Prediction of Non-Alcoholic Fatty Liver Disease development through single nucleotide polymorphism analysis in patients with obesity types II and III

Ismael Prieto Sánchez

TREBALL FINAL DE GRAU BIOTECNOLOGIA

Tutor acadèmic: Ana Fernández Bravo, ana.fernandez@urv.cat, Biology, Ciències Mèdiques Bàsiques, ana.fernandez@urv.cat

En cooperació amb: Unitat de Recerca Biomèdica de l'Institut d'Investigació Sanitària Pere Virgili
Supervisor/s: Jorge Joven Maried, jorge.joven@urv.cat
Helena Castañé Vilafranca, helena.castane@iispv.cat

Agost 2023

Jo, “Ismael Prieto Sánchez”, amb DNI “75926384-A”, sóc coneixedor de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guía per a estudiants (aprovada el juliol 2017) (http://www.urv.cat/ca/vida-campus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquet TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 31 de Agost de 2023

(signatura)

PRIETO SANCHEZ ISMAEL - 75926384A

Firmado digitalmente por PRIETO SANCHEZ ISMAEL - 75926384A
Fecha: 2023.08.31 08:09:56 +02'00'

Table of Contents

1 Data of the center

This study was conducted in 2023 at the Biomedical Research Unit (URB) in the Hospital Universitari Sant Joan de Reus. It was supervised by Dr. Jorge Joven and predoctoral researcher Helena Castañé, with academic guidance provided by Dr. Ana Fernández. The URB is situated within the Hospital Universitari Sant Joan de Reus premises and functions as a constituent unit of the Institut d'Investigació Sanitària Pere Virgili (IISPV).

The URB studies non-communicable diseases, such as morbid obesity, liver diseases, cardiovascular diseases, and cancer. The unit aims to understand how metabolism, oxidation, and inflammation are related to these diseases. Specifically, research efforts have been dedicated to understanding the role of inflammation and oxidation markers in obesity-related pathologies and investigating the nature and origins of metabolic disorders.

2 Abstract

Background: Genome-wide Association Studies (GWAS) have revolutionized genetic exploration, providing crucial insights into complex disease mechanisms. This newfound genetic understanding holds immense promise for enhancing diagnostics and treatment strategies. This progress has proven especially invaluable in Non-Alcoholic Fatty Liver Disease (NAFLD). GWAS approaches have revealed a multitude of Single Nucleotide Polymorphisms (SNPs) associated with NAFLD. However, the escalating prevalence of NAFLD driven by obesity underlines a pressing healthcare challenge. NAFLD's progression to severe stages like Non-Alcoholic Steatohepatitis (NASH) and advanced liver diseases emphasizes the urgent need for non-invasive diagnostic tools, early detection methods, and treatment strategies.

Methods: The OpenArray platform was utilized to genotype 25 specific SNPs from plasma samples. The study cohort comprised 1719 individuals, meticulously divided into the control group (n = 415) and the severe obesity group (n = 1313). The latter encompassed three subgroups: NAFL (n = 137), NASH (n = 532), and individuals without NAFLD (n = 405). We subsequently conducted an extensive analysis to evaluate the correlation, significance, and strength of the identified SNPs across the various groups while assessing their predictive power for disease progression.

Objective: Determine if genetic variants identified through GWAS can serve as predictive indicators of NAFLD development in patients with obesity types II and III.

Results: Four significant SNPs (rs58542926, rs1800629, rs8107974, and rs5982) were discovered to be correlated with disease progression within the subset of severely obese patients. Among these, rs5982 (*F13A1*) exhibited dual significance in both NAFLD and obesity, with a stronger association observed for the disease. Notably, rs58542926 (*TM6SF2*) significantly affected steatosis scores, while rs1800629 (*TNFA*) correlated with inflammation scores. All identified SNPs presented weak associations and held little predictive power for disease progression.

Conclusions: The identified SNPs showed significant correlations with NAFLD progression in severely obese patients with obesity types II and III. However, these associations do not translate into robust predictive power for NAFLD progression assessment.

Keywords: NAFLD, NASH, Genetic variants, SNPs, Obesity

3 Abbreviations

ALT: Alanine Aminotransferase

ARBs: Angiotensin II Receptor Blockers

AST: Aspartate Aminotransferase

BBs: Beta Blockers

BMI: Body Mass Index

CBC: Complete Blood Count

CCBs: Calcium channel blockers

DBP: Diastolic Blood Pressure

GGT: Gamma-Glutamyl Transferase

HDL-C: High-Density Lipoprotein Cholesterol

HC: Hip Circumference

HR: Heart Rate

LDL-C: Low-Density Lipoprotein Cholesterol

MetS: Metabolic Syndrome

NASH: Nonalcoholic Steatohepatitis

NAFL: Non-Alcoholic Fatty Liver

NAFLD: Non-Alcoholic Fatty Liver Disease

NAS: NAFLD Activity Score

OOB: Out-of-bag

PCR: Polymerase Chain Reaction

PNPLA3: Patatin-like phospholipase domain-containing protein 3

RBCLB: Red Blood Cell Lysis Buffer

SAF: Steatosis Activity Fibrosis

SBP: Systolic Blood Pressure

SNP: Single Nucleotide Polymorphism

SNV: Single Nucleotide Variant

SUGP1: SURP And G-Patch Domain Containing 1

TC: Total Cholesterol

T2DM: Type 2 Diabetes Mellitus

VLDL-C: Very-Low-Density Lipoprotein Cholesterol

WC: Waist Circumference

WGS: Whole Genome Sequencing

4 Introduction

4.1 Genetics and its Applications

Genetics is the branch of biology responsible for the study of genes. This study can involve investigating mechanisms of inheritance, variability, and how genes determine the characteristics of organisms.

One of the most promising applications of genetics is the study of genetic variation to understand its role in diseases. Identifying genetic risk factors provides new possibilities for the prevention and early detection of diseases (Germain et al., 2021). Knowing the genetic variants associated with the disease also allows us to understand the pathways involved in its pathogenesis and identify potential pharmacological targets (Lappalainen & Macarthur, 2021).

Identifying genetic risk factors was only possible with the advances made in recent decades in understanding the structure and function of genes (Horton & Lucassen, 2019). These advances have allowed the development of new DNA manipulation and sequencing technologies, opening new possibilities in identifying and understanding genetic variability and its relationship to diseases.

4.2 Genetic Variations and Their Role in Disease

The central idea behind predicting genetic diseases is that affected individuals have an excess of pathogenic DNA variants compared to a control group unaffected by the disease. The closer the DNA variant is related to the disease, the higher its impact; therefore, it becomes more relevant for diagnosis and therapeutic research.

Although these pathogenic DNA variants may follow a simple Mendelian logic, where one allele is responsible for the disease, most diseases are not determined by a single gene or allele but by multiple genetic variants that can also be related to environmental factors, making their prediction difficult (Di Taranto et al., 2020).

4.2.1 Types of Genetic Variations in Humans

Multiple types of mutations in humans vary in the size of affected base pairs and their frequency of occurrence. Table 1 summarizes the various genetic variants in humans, including their size in base pairs and other characteristics.

Table 1. Genetics variations comparison (Eichler, 2019).

Class	Size of Variant (bp)	No. per Genome	Size of Region Affected (Mbp)	Percent of Genome
Single-nucleotide variants	1	4,000,000-5,000,000	4-5	0.078
Insertion-deletions	1-49	700,000-800,000	3-5	0.069
Structural variants	>50	23,000-28,000	10-12	0.19
Inversions	>50	153	23	0.397
Multi-copy-number variants	>1000	500	12-15	0.232

Single nucleotide variants (SNVs) are the most common genetic variants (Zou et al., 2020) but despite their high frequency SNVs usually do not have a strong effect due to their limited impact on gene function. In contrast, structural variants such as insertions, deletions, inversions, and translocations that affect at least 50 base pairs can have a more significant impact on gene function and, therefore, on the characteristics of an organism.

4.2.2 Genome-wide association studies

Genome-wide association studies, or GWAS, test hundreds of thousands to millions of genetic variants in the genomes of many individuals to identify genotype-phenotype associations (Ishigaki, 2022).

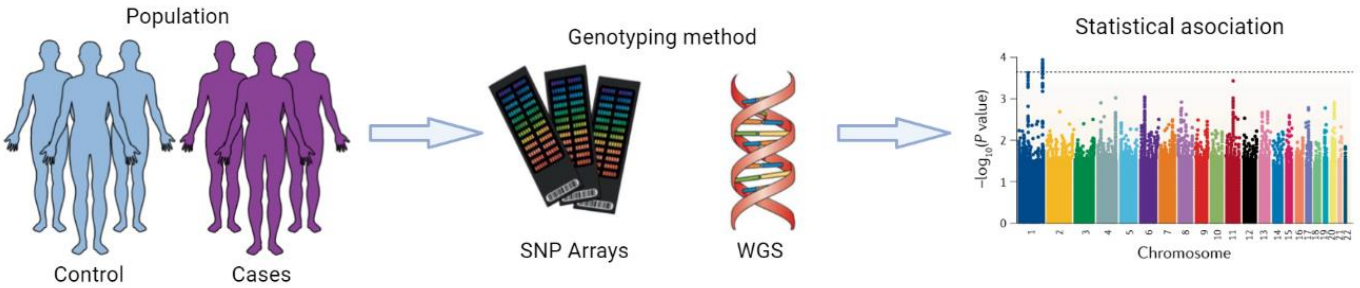

Figure 1. GWAS Process.

Figure 1 illustrates that the process involves using a group of patients who exhibit the characteristic or disease of interest and a control group. Researchers genotype hundreds of thousands of genetic variants in these individuals using SNP arrays or whole genome

sequencing (WGS) methods. Statistical analyses are then performed on the results to identify associations between genotype and phenotype (Tam et al., 2019).

GWAS are very effective in identifying these variant-phenotype associations, allowing the identification of risk loci for many diseases, including cancers, type 2 diabetes mellitus, and schizophrenia, among others. Additionally, these associations can lead to the discovery of new biological mechanisms. For example, in Crohn's disease, the role of autophagy was not known before multiple SNPs were associated with disease risk in a GWAS (Fritz et al., 2011).

One of the main limitations of GWAS is that the genetic components related to the risk of human diseases typically consist of many variants with minor effects, which would require study sizes that are too large to imply a significant relationship (Speakman et al., 2018). Additionally, although these genetic variants are identified, they usually only explain a small fraction of the heritability of the disease. Despite there being some hypotheses to justify this lost heritability, no studies confirm them (Manolio et al., 2009).

4.2.3 Importance of SNPs

Although their impact may be minor, single nucleotide variants are frequently utilized in genetic studies, specifically those known as SNPs or single nucleotide polymorphisms (Guan et al., 2022). These variations occur at a specific position in the genome and are present in at least 1% of the population. They are essential because they are easy and cheap to genotype, unlike other variations, such as structural variants, which may require more sophisticated sequencing methods (Gong et al., 2021). Additionally, although they may have a minor effect individually, this effect can increase when they are combined (Shastry, 2009). For example, genetic studies can identify hundreds or thousands of SNPs whose combined effect may be significant for a disease or trait.

4.3 Genetic Testing and Disease

The impact of genotypes on diseases is an increasingly relevant topic in medicine and biomedical research. The ability to predict and understand the genetic risk associated with a disease is a fundamental step in improving disease prevention, diagnosis, and treatment. Genetic testing can diagnose several diseases, and some can be predicted with 100% accuracy. An example of this is Huntington's disease, caused by a mutation in the huntingtin

gene. If a person has the mutation, then it is practically certain that they will develop the disease during their lifetime (Stoker et al., 2022).

On the other hand, some diseases can be detected through genetic testing, but having the genetic variant does not imply that the individual will develop the disease. *BRCA1* and *BRCA2* are examples of genes that, when mutated, can be present in a subtype of hereditary breast cancer. Genetic testing can identify the presence of these mutations but cannot predict with certainty that a person will develop the disease (Saleem et al., 2018). However, early detection of these mutations can allow for proper surveillance and preventive treatment. Genetic study and detection of genetic variants have also allowed for a better understanding of the underlying biological mechanisms of many diseases. For example, identifying the *CFTR* gene mutation as the cause of cystic fibrosis has led to a better understanding of how the disease works and how an effective treatment can be developed (Bienvenu et al., 2020).

Similarly, the study of genetic variants associated with diseases such as type 2 diabetes and Alzheimer's disease has led to a greater understanding of the underlying biological mechanisms of these diseases. However, many diseases are still not well understood from a genetic standpoint, which limits our ability to predict, diagnose, and treat these diseases. Non-alcoholic fatty liver disease (NAFLD) is an example of a disease that is not yet fully understood from a genetic perspective.

4.3.1 NAFLD and NASH

NAFLD, or nonalcoholic fatty liver disease, is a metabolic disease characterized by the accumulation of lipids in the liver. NAFLD encompasses a spectrum of liver diseases with causes unrelated to alcohol consumption. This spectrum ranges from hepatic steatosis or NAFL, where a minimal threshold of 5% of hepatocytes contains fat droplets (Lonardo et al., 2017), to more advanced forms, where the accumulation of lipids leads to lipotoxicity and progresses to nonalcoholic steatohepatitis (NASH), and finally to hepatocellular carcinoma (Nassir, 2022). NASH is a more developed form of NAFL that not only presents with steatosis but also with inflammation, cellular damage, and fibrosis in the liver.

NAFLD has a global prevalence of 24% and is strongly related to obesity and type 2 diabetes. The prevalence reaches 70% in overweight individuals and exceeds 90% in cases of morbid obesity (Nassir, 2022). Other diseases related to obesity and insulin resistance increase the prevalence of NAFLD, such as polycystic ovary syndrome (Manikat & Nguyen, 2023).

4.3.1.1 Impact of NAFLD

The growth of NAFLD cases that parallels the global increase in obesity constitutes a new epidemic in the field of chronic liver disease. Currently, the global prevalence of NAFLD is 24%, and in Europe, it is 23% (Younossi et al., 2018). This prevalence has steadily increased in recent years and may continue. An increase in NAFLD cases has also led to similar growth in its more advanced form, NASH, which has been established as the second most common indication for liver transplantation in the USA (Anstee et al., 2013). This increase in the prevalence of both NAFL and NASH, along with their severe medical consequences, causes increased medical costs directly attributable to the disease. The cost is estimated to exceed 35 thousand million euros in multiple European countries, such as France, Germany, Italy, and the United Kingdom (Younossi et al., 2018).

4.3.1.2 Treatments

Currently, there is no approved pharmacological treatment or surgery for NASH. The current approach is mainly based on lifestyle modifications, such as diet, physical activity, and exercise. However, when patients are unable to make these lifestyle changes or are in advanced stages of the disease, pharmacological treatments aimed at treating aspects such as inflammation and fibrosis are used (Raza et al., 2021).

These drugs used are not a cure for the disease and are of limited use. Therefore, there is a great need for the development of effective and safe drugs for NASH.

4.3.1.3 Diagnosis

A liver biopsy is currently the gold-standard for the diagnosis and prognosis of NASH. However, this procedure is expensive and invasive, with a high risk of complications (Kleiner et al., 2005). Other non-invasive methods have been suggested for diagnosis, including multiple imaging techniques such as MRI and CT scans. The problem with these techniques is that they lack standardization and validation and are less accurate than a liver biopsy (Piazzolla & Mangia, 2020). In addition, imaging techniques may not always suit morbidly obese people.

Due to the high cost and invasiveness of the current diagnostic technique, alternatives that are less invasive and help to determine the disease are being sought. Thus, one possible aid in this process would be predicting predisposition to decide whether performing this biopsy would be necessary.

4.3.1.4 Identification of Genetic Variants Associated with NAFL and NASH

Although the etiology of NAFL and its progression to NASH is not yet fully understood, it has been suggested that there may be a genetic predisposition for its development (Trépo & Valenti, 2020). Several genome-wide association studies have been carried out to identify genetic variants related to this disease. These studies have produced promising results that suggest that specific genetic variants may be associated with an increased risk of developing NAFLD.

One of these studies, "Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterized cohort" (Anstee et al., 2020) focused on identifying genetic variants associated with the development of NAFLD.

The study included a sample of 1,482 patients with NAFLD and 17,781 controls of European ancestry, who were genotyped on over 7 million common genetic variants. Cases of NAFLD were classified based on their degree of hepatic fibrosis and the presence or absence of NASH.

The study's results revealed multiple significant loci associated with the study of NAFLD compared to healthy patients. Additionally, more loci were found to be significant in relation to various NAFLD characteristics, including NASH progression and fibrosis.

Table 2. Summary of top findings related to NAFLD characteristics.

SNP	Chromosome	Gene	Phenotype	n	p value
rs738409	22	*PNPLA3*	Steatosis	1,469	2.3×10^{-9}
rs62021874	15	*PYGO1*	Steatosis	1,469	8.16×10^{-8}
rs11858624	15	*PYGO1*	Steatosis	1,469	1.64×10^{-7}
rs738409	22	*PNPLA3*	Fibrosis	1,481	7.58×10^{-11}
rs738409	22	*PNPLA3*	NAS Score	1,467	8.78×10^{-9}

As shown by Table 2, among these genetic variants, the SNP rs738409 in the *PNPLA3* gene was found to have a significant effect on multiple NAFLD characteristics like steatosis, fibrosis, and NAFLD Activity Score (NAS). *PNPLA3* encodes patatin-like phospholipase domain-containing protein 3 (*PNPLA3*), which is expressed in the liver and linked to lipid metabolism. The rs738409 variant was associated with an increased risk of NAFLD and steatohepatitis. It was found to increase the activity of the *PNPLA3* protein, resulting in the accumulation of lipids in the liver and disease progression.

5 Hypothesis and Objectives

The accessibility of genotyping techniques, coupled with the emergence of GWAS, has transformed the field of genetics. Identifying and exploring genetic traits has become increasingly crucial in revealing insights into disease mechanisms. This surge in genetic understanding not only helps unravel complex disorders' complex pathways but also opens the possibility of innovative approaches to early diagnosis and treatment targets. In the context of NAFLD, these genetic advancements have proved especially beneficial. Research utilizing tools like GWAS has unveiled numerous SNPs linked to the disease. These genetic markers offer a deeper understanding of NAFLD's genetic landscape.

Despite the advancements in genetics and understanding, the prevalence of NAFLD is rising due to the increase in obesity rates, posing a significant healthcare challenge. Due to disease progression to NASH, cirrhosis, and hepatocellular carcinoma, many NAFLD patients die. The increasing disease prevalence and high mortality rates highlight the need for less invasive diagnostic tools to detect NAFLD early, identify at-risk individuals, and develop treatment strategies.

We hypothesize that identifying genetic variants previously delineated through GWAS could offer a valuable means to predict the NAFLD development in patients with obesity types II and III. What sets our study apart is its unique focus on a subset of severely obese patients, marked by the distinctive characteristics of obesity types II and III, in contrast to the previous studies that only compared control groups with NAFLD patients.

The main objectives of this study are:

1. Determine the correlation of SNPs with disease development and obesity.
    a. Analyze the genotyping data to compare genetic profiles for the following scenarios:
        i. Compare the genetic profiles between the control group and the severe obesity group.
        ii. Compare the genetic profiles among patients within the severe obesity group diagnosed with NAFL, NASH, and non-NAFLD.
        iii. Compare the genetic profiles among patients within the severe obesity group separated by their histological scores.
    b. Based on the analysis data, identify SNPs related to the disease development.

        i. Differentiate these disease-related SNPs from those associated with obesity and evaluate their effect on the disease, considering their effect on the histological scores.

2. Assess genetic predictive power and model performance.
    a. Evaluate predictive potential based on association strength of disease-related SNPs.
    b. Train and test a Random Forest Model with the data of the disease-related SNPs to evaluate their predictive power.

6 Materials and Methods

6.1 Study Design

The experiment included a total of 1719 participants. All participants were required to be at least 18 years old. The study's exclusion criteria included evidence of severe illnesses, chronic inflammations, or infectious diseases. These participants were divided into a control or severe obesity group as shown by Figure 2.

The control group consisted of individuals that were previously recruited and whose samples were provided by the bank of biological samples of the IISPV.

The participants with morbid obesity had a BMI ≥ 35 kg/m² and met the criteria to undergo bariatric surgery at Sant Joan Hospital (Reus, Spain). The morbidly obese patients were further stratified by their Steatosis, Activity, Fibrosis (SAF) score (Bedossa et al., 2012) in patients with no evidence of NAFLD, those diagnosed with NAFL, and those with NASH.

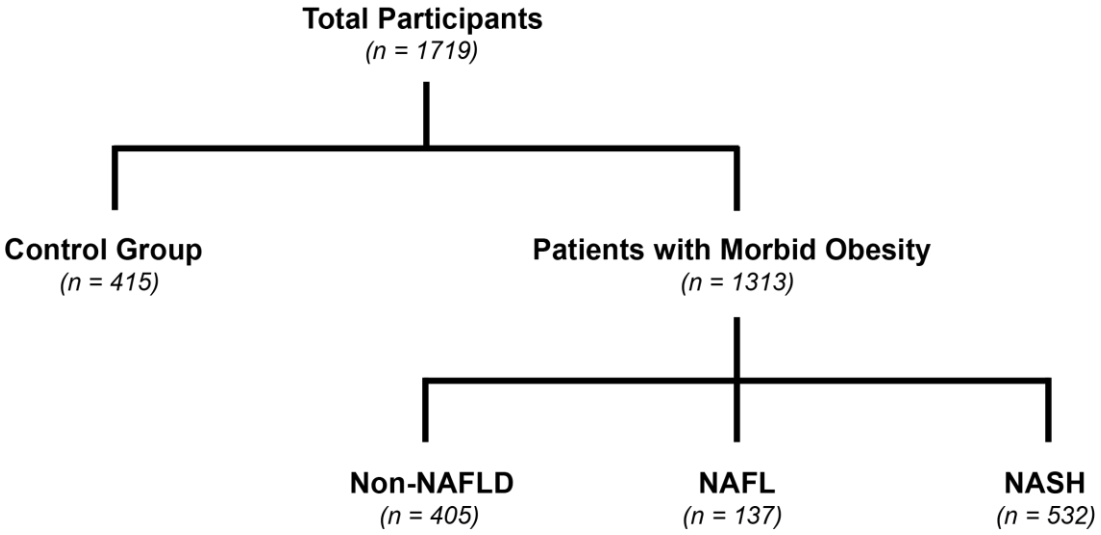

Figure 2. Diagram illustrating participant distribution in the study.

6.2 Sampling

Blood samples were procured from obese patients before bariatric surgical intervention. These samples were carefully drawn into tubes and subjected to centrifugation at 2500rpm for 15 minutes, all performed at a temperature of 4ºC, to acquire plasma for DNA extraction and serum for biochemical analysis.

During the bariatric surgical procedure, liver specimens were biopsied. These samples were fixed in formaldehyde for 24 hours. The tissues were paraffin-embedded for subsequent histological examinations.

6.3 Biochemical Analysis

Serum samples underwent routinary biochemical assessments. Using an automatic analyzer (COBAS 8000, Roche Farma, Basel, Switzerland), we determined the following parameters: glucose, insulin, total triglycerides, total cholesterol, HDL-cholesterol, LDL-cholesterol, VLDL-cholesterol, ALT, AST, and GGT. These analyses followed established protocols from Laboratori de Referència del Camp de Tarragona i Terres de l'Ebre.

6.4 Histological Analysis

The paraffin-embedded specimens were sectioned into slices of 2 µm thickness using a microtome. Two distinct slices were procured from each patient's sample. After dewaxing, two distinct staining methods were applied: Masson's Trichrome and Hematoxylin stain.

An experienced hepatologist determined histological characteristics using a Nikon Eclipse microscope and following Kleiner's framework (Kleiner et al., 2005). Steatosis, ballooning, and lobular inflammation were determined using the Hematoxylin-stained slices, while we observed fibrotic degree from Masson's Trichrome-stained section.

6.4.1 SAF Score

In our study, we deliberately chose to utilize the SAF score instead of the NAFLD Activity Score (NAS) to evaluate and classify individuals with NAFLD. The SAF score's ability to clearly distinguish between NAFLD and NASH was a critical factor behind this decision, eliminating uncertainties. The SAF score provides a distinct classification (Bedossa et al., 2012), unlike the NAS score, which can sometimes leave a diagnostic gray area (Kleiner et al., 2005).

The calculation of the SAF score involves assessing three key components:

1. Steatosis (S): Steatosis represents the extent of fat accumulation within liver cells. This attribute is measured on a scale that goes from 0 to 3.
2. Activity (A): The activity component combines hepatocellular ballooning and lobular inflammation.

3. Fibrosis (F): Fibrosis denotes the presence and severity of scar tissue formation within the liver. The scoring system ranges from 0 to 4.

To be diagnosed with NAFLD, a patient must have at least mild fat accumulation with a minimum steatosis score of 1. To be considered as NASH, a patient must have an activity score of 2 or higher.

6.4.2 DNA Extraction

We used a buffy coat extraction protocol to isolate DNA from plasma samples through multiple steps. Four mL of the buffy coat were mixed with 38 mL of red blood cell lysis buffer (RBCLB) and incubated for 20 min at room temperature. Next, samples were centrifuged for 5 min at 400g, and the supernatant was removed. Then, the pellet was resuspended again in 20 mL of RBCLB to ensure the lysis of erythrocytes. After centrifugation, the pellet was resuspended with 10 mL of cell lysis buffer (Qiagen) and shaken gently overnight at room temperature. Next, the samples were cooled, and a protein precipitation solution (Qiagen) was added. After 20 min of incubation, the samples were centrifuged for 5 min at 400g. The supernatant was collected, and cooled isopropanol was added to precipitate the DNA. After washing twice with ethanol 70% the diluted DNA was gently shaken with nuclease-free water for seven days. Finally, NanoDrop (Thermofisher, Massachusetts, USA) was used to quantify the DNA, and a 260nm/280nm ratio was used to ensure the quality of DNA.

6.5 Genotyping

We employed OpenArray technology (Thermofisher) to perform genotyping analysis on 25 SNPs (Table 3, Figure 3) across a cohort of 1719 participants in our study.

DNA samples at 30 ng/uL were sent to a specialized facility equipped with the necessary instrumentation to perform the genotyping process. Briefly, a first 384-well plate was prepared combining the genomic DNA and master mix. This mixture was loaded into the OpenArray plate using the OpenArray AutoLoader instrument (Thermofisher). The case of the array was sealed and introduced to a QuantStudio 12K qPCR system.

After loading the array, the polymerase chain reaction (PCR) amplification process was initiated concurrently with the hybridization step. This joint process enabled the targeted DNA sequences containing the SNPs of interest to be amplified while simultaneously undergoing hybridization with allele-specific probes. These probes, designed to be

complementary to the SNP variants under investigation, were structured to fluoresce upon binding to the corresponding DNA sequence. Measuring the emitted fluorescence signals from the probes on the array allowed the genotyping instrument to determine the specific SNP alleles present in each participant's DNA sample once the amplification and hybridization were completed.

Table 3. List of genotyped SNPs with gene symbols and assay names.

SNP	Gene	Assay Name
rs3774261	*ADIPOQ*	C__27479710_10
rs1919127	*C2orf16*	C__12095077_10
rs1761667	*CD36*	C__8314999_10
rs5982	*F13A1*	C___8786720_10
rs17817449	*FTO*	C__34511515_10
rs8050136	*FTO*	C__2031259_10
rs9930506	*FTO*	C__29819994_10
rs9939609	*FTO*	C__30090620_10
rs1260326	*GCKR*	C___2862880_1_
rs780094	*GCKR*	C___2862873_10
rs13118664	*HSD17B13*	C__11556153_10
rs9992651	*HSD17B13*	C__26031528_10
rs12077210	*LEPR*	C__30697741_20
rs5764455	*PARVB*	C__26640554_10
rs10500212	*PBX4*	C__30210566_20
rs738409	*PNPLA3*	C_____7241_10
rs738409d	*PNPLA3*	C__7514879_10
rs738409d2	*PNPLA3*	C___7241_10
rs11858624	*PYGO1*	C___2100841_10
rs62021874	*PYGO1*	C__89205479_10
rs4880	*SOD2*	C___8709053_10
rs8107974	*SUGP1*	C__29427528_10
rs58542926	*TM6SF2*	C__89463510_10
rs1800629	*TNFA*	C__7514879_10
rs17216588	None	C__33703647_10
rs2943634	None	C___2862880_1_
rs6006473	None	C___2520586_20
rs6982502	None	C__29827528_10

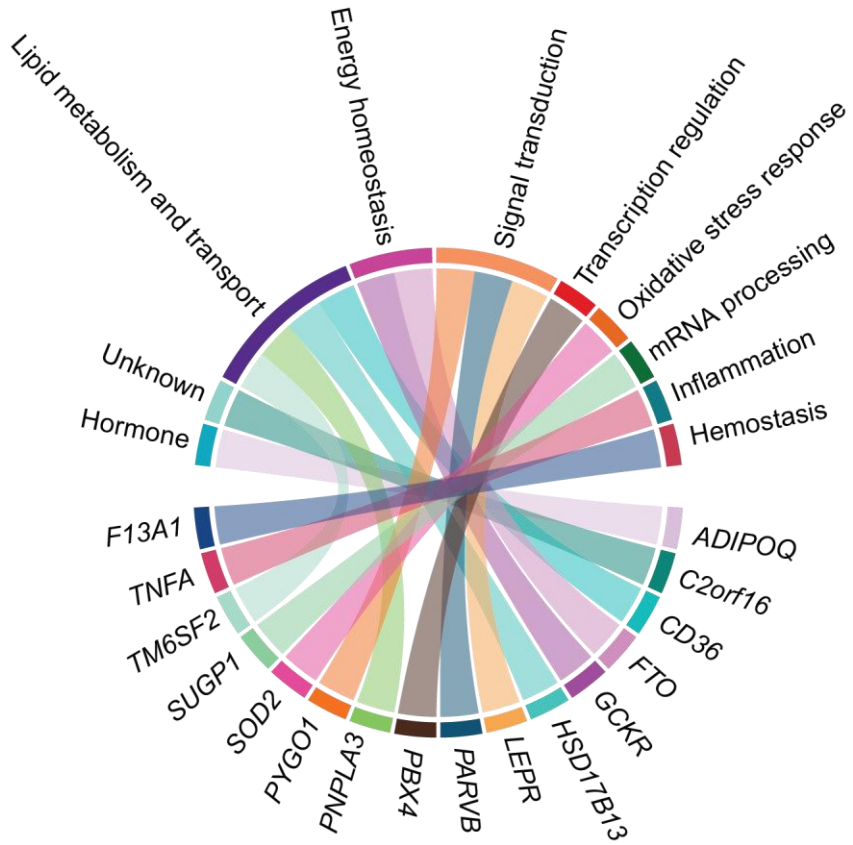

Figure 3. Chord diagram of genotyped SNPs and their biological functions.

6.6 Data Analysis

The data analysis was conducted using RStudio version 2023.06.01+524. This section provides an overview of the tools and techniques employed for the various stages of the study. We considered differences as significative for values of $p < 0.05$. Refer to the annex for a more detailed understanding of the code used.

6.6.1 Population Study

We utilized the 'tableone' package in RStudio (Kazuki Yoshida & Alexander Bartel, 2022) to generate statistical summaries, frequencies, percentages, and significance levels for the population study.

6.6.2 Correlation Study

For the significance assessment of SNPs, we conducted chi-square tests using the 'tableone' package in RStudio. Specifically, we investigated the significance of each genotyped SNP at differentiating selected groups using both standard and one-hot encoded representations.

In cases where there were significant differences in SNPs between the studied groups, we conducted further analysis to measure the level of correlation. Specifically, we employed the Cramer's V statistic, calculated using the 'vcd' package in RStudio (Meyer D et al., 2023).

6.6.3 Prediction Study

To evaluate the predictive capability of SNPs in distinguishing groups, we trained a Random Forest model on our dataset and assessed its classification performance. For this task, we utilized the 'randomForest' package in RStudio (Andy Liaw & Matthew Wiener, 2002). This package provided:

- Essential functionality for training the model.
- Conducting testing.
- Obtaining key measures such as the out-of-bag error.

6.6.4 Construction of Plots

To demonstrate the importance of SNPs in distinguishing patient groups, bar plots were created using the 'ggplot2' package (Hadley Wickham, 2016).

7 Results

7.1 Population Study

Table 4 presents the characteristics of two distinct groups: a control group of volunteers and a group of patients with severe obesity.

Noteworthy differences in clinical characteristics could be observed between the two groups. Expected variables such as BMI, hip circumference, and waist circumference showed significant disparities, with the control group exhibiting smaller mean values than the other group. Additionally, both groups had a higher percentage of women; however, this proportion was significantly elevated in the patients with severe obesity.

Furthermore, measures of heart rate, systolic blood pressure, and diastolic blood pressure exhibited significant distinctions, with lower values observed in the control group. Significative differences in comorbidities were also identified, with patients with obesity showing a higher frequency of conditions such as type 2 diabetes mellitus, hypertension, dyslipidemia, and metabolic syndrome.

The analysis of biochemical variables revealed significant differences across all parameters. Notably, the severe obesity group had lower HDL, LDL, and total cholesterol values while exhibiting higher values for VLDL cholesterol. Moreover, glucose, insulin, triglycerides, and HOMA-IR levels were significantly lower in the control group.

In terms of transaminases, the group with severe obesity had significantly higher values for all variables.

Finally, regarding medication usage, the control group showed significantly lower frequencies of insulin and diuretics, while no difference was observed for sulfonylureas.

Table 4. Comparison of characteristics between control group and patients with obesity.

	Control Group (n = 415)	**Patients with obesity** (n = 1,398)	**p-value**
Clinical characteristics			
Sex, woman [n (%)]	231 (56.9)	934 (72.2)	<0.001
Age, years	47.5 ± 15.0	48.3 ± 10.6	0.198
BMI, kg/m^2	27.2 ± 5.0	44.8 ± 6.5	<0.001
HR, bpm	74.6 ± 9.3	76.9 ± 13.6	0.005
SBP, mmhg	127.7 ± 19.1	130.4 ± 19.7	0.025
DBP, mmhg	79.1 ± 13.4	77.0 ± 13.6	0.015
HC, cm	102.0 ± 10.5	136.2 ± 15.0	<0.001
WC, cm	89.6 ± 14.3	130.3 ± 15.6	<0.001
T2DM [n (%)]	26 (6.3)	328 (23.6)	<0.001
Hypertension [n (%)]	62 (15.0)	536 (38.5)	<0.001
Dyslipidemia [n (%)]	36 (8.7)	302 (21.7)	<0.001
MetS [n (%)]	0 (0)	502 (36.1)	<0.001
Biochemical variables			
TC, mmol/L	5.2 ± 0.9	4.2 ± 1.0	<0.001
HDL-C, mmol/L	1.4 ± 0.4	1.1 ± 0.4	<0.001
LDL-C, mmol/L	3.2 ± 0.9	2.6 ± 1.0	<0.001
VLDL-C, mmol/L	0.5 ± 0.3	0.8 ± 0.4	<0.001
Glucose, mmol/L	4.9 ± 1.1	7.5 ± 3.1	<0.001
Insulin, pmol/L	64.6 ± 108.3	110.0 ± 208.5	<0.001
Triglycerides, mmol/L	1.2 ± 0.8	1.7 ± 1.0	<0.001
HOMA-IR	2.1 ± 4.5	6.1 ± 13.8	<0.001
Transaminases			
ALT, μKat/L	0.3 ± 0.2	0.7 ± 0.4	<0.001
AST, μKat/L	0.4 ± 1.8	0.6 ± 0.5	<0.001
GGT, μKat/L	0.3 ± 0.4	0.5 ± 0.5	<0.001
Medical Treatments			
Sulfonylureas [n (%)]	10 (2.4)	24 (1.7)	0.483
Insulin [n (%)]	6 (1.4)	87 (6.3)	<0.001
Diuretics [n (%)]	20 (4.8)	136 (9.8)	0.002

ALT: Alanine Aminotransferase; AST: Aspartate Aminotransferase; BMI: Body Mass Index; DBP: Diastolic Blood Pressure; GGT: Gamma-Glutamyl Transferase; HDL-C: High-Density Lipoprotein Cholesterol; HC: Hip Circumference; HR: Heart Rate; LDL-C: Low-Density Lipoprotein Cholesterol; MetS: Metabolic Syndrome; SBP: Systolic Blood Pressure; T2DM: Type 2 Diabetes Mellitus; TC: Total Cholesterol; VLDL-C: Very-Low-Density Lipoprotein Cholesterol; WC: Waist Circumference.

7.1.1 The genetic background of obesity

Table 5. Significant SNP Correlations for comparison between control and severe obesity groups.

SNP	Gene	Biological Function	p-value
rs9939609	*FTO*	Energy homeostasis	<0.001
rs9930506	*FTO*	Energy homeostasis	<0.001
rs8050136	*FTO*	Energy homeostasis	<0.001
rs17817449	*FTO*	Energy homeostasis	<0.001
rs5982	*F13A1*	Hemostasis	0.02
rs11858624	*PYGO1*	Signal transduction	0.03
rs62021874	*PYGO1*	Signal transduction	0.03

Only correlations with a p-value less than 0.05 are included, indicating statistical significance.

After analyzing the correlations among the chosen 25 SNPs in both our control and severe obesity groups, our investigation revealed the presence of significant associations across seven SNPs. Table 5 shows all the SNPs found to be linked to severe obesity. These SNPs influence genes *FTO, PYGO1*, and *F13A1*.

7.2 Severe Obesity Study

Table 6 presents the characteristics of different subgroups of the patients with obesity group, categorized based on the SAF score (Bedossa et al., 2012): patients with obesity without NAFLD, patients with NAFL, and patients with NASH.

In terms of clinical characteristics, some variables showed no significant differences among the groups, including systolic and diastolic blood pressure, and metabolic syndrome. The percentage of female patients was significantly lower in both the NAFL and NASH groups compared to Non-NAFLD patients. However, the percentage of patients diagnosed with type 2 diabetes mellitus was significantly higher in the NAFL and NASH groups. The rest of clinical characteristics were only significantly different between Non-NAFLD and NASH groups. Patients of the NASH group showed a significantly higher incidence of hypertension and dyslipidemia compared to Non-NAFLD patients. Significant differences between both groups were also found for age, BMI and heart rate, where the Non-NAFLD group had lower values.

Regarding the biochemical variables, no significant differences were found for total or LDL cholesterol. HOMA-IR and insulin levels were significantly higher in the NASH group compared with Non-NAFLD group. Also, Non-NAFLD patients showed significantly lower

levels of VLDL cholesterol, glucose and triglycerides than NAFL and NASH patients. However, no differences are found between NAFL and NASH patients for any biochemical variables.

All transaminases showed significantly differences, with Non-NAFLD patients having lower values than NAFL and NASH patients.

Analysis of medical treatments revealed that while most treatments showed significant differences between Non-NAFLD and NASH groups, with Non-NAFLD patients having lower incidence of the treatments (except for ARBs where the roles were reversed), only biguanides exhibited a significant difference between NAFL and NASH patients with NASH patients having a significantly higher frequency of the treatment. In regards of diuretics and beta blockers, no disparities between the groups were found.

Lastly, regarding histological parameters, as expected, variables related to categorizing the groups (steatosis, inflammation, and ballooning) exhibited significant differences in all comparisons. However, fibrosis did not show significant differences when comparing Non-NAFLD and NAFL groups.

Table 6. Comparison of characteristics between patients with obesity that did not suffer from NAFLD, patients with NAFL and patients with NASH.

	Non-NAFLD (n = 405)	**NAFL** (n = 137)	**NASH** (n = 532)	**p-value**
Clinical characteristics				
Sex, woman [n (%)]	330 (81.5)	87 (64.0)	360 (67.8)	a, b
Age, years	47.3 ± 11.0	48.5 ± 10.2	49.0 ± 10.3	b
BMI, kg/m^2	44.0 ± 6.6	44.6 ± 6.1	45.3 ± 6.3	b
HR, bpm	74.8 ± 11.3	74.0 ± 12.2	78.4 ± 14.7	b
SBP, mmhg	130.1 ± 20.0	132.2 ± 18.6	130.8 ± 20.4	
DBP, mmhg	76.7 ± 12.6	76.6 ± 13.9	77.3 ± 13.9	
HC, cm	134.3 ± 14.2	136.2 ± 13.6	137.6 ± 15.6	b
WC, cm	127.5 ± 15.9	128.7 ± 17.3	133.1 ± 13.8	b, c
T2DM [n (%)]	65 (16.0)	38 (27.7)	180 (33.8)	a, b
Hypertension [n (%)]	141 (34.8)	56 (40.9)	262 (49.2)	b
Dyslipidemia [n (%)]	65 (16.0)	31 (22.6)	161 (30.3)	b
MetS [n (%)]	148 (36.5)	61 (44.5)	217 (40.8)	

Biochemical variables				
TC, mmol/L	4.1 ± 0.9	4.1 ± 0.9	4.2 ± 1.0	
HDL-C, mmol/L	1.1 ± 0.3	1.0 ± 0.4	1.0 ± 0.4	b
LDL-C, mmol/L	2.5 ± 0.8	2.5 ± 0.9	2.5 ± 1.1	
VLDL-C, mmol/L	0.6 ± 0.3	0.8 ± 0.6	0.8 ± 0.4	a, b
Glucose, mmol/L	6.9 ± 2.9	7.6 ± 2.8	7.8 ± 3.2	a, b
Insulin, pmol/L	90.9 ± 123.0	98.9 ± 91.5	128.7 ± 290.6	b
Triglycerides, mmol/L	1.5 ± 0.7	1.9 ± 1.4	1.8 ± 1.0	a, b
HOMA-IR	4.7 ± 8.6	5.6 ± 8.3	7.6 ± 19.1	b
Transaminases				
ALT, μKat/L	0.5 ± 0.3	0.7 ± 0.4	0.8 ± 0.5	a, b
AST, μKat/L	0.5 ± 0.4	0.7 ± 0.4	0.7 ± 0.5	a, b
GGT, μKat/L	0.4 ± 0.5	0.5 ± 0.6	0.6 ± 0.6	a, b
Medical treatments				
Sulfonylureas [n (%)]	3 (0.7)	4 (2.9)	15 (2.8)	b
Insulin [n (%)]	17 (4.2)	12 (8.8)	44 (8.3)	b
Diuretics [n (%)]	40 (9.9)	16 (11.7)	59 (11.1)	
CCBs [n (%)]	15 (3.7)	7 (5.1)	38 (7.1)	b
Biguanides [n (%)]	42 (10.4)	20 (14.6)	142 (26.7)	b, c
BBs [n (%)]	13 (3.2)	1 (0.7)	22 (4.1)	
ARBs [n (%)]	24 (5.9)	12 (8.8)	64 (2.0)	b
GLP-1Ras [n (%)]	7 (1.7)	3 (2.2)	25 (4.7)	b
Histological parameters				
Steatosis [n (%)]				a, b, c
<5%	405 (100.0)	0 (0.0)	0 (0.0)	
5-33%	0 (0.0)	100 (73.0)	266 (50.0)	
34-66%	0 (0.0)	35 (25.5)	184 (34.6)	
>66%	0 (0.0)	2 (1.5)	82 (15.4)	
Inflammation [n (%)]				a, b, c
No foci	88 (21.7)	46 (33.6)	8 (1.5)	
<2 foci per 200 field	223 (55.1)	91 (66.4)	241 (45.3)	
2-4 foci per 200 field	78 (19.3)	0 (0.0)	232 (43.6)	
>4 foci per 200 field	16 (4.0)	0 (0.0)	51 (9.6)	
Ballooning [n (%)]				a, b, c
None	214 (52.8)	121 (88.3)	77 (14.5)	
Few cells	127 (31.4)	16 (11.7)	251 (47.2)	
Many cells	64 (15.8)	0 (0.0)	204 (38.3)	
Fibrosis [n (%)]				b, c
None	59 (14.9)	23 (16.8)	28 (5.4)	
Perisinusoidal or periportal	156 (39.5)	61 (44.5)	154 (29.7)	
Perisinusoidal and periportal	145 (36.7)	43 (31.4)	227 (43.7)	
Bridging fibrosis or cirrhosis	35 (8.9)	10 (7.3)	105 (22.2)	

P value < 0.05 is indicated as (a) Non-NAFLD vs NAFLD, (b) Non-NAFLD vs NASH, (c) NAFLD vs NASH. ALT: Alanine Aminotransferase; ARBs: Angiotensin II Receptor Blockers; AST: Aspartate Aminotransferase; BBs: Beta Blockers; BMI: Body Mass Index; CCBs: Calcium Channel Blockers; CBC: Complete Blood Count; DBP: Diastolic Blood Pressure; GGT: Gamma-Glutamyl Transferase; HDL-C: High-Density Lipoprotein Cholesterol; HC: Hip Circumference; HR: Heart Rate; LDL-C: Low-Density Lipoprotein Cholesterol; MetS: Metabolic Syndrome; SBP: Systolic Blood Pressure; T2DM: Type 2 Diabetes Mellitus; TC: Total Cholesterol; VLDL-C: Very-Low-Density Lipoprotein Cholesterol; WC: Waist Circumference; GLP-1Ras: GLP-1 Receptor Antagonists.

7.3 The genetic background of NAFLD and histological scores

We analyzed the associations of the selected 25 SNPs comparing subsets within the severe obesity cohort, including individuals with NAFL, NASH, and those without NAFLD. Moreover, we analyzed the associations of the SNPs in the severe obesity group based on histological scores, encompassing parameters like steatosis, fibrosis, ballooning, and inflammation.

Figure 4 visually represents the correlations between the identified 25 SNPs and the three patient groups: NAFL, NASH, and Non-NAFLD. We examined the frequency of each SNP combination within the patient groups to discern any statistically significant associations.

Our analysis revealed four SNPs (rs58542926, rs1800629, rs8107974, and rs5982) that exhibited statistically significant correlations with the different patient groups. Moreover, SNP rs5982 was also found significantly correlated with the presence of obesity as shown by Table 5.

Notably, for all statistically significant tests, we observed that the most substantial difference in frequency consistently occurred between the patients with obesity and one of the other groups.

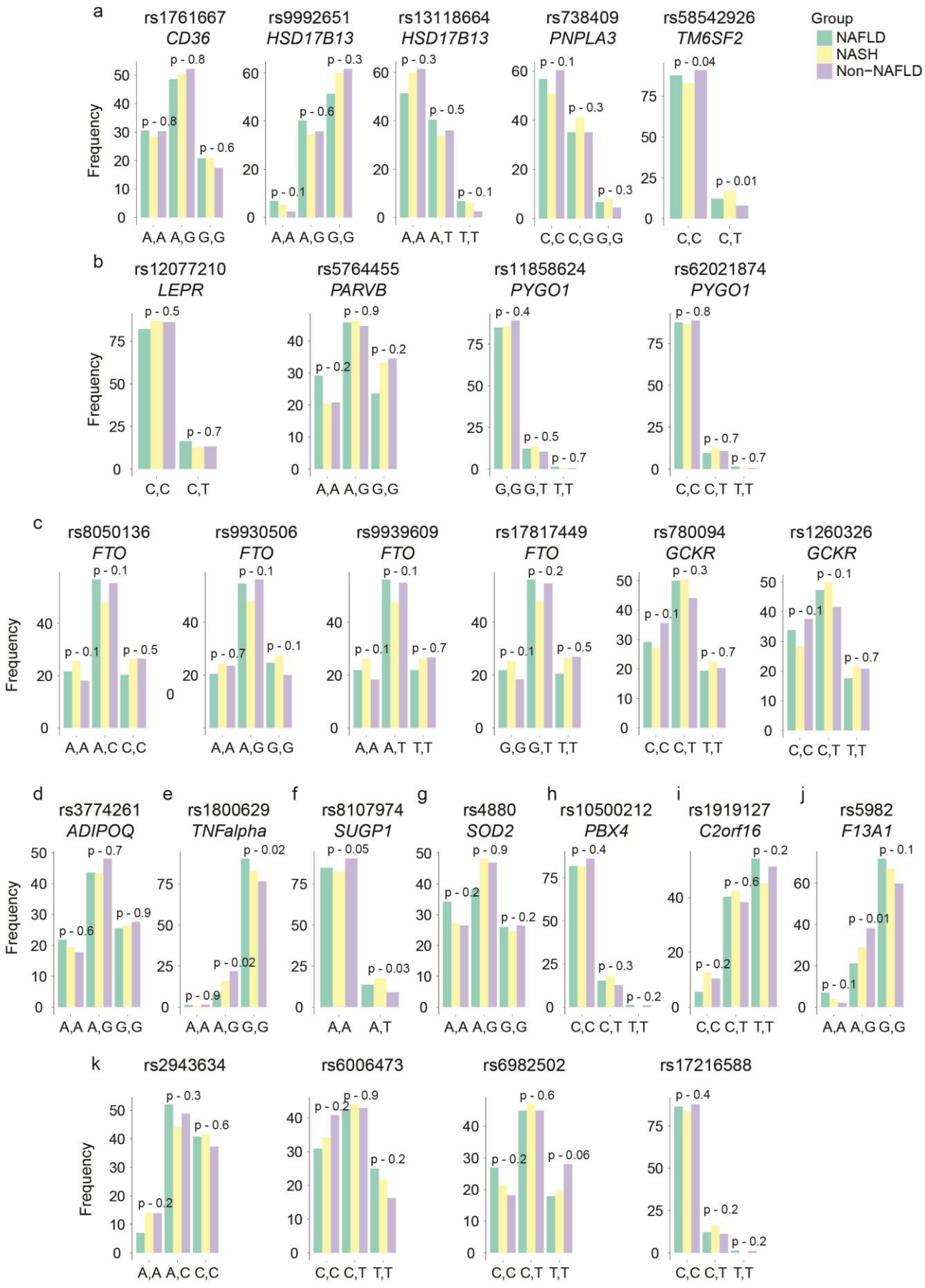
a
rs1761667
CD36
rs9992651
HSD17B13
rs13118664
HSD17B13
rs738409
PNPLA3
rs58542926
TM6SF2
Group
NAFLD
NASH
Non−NAFLD
Frequency
A,A A,G G,G
C,C C,G G,G
C,C C,T
b
rs12077210
LEPR
rs5764455
PARVB
rs11858624
PYGO1
rs62021874
PYGO1
G,G G,T T,T
c
rs8050136
FTO
rs9930506
FTO
rs9939609
FTO
rs17817449
FTO
rs780094
GCKR
rs1260326
GCKR
A,A A,C C,C
A,A A,T T,T
C,C C,T T,T
d
rs3774261
ADIPOQ
e
rs1800629
TNFalpha
f
rs8107974
SUGP1
g
rs4880
SOD2
h
rs10500212
PBX4
i
rs1919127
C2orf16
j
rs5982
F13A1
A,A
k
rs2943634
rs6006473
rs6982502
rs17216588

Figure 4. Correlation between SNPs and Patient Groups (NAFL, NASH, Non-NAFLD). The 25 grouped bar plots examine the correlation between SNPs and their frequencies within three patient groups. Each graph is dedicated to a specific SNP and, if applicable, includes the accompanying gene name in the title. For every SNP graph, the x-axis displays the different combinations of alleles for the SNP. The y-axis of each graph indicates the frequency of the SNP combination within each patient group. For every group of bars, there are p-values available that help determine the statistical significance of the SNP associations. These p-values assess whether a specific combination of SNP is related to being part of a particular patient group such as NAFL, NASH, or Non-NAFLD. The SNPs are arranged in the following order, classified based on the biological process of the gene they are associated with: a - Lipid metabolism and transport, b - Signal transduction, c - Energy homeostasis, d – Hormone, e – Inflammation, f - mRNA processing, g - Oxidative stress response, h - Transcription regulation, i - Unknown function, j – Hemostasis, k - Not within a gene.

Table 7. Significant SNP Correlations for Steatosis, Ballooning, and Inflammation.

SNP	Gene	Biological Function	p-value
Steatosis			
rs58542926	*TM6SF2*	Lipid metabolism and transport	0.02
rs11858624	*PYGO1*	Signal transduction	0.009
rs62021874	*PYGO1*	Signal transduction	0.02
rs1800629	*TNFA*	Cytokine	0.03
Ballooning			
rs1919127	*C2orf16*	Unknown	0.03
Inflammation			
rs12077210	*LEPR*	Signal transduction	0.01
rs8050136	*FTO*	Energy homeostasis	0.02
rs9939609	*FTO*	Energy homeostasis	0.03

Only correlations with a p-value less than 0.05 are included, indicating statistical significance.

Table 7 shows the significant SNPs found in comparing the histological scores. SNPs rs58542926 and rs1800629, found to correlate with the diagnosis of NAFLD in the severe obesity group, are also found to affect the steatosis score of the patients. SNPs rs11858624, rs62021874, rs8050136, and rs9939609 affecting *PYGO1* and *FTO* genes, respectively, were found to correlate with the presence of obesity in the control and severe obesity groups comparison. The table shows that the SNPs affecting the *PYGO1* gene affected the patients' steatosis score, while the SNPs affecting the *FTO* gene affected the inflammation score.

7.3.1 SNPs predictive power

Upon examining the associations between selected SNPs and comparative groups, we evaluated the robustness of these correlations. This assessment aimed to ascertain the potential predictive capacity that these specific SNPs might possess.

Table 8 displays the strength of significant correlations between specific SNPs and multiple patient groups. Each row of the table presents essential information, including the SNP identifier, associated gene, biological function, and the corresponding correlation strength measured by Cramer's V. Cramer's V is a statistic that ranges from 0 to 1, with values closer to 1, indicating stronger correlations.

Table 8. Strength of Correlation (Cramer's V) between SNPs and groups for control (control and severe obesity groups), diagnosis (Non-NAFLD, NAFL and NASH), steatosis, ballooning, and inflammation.

SNP	Gene	Biological Function	Strength
Control			
rs9939609	*FTO*	Energy homeostasis	0.13
rs9930506	*FTO*	Energy homeostasis	0.14
rs8050136	*FTO*	Energy homeostasis	0.14
rs17817449	*FTO*	Energy homeostasis	0.13
rs5982	*F13A1*	Hemostasis	0.08
rs11858624	*PYGO1*	Signal transduction	0.07
rs62021874	*PYGO1*	Signal transduction	0.07
Diagnosis			
rs58542926	*TM6SF2*	Lipid metabolism and transport	0.12
rs1800629	*TNFA*	Cytokine	0.11
rs8107974	*SUGP1*	mRNA processing	0.11
rs5982	*F13A1*	Hemostasis	0.12
Steatosis			
rs58542926	*TM6SF2*	Lipid metabolism and transport	0.12
rs11858624	*PYGO1*	Signal transduction	0.12
rs62021874	*PYGO1*	Signal transduction	0.13
rs1800629	*TNFA*	Cytokine	0.12
Ballooning			
rs1919127	*C2orf16*	Unknown	0.11
Inflammation			
rs12077210	*LEPR*	Signal transduction	0.13
rs8050136	*FTO*	Energy homeostasis	0.12
rs9939609	*FTO*	Energy homeostasis	0.12

It is evident that the SNPs investigated in this study exhibited relatively weak correlations with the patient groups. The calculated correlation strengths consistently fall within the range of 0.07 to 0.14.

Furthermore, the SNP rs5982, which demonstrated a notable association with obesity and NAFLD development, exhibited a stronger correlation, as evidenced in Table 8, in the context of NAFLD development. This strengthened association underscores that the link between this SNP and the disease's onset is not solely attributed to its connection with obesity.

7.3.2 Selected SNPs do not serve as predictive indicators of NAFLD status

We employed a Random Forest approach using R and the randomForest package to investigate whether the genetic markers found to be significantly correlated with NAFLD development could reliably predict disease progression.

We used a random data set of 70% of the total data to train the Random Forest model. Once the model was trained, we tested its predictive capability with the remaining data not used during training. By utilizing this technique, we sought to evaluate the predictive strength of SNPs in distinguishing between morbidly obese patients without disease, those with NAFL, and those with NASH.

Upon analyzing the performance of the Random Forest model, we observed that, despite significant correlations between SNPs and the three patient groups, their predictive power fell short. The model's error rate remained high, consistently hovering around 49%.

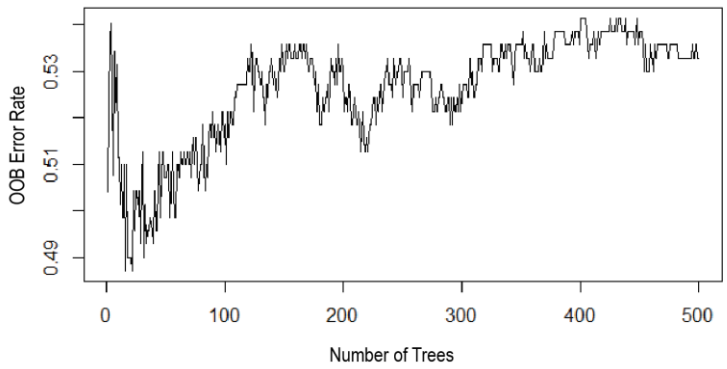

Figure 5. Out-of-Bag Error Rate of Random Forest Model with Different Number of Decision Trees for SNP-Based Disease Prediction in Morbidly Obese Patients with NAFL and NASH.

Figure 5 illustrates the Random Forest model's out-of-bag (OOB) error rate as a function of the number of decision trees used in the ensemble. The OOB error rate is a metric commonly used in Random Forest analysis and represents the prediction error on each data point using only the trees that were not trained on that particular data point. It is a proxy for the model's performance on unseen data where lower error rates indicate a more robust predictive model.

7.4 Discussion

NAFL and NASH are significant global health challenges, with obesity emerging as a notable risk factor (Nassir, 2022). It is interesting to note that in cases of obesity, not all individuals experience the same disease progression. Some remain unaffected, while others may develop NAFL, and a subset may progress to the more severe stage of NASH. The variation in the disease suggests that the development and progression of NAFLD involve a complex interaction of different factors (Pouwels et al., 2022). It is becoming apparent that a "multiple-hit" pathogenic model influences disease susceptibility and progression (Buzzetti et al., 2016).

In this context, the study of SNPs holds the potential to illuminate the genetic variations that contribute to an individual's predisposition to NAFLD and NASH. Ultimately, the study of SNPs can facilitate advancements in diagnosis and treatment.

Our analysis has revealed a multitude of SNPs that show correlations with the diagnosis (NAFL, NASH, or non-NAFLD) and the histological scores within the severe obesity group. Furthermore, we have identified SNPs that exhibit correlations when comparing the severe obesity group with the control group. These findings help us study the impact of SNPs on disease progression and identify those contributing to NAFLD, regardless of obesity.

We identified significant correlations between the C > T allele change of SNP rs58542926, and NAFLD development. The T allele's presence exhibited a higher frequency in the NASH group followed by the NAFL group. Furthermore, the T allele was notably more frequent in individuals presenting higher steatosis scores. This finding suggests a potential link between the T allele and increased hepatic fat accumulation.

Another study on rs58542926 (Dongiovanni et al., 2014) explored the impact of the SNP on various aspects of NAFLD. In congruence with our observations, this study reported a significant association between the rs58542926 C > T allele change and the development of NASH. Furthermore, both our study and (Dongiovanni et al., 2014) identified correlations between the T allele and higher histological steatosis scores. However, it is noteworthy that our findings diverge from (Dongiovanni et al., 2014) in certain aspects. While (Dongiovanni et al., 2014) highlighted correlations between the rs58542926 variant and NASH development and steatosis scores, our study did not yield significant associations with fibrosis, ballooning, and inflammation. One plausible explanation for this divergence could lie in the cascading effects of the observed steatosis correlations. The pronounced influence

of the rs58542926 variant on hepatic fat accumulation, evident in both studies, could act as a catalyst for subsequent processes. Elevated steatosis, driven by the genetic variant, may set the stage for secondary events such as inflammation and fibrosis. This cascade of events, albeit significant in its impact, might not manifest as a correlation in our study due to factors such as sample size limitations.

The SNP rs58542926 is located in the *TM6SF2* gene. This gene association with NAFLD has been observed by a several studies (Xu et al., 2019). These studies suggest a potential involvement of *TM6SF2* in regulating lipid dynamics within the liver showing significance in lipid metabolism and VLDL lipidation. VLDLs are vital in lipid secretion into the circulation. VLDLs undergo a complex process, moving from the endoplasmic to the Golgi reticulum, where they undergo essential modifications before being transported to the plasma membrane and released into the circulatory system (Tiwari & Siddiqi, 2012). Disruptions in VLDL secretion pathways can lead to lipid accumulation within hepatocytes and, therefore, to NAFLD development (Wang et al., 2016). An in-depth investigation (Luo et al., 2022) explores the function of *TM6SF2* in VLDL lipidation concluding that loss of functional *TM6SF2* can contribute to hepatic steatosis by affecting VLDL secretion.

Luo et al.'s research delves into the implications of *TM6SF2* on this process. They propose that functional loss of *TM6SF2* could hinder the proper lipidation of VLDL particles. Consequently, this disruption could produce smaller-sized VLDL particles that struggle to be secreted into the bloodstream, leading to the accumulation of fat within hepatocytes. Our study aligns with this mechanism, as we observed a correlation between the rs58542926 variant and both NAFLD development and fat accumulation within hepatocytes.

The SNP rs1800629 G > A, located in the upstream region of the *TNFA* gene, encodes for a pro-inflammatory cytokine involved in signaling events that lead to necrosis and apoptosis (Idriss & Naismith, 2000). Our results showed a significative correlation for the SNP and NAFLD development. The A allele was present with a higher frequency in the non-NAFLD group followed by the NASH group. Furthermore, the A allele was notably more frequent in individuals presenting lower steatosis scores.

While the SNP rs1800629 has been extensively investigated for its association with various conditions, such as cardiomyopathy (Y. Zhang et al., 2018), its direct link to NAFL progression to NASH remains relatively underexplored in the existing literature. It has been established (Astarini et al., 2022) that the G > A change in the SNP results in higher *TNFA*

expression. Also, a heightened *TNFA* circulation has been correlated with the progression of NAFL to NASH (Potoupni et al., 2021). However, we encountered some intriguing results that deviate from the anticipated associations. While the literature has suggested that the G > A change in the SNP leads to higher *TNFA* expression and is correlated with inflammation and the progression of NAFL to NASH (Kurbatova et al., 2017), our findings did not find a correlation with hepatic lobular inflammation, a key function of the gene.

One notable factor that emerged from our analysis is the inclusion of non-NAFLD patients in our control group. These patients lack steatosis, a defining characteristic of NAFLD, and their presence has potentially influenced our results. The absence of steatosis in our control group may have contributed to the observation that the A allele of the SNP correlates with decreased steatosis scores in the liver.

Moreover, although we did not observe any correlation with the inflammation score, we found significant differences in the inflammation analysis when we removed the non-NAFLD patient group. Patients with a higher presence of inflammation had the allele A. This contrasted with the anticipated pro-inflammatory role of *TNFA* associated with the A allele.

Additionally, our analysis revealed significant correlations between the A > T allele change of SNP rs8107974 and NAFLD progression, where the frequency of the T allele was higher in the NASH and NAFLD groups.

SNP rs8107974 is located in the genomic region of the SURP and G patch domain containing 1 gene (*SUGP1*). *SUGP1* encodes a protein involved in pre-mRNA splicing by participating in the assembly of the spliceosome. The protein contains distinct functional domains, including the SURP and G patch domains, which are known to play essential roles in RNA binding.

Although no studies have investigated the association between the *SUGP1* gene and the NAFLD spectrum, multiple investigations have explored its association with of cancer (J. Zhang et al., 2022). A recent study (Kim et al., 2016) has revealed a significant link between *SUGP1* and the regulation of cholesterol metabolism. The study showed that reducing *SUGP1* expression using a targeted knockdown approach resulted in a heightened cholesterol intake alongside a disrupted secretion of LDL particles.

The significant increase in the frequency of the T allele for the SNP rs8107974 in both NASH and NAFLD groups suggests that the heightened risk associated with this allele could be due to a reduction in the expression or function of the *SUGP1* protein.

SNP rs5982 G > A allele change was also found significantly correlated with NAFLD development. Specifically, the G allele was shown to be present with a higher frequency in the NAFLD group followed by the NASH group. Our analysis also showed a significant correlation between the SNP and the presence of obesity but given that the strength of the association for the obesity presence is weaker than the strength of the correlation with NAFLD progression it is unlikely that the later association is caused by the first.

The SNP rs5982 is found within the gene *F13A1*, which encodes the alpha subunit of coagulation factor XIII. Factor XIII is a pivotal clotting factor essential for the final stages of blood clot formation. Research has explored *F13A1*'s potential relevance to obesity beyond its role in blood clotting. A recent study on human weight gain (Kaartinen et al., 2021) revealed significant increases in *F13A1* expression in adipose tissue in individuals with higher body weight. However, the study found no significant correlation between *F13A1* and liver fat accumulation.

Even though correlations were observed between disease development and the presence of specific SNPs within the severe obesity group, our results showed that the strength of these associations is notably weak. This observation is further validated by the substantial error rates encountered in the trained Random Forest model when attempting to predict a patient's disease status (NAFL, NASH, or non-NAFLD) using the identified SNPs. Our research suggests that relying solely on SNPs may not accurately predict NAFLD outcomes due to limitations in their association strength.

In contrast, existing endeavors to predict disease development have leaned on a broader array of characteristics, such as triglyceride levels, glucose levels, and BMI. These composite predictors have demonstrated high success rates, reaching up to 90% accuracy in some instances (Peng et al., 2023). A combination of genetic markers and clinical traits could further enhance this predictive accuracy, overcoming the limitations of using genetic markers or clinical characteristics alone.

Moreover, our analysis focused solely on a subset of four SNPs correlated with disease development out of the 25 initially proposed candidates. Expanding the scope of correlated SNPs could also enhance the predictive potential of our model.

8 Conclusion

Our genotyping data analysis has uncovered multiple SNPs distinctly associated with disease and obesity, aligning with our first objective of determining genetic correlations. Four significant SNPs have emerged as key players in disease progression among our subset of severely obese patients. Specifically, rs58542926, rs1800629, rs8107974, and rs5982 have showcased a correlation with disease presence. Among them, rs5982 showed a dual significance in both disease and obesity. However, given that it presented a stronger association with the disease, we can conclude that the effect of the SNP on the disease is not entirely related to its role in obesity.

Our analysis of the correlation of these SNPs based on histological scores has also given us more profound insights into disease mechanisms. Notably, rs58542926 showed an effect over steatosis score between severe obesity participants. Moreover, rs1800629 analysis revealed an effect in the inflammation score comparing participants with NAFLD and NASH.

For our second objective, we analyzed the strength of associations of the SNPs that we found correlated. While offering disease insights, the SNPs demonstrated limited predictive power, showing low strength values. Moreover, the training and testing of a random forest model presented high error rates, making it unviable for predicting the liver condition in patients with severe obesity.

In conclusion, although SNPs help uncover disease mechanisms, they lack strong predictive power for liver conditions in obesity types II and III. This highlights the need for a more robust set of SNPs with stronger disease correlations and the combination of clinical characteristics to further enhance the prediction.

9 Acknowledgements

I want to express my sincere gratitude to my academic tutor, Dr. Ana Fernández Bravo, for her invaluable assistance and unwavering support throughout the journey of working on my bachelor's thesis.

I am deeply grateful to my laboratory team, especially my supervisor, Dr. Jorge Joven Maried. He gave me the opportunity to join his research and provided exceptional guidance. His efforts instilled in me a profound appreciation for continuous learning and improvement.

I also want to extend my heartfelt appreciation to the predoctoral researcher, Helena Castañé Vilafranca. Her mentorship, patience and constant support were invaluable during my internship and throughout the entire process of realizing my bachelor's thesis.

In addition, I would like to express my heartfelt gratitude to my family and friends. Their unwavering support, encouragement, and belief in my abilities have been a driving force throughout this academic endeavor.

A special note of gratitude goes to the individuals who participated as patients in this research and made this bachelor thesis possible.

10 References

Andy Liaw, & Matthew Wiener. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Anstee, Q. M., Darlay, R., Cockell, S., Meroni, M., Govaere, O., Tiniakos, D., Burt, A. D., Bedossa, P., Palmer, J., Liu, Y. L., Aithal, G. P., Allison, M., Yki-Järvinen, H., Vacca, M., Dufour, J. F., Invernizzi, P., Prati, D., Ekstedt, M., Kechagias, S., … Daly, A. K. (2020). Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort☆. *Journal of Hepatology*, *73*(3), 505–515. https://doi.org/10.1016/j.jhep.2020.04.003

Anstee, Q. M., Targher, G., & Day, C. P. (2013). Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 10, Issue 6, pp. 330–344). https://doi.org/10.1038/nrgastro.2013.41

Astarini, F. D., Ratnasari, N., & Wasityastuti, W. (2022). Update on Non-Alcoholic Fatty Liver Disease-Associated Single Nucleotide Polymorphisms and Their Involvement in Liver Steatosis, Inflammation, and Fibrosis: A Narrative Review. *Iranian Biomedical Journal*, *26*(4), 252–268. https://doi.org/10.52547/ibj.3647

Bedossa, P., Poitou, C., Veyrie, N., Bouillot, J. L., Basdevant, A., Paradis, V., Tordjman, J., & Clement, K. (2012). Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. *Hepatology*, *56*(5), 1751–1759. https://doi.org/10.1002/hep.25889

Bienvenu, T., Lopez, M., & Girodon, E. (2020). Molecular diagnosis and genetic counseling of cystic fibrosis and related disorders: New challenges. In *Genes* (Vol. 11, Issue 6, pp. 1–16). MDPI AG. https://doi.org/10.3390/genes11060619

Buzzetti, E., Pinzani, M., & Tsochatzis, E. A. (2016). The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). *Metabolism: Clinical and Experimental*, *65*(8), 1038–1048. https://doi.org/10.1016/j.metabol.2015.12.012

Di Taranto, M. D., Giacobbe, C., & Fortunato, G. (2020). Familial hypercholesterolemia: A complex genetic disease with variable phenotypes. In *European Journal of Medical Genetics* (Vol. 63, Issue 4). Elsevier Masson SAS. https://doi.org/10.1016/j.ejmg.2019.103831

Dongiovanni, P., Petta, S., Maglio, C., Fracanzani, A. L., Pipitone, R., Mozzi, E., Motta, B. M., Kaminska, D., Rametta, R., Grimaudo, S., Pelusi, S., Montalcini, T., Alisi, A., Maggioni, M., K€ Arj€, V., Bor, J., K€ Akel€, P., Marco, V. Di, Xing, C., … Valenti, L. (2014). *Transmembrane 6 Superfamily Member 2 Gene Variant Disentangles Nonalcoholic Steatohepatitis From Cardiovascular Disease*. https://doi.org/10.1002/hep.27490/suppinfo

Eichler, E. E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *New England Journal of Medicine*, *381*(1), 64–74. https://doi.org/10.1056/nejmra1809315

Fritz, T., Niederreiter, L., Adolph, T., Blumberg, R. S., & Kaser, A. (2011). Crohn's disease: NOD2, autophagy and ER stress converge. In *Gut* (Vol. 60, Issue 11, pp. 1580–1588). https://doi.org/10.1136/gut.2009.206466

Germain, D. P., Moiseev, S., Suárez-Obando, F., Al Ismaili, F., Al Khawaja, H., Altarescu, G., Barreto, F. C., Haddoum, F., Hadipour, F., Maksimova, I., Kramis, M., Nampoothiri, S., Nguyen, K. N., Niu, D. M., Politei, J., Ro, L. S., Vu Chi, D., Chen, N., & Kutsev, S. (2021). The benefits and challenges of family genetic testing in rare genetic diseases—lessons from Fabry disease. In *Molecular Genetics and Genomic Medicine* (Vol. 9, Issue 5). John Wiley and Sons Inc. https://doi.org/10.1002/mgg3.1666

Gong, T., Hayes, V. M., & Chan, E. K. F. (2021). Detection of somatic structural variants from short-read next-generation sequencing data. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa056

Guan, B., Zhao, Y., Yin, Y., & Li, Y. (2022). Detecting Disease-Associated SNP-SNP Interactions Using Progressive Screening Memetic Algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 878–887. https://doi.org/10.1109/TCBB.2020.3019256

Hadley Wickham. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*.

Horton, R. H., & Lucassen, A. M. (2019). Recent developments in genetic/genomic medicine. In *Clinical Science* (Vol. 133, Issue 5, pp. 697–708). Portland Press Ltd. https://doi.org/10.1042/CS20180436

Idriss, H. T., & Naismith, J. H. (2000). TNFα and the TNF receptor superfamily: Structure-function relationship(s). *Microscopy Research and Technique*, *50*(3), 184–195. https://doi.org/10.1002/1097-0029(20000801)50:3<184::AID-JEMT2>3.0.CO;2-H

Ishigaki, K. (2022). Beyond GWAS: from simple associations to functional insights. In *Seminars in Immunopathology* (Vol. 44, Issue 1, pp. 3–14). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00281-021-00894-5

Kaartinen, M. T., Arora, M., Heinonen, S., Hang, A., Barry, A., Lundbom, J., Hakkarainen, A., Lundholm, N., Rissanen, A., Kaprio, J., & Pietiläinen, K. H. (2021). F13A1 transglutaminase expression in human adipose tissue increases in acquired excess weight and associates with inflammatory status of adipocytes. *International Journal of Obesity*, *45*(3), 577–587. https://doi.org/10.1038/s41366-020-00722-0

Kazuki Yoshida, & Alexander Bartel. (2022). *tableone: Create "Table 1" to Describe Baseline Characteristics with or without Propensity Score Weights*. Https://CRAN.R-Project.Org/Package=tableone.

Kim, M. J., Yu, C. Y., Theusch, E., Naidoo, D., Stevens, K., Kuang, Y. L., Schuetz, E., Chaudhry, A. S., & Medina, M. W. (2016). SUGP1 is a novel regulator of cholesterol metabolism. *Human Molecular Genetics*, *25*(14), 3106–3116. https://doi.org/10.1093/hmg/ddw151

Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., Ferrell, L. D., Liu, Y. C., Torbenson, M. S., Unalp-Arida, A., Yeh, M., McCullough, A. J., & Sanyal, A. J. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*, *41*(6), 1313–1321. https://doi.org/10.1002/hep.20701

Kurbatova, I. V., Topchieva, L. V., & Dudanova, O. P. (2017). Gene TNF Polymorphism -308G>A (rs1800629) and Its Relationship with the Efficiency of Ursodeoxycholic Acid Therapy in Patients with Nonalcoholic Stetohepatitis. *Bulletin of Experimental Biology and Medicine*, *164*(2), 181–185. https://doi.org/10.1007/s10517-017-3953-1

Lappalainen, T., & Macarthur, D. G. (2021). *From variant to function in human disease genetics*. https://www.science.org

Lonardo, A., Nascimbeni, F., Maurantonio, M., Marrazzo, A., Rinaldi, L., & Adinolfi, L. E. (2017). Nonalcoholic fatty liver disease: Evolving paradigms. In *World Journal of*

Gastroenterology (Vol. 23, Issue 36, pp. 6571–6592). Baishideng Publishing Group Co. https://doi.org/10.3748/wjg.v23.i36.6571

Luo, F., Oldoni, F., & Das, A. (2022). TM6SF2: A Novel Genetic Player in Nonalcoholic Fatty Liver and Cardiovascular Disease. In *Hepatology Communications* (Vol. 6, Issue 3, pp. 448–460). John Wiley and Sons Inc. https://doi.org/10.1002/hep4.1822

Manikat, R., & Nguyen, M. H. (2023). NAFLD and Non-Liver Comorbidities. *Clinical and Molecular Hepatology*. https://doi.org/10.3350/cmh.2022.0442

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. In *Nature* (Vol. 461, Issue 7265, pp. 747–753). https://doi.org/10.1038/nature08494

Meyer D, Zeileis A, & Hornik K. (2023). *vcd: Visualizing Categorical Data_. R package version 1.4-11*. Https://CRAN.R-Project.Org/Package=vcd.

Nassir, F. (2022). NAFLD: Mechanisms, Treatments, and Biomarkers. In *Biomolecules* (Vol. 12, Issue 6). MDPI. https://doi.org/10.3390/biom12060824

Peng, H., Pan, L., Ran, S., Wang, M., Huang, S., Zhao, M., Cao, Z., Yao, Z., Xu, L., Yang, Q., & Lv, W. (2023). Prediction of MAFLD and NAFLD using different screening indexes: A cross-sectional study in U.S. adults. *Frontiers in Endocrinology*, *14*. https://doi.org/10.3389/fendo.2023.1083032

Piazzolla, V. A., & Mangia, A. (2020). Noninvasive diagnosis of NAFLD and NASH. In *Cells* (Vol. 9, Issue 4). MDPI. https://doi.org/10.3390/cells9041005

Potoupni, V., Georgiadou, M., Chatzigriva, E., Polychronidou, G., Markou, E., Zapantis Gakis, C., Filimidou, I., Karagianni, M., Anastasilakis, D., Evripidou, K., Ftergioti, A., Togkaridou, M., Tsaftaridis, N., Apostolopoulos, A., & Polyzos, S. A. (2021). Circulating tumor necrosis factor-α levels in non-alcoholic fatty liver disease: A systematic review and a meta-analysis. *Journal of Gastroenterology and Hepatology*, *36*(11), 3002–3014. https://doi.org/10.1111/jgh.15631

Pouwels, S., Sakran, N., Graham, Y., Leal, A., Pintar, T., Yang, W., Kassir, R., Singhal, R., Mahawar, K., & Ramnarain, D. (2022). Non-alcoholic fatty liver disease (NAFLD): a

review of pathophysiology, clinical management and effects of weight loss. In *BMC Endocrine Disorders* (Vol. 22, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12902-022-00980-1

Raza, S., Rajak, S., Upadhyay, A., Tewari, A., & Sinha, R. A. (2021). *Current treatment paradigms and emerging therapies for NAFLD/NASH*.

Saleem, M., Ghazali, M. B., Wahab, M. A. M. A., Yusoff, N. M., Mahsin, H., Seng, C. E., Khalid, I. A., Rahman, M. N. G., & Yahaya, B. H. (2018). *The BRCA1 and BRCA2 Genes in Early-Onset Breast Cancer Patients* (pp. 1–12). https://doi.org/10.1007/5584_2018_147

Shastry, B. S. (2009). SNPs: impact on gene function and phenotype. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 578, pp. 3–22). https://doi.org/10.1007/978-1-60327-411-1_1

Speakman, J. R., Loos, R. J. F., O'Rahilly, S., Hirschhorn, J. N., & Allison, D. B. (2018). GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity. In *International Journal of Obesity* (Vol. 42, Issue 8, pp. 1524–1531). Nature Publishing Group. https://doi.org/10.1038/s41366-018-0147-5

Stoker, T. B., Mason, S. L., Greenland, J. C., Holden, S. T., Santini, H., & Barker, R. A. (2022). Huntington's disease: diagnosis and management. In *Practical neurology* (Vol. 22, Issue 1, pp. 32–41). NLM (Medline). https://doi.org/10.1136/practneurol-2021-003074

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. In *Nature Reviews Genetics* (Vol. 20, Issue 8, pp. 467–484). Nature Publishing Group. https://doi.org/10.1038/s41576-019-0127-1

Tiwari, S., & Siddiqi, S. A. (2012). Intracellular trafficking and secretion of VLDL. In *Arteriosclerosis, Thrombosis, and Vascular Biology* (Vol. 32, Issue 5, pp. 1079–1086). https://doi.org/10.1161/ATVBAHA.111.241471

Trépo, E., & Valenti, L. (2020). Update on NAFLD genetics: From new variants to the clinic. In *Journal of Hepatology* (Vol. 72, Issue 6, pp. 1196–1209). Elsevier B.V. https://doi.org/10.1016/j.jhep.2020.02.020

Wang, Y., Liu, L., Zhang, H., Fan, J., Zhang, F., Yu, M., Shi, L., Yang, L., Lam, S. M., Wang, H., Chen, X., Wang, Y., Gao, F., Shui, G., & Xu, Z. (2016). Mea6 controls VLDL transport through the coordinated regulation of COPII assembly. *Cell Research*, *26*(7), 787–804. https://doi.org/10.1038/cr.2016.75

Xu, M., Li, Y., Zhang, S., Wang, X., Shen, J., & Zhang, S. (2019). Interaction of TM6SF2 E167K and PNPLA3 I148M variants in NAFLD in northeast China. *Annals of Hepatology*, *18*(3), 456–460. https://doi.org/10.1016/j.aohep.2018.10.005

Younossi, Z., Anstee, Q. M., Marietti, M., Hardy, T., Henry, L., Eslam, M., George, J., & Bugianesi, E. (2018). Global burden of NAFLD and NASH: Trends, predictions, risk factors and prevention. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 15, Issue 1, pp. 11–20). Nature Publishing Group. https://doi.org/10.1038/nrgastro.2017.109

Zhang, J., Huang, J., Xu, K., Xing, P., Huang, Y., Liu, Z., Tong, L., & Manley, J. L. (2022). DHX15 is involved in SUGP1-mediated RNA missplicing by mutant SF3B1 in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(49). https://doi.org/10.1073/pnas.2216712119

Zhang, Y., Cao, Y., Xin, L., Gao, N., & Liu, B. (2018). Association between rs1800629 polymorphism in tumor necrosis factor-ά gene and dilated cardiomyopathy susceptibility Evidence from case-control studies. In *Medicine (United States)* (Vol. 97, Issue 50). Lippincott Williams and Wilkins. https://doi.org/10.1097/MD.0000000000013386

Zou, H., Wu, L. X., Tan, L., Shang, F. F., & Zhou, H. H. (2020). Significance of Single-Nucleotide Variants in Long Intergenic Non-protein Coding RNAs. In *Frontiers in Cell and Developmental Biology* (Vol. 8). Frontiers Media S.A. https://doi.org/10.3389/fcell.2020.00347

11 Self-Assessment

My understanding of laboratory operations and research processes was limited when I joined the Biomedical Research Unit (URB). However, throughout my time here, I have gained invaluable insights. My experience at URB has not only equipped me with practical skills in laboratory organization, teamwork, and quality analysis execution but has also highlighted the significance of effective data management. I have learned how to meticulously handle data, from its collection and storage to its processing and presentation of results.

Moreover, collaborating with my peers within the team has emphasized the importance of open communication and idea exchange. Engaging in discussions, presenting my ideas, and receiving constructive feedback have enhanced my research approach and taught me the value of diverse perspectives.

As a newcomer to the research world, the knowledge and skills I've gained at the URB have proven immensely valuable and transformative. Armed with a deeper understanding, I am excited to embark on future research projects.

Thank you for taking the time to read my bachelor's thesis.

12 Annex

R Notebook

Libraries

```
library(readxl)
library(here)
library(openxlsx)
library(tableone)
library(vcd)
library(lavaan)
library(dplyr)
library(randomForest)
library(ggplot2)
library(tidyr)
```

Import data

```
full_data <- read_excel(here("Databases", "Complete_data.xlsx"))
full_data_allele <- read_excel(here("Databases", "Complete_data_allele.x
lsx"))
full_data_genotype <- read_excel(here("Databases", "Complete_data_encode
d.xlsx"))

column_names <- colnames(full_data_allele)
column_index <- which(column_names == "GroupEOM")
column_names[column_index] <- "Group"
colnames(full_data_allele) <- column_names

column_names <- colnames(full_data_genotype)
column_index <- which(column_names == "GroupEOM")
column_names[column_index] <- "Group"
colnames(full_data_genotype) <- column_names
```

Descriptive Analysis

Function Definition

```
write_descriptive <- function(df, dependent, factorVar) {
  # Create table one
  additional_factorVars <- colnames(df)[grep("^X|^rs", colnames(df))]
  # Combine additional factorVars with the original factorVar
  all_factorVars <- c(factorVar, additional_factorVars)
  one_group <- CreateTableOne(strata = dependent, data = df, factorVars
= all_factorVars)

  # Generate the first result file
  result <- print(one_group, printToggle = FALSE)
  write.csv(result, file = here("ResultFiles/Descriptive", paste("descri
ptive_", dependent, ".csv", sep = "")))
```

```r
  # Generate the second result file
  clean_data <- read.csv(here("ResultFiles/Descriptive", paste("descript
ive_", dependent, ".csv", sep = "")))
  # Get data for only snps
  clean_data <- clean_data[grepl("^rs|^X", clean_data[, 1]), ]
  colnames(clean_data)[1] <- "variable"
  clean_data <- clean_data[, c("variable", "p")]
  write.xlsx(clean_data, here("ResultFiles/Descriptive", paste("descript
ive_snp_", dependent, ".xlsx", sep = "")))
}

generate_df <- function(df, dependent, factorVar) {
  write_descriptive(df, dependent, factorVar)
  significatives <- read.xlsx(here("ResultFiles/Descriptive", paste("des
criptive_snp_", dependent, ".xlsx", sep = "")))
  significatives$p <- as.numeric(significatives$p)
  selected_variables <- subset(significatives, p < 0.05 | is.na(p))$vari
able
  selected_variables <- sapply(strsplit(selected_variables, " "), functi
on(x) x[1])
  return(df[unlist(c(selected_variables, dependent))])
}

get_strength <- function(df, dependent, factorVar, folder) {
  significatives <- generate_df(df, dependent, factorVar)
  result_df <- data.frame( "Variable" = character(0), "Strength" = numer
ic(0))
  # Repeat for all columns
  for (col_name in names(significatives)[1:(ncol(significatives) - 1)])
{
    complete_rows <- complete.cases(significatives[[col_name]], signific
atives[[dependent]])
    complete <- significatives[complete_rows, ]
    complete <- complete[, c(col_name, dependent)]
    cont_table <- table(complete)
    assoc_result <- assocstats(cont_table)
    cramer_v <- assoc_result[[5]]
    result_df[nrow(result_df) + 1,] <- list(col_name, cramer_v)
  }
  write.xlsx(result_df, here(paste("ResultFiles/Association_Strength/",
folder, sep = ""), paste(dependent, ".xlsx", sep = "")))
}


# Decide factor Var
factorVarComplete = c("Group",
  # Antropometrics
  "Sex",
  # Habits
  "Smoking", "Drinking", "Drugs",
  # Family history
  "Unknown_FH", "No_family_history", "Neoplasia_FH", "Obesity_FH", "T2DM
_FH", "CVD_FH", "HT_FH", "DLP_FH",
```

```
  # Surgery
  "BS_type", "VLCD", "VLCD_type",
  # Comorbidities
  "Cancer", "T2DM", "HT", "DLP", "No_comorb", "Metabolic_syn", "OSA", "C
PAP", "Depression", "Anxiety", "Hipotiroidism", "Hipertiroidism", "Chole
cystectomy", "CAD", "Hepatitis_A", "Hepatitis_B", "Hepatitis_C", "Hepati
tis_inf",
  # Drugs
  "Vasodilators",  "Unknown", "Tiazolidinadonas", "Sulfonylureas", "Seda
tive", "Other_antidep", "Other_lipid_mod", "Insulina_drug", "Inh_seroton
in", "Inh_monoamines", "Inh_SGLT2", "Inh_HMG", "Horm_tiroid", "Inh_DPP4"
, "Fibrates", "Diuretics", "Ca_block", "Biguanides", "Beta_block", "Anti
psic", "Antidep", "ARH2",  "Anxiolytic", "AngII_block", "AR_GLP1", "Alph
a_block", "ACE_inh", "Statines",
  # Liver
  "Liver_sample", "Steatosis_score", "Inflammation", "Ballooning",
  "Fibrosis", "NAS", "NAS_qual", "NAS_qual_quentin", "SAF")

factorVarAllele = c("Group",
  # Antropometrics
  "SexMan",
  # Habits
  "Smoking", "Drinking", "Drugs",
  # Family history
  "Unknown_FH", "No_family_history", "Neoplasia_FH", "Obesity_FH", "T2DM
_FH", "CVD_FH", "HT_FH", "DLP_FH",
  # Surgery
  "BS_type", "VLCD", "VLCD_type",
  # Comorbidities
  "Cancer", "T2DM", "HT", "DLP", "No_comorb", "Metabolic_syn", "OSA", "C
PAP", "Depression", "Anxiety", "Hipotiroidism", "Hipertiroidism", "Chole
cystectomy", "CAD", "Hepatitis_A", "Hepatitis_B", "Hepatitis_C", "Hepati
tis_inf",
  # Drugs
  "Vasodilators",  "Unknown", "Tiazolidinadonas", "Sulfonylureas", "Seda
tive", "Other_antidep", "Other_lipid_mod", "Insulina_drug", "Inh_seroton
in", "Inh_monoamines", "Inh_SGLT2", "Inh_HMG", "Horm_tiroid", "Inh_DPP4"
, "Fibrates", "Diuretics", "Ca_block", "Biguanides", "Beta_block", "Anti
psic", "Antidep", "ARH2",  "Anxiolytic", "AngII_block", "AR_GLP1", "Alph
a_block", "ACE_inh", "Statines",
  # Liver
  "Liver_sample", "Steatosis_score", "Inflammation", "Ballooning",
  "Fibrosis", "NAS", "NAS_qual", "NAS_qual_quentin", "SAF")
```

Data Gathering

```
# Decide data to use
data <- full_data_genotype
factorVar <- factorVarAllele
data <- subset(data, !(SAF == "Non-FLD" | is.na(SAF)))
data <- full_data
write_descriptive(data, "Group", factorVar)

# Generate descriptive files for each dependent variable
```

```r
write_descriptive(data, "NAS_qual_quentin", factorVar)
write_descriptive(data, "NAS_qual", factorVar)
write_descriptive(data, "NAS", factorVar)
write_descriptive(data, "Steatosis_score", factorVar)
write_descriptive(data, "Ballooning", factorVar)
write_descriptive(data, "Inflammation", factorVar)
write_descriptive(data, "Fibrosis", factorVar)
write_descriptive(data, "SAF", factorVar)
write_descriptive(data, "Group", factorVar)

iterate_strength  <- function(df, factorVar, folder) {
  dependents <- c("NAS_qual_quentin", "SAF", "NAS_qual", "NAS", "Steatos
is_score", "Ballooning", "Inflammation", "Fibrosis", "Group")
  for (dependent in dependents) {
    get_strength(df, dependent, factorVar, folder)
  }
}

iterate_strength(full_data_genotype, factorVarAllele, "Genotype")
iterate_strength(full_data_allele, factorVarAllele, "Alleles")
iterate_strength(full_data, factorVarComplete, "Normal")
```

Graph Generation

Data Preparation

```r
df <- full_data
factorVar <- factorVarComplete
dependent <- "SAF"

# Create table one
additional_factorVars <- colnames(df)[grep("^X|^rs", colnames(df))]
# Combine additional factorVars with the original factorVar
all_factorVars <- c(factorVar, additional_factorVars)
one_group <- CreateTableOne(strata = dependent, data = df, factorVars =
all_factorVars)

# Generate the first result file
result <- print(one_group, printToggle = FALSE)
write.csv(result, file = here("ResultFiles/Descriptive", paste("descript
ive_", dependent, ".csv", sep = "")))

# Generate the second result file
clean_data <- read.csv(here("ResultFiles/Descriptive", paste("descriptiv
e_", dependent, ".csv", sep = "")))
# Find the index of the first row that meets the condition
first_index <- min(which(grepl("^rs|^X", clean_data[, 1])))

# Keep all rows after the first row that meets the condition
clean_data <- clean_data[first_index:nrow(clean_data), ]
colnames(clean_data)[1] <- "variable"
clean_data <- clean_data[, c(1,4,2,3)]
```

```
# Initialize variables
last_rs_number <- ""
modified_variable <- character(nrow(clean_data))

# Iterate over each row
for (i in 1:nrow(clean_data)) {
  current_value <- clean_data$variable[i]

  if (grepl("^rs", current_value)) {
    # Update last_rs_number
    last_rs_number <- current_value

    # Remove row if it starts with "rs"
    modified_variable[i] <- ""
  } else {
    # Append last_rs_number to the row
    modified_variable[i] <- paste(last_rs_number, current_value, sep = "
")
  }
}

# Assign modified values back to the clean_data dataframe
clean_data$variable <- modified_variable


clean_data$variable <- modified_variable
clean_data <- clean_data[!clean_data$variable == "", ]

clean_data$variable <- gsub("\\(\\%\\)", "", clean_data$variable)  # Rem
ove "(%)"
clean_data$variable <- gsub("\\s+", " ", clean_data$variable)  # Replace
multiple whitespaces with a single whitespace

# Extract number between parentheses in Non.FLD column
clean_data$Non.FLD <- as.numeric(gsub(".*\\((.*?)\\).*", "\\1", clean_da
ta$Non.FLD))

# Extract number between parentheses in NAFLD column
clean_data$NAFLD <- as.numeric(gsub(".*\\((.*?)\\).*", "\\1", clean_data
$NAFLD))

# Extract number between parentheses in NASH column
clean_data$NASH <- as.numeric(gsub(".*\\((.*?)\\).*", "\\1", clean_data$
NASH))

clean_data <- transform(clean_data,
                        snp = gsub("^([a-zA-Z]+\\d+).*", "\\1", variable
),
                        genotype = gsub("^[a-zA-Z]+\\d+\\s+(.*)", "\\1",
variable))
clean_data <- clean_data[, !(names(clean_data) %in% "variable")]
clean_data <- clean_data[!(clean_data$genotype %in% c("NOAMP,NOAMP", "UN
D,UND", "NAS_qual_quentin = 1 ", "SAF ", "NAFLD", "NASH", "Non-FLD")), ]
```

```r
# Assuming your dataframe is named "df"
clean_data <- subset(clean_data, Non.FLD + NAFLD + NASH > 1.5)
genes <- read_excel(here("Databases", "genes.xlsx"))

clean_data$gene <- NA  # Initialize the 'gene' column with NA

# Loop through each row of clean_data
for (i in seq_len(nrow(clean_data))) {
  snp_value <- clean_data$snp[i]  # Get the SNP value from clean_data
  matching_row <- genes[genes$snp == snp_value, ]  # Find the matching r
ow in genes
  if (nrow(matching_row) > 0) {
    clean_data$gene[i] <- matching_row$gene  # Assign the 'gen' value to
'gene' column in clean_data
  }
}
```

Graph Plotting

```r
rsNumber <- "rs5982"  # Replace with the desired rsNumber

generate_plot <- function(clean_data, rsNumber, gene_values) {

# Subset the data for the given rsNumber
subset_data <- clean_data[clean_data$snp == rsNumber, ]

# Create the new dataframe
new_dataframe <- data.frame(Genotype = subset_data$genotype,
                            Group = rep(c("Non.FLD", "NAFLD", "NASH"), e
ach = nrow(subset_data)),
                            Frequency = c(subset_data$Non.FLD, subset_da
ta$NAFLD, subset_data$NASH))

new_dataframe$Group <- ifelse(new_dataframe$Group == "Non.FLD", "Non-NAF
LD", new_dataframe$Group)


# Update the theme settings
my_plot = ggplot(new_dataframe, aes(fill=Group, y=Frequency, x=Genotype)
) +
  geom_bar(position='dodge', stat='identity') +

  #theme_minimal() +
  theme(#panel.grid = element_blank(),     # Remove background grid line
s
        axis.line.y = element_line(color = 'black'),
        panel.background = element_blank(),  # Remove background fill
        #axis.line.x = element_blank(),  # Remove the x-axis line
        #axis.ticks = element_blank(),     # Remove axis ticks
        axis.text.x = element_text(color = 'black'),  # Set color for x-
axis tick labels
    axis.text.y = element_text(color = 'black'),  # Set color for y-axis
```

```
tick labels
    plot.title = element_text(hjust = 0.5, size = 12)) +
  labs(x=NULL, y='Frequency', title=paste(rsNumber, gene_values, sep = "
- ")) +
  scale_fill_manual('Group', values=c('#8bd3c7', '#ffffaf', '#beb9db'))



ggsave(
  here("Plots/Descriptive", paste(rsNumber, "pdf", sep = ".")),
  plot = my_plot,
  device = "pdf",
  height = 3.5,
  width = 3.5
)

}


# Get unique values in the snp column of clean_data
unique_snps <- unique(clean_data$snp)

# Loop through each unique snp value and access the corresponding gene v
alues
for (snp_value in unique_snps) {
  gene_values <- clean_data$gene[clean_data$snp == snp_value]

  # Print the snp_value and corresponding gene_values
  generate_plot(clean_data, snp_value, gene_values)
}
```

Random Forest

```
columns_to_select <- c("SAF", "rs58542926C.C", "rs58542926C.T", "rs18006
29A.G", "rs1800629G.G", "rs8107974A.A", "rs8107974A.T", "rs5982A.G", "rs
11858624G.G", "rs11858624G.T", "rs62021874C.C", "rs62021874C.T", "rs1919
127T.T", "rs12077210C.C", "rs12077210C.T", "rs8050136C.C", "rs9939609A.A
", "rs9939609T.T")

selected_columns <- full_data_genotype[, columns_to_select]
selected_columns <- na.omit(selected_columns, cols = "SAF")
selected_columns <- as.data.frame(lapply(selected_columns, as.factor))

set.seed(123) # For reproducibility
train_indices <- sample(1:nrow(selected_columns), 0.7 * nrow(selected_co
lumns)) # 70% for training
train_data <- selected_columns[train_indices, ]
test_data <- selected_columns[-train_indices, ]

dependent_variable_name <- "SAF"
independent_variable_names <- c("rs58542926C.C", "rs58542926C.T", "rs180
0629A.G", "rs1800629G.G", "rs8107974A.A", "rs8107974A.T", "rs5982A.G")
```

```r
rf_model <- randomForest(SAF ~ ., data = train_data, mtry = 3,
                         importance = TRUE, na.action = na.omit)


# Step 4: Evaluate the model's performance (you can use different evalua
tion metrics based on your problem)
# For example, if it's a regression problem, you can use mean squared er
ror (MSE)
# If it's a classification problem, you can use accuracy, precision, rec
all, F1-score, etc.

# For regression:
predictions <- predict(rf_model, newdata = test_data)


# Calculate accuracy
accuracy <- sum(predictions == test_data$SAF) / nrow(test_data)

# Display the accuracy
print(paste("Accuracy:", accuracy))

plot(rf_model$err.rate[,1], type = "l")

my_plot <- ggplot(data = data.frame(x = seq_along(rf_model$err.rate[,1])
, y = rf_model$err.rate[,1])) +
  geom_line(aes(x, y)) +
  labs(x = "X-axis Label", y = "Y-axis Label", title = "My RF Model Erro
r Rate Plot")

# Save the plot as a PDF file
ggsave("rf_model_plot.pdf", plot = my_plot, width = 8, height = 6)  # Ad
just width and height as needed


plot(rf_model)
```