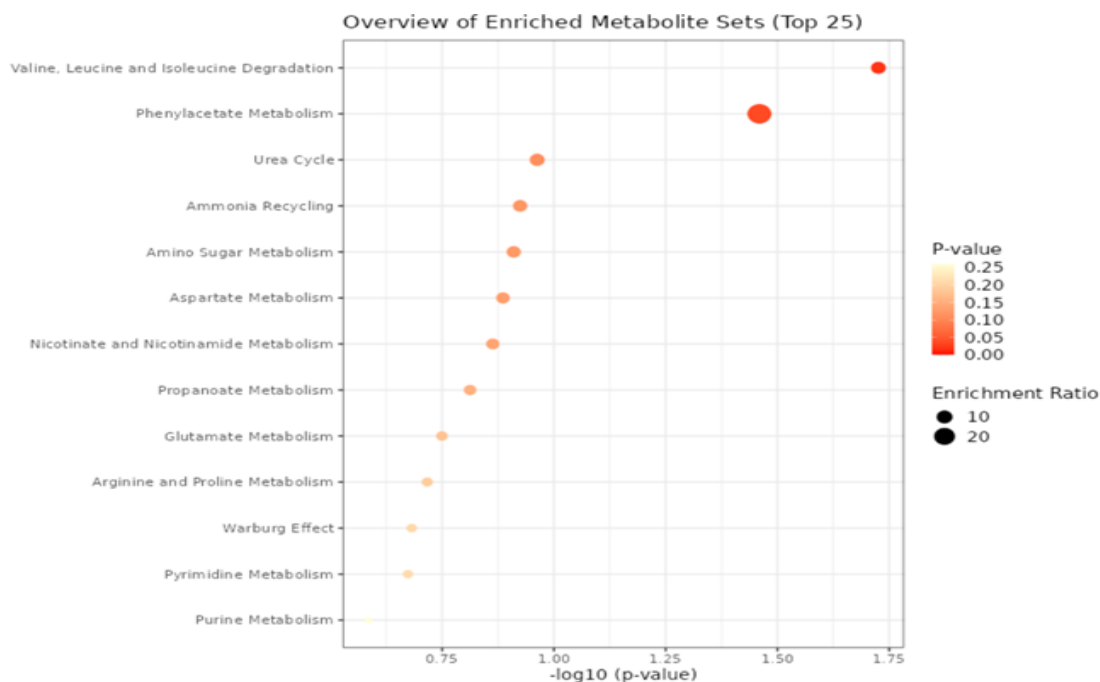


Figure 4.5: Summary Plot for Over Representation Analysis (ORA) from MSEA analysis using MetaboAnalyst 5.0. In the Y-axis, the top metabolite sets are reported; in the X-axis, the $-\log_{10}(\text{p-value})$ was indicated.

In the pathways analysis, the Homo sapiens KEGG pathway library and Fisher's Exact Test method were used. The results are summarised in Table 4.1, in which is reported that 2 pathways show an $\text{FDR} < 0.05$, the Aminoacyl-tRNA biosynthesis ($\text{FDR} = 6.82 \times 10^{-5}$) and Valine, leucine and isoleucine biosynthesis ($\text{FDR} = 5.85 \times 10^{-3}$). (Figures 4.6 and 4.7).

	Total	Expected	Hits	Raw p	-log10(p)	Holm adjust	FDR	Impact
Aminoacyl-tRNA biosynthesis	48	0.12	4	8.12E-07	6.09E+00	6.82E-05	6.82E-05	0.00
Valine, leucine and isoleucine biosynthesis	8	0.02	2	1.39E-04	3.86E+00	1.16E-02	5.85E-03	0.00
Valine, leucine and isoleucine degradation	40	0.10	2	3.77E-03	2.42E+00	3.09E-01	1.06E-01	0.00
D-Glutamine and D-glutamate metabolism	6	0.02	1	1.54E-02	1.81E+00	1.00E+00	2.59E-01	0.00
Nitrogen metabolism	6	0.02	1	1.54E-02	1.81E+00	1.00E+00	2.59E-01	0.00
Arginine biosynthesis	14	0.04	1	3.57E-02	1.45E+00	1.00E+00	4.99E-01	0.00
Pantothenate and CoA biosynthesis	19	0.05	1	4.82E-02	1.32E+00	1.00E+00	5.78E-01	0.00
Alanine, aspartate and glutamate metabolism	28	0.07	1	7.04E-02	1.15E+00	1.00E+00	7.39E-01	0.11
Glyoxylate and dicarboxylate metabolism	32	0.08	1	8.01E-02	1.10E+00	1.00E+00	7.41E-01	0.00
Arginine and proline metabolism	38	0.10	1	9.46E-02	1.02E+00	1.00E+00	7.41E-01	0.08
Pyrimidine metabolism	39	0.10	1	9.70E-02	1.01E+00	1.00E+00	7.41E-01	0.00
Purine metabolism	65	0.17	1	1.58E-01	8.02E-01	1.00E+00	1.00E+00	0.00

Table 4.1: The table shows the detailed results from the pathway analysis. Since many pathways were tested at the same time, the statistical p values from enrichment analysis are further adjusted for multiple tests. In particular, the Total is the total number of compounds in the pathway; the Hits is the actually matched number from the uploaded data; the Raw p is the original p value calculated from the enrichment analysis; the Holm p is the p-value adjusted by Holm-Bonferroni method; the FDR p is the p-value adjusted using False Discovery Rate; the Impact is the pathway impact value calculated from pathway topology analysis.

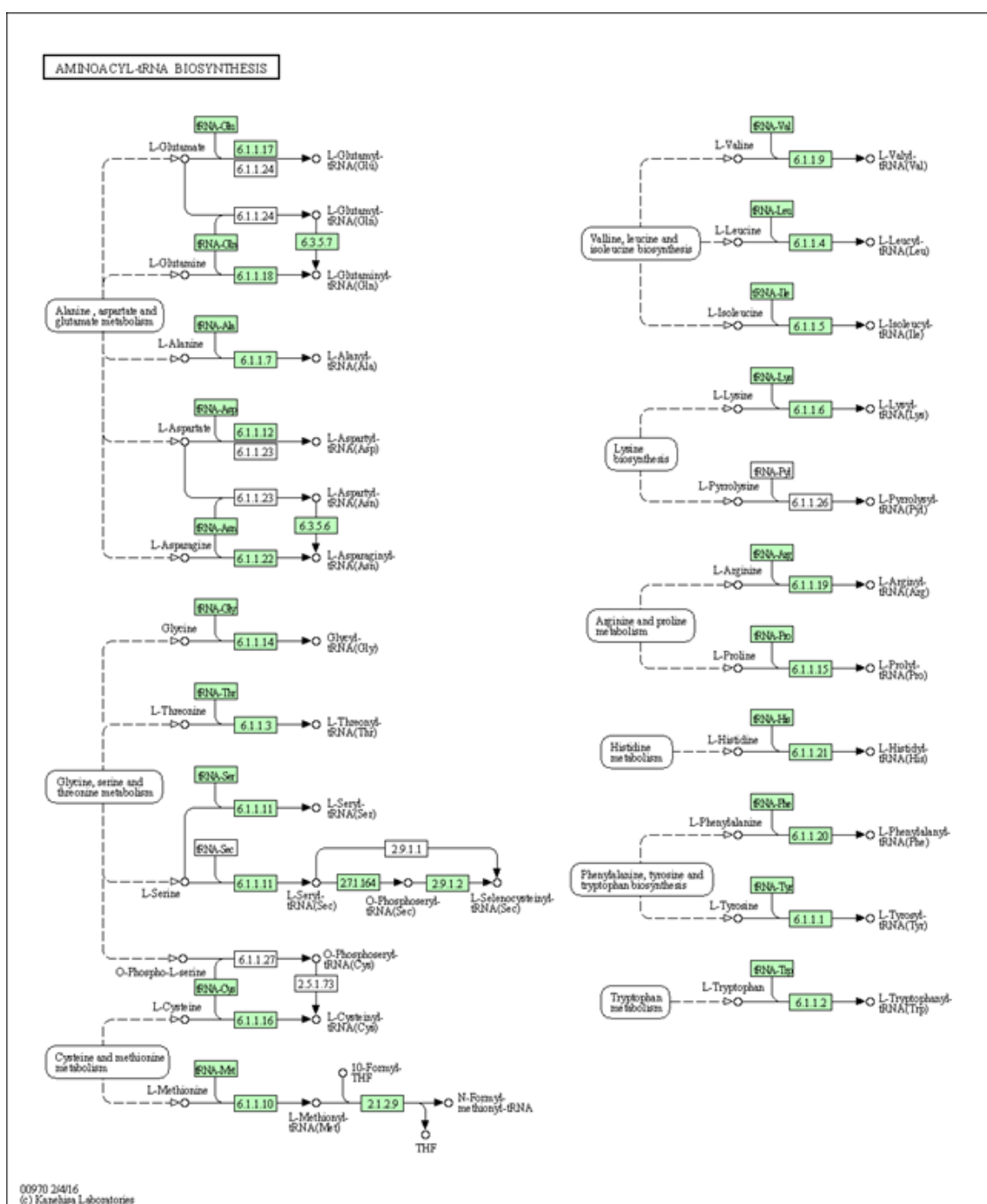


Figure 4.6: The graphic representations of the Aminoacyl-tRNA biosynthesis in the Homo sapiens (human) KEGG database.

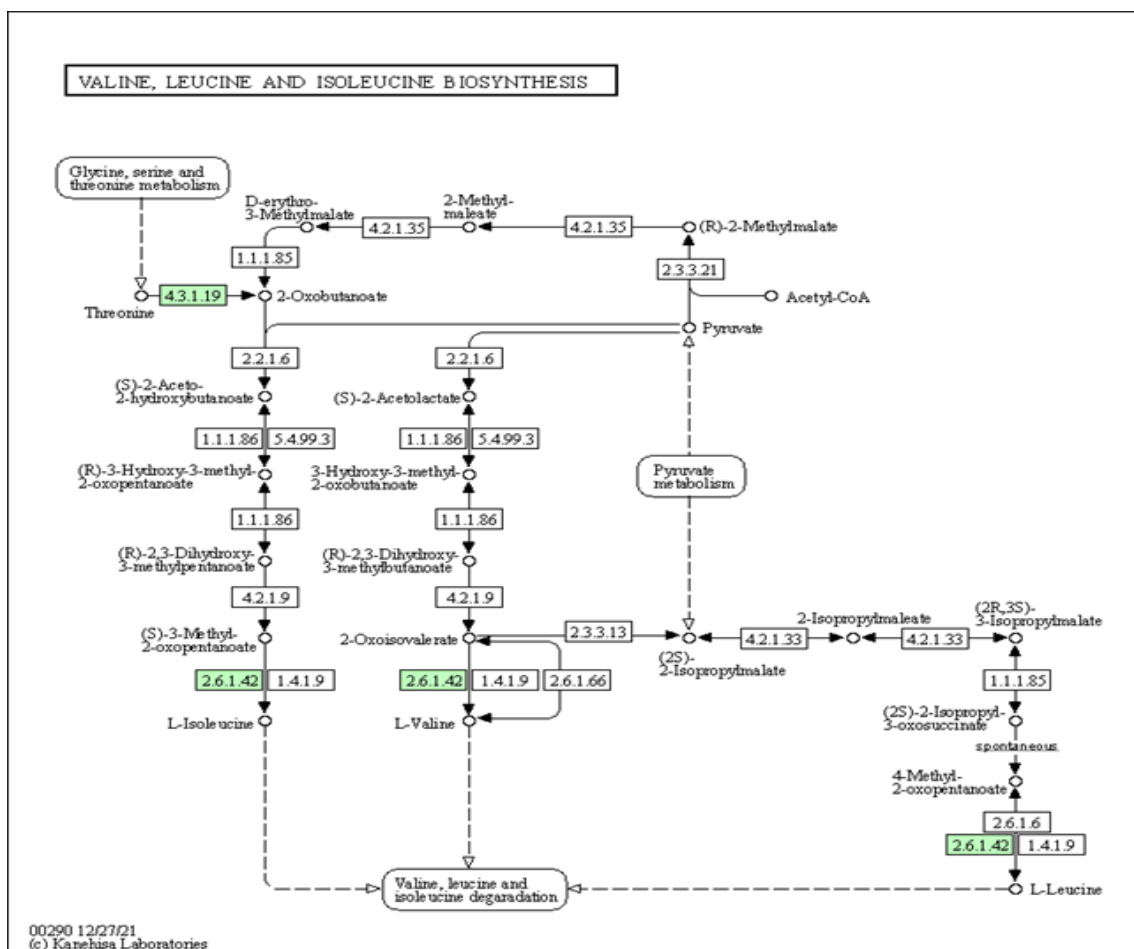


Figure 4.7: The graphic representations of the Valine, leucine and isoleucine biosynthesis - Homo sapiens (human) KEGG database.

4.4 Discussion

In this study, serum samples of 154 very well histologically NAFLD patients were used to perform an untargeted metabolomics analysis to identify metabolites that are differentially expressed in NAFL or NASH patients. A total of 380 small molecules were identified, analyzing all samples. Although in the literature, most studies on the metabolome are conducted on plasma [142] [143], numerous experiments have shown that plasma and serum offer similar analytical opportunities, suggesting that studies performed on plasma can be compared with those on serum and vice versa [144].

Here, a differential expression analysis was performed to find metabolites specifically associated with NASH or NAFL conditions. Following adjustment for age, gender and BMI, 34 metabolites were observed to be differentially expressed according to the diagnosis of NASH (nominal p-value<0.05), but no small molecules emerged with a significant FDR value.

To give robustness to the study, a cross-validation analysis (splitting the dataset in training (80%) and test set (20%) using a 10-times repeated k-fold validation) was performed. The study was focused on the 8 metabolites, including L-isoleucine, Cholest-5-en-3-ol,

(3 α)-, L-Valine, L-proline, Norleucine, Dehydroabiatic acid, L-Glutamine and Pyrogallol emerged to be among the top significant deregulated metabolites at least 8 times out 10 and with a $1.3 < FC < 0.769$.

Among these 8 metabolites, L-isoleucine (Mean FC=1.4; p-value=0.014), Cholest-5-en-3-ol, (3 α)- (Mean FC=1.5; p-value=0.004), L-Valine (Mean FC=1.31; p-value=0.002), L-proline (Mean FC=1.33; p-value=0.02) and Norleucine (Mean FC=1.32; p-value=0.02) resulted to be more abundant among NASH patients compared to NAFL ones, on the other hand Dehydroabiatic acid (Mean FC=0.5; p-value=0.005), L-Glutamine (Mean FC=0.66; p-value=0.02) and Pyrogallol (Mean FC=0.5; p-value=0.02) emerged as more abundant in NAFL patients compared to NASH individuals.

The higher abundance of the 4 small molecules in NASH individuals confirmed what already emerged in the literature indeed several studies on plasma or serum metabolites showed the associations between an increase of L-Valine, L-Isoleucine and L-Proline with more advanced stages of NAFLD [145] [146] [147].

As already reported, the increase in essential amino acids suggests a higher rate of protein buildup, which can be indicative of an increased rate of catabolic processes, which, in turn, may be related to inflammation and disease progression. This increase in the more advanced stage of the disease suggests that these changes may be a late event in the evolution of NASH and may be modulated by other factors, such as cytokines and inflammation, in addition to insulin resistance [148].

In addition, among the branched-chain amino acids, isoleucine has been associated with gluconeogenesis, insulin resistance and increased risk of developing type 2 diabetes, as well as cardiometabolic risk [149] [150].

Interestingly, the lower abundance of Dehydroabiatic acid in the advanced NAFLD stage was found confirmation in the literature. This is a natural compound obtained from pinewood, and is a PPAR α/γ dual agonist that reduces lipid accumulation and promotes mitochondrial β -oxidation. Indeed, in a recent in vitro study, it was demonstrated that as a dual-PPAR- α/γ and PPAR- γ partial agonist, it can attenuate Insulin Resistance and hepatic steatosis induced by the HFD-consumption in mice [151].

Regarding the enrichment analysis, despite the limited number of metabolites used for the analyses performed with the MetaboAnalyst tool, interesting results were observed. The pathways enrichment analysis highlighted that the 8 metabolites are mainly associated with the aminoacyl-tRNA biosynthesis pathway. This finding is in accordance with data present in literature; indeed, in a mice model study, it was demonstrated that this pathway, together with nitrogen metabolism, lipid metabolism, glyoxylate and dicarboxylate metabolism, and amino metabolism, were disturbed during progression from control to NAFL and from NAFL to NASH [152].

In summary, evidence emerged from this work confirmed several hypotheses that recently emerged from different clinical or in vitro studies.

Together, all data obtained from this study indicated that serum untargeted metabolomics could represent an appropriate and useful instrument to discover the metabolomic profiles that are involved in the different forms of NAFLD. In addition, exploring the human metabolome is precious for understanding pathophysiological processes and also searching for new diagnostic and prognostic biomarkers that can accelerate drug discovery and assess the impact of lifestyle and other factors.

Despite several strengths, including the use of a histologically and clinically well-characterized cohort, this is still a study to be considered preliminary due to a series of limitations.

First of all, the small number of patients analyzed and the important imbalance between the two categories analyzed. Another limitation is the lack of prospective design of the study. The disease progression was evaluated among different individuals recruited at different disease stages. Following the same patients during the disease progression could clearly evidence the mechanism involved in these events.

Chapter 5

Conclusions

In this thesis, different aspects of NAFLD were studied in a single cohort of patients. Thanks to the use of the NGS and mass spectrometry (MS) tools, it was possible to investigate and obtain a large amount of data to perform comprehensive discovery analyses. In Chapter 2, the sequencing of the entire exome of 154 NAFLD patients, including 108 histologically NASH diagnosed, revealed novel uncommon potential risk alleles emerged to be mainly correlated with the aggressive form of the disease, suggesting the presence of different genetic components implicated in the different stages. In addition, starting from the different genetic landscape among the NAFL and NASH patients, it was possible to identify several pathways, including the Cilium and Lysosome, already described to be involved in the NAFLD pathogenesis, emerging to be mostly correlated with the NASH. In Chapter 3, the different EVs-MiRNAs profiles of 228 patients were investigated according to the NAFLD stage or to the fibrosis grade. A panel composed of 8 signals, significantly deregulated, emerged from the two different investigations. This preliminary study provides further evidence of the involvement of several miRNAs in the pathophysiological mechanisms underlying NAFLD and suggests some of them not only as prognosis biomarkers but also as possible candidates for the development of miRNA-targeting drugs for the treatment of NASH and liver fibrosis.

In Chapter 4, an untargeted metabolomics analysis was conducted to find the different molecules characterizing the NAFL and NASH stages. Using this approach, 380 metabolites were identified and among them, after a cross-validation analysis applied to give robustness to the differential expression analysis, 8 emerged as the mainly deregulated among the NAFL and NASH patients. This strategy could be used as a starting point; indeed, the different metabolite profiles could be used to discover metabolic pathways that are mainly disturbed in the NAFLD progression.

All these data individually represent important information useful to better identify different mechanisms underlying the pathophysiology of NAFLD.

According to our knowledge, this is one of the first studies in which a cohort of NAFLD patients, very well histologically characterised, was used in omics studies to investigate the differences between NAFL and NASH. Indeed, due to the invasive approach (liver biopsy) necessary for a clear NASH diagnosis and not always available, normally, omics studies on NAFLD, including patients affected by different forms of the pathology and investigate the differences comparing them with healthy individuals or patients affected by other conditions. Here, omics data were used to evidence the differences between NAFL and NASH patients, suggesting that different components are involved in the progression

of the disease.

This is also one of the first studies in which three different omics data types were collected on the same patients. This aspect is of particular interest because of multi-dimensional approaches on which, in recent years, the researchers have focused on needing samples with all the omics data measured in the same individuals. In this study, the main future prospectives will be first to investigate the correlation between genetics, metabolomics and MiRNAs expression levels in order to understand if and how each aspect affects the others, and then the goal will be to create a prediction model that includes all the different collected information, useful to stratify patients according to the risk.

Integration of multi-omics data providing information from different layers seems to be promising to understand the complex biology systematically. Integrated approaches combine individual omics data, in a sequential or simultaneous manner, to understand the interplay of molecules. They help in assessing the flow of information from one omics level to the other and thus help in bridging the gap from genotype to phenotype. There are two types of integration, the horizontal integration, which studies the same omics across different groups of samples or the vertical integration, which examines multiple omics variables on the same samples. The integration of multi-omics results collected from the same set of samples is a type of analysis searching for correlations across multiple datasets, to discover patterns of co-varying features, and thus to help understand the implicated dysregulated mechanisms in the disease sample set. The integrative analysis of multiomics data collected from the same samples can significantly contribute to the personalised and precision medicine. As successfully used in a recent paper [153], a possible approach to integrate data could be the software MOFA (Multi-Omics Factor Analysis). It is a factor analysis model that provides a general framework for the integration of multi-omic data sets in an unsupervised fashion. It is a versatile and statistically rigorous generalization of principal component analysis to multi-omics data. Given several data matrices with measurements of multiple -omics data types on the same or on overlapping sets of samples, MOFA infers an interpretable low-dimensional representation in terms of a few latent factors. These factors represent the driving sources of variation across data modalities, thus facilitating the identification of disease subgroups. Indeed, MOFA is able to integrate different molecular datasets by generating independent continuous variables, named latent factors, that explain most variation from the joint datasets.

As an explorative analysis, using the Metaboanalyst 5.0 tool, it was possible to investigate the gene-metabolite interaction networks. In this approach, the chemical and human gene associations were extracted from STITCH, in which most of the associations are based on co-mentions highlighted in PubMed, including reactions from similar chemical structures and similar molecular activities. As reported in Figure 5.1, using the 50 genes resulted to be mutated only in NASH patients and involved in the metabolic pathways and the 8 molecules emerged to be deregulated from the metabolomics analysis, several correlations were found, suggesting the data integration approach as a powerful tool, not only for patient stratification but it could be useful also for the identification of new therapeutic targets and for the development of personalized medicine.

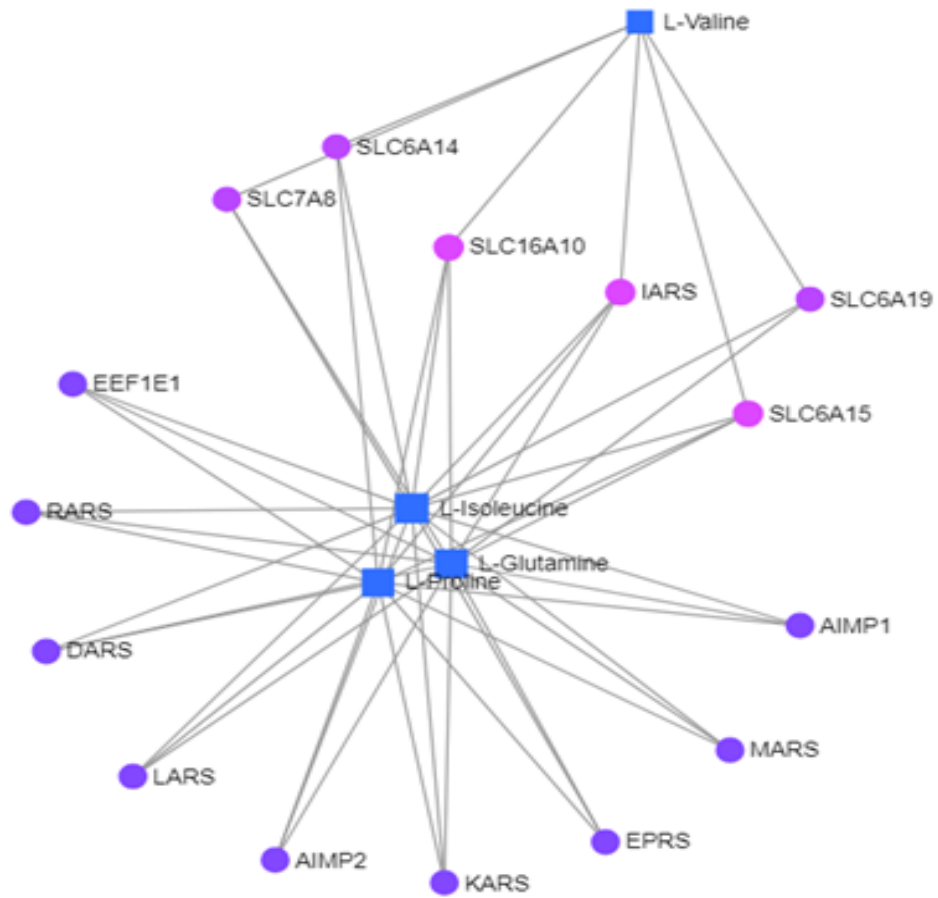


Figure 5.1: Gene-metabolite interaction network using MetaboAnalyst 5.0. The 50 genes mutated only in NASH patients and involved in the metabolic pathways according KEGG database and the 8 metabolites deregulated among NASH and NAFL patients were used to perform the analysis.

Aknowledgments

I would like to express my sincere gratitude to the people who shared with me these 4 years of activity and collaborated to this work.

First and foremost, I would like to thank Professor Giuseppe Matullo, PI of the Research Unit of Genomic Variability, Complex Diseases and Population Medicine and supervisor of this thesis, for believing in me and accepting me into his research group, for his precious counselling and immense wisdom.

I would like to thank all my colleagues part of this incredible team, Clara Viberti, Giovanni Birolo, Serena Aneli, Angelo Savoca, Carla Debernardi, Elton Jalis Herman, Miriam Rosselli, Cecilia Di Primio, all of them gave important contribution to this work. A special thanks to Alessandra Allione, Alessia Russo and Elisabetta Casalone who guided me and helped me throughout the doctoral activity.

I thank Professor Elisabetta Bugianesi, PI of the Division of Gastroenterology and Hepatology of the Department of Medical Sciences, Gian Paolo Caviglia and Chiara Rosso for their precious collaboration, counselling, and inspiration.

Finally, I would like to thank Marcello Manfredi of Department of Translational Medicine of University of Piemonte Orientale for the collaboration.

Bibliography

- [1] Day CP. Natural history of NAFLD: remarkably benign in the absence of cirrhosis. *Gastroenterology* 2005; 29:375- 8.
- [2] Stefano Bellentani, Mariano Marino; Epidemiology and natural history of non-alcoholic fatty liver disease (NAFLD); *Annals of Hepatology*, Volume 8; Issue1;2009.
- [3] Younossi ZM, Golabi P, Paik JM, Henry A, Van Dongen C, Henry L. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. *Hepatology*. 2023 Apr 1;77(4):1335-1347. doi: 10.1097/HEP.0000000000000004. Epub 2023 Jan 3. PMID: 36626630; PMCID: PMC10026948.
- [4] Estes C, Anstee QM, Arias-Loste MT, Bantel H, Bellentani S, Caballeria J, et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016-2030. *J Hepatol* 2018;69:896-904.
- [5] Estes C, Anstee QM, Arias-Loste MT, et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016-2030. *J Hepatol* 2018;69:896-904.
- [6] Alexander M, Loomis AK, van der Lei J, et al. Risks and clinical predictors of cirrhosis and hepatocellular carcinoma diagnoses in adults with diagnosed NAFLD: real-world study of 18 million patients in four European cohorts. *BMC Med* 2019;17:95
- [7] Bacon BR, Farahvash MJ, Janney CG, Neuschwander-Tetri BA. Nonalcoholic steatohepatitis: an expanded clinical entity. *Gastroenterology*. 1994;107(4):1103-9
- [8] Targher, G.; Day, C.P.; Bonora, E. Risk of cardiovascular disease in patients with nonalcoholic fatty liver disease. *N. Engl. J. Med.* 2010, 363, 1341-1350
- [9] Kim, S.U.; Song, D.; Heo, J.H.; Yoo, J.; Kim, B.K.; Park, J.Y.; Kim, D.Y.; Ahn, S.H.; Kim, K.J.; Han, K.H.; et al., Liver fibrosis assessed with transient elastography is an independent risk factor for ischemic stroke. *Atherosclerosis* 2017, 260, 156-162.
- [10] Chang, Y.; Ryu, S.; Sung, E.; Woo, H.Y.; Oh, E.; Cha, K.; Jung, E.; Kim, W.S. Nonalcoholic fatty liver disease predicts chronic kidney disease in nonhypertensive and nondiabetic Korean men. *Metabolism* 2008, 57, 569-576.
- [11] Guo X, Yin X, Liu Z, Wang J. Non-Alcoholic Fatty Liver Disease (NAFLD) Pathogenesis and Natural Products for Prevention and Treatment. *Int J Mol Sci*.

- 2022 Dec 7;23(24):15489. doi: 10.3390/ijms232415489. PMID: 36555127; PMCID: PMC9779435.
- [12] Guo X, Yin X, Liu Z, Wang J. Non-Alcoholic Fatty Liver Disease (NAFLD) Pathogenesis and Natural Products for Prevention and Treatment. *Int J Mol Sci.* 2022 Dec 7;23(24):15489.
 - [13] Day, C.P.; James, O.F. Steatohepatitis: A tale of two 'hits'? *Gastroenterology* 1998
 - [14] Lonardo, A.; Nascimbeni, F.; Maurantonio, M.; Marrazzo, A.; Rinaldi, L.; Adinolfi, L.E. Nonalcoholic fatty liver disease: Evolving paradigms. *World J. Gastroenterol.* 2017, 23, 6571-6592.
 - [15] Bugianesi, E.; Moscatiello, S.; Ciaravella, M.F.; Marchesini, G. Insulin resistance in nonalcoholic fatty liver disease. *Curr. Pharm Des.* 2010, 16, 1941-1951.
 - [16] Jacome-Sosa, M.M.; Parks, E.J. Fatty acid sources and their fluxes as they contribute to plasma triglyceride concentrations and fatty liver in humans. *Curr. Opin. Lipidol.* 2014, 25, 213-220.
 - [17] Caturano, A.; Acierno, C.; Nevola, R.; Pafundi, P.C.; Galiero, R.; Rinaldi, L.; Salvatore, T.; Adinolfi, L.E.; Sasso, F.C. Non-Alcoholic Fatty Liver Disease: From Pathogenesis to Clinical Impact. *Processes* 2021, 9, 135. <https://doi.org/10.3390/pr9010135>.
 - [18] Sabio, G.; Das, M.; Mora, A.; Zhang, Z.; Jun, J.Y.; Ko, H.J.; Barrett, T.; Kim, J.K.; Davis, R.J. A stress signaling pathway in adipose tissue regulates hepatic insulin resistance. *Science* 2008, 322, 1539-1543.
 - [19] Taniguchi, C.M.; Emanuelli, B.; Kahn, C.R. Critical nodes in signaling pathways: Insights into insulin action. *Nat. Rev. Mol. Cell Biol.* 2006, 7, 85-96.
 - [20] Hanlon CL, Yuan L. Nonalcoholic Fatty Liver Disease: The Role of Visceral Adipose Tissue. *Clin Liver Dis (Hoboken).* 2022 Jan 28;19(3):106-110. doi: 10.1002/cld.1183. PMID: 35355841; PMCID: PMC8958250.
 - [21] Pessayre, D.; Fromenty, B. NASH: A mitochondrial disease. *J. Hepatol.* 2005, 42, 928-940.
 - [22] Wang, J.; He, W.; Tsai, P.J.; Chen, P.H.; Ye, M.; Guo, J.; Su, Z. Mutual interaction between endoplasmic reticulum and mitochondria in nonalcoholic fatty liver disease. *Lipids Health Dis.* 2020, 19, 1-19
 - [23] Paradies, G.; Paradies, V.; Ruggiero, F.M.; Petrosillo, G. Oxidative stress, cardiolipin and mitochondrial dysfunction in nonalcoholic fatty liver disease. *World J. Gastroenterol.* 2014, 20, 14205-14218.
 - [24] Cusi, K. Nonalcoholic fatty liver disease in type 2 diabetes mellitus. *Curr. Opin. Endocrinol. Diabetes Obes.* 2009, 16, 141-149.
 - [25] Spencer, M.D.; Hamp, T.J.; Reid, R.W.; Fischer, L.M.; Zeisel, S.H.; Fodor, A.A. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology* 2011, 140, 976-986.

- [26] Tremaroli, V.; Backhed, F. Functional interactions between the gut microbiota and host metabolism. *Nature* 2012, 489, 242-249.
- [27] Anstee, Q.M.; Day, C.P. The genetics of NAFLD. *Nat. Rev. Gastroenterol. Hepatol.* 2013, 10, 645-655.
- [28] Cheung, O.; Puri, P.; Eicken, C.; Contos, M.J.; Mirshahi, F.; Maher, J.W.; Kellum, J.M.; Min, H.; Luketic, V.A.; Sanyal, A.J. Nonalcoholic steatohepatitis is associated with altered hepatic MicroRNA expression. *Hepatology* 2008, 48, 1810-1820.
- [29] Brown GT, Kleiner DE. Histopathology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Metabolism*. 2016 Aug;65(8):1080-6. doi: 10.1016/j.metabol.2015.11.008. Epub 2015 Dec 2. PMID: 26775559; PMCID: PMC4889547.
- [30] Younossi, Zobair M.*; Koenig, Aaron B.; Abdelatif, Dinan; Fazel, Yousef; Henry, Linda; Wymer, Mark. Global epidemiology of nonalcoholic fatty liver disease?Meta?analytic assessment of prevalence, incidence, and outcomes. *Hepatology* 64(1):p 73-84, July 2016. — DOI: 10.1002/hep.28431.
- [31] Mayo R, Crespo J, Martínez-Arranz I, et al. Metabolomic-based noninvasive serum test to diagnose nonalcoholic steatohepatitis: results from discovery and validation cohorts. *Hepatol Commun* 2018;2:807-20.
- [32] Sanna C, Rosso C, Marietti M, et al. Non alcoholic fatty liver disease and extrahepatic cancers. *Int J Mol Sci* 2016; 10.3390/ijms17050717
- [33] <http://omics.org/>
- [34] Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials; Board on Health Care Services; Board on Health Sciences Policy; Institute of Medicine; Micheel CM, Nass SJ, Omenn GS, editors. Washington (DC): National Academies Press (US); 2012 Mar 23.
- [35] Dai X, Shen L. Advances and Trends in Omics Technology Development. *Front Med (Lausanne)*. 2022 Jul 1;9:911861. doi: 10.3389/fmed.2022.911861. PMID: 35860739; PMCID: PMC9289742.
- [36] Rogers, Kara. "omics". *Encyclopedia Britannica*, 9 Dec. 2022, <https://www.britannica.com/science/omics>. Accessed 17 September 2023.
- [37] Sohag MMH, Raqib SM, Akhmad SA. OMICS approaches in cardiovascular diseases: a mini review. *Genomics Inform.* 2021 Jun;19(2):e13. doi: 10.5808/gi.21002. Epub 2021 Jun 30. PMID: 34261298; PMCID: PMC8261269
- [38] Rai V, Mukherjee R, Ghosh AK, Routray A, Chakraborty C. "Omics" in oral cancer: New approaches for biomarker discovery. *Arch Oral Biol.* 2018 Mar;87:15-34. doi: 10.1016/j.archoralbio.2017.12.003. Epub 2017 Dec 8. PMID: 29247855.
- [39] Ullah I, Yang L, Yin FT, Sun Y, Li XH, Li J, Wang XJ. Multi-Omics Approaches in Colorectal Cancer Screening and Diagnosis, Recent Updates and Future Perspectives. *Cancers (Basel)*. 2022 Nov 11;14(22):5545. doi: 10.3390/cancers14225545. PMID: 36428637; PMCID: PMC9688479.

- [40] Hampel H, Nisticò R, Seyfried NT, Levey AI, Modeste E, Lemerrier P, Baldacci F, Toschi N, Garaci F, Perry G, Emanuele E, Valenzuela PL, Lucia A, Urbani A, Sancesario GM, Mapstone M, Corbo M, Vergallo A, Lista S; Alzheimer Precision Medicine Initiative (APMI). Omics sciences for systems biology in Alzheimer's disease: State-of-the-art of the evidence. *Ageing Res Rev.* 2021 Aug;69:101346. doi: 10.1016/j.arr.2021.101346. Epub 2021 Apr 27. PMID: 33915266.
- [41] Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* 18, 83 (2017). <https://doi.org/10.1186/s13059-017-1215-1>
- [42] Diatchenko L, Parisien M, Jahangiri Esfahani S, Mogil JS. Omics approaches to discover pathophysiological pathways contributing to human pain. *Pain.* 2022 Nov 1;163(Suppl 1):S69-S78. doi: 10.1097/j.pain.0000000000002726. Epub 2022 Jul 1. PMID: 35994593; PMCID: PMC9557800.
- [43] Pirola CJ, Sookoian S. Multiomics biomarkers for the prediction of nonalcoholic fatty liver disease severity. *World J Gastroenterol.* 2018 Apr 21;24(15):1601-1615. doi: 10.3748/wjg.v24.i15.1601. PMID: 29686467; PMCID: PMC5910543.
- [44] Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, Boerwinkle E, Cohen JC, Hobbs HH. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 2008; 40: 1461-1465, PMID: 18820647 DOI: 10.1038/ng.257
- [45] Sookoian S, Castaño GO, Burgueño AL, Gianotti TF, Rosselli MS, Pirola CJ. A nonsynonymous gene variant in the adiponutrin gene is associated with nonalcoholic fatty liver disease severity. *J Lipid Res* 2009; 50: 2111-2116, PMID: 19738004 DOI: 10.1194/jlr.P900013-JLR200.
- [46] Sookoian S, Pirola CJ. Meta-analysis of the influence of I148M variant of patatin-like phospholipase domain containing 3 gene (PNPLA3) on the susceptibility and histological severity of nonalcoholic fatty liver disease. *Hepatology* 2011; 53: 1883-1894, PMID: 21381068 DOI: 10.1002/hep.24283.
- [47] Kozlitina J, Smagris E, Stender S, Nordestgaard BG, Zhou HH, Tybjærg-Hansen A, Vogt TF, Hobbs HH, Cohen JC. Exomewide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 2014; 46: 352-356, PMID: 24531328 DOI: 10.1038/ng.2901.
- [48] Bale G, Vishnubhotla RV, Mitnala S, Sharma M, Padaki RN, Pawar SC, Duvvur RN. Whole-Exome Sequencing Identifies a Variant in Phosphatidylethanolamine N-Methyltransferase Gene to be Associated With Lean-Nonalcoholic Fatty Liver Disease. *J Clin Exp Hepatol.* 2019 Sep-Oct;9(5):561-568. doi: 10.1016/j.jceh.2019.02.001. Epub 2019 Feb 11. PMID: 31695245; PMCID: PMC6823660.
- [49] Kleinstein SE, Rein M, Abdelmalek MF, Guy CD, Goldstein DB, Mae Diehl A, Moylan CA. Whole-Exome Sequencing Study of Extreme Phenotypes of NAFLD. *Hepatology Commun.* 2018 Sep 5;2(9):1021-1029. doi: 10.1002/hep4.1227. PMID: 30202817; PMCID: PMC6128233.

- [50] Nobili V, Donati B, Panera N, Vongsakulyanon A, Alisi A, Dallapiccola B, Valenti L. A 4-polymorphism risk score predicts steatohepatitis in children with nonalcoholic fatty liver disease. *J Pediatr Gastroenterol Nutr* 2014; 58: 632-636, PMID: 24345846 DOI: 10.1097/MPG.0000000000000279.
- [51] Zhou Y, Oresic M, Leivonen M, Gopalacharyulu P, Hyysalo J, Arola J, Verrijken A, Francque S, Van Gaal L, Hyötyläinen T, Yki-Järvinen H. Noninvasive Detection of Nonalcoholic Steatohepatitis Using Clinical Markers and Circulating Levels of Lipids and Metabolites. *Clin Gastroenterol Hepatol* 2016; 14: 1463-1472. e6, PMID: 27317851 DOI: 10.1016/j.cgh.2016.05.046.
- [52] Ahrens M, Ammerpohl O, von Schönfels W, Kolarova J, Bens S, Itzel T, Teufel A, Herrmann A, Brosch M, Hinrichsen H, Erhart W, Egberts J, Sipos B, Schreiber S, Häslér R, Stickel F, Becker T, Krawczak M, Röcken C, Siebert R, Schafmayer C, Hampe J. DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery. *Cell Metab* 2013; 18: 296-302, PMID: 23931760 DOI: 10.1016/j.cmet.2013.07.004.
- [53] Nikolaos Perakakis, Konstantinos Stefanakis, Christos S. Mantzoros, The role of omics in the pathophysiology, diagnosis and treatment of non-alcoholic fatty liver disease, *Metabolism*, Volume 111, Supplement, 2020, 154320, ISSN 0026-0495.
- [54] Puri P, Baillie RA, Wiest MM, Mirshahi F, Choudhury J, Cheung O, et al. A lipidomic analysis of nonalcoholic fatty liver disease. *Hepatology*. 2007;46:1081-90. <https://doi.org/10.1002/hep.21763>.
- [55] Musso G, Cassader M, Paschetta E, Gambino R. Bioactive lipid species and metabolic pathways in progression and resolution of nonalcoholic steatohepatitis. *Gastroenterology*. 2018;155:282-302 e8, <https://doi.org/10.1053/j.gastro.2018.06.031>.
- [56] Rodriguez-Gallego E, Guirro M, Riera-Borrull M, Hernandez-Aguilera A, Marine Casado R, Fernandez-Arroyo S, et al. Mapping of the circulating metabolome reveals alpha-ketoglutarate as a predictor of morbid obesity-associated non-alcoholic fatty liver disease. *Int J Obes (Lond)*. 2015;39:279-87. <https://doi.org/10.1038/ijo.2014.53>.
- [57] Polyzos SA, Kang ES, Boutari C, Rhee EJ, Mantzoros CS. Current and emerging pharmacological options for the treatment of nonalcoholic steatohepatitis. *Metabolism*. 2020;154203.
- [58] Neuschwander-Tetri BA, Loomba R, Sanyal AJ, Lavine JE, Van Natta ML, Abdelmalek MF, et al. Farnesoid X nuclear receptor ligand obeticholic acid for non-cirrhotic, non-alcoholic steatohepatitis (FLINT): a multicentre, randomised, placebo-controlled trial. *Lancet*. 2015;385:956-65. [https://doi.org/10.1016/S0140-6736\(14\)61933-4](https://doi.org/10.1016/S0140-6736(14)61933-4).
- [59] Genetic contributions to NAFLD: leveraging shared genetics to uncover systems biology; *nature* 2020.
- [60] Genetic and epigenetic disease modifiers: non-alcoholic fatty liver disease (NAFLD) and alcoholic liver disease (ALD), 2021.

- [61] Browning JD, Kumar KS, Saboorian MH, et al. Ethnic differences in the prevalence of cryptogenic cirrhosis. *Am J Gastroenterol* 2004;99:292-8.
- [62] Romeo S, Kozlitina J, Xing C, et al. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet.* 2008;40:1461-146
- [63] Sookoian S, Castano GO, Scian R, et al. Genetic variation in transmembrane 6 superfamily member 2 and the risk of nonalcoholic fatty liver disease and histological disease severity. *Hepatology.*
- [64] Ma Y, Belyaeva OV, Brown PM, et al. 17-Beta Hydroxysteroid Dehydrogenase 13 Is a Hepatic Retinol Dehydrogenase Associated With Histological Features of Non-alcoholic Fatty Liver Disease. *Hepatology.* 2019;69:1504-1519
- [65] Teo K, Abeysekera KWM, Adams L, et al. rs641738C>T near MBOAT7 is associated with liver fat, ALT and fibrosis in NAFLD: A meta-analysis. *J Hepatol.* 2021;74:20-30. [PubMed: 32882372].
- [66] Santoro N, Zhang CK, Zhao H, et al. Variant in the glucokinase regulatory protein (GCKR) gene is associated with fatty liver in obese children and adolescents. *Hepatology.* 2012;55:781-789.
- [67] Du X, DeForest N, Majithia AR. Human Genetics to Identify Therapeutic Targets for NAFLD: Challenges and Opportunities. *Front Endocrinol (Lausanne).* 2021 Dec.
- [68] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015 May;17(5):405-24.
- [69] Ge SX, Jung D & Yao R, *Bioinformatics* 36:2628-2629, 2020.
- [70] Kong X, Bogyo K, Kapoor S, et al. Challenges associated with diagnostic exome sequencing in liver diseases. *medRxiv*; 2022.
- [71] Yang CF, Wang CH, Siong H'ng W, Chang CP, Lin WD, Chen YT, Wu JY, Tsai FJ. Filamin B Loss-of-Function Mutation in Dimerization Domain Causes Autosomal-Recessive Spondylometaphyseal Dysplasia with Rib Anomalies. *Hum Mutat.* 2017 May;38(5):540-547. doi: 10.1002/humu.23186. Epub 2017 Feb 27.
- [72] Length of Variable Numbers of Tandem Repeats in the Carboxyl Ester Lipase (CEL) Gene May Confer Susceptibility to Alcoholic Liver Cirrhosis but Not Alcoholic Chronic Pancreatitis; 2016.
- [73] Idalsoaga F, Kulkarni AV, Mousa OY, Arrese M, Arab JP. Non-alcoholic Fatty Liver Disease and Alcohol-Related Liver Disease: Two Intertwined Entities. *Front Med (Lausanne).* 2020 Aug 20;7:448.
- [74] Bikle, D.D. Vitamin D: Newer Concepts of Its Metabolism and Function at the Basic and Clinical Level. *J. Endocr. Soc.* 2020.

- [75] Vitamin D-Related Genetic Variations and Nonalcoholic Fatty Liver Disease: A Systematic Review, 2022.
- [76] Distinct Hepatic Gene-Expression Patterns of NAFLD in Patients With Obesity, 2022.
- [77] COL6A3-derived endotrophin links reciprocal interactions among hepatic cells in the pathology of chronic liver disease, 2018.
- [78] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglou, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, Tiffany J Callahan, Christopher G Chute, Johanna L Est, Peter D Galer, Shiva Ganesan, Matthias Griesse, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L Harris, Michael J Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A McMurry, Jillian A Miller, Monica C Munoz-Torres, Rebecca L Peters, Christina K Rapp, Ana M Rath, Shahmir A Rind, Avi Z Rosenberg, Michael M Segal, Markus G Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J Mungall, Melissa A Haendel, Peter N Robinson, The Human Phenotype Ontology in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D1207?D1217, <https://doi.org/10.1093/nar/gkaa1043>.
- [79] Interrogation of selected genes influencing serum LDL Cholesterol levels in patients with well-characterized NAFLD, 2021.
- [80] LRRK2 Regulates CPT1A to Promote β -Oxidation in HepG2 Cells.
- [81] Abnormal metabolic processes involved in the pathogenesis of nonalcoholic fatty liver disease (Review) MINGMEI SHAO, ZIXIANG YE, YANHONG QIN and TAO WU, 2020.
- [82] Targeting the lysosome: Mechanisms and treatments for nonalcoholic fatty liver disease; 2022.
- [83] Lysosomal acid lipase activity and liver fibrosis in the clinical continuum of non-alcoholic fatty liver disease; 2019.
- [84] Cell Autophagy in NASH and NASH-Related Hepatocellular Carcinoma; 2022.
- [85] Lysosomal-associated protein transmembrane 5 ameliorates non-alcoholic steatohepatitis by promoting the degradation of CDC42 in mice, 2023.
- [86] Masyuk AI, Marinelli RA, LaRusso NF. Water transport by epithelia of the digestive tract. *Gastroenterology*. 2002; 122:545?562. PubMed: 11832467.
- [87] Afroze S, Meng F, Jensen K, McDaniel K, Rahal K, Onori P, Gaudio E, Alpini G, Glaser SS. The physiological roles of secretin and its receptor. *Annals of translational medicine*. 2013; 1:29.
- [88] THE CHOLANGIOCYTE PRIMARY CILIUM IN HEALTH AND DISEASE, 2018.

- [89] Sorrentino, P.; Tarantino, G.; Perrella, A.; Micheli, P.; Perrella, O.; Conca, P. A clinical-morphological study on cholestatic presentation of nonalcoholic fatty liver disease. *Dig. Dis. Sci.* 2005, 50, 1130?1135.
- [90] Mechanisms of cholangiocyte responses to injury, 2018.
- [91] Feedback Signaling between Cholangiopathies, Ductular Reaction, and Non-Alcoholic Fatty Liver Disease.
- [92] Dhamija E, Paul SB, Kedia S. Non-alcoholic fatty liver disease associated with hepatocellular carcinoma: An increasing concern. *Indian J Med Res.* 2019 Jan;149(1):9-17.
- [93] Kawada N, Imanaka K, Kawaguchi T, Tamai C, Ishihara R, Matsunaga T, et al. Hepatocellular carcinoma arising from non-cirrhotic nonalcoholic steatohepatitis. *J Gastroenterol.* 2009;44:1190?4.
- [94] Mohr AM, Mott JL. Overview of microRNA biology. *Semin Liver Dis* 2015;35:3-11.
- [95] doi: 10.1161/ATVBAHA.112.300144.
- [96] Wang X, He Y, Mackowiak B, Gao B. MicroRNAs as regulators, biomarkers and therapeutic targets in liver diseases. *Gut* 2021;70:784-795.
- [97] Hernández A, Arab JP, Reyes D, Lapitz A, Moshage H, Bañales JM, et al. Extracellular Vesicles in NAFLD/ALD: From Pathobiology to Therapy. *Cells* 2020;9:817.
- [98] Fang Z, Dou G, Wang L. MicroRNAs in the Pathogenesis of Nonalcoholic Fatty Liver Disease. *Int J Biol Sci* 2021;17:1851-1863. Tadokoro T, Morishita A, Masaki T. Diagnosis and Therapeutic Management of Liver Fibrosis by MicroRNA. *Int J Mol Sci* 2021;22:8139.
- [99] Keller A, Rounge T, Backes C, Ludwig N, Gislefoss R, Leidinger P, et al. Sources to variability in circulating human miRNA signatures. *RNA Biol* 2017;14:1791-1798.
- [100] Peters LJF, Biessen EAL, Hohl M, Weber C, van der Vorst EPC, Santovito D. Small Things Matter: Relevance of MicroRNAs in Cardiovascular Disease. *Front Physiol.* 2020 Jul 7;11:793.
- [101] MicroRNA Analysis Using Next-Generation Sequencing, Monika Kolanowska, Anna Kubiak, Krystian Ja?d?ewski & Anna Wójcicka 2018.
- [102] Fang Z, Dou G, Wang L. MicroRNAs in the Pathogenesis of Nonalcoholic Fatty Liver Disease. *Int J Biol Sci* 2021;17:1851-1863.
- [103] Wang J, Yue BL, Huang YZ, Lan XY, Liu WJ, Chen H. Exosomal RNAs: Novel Potential Biomarkers for Diseases-A Review. *Int J Mol Sci.* 2022;23(5):2461. Published 2022 Feb 23. doi:10.3390/ijms23052461.
- [104] Skowronova et al
- [105] Martin M.; Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011;17:3. doi: 10.14806/ej.17.1.200

- [106] Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, et al. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Res* 2014;42:e133.
- [107] Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 2020;36:2628-2629.
- [108] Ulgen E, Ozisik O, Sezerman OU. pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. *Front Genet* 2019;10:858.
- [109] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
- [110] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- [111] Jopling C. Liver-specific microRNA-122: Biogenesis and function. *RNA Biol* 2012;9:137-142.
- [112] Esau C, Davis S, Murray SF, Yu XX, Pandey SK, Pear M, et al. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. *Cell Metab* 2006;3:87-98
- [113] Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* 2005;309:1577-1581.
- [114] Wang S, Qiu L, Yan X, Jin W, Wang Y, Chen L, et al. Loss of microRNA 122 expression in patients with hepatitis B enhances hepatitis B virus replication through cyclin G(1) -modulated P53 activity. *Hepatology* 2012;55:730-741
- [115] Nakao K, Miyaaki H, Ichikawa T. Antitumor function of microRNA-122 against hepatocellular carcinoma. *J Gastroenterol* 2014;49:589-593.
- [116] Pirola CJ, Fernández Gianotti T, Castaño GO, Mallardi P, San Martino J, et al. Circulating microRNA signature in non-alcoholic fatty liver disease: from serum non-coding RNAs to liver histology and disease pathogenesis. *Gut* 2015;64:800-812.
- [117] Caviglia GP, Abate ML, Gaia S, Petrini E, Bosco C, Olivero A, et al. Risk of hepatocellular carcinoma in HBV cirrhotic patients assessed by the combination of miR-122, AFP and PIVKA-II. *Panminerva Med* 2017;59:283-289
- [118] Lanford RE, Hildebrandt-Eriksen ES, Petri A, Persson R, Lindow M, Munk ME, et al. Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science* 2010;327:198-201
- [119] Wang L, Tang W, Yan S, Zhou L, Shen T, Huang X, et al. Efficient delivery of miR-122 to regulate cholesterol metabolism using a non-covalent peptide-based strategy. *Mol Med Rep* 2013;8:1472-1478.
- [120] Chai C, Rivkin M, Berkovits L, Simerzin A, Zorde-Khvaleyevsky E, Rosenberg N, et al. Metabolic Circuit Involving Free Fatty Acids, microRNA 122, and Triglyceride Synthesis in Liver and Muscle Tissues. *Gastroenterology* 2017;153:1404-1415.

- [121] Chai C, Cox B, Yaish D, Gross D, Rosenberg N, Amblard F, et al. Agonist of RORA Attenuates Nonalcoholic Fatty Liver Progression in Mice via Up-regulation of MicroRNA 122. *Gastroenterology* 2020;159:999-1014.e9.
- [122] Davoodian N, Lotfi AS, Soleimani M, Mowla SJ. MicroRNA-122 overexpression promotes hepatic differentiation of human adipose tissue-derived stem cells. *J Cell Biochem* 2014;115:1582-1593.
- [123] Watt AJ, Garrison WD, Duncan SA. HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology* 2003;37:1249-1253.
- [124] Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, et al. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* 2004; 432:226-230.
- [125] Lei L, Zhou C, Yang X, Li L. Down-regulation of microRNA-375 regulates adipokines and inhibits inflammatory cytokines by targeting AdipoR2 in non-alcoholic fatty liver disease. *Clin Exp Pharmacol Physiol* 2018;45:819-831.
- [126] Pirola CJ, Fernández Gianotti T, Castaño GO, Mallardi P, San Martino J, et al. Circulating microRNA signature in non-alcoholic fatty liver disease: from serum non-coding RNAs to liver histology and disease pathogenesis. *Gut* 2015;64:800-812.
- [127] Lawan A, Bennett AM. Mitogen-Activated Protein Kinase Regulation in Hepatic Metabolism. *Trends Endocrinol Metab* 2017;28:868-878.
- [128] Nguyen-Lefebvre AT, Selzner N, Wrana JL, Bhat M. The hippo pathway: A master regulator of liver metabolism, regeneration, and disease. *FASEB J* 2021;35:e21570.
- [129] Anstee QM, Darlay R, Cockell S, Meroni M, Govaere O, Tiniakos D, et al. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort. *J Hepatol* 2020;73:505-515
- [130] Govaere O, Cockell S, Tiniakos D, Queen R, Younes R, Vacca M, et al. Transcriptional profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis. *Sci Transl Med* 2020;12:eaba4448.
- [131] Donnelly, K. L. et al. Sources of fatty acids stored in liver and secreted via lipoproteins in patients with nonalcoholic fatty liver disease. *J. Clin. Invest.* 115, 1343-1351 (2005)
- [132] Westerbacka, J. et al. Splanchnic balance of free fatty acids, endocannabinoids, and lipids in subjects with nonalcoholic fatty liver disease. *Gastroenterology* 139, 1961-1971.e1 (2010)
- [133] Barr, J. et al. Obesity-dependent metabolic signatures associated with nonalcoholic fatty liver disease progression. *J. Proteome Res.* 11, 2521-2532 (2012)
- [134] Gaggini, M. et al. Altered amino acid concentrations in NAFLD: impact of obesity and insulin resistance. *Hepatology* 67, 145-158 (2018)

- [135] European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD) & European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. *J. Hepatol.* 64, 1388-1402 (2016).
- [136] Rello J., van Engelen T. S. R., Alp E., Calandra T., Cattoir V., Kern W. V., et al. (2018). Towards precision medicine in sepsis: A position paper from the European society of clinical microbiology and infectious diseases. *Clin. Microbiol. Infect.* 24 (12), 1264-1272. 10.1016/j.cmi.2018.03.011
- [137] Cajka, T. & Fiehn, O. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Anal. Chem.* 88, 524-545 (2016).
- [138] Shao M, Lu Y, Xiang H, Wang J, Ji G, Wu T. Application of metabolomics in the diagnosis of non-alcoholic fatty liver disease and the treatment of traditional Chinese medicine. *Front Pharmacol.* 2022 Aug 25;13:971561. doi: 10.3389/fphar.2022.971561. PMID: 36091827; PMCID: PMC9453477.
- [139] Barberis E, Timo S, Amede E, Vanella VV, Puricelli C, Cappellano G, Raineri D, Cittone MG, Rizzi E, Pedrinelli AR, Vassia V, Casciaro FG, Priora S, Nerici I, Galbiati A, Hayden E, Falasca M, Vascetto R, Sainaghi PP, Dianzani U, Rolla R, Chiocchetti A, Baldanzi G, Marengo E, Manfredi M. Large-Scale Plasma Analysis Revealed New Mechanisms and Molecules Associated with the Host Response to SARS-CoV-2. *Int J Mol Sci.* 2020 Nov 16;21(22):8623. doi: 10.3390/ijms21228623. PMID: 33207699; PMCID: PMC7696386.
- [140] www.metaboanalyst.org
- [141] Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucl. Acids Res.* 37, W652-660.
- [142] Ji M, Jo Y, Choi SJ, Kim SM, Kim KK, Oh BC, Ryu D, Paik MJ, Lee DH. Plasma Metabolomics and Machine Learning-Driven Novel Diagnostic Signature for Non-Alcoholic Steatohepatitis. *Biomedicines.* 2022 Jul 11;10(7):1669. doi: 10.3390/biomedicines10071669. PMID: 35884973; PMCID: PMC9312563
- [143] Kalhan SC, Guo L, Edmison J, Dasarathy S, McCullough AJ, Hanson RW, Milburn M. Plasma metabolomic profile in nonalcoholic fatty liver disease. *Metabolism.* 2011 Mar;60(3):404-13. doi: 10.1016/j.metabol.2010.03.006. Epub 2010 Apr 27. PMID: 20423748; PMCID: PMC2950914.
- [144] Kiseleva O, Kurbatov I, Ilgisonis E, Poverennaya E. Defining Blood Plasma and Serum Metabolome by GC-MS. *Metabolites.* 2021 Dec 24;12(1):15. doi: 10.3390/metabo12010015. PMID: 35050137; PMCID: PMC8779220.
- [145] Plasma Amino Acids in NAFLD Patients with Obesity Are Associated with Steatosis and Fibrosis: Results from the MAST4HEALTH Study, *Metabolites* 2023.
- [146] Gaggini, Melania; Carli, Fabrizia; Rosso, Chiara; Buzzigoli, Emma; Marietti, Milena; Della Latta, Veronica; Ciociaro, Demetrio; Abate, Maria Lorena; Gambino, Roberto; Cassader, Maurizio; Bugianesi, Elisabetta*; Gastaldelli, Amalia*.

Altered amino acid concentrations in NAFLD: Impact of obesity and insulin resistance. *Hepatology* 67(1):p 145-158, January 2018. — DOI: 10.1002/hep.29465

- [147] Masarone M, Troisi J, Aglitti A, Torre P, Colucci A, Dallio M, Federico A, Balsano C, Persico M. Untargeted metabolomics as a diagnostic tool in NAFLD: discrimination of steatosis, steatohepatitis and cirrhosis. *Metabolomics*. 2021 Jan 16;17(2):12. doi: 10.1007/s11306-020-01756-1. PMID: 33458794.
- [148] Kalhan, S. C. (2009). Fatty acids, insulin resistance, and protein metabolism. *The Journal of Clinical Endocrinology and Metabolism*, 94(8), 2725-2727. <https://doi.org/10.1210/jc.2009-1235>.
- [149] Zhao, X., Han, Q., Liu, Y., Sun, C., Gang, X., & Wang, G. (2016). The relationship between branched-chain amino acid related metabolomic signature and insulin resistance: A systematic review. *Journal of Diabetes Research*.
- [150] Asghari, G., Farhadnejad, H., Teymoori, F., Mirmiran, P., Tohidi, M., & Azizi, F. (2018). High dietary intake of branched-chain amino acids is associated with an increased risk of insulin resistance in adults. *Journal of Diabetes*, 10(5), 357-364.
- [151] Z. Xie, G. Gao, H. Wang, E. Li, Y. Yuan, J. Xu, Z. Zhang, P. Wang, Y. Fu, H. Zeng, J. Song, C. Holscher, H. Chen, Dehydroabietic acid alleviates high fat diet-induced insulin resistance and hepatic steatosis through dual activation of PPAR- γ and PPAR- α , *Biomed. Pharm.* 127 (2020), 110155 <https://doi.org/10.1016/j.biopha.2020.110155>
- [152] Zhu N, Huang S, Zhang Q, Zhao Z, Qu H, Ning M, Leng Y, Liu J. Metabolomic Study of High-Fat Diet-Induced Obese (DIO) and DIO Plus CCl4-Induced NASH Mice and the Effect of Obeticholic Acid. *Metabolites*. 2021 Jun 10;11(6):374. doi: 10.3390/metabo11060374. PMID: 34200685; PMCID: PMC8230384.
- [153] Mangiante et al *Nat Genet* 55, 607-618 (2023). <https://doi.org/10.1038/s41588-023-01321-1>

List of publications

- Cugliari G, **Catalano C**, Guarrera S, Allione A, Casalone E, Russo A, Grosso F, Ferrante D, Viberti C, Aspesi A, Sculco M, Pirazzini C, Libener R, Mirabelli D, Magnani C, Dianzani I, Matullo G. DNA Methylation of FKBP5 as Predictor of Overall Survival in Malignant Pleural Mesothelioma. *Cancers (Basel)*. 2020 Nov 21;12(11):3470. doi: 10.3390/cancers12113470. PMID: 33233407; PMCID: PMC7700347
- Cugliari G, Allione A, Russo A, **Catalano C**, Casalone E, Guarrera S, Grosso F, Ferrante D, Sculco M, La Vecchia M, Pirazzini C, Libener R, Mirabelli D, Magnani C, Dianzani I, Matullo G. New DNA Methylation Signals for Malignant Pleural Mesothelioma Risk Assessment. *Cancers (Basel)*. 2021 May 27;13(11):2636. doi:10.3390/cancers13112636. PMID: 34071989; PMCID: PMC8199167.
- Allione A, Viberti C, Cotellessa I, **Catalano C**, Casalone E, Cugliari G, Russo A, Guarrera S, Mirabelli D, Sacerdote C, Gentile M, Eichelmann F, Schulze MB, Harlid S, Eriksen AK, Tjønneland A, Andersson M, Dollé MET, Van Puyvelde H, Weiderpass E, Rodriguez-Barranco M, Agudo A, Heath AK, Chirlaque MD, Truong T, Dragic D, Severi G, Sieri S, Sandanger TM, Ardanaz E, Vineis P, Matullo G. Blood cell DNA methylation biomarkers in preclinical malignant pleural mesothelioma: The EPIC prospective cohort. *Int J Cancer*. 2022 Oct 28. doi: 10.1002/ijc.34339. Epub ahead of print. PMID: 36305648.
- Casalone, E.; Birolo, G.; Pardini, B.; Allione, A.; Russo, A.; **Catalano C.**; Mencoboni, M.; Ferrante, D.; Magnani, C.; Sculco, M.; Dianzani, I.; Grosso, F.; Mirabelli, D.; Filiberti, R.A.; Rena, O.; Sacerdote, C.; Rodriguez-Barranco, M.; Smith-Byrne, K.; Panico, S.; Agnoli, C.; Johnson, T.; Kaaks, R.; Tumino, R.; Huerta, J.M.; Riboli, E.; Heath, A.K.; Trobajo-Sanmartín, C.; Schulze, M.B.; Saieva, C.; Amiano, P.; Agudo, A.; Weiderpass, E.; Vineis, P.; Matullo, G. Serum Extracellular Vesicle-Derived microRNAs as Potential Biomarkers for Pleural Mesothelioma in a European Prospective Study. *Cancers* 2023, 15, 125.
- **Catalano C** et al. Whole exome sequencing to identify genes predisposing to the progression of non-alcoholic fatty liver (NAFL). Manuscript in preparation
- **Catalano C** et al. Untargeted analysis to identify metabolites related to non-alcoholic fatty liver disease (NAFLD) and non-alcoholic steatohepatitis (NASH) condition. Manuscript in preparation